Navigating the New Era of Grid Demand: Large Load Forecasting

A review of challenges and methodologies in forecasting large data centre loads for utilities and system operators

Dr Diptargha Chakravorty
Head of Grid Analysis and Planning (UK)
Siemens Energy

International Workshop on Managing Global Energy Demand of AI Data Centres

ESIG Large Load Task Force

September 2025

Imperial College London

Topics to cover...

1. Introduction: The New Energy Landscape
2. Key Challenges in Forecasting Large Loads
3. Data Analysis: Scale & Uncertainty
4. Forecasting Techniques : Utility and RTO Methodologies
5. Data Centre Flexibility
6. Conclusion & Key Trends

The Challenge of Unprecedented Growth

Large loads are requesting grid interconnection at an unprecedented speed and scale, moving the US out of an era of flat power demand.

Key Drivers

- **Data centres**- Driven by cloud migration and AI deployment
- **Hydrogen production**- Clean energy transition
- **Industrial electrification**- Manufacturing and processing facilities
- **Large-scale EV charging**- Fleet and commercial charging hubs

Scale

- Individual projects can be **GW-scale,** with rapid ramp-up times

What's at Stake?

- **Grid Reliability:** Inaccurate forecasts threaten ability to meet demand
- **Infrastructure Investment**: Risk of overbuilding (stranded assets) or underbuilding
- **Resource Adequacy**: Sufficient generation to meet forecasted load

> US power demand is projected to grow up to **3.1% annually** through 2040, a significant shift from the **0.4% annual growth** seen from 2010-2022

Categorisation of Large Electrical Loads

Large loads are diverse, with different operational characteristics, flexibility potential, and impacts on the grid. Most definitions consider a "large load" to be greater than 50-75 MW.

Category	Components	Key Characteristics
Data Centres & Computational	Servers (CPU/GPU), cooling systems, UPS, backup generators. Includes AI training and crypto mining.	High, consistent 24/7 load; very high load factor (85%+). Cooling load is weather-sensitive.
Industrial	Furnaces, mills, pumps, compressors, electrolysers. Sectors include mining, metals, semiconductors, and chemicals.	Load profiles are generally well-understood, but electrification of industrial heat is a new driver.
Hydrogen Production	Electrolyser stacks (PEM, Alkaline), compressors, storage systems.	Can be a highly flexible load, able to ramp up or down quickly depending on technology and storage availability.
Transportation	DC fast chargers (for fleets). Includes electrified rail, ports, and airports.	High uncertainty due to dependency on fleet schedules and charging behaviour.

PEM – Proton Exchange Membrane

Focus on Data Centres: A Primary Growth Driver

Data centres are a dominant force in the new load growth, driven by cloud migration and the rapid rollout of Generative AI.

Global & US Demand

- **Global Demand:** Estimated to more than triple to at least **171 GW** by 2030
- **US Demand:** Data centre power demand forecast to grow **160%** by 2030
- Potentially consuming **8% of total US power** by 2030

AI's Impact

- A single ChatGPT query requires nearly **10 times more electricity** than a Google search
- AI workloads are driving a shift from CPUs to more power-intensive **GPUs and ASICs**
- Rack power density is increasing from a typical 5-15 kW to **40-300 kW** for AI training
- High power density requires innovations in **liquid cooling** technology

> The energy intensity of AI workloads is fundamentally transforming data centre design, power requirements, and utility forecasting models

ASIC – Application Specific Integrated Circuit

Data Centre Types: Business Models

The ownership model influences operational decisions, contractual obligations, and the potential for providing grid flexibility.

Ownership Model	Description	IT User Role	Data Centre Operator Role
Self-Owned / Enterprise	The end-user owns and operates the data centre for its own purposes.	Owns DC, racks; arranges maintenance.	(Same as user)
Co-location	An operator builds the facility and leases space, power, and cooling to multiple tenants.	Owns and manages their own servers within the leased space.	Owns DC; provides power, cooling, security.
Hosted	A provider owns the data centre and leases its servers and storage to customers.	Manages workloads on leased hardware.	Owns DC and all hardware; arranges maintenance.
Cloud	Hyperscale providers own and operate massive data centres, offering services (e.g., computing, storage) to third parties via SLAs.	Accesses services, does not manage physical hardware.	Owns and operates all infrastructure.

Data Centre Types: Technical Characteristics

Workload and reliability requirements are key technical differentiators that impact a data centre's energy profile and flexibility.

Classification by Workload:

Batch Processes
Discrete, schedulable tasks like AI model training. Potentially tolerant to delay.

Interactive Workloads
User-driven processes like e-commerce. Less delay tolerant.

Real-time Systems
"Always-on" services like IoT controls or online gaming. Mission-critical and sensitive to delays.

Streaming
High-throughput, low-latency data flows for services like video conferencing. Delay sensitive.

Classification by Reliability (Uptime Tiers):

Tier 1
Basic capacity with a single power path, no redundancy.

Tier 2
Redundant power and cooling components (N+1).

Tier 3
Concurrently maintainable with multiple power paths (2N).

Tier 4
Fully fault-tolerant, with redundancy for every component (2N+1).

Key Forecasting Challenges: Overview

Major uncertainties have emerged in the growth of new types of load. Four critical challenge areas demand new approaches:

1. **Uncertainty of Customer-Provided Data**

   Developers provide unreliable estimates of capacity needs and timeline projections

2. **The Impact of Artificial Intelligence (AI)**

   AI servers require 3-4x more energy, transforming load profiles and growth patterns

3. **Data Scarcity & Double-Counting**

   Limited historical precedent and risk of counting loads twice in forecasts as customers may apply in multiple regions

4. **Geographic Concentration**

   New loads cluster in areas with critical infrastructure, creating pockets of intense grid stress

Data Confidentiality

Customer privacy limits the information available to forecasters, hindering efforts to create accurate load profiles and share data between utilities.

> Traditional forecasting methods are insufficient for this new reality

Challenge 1: Uncertainty of Customer-Provided Data

Forecast inputs are often overestimated or unreliable due to developers' optimistic projections.

Capacity vs. Actual Demand

- Significant difference between requested **capacity** and metered **demand**
- Colocation data centers build capacity years before full utilization
- Actual load materializes much more slowly than projected

Schedule Delays

- Securing tenants creates uncertainty in timing
- Permitting processes often take longer than anticipated
- High likelihood of project schedule slips that must be accounted for in forecasts

Challenge 2: The Impact of Artificial Intelligence (AI)

AI workloads are accelerating energy needs:

- **AI servers require 3-4 times more energy than traditional servers**
- **Could accelerate how quickly a data centre reaches its full load**
- **Load factor uncertainty may increase ultimate demand but reduce overall load factors**

Challenge 3: Data Scarcity and Double-Counting Risk

Limited historical precedence for new load types makes it hard to generalise.

Few Data Points

Limited historical samples for AI and data center load profiles

Evolving Development Modes

Rapidly changing technology and deployment strategies

Double-Counting Risk

Counting loads both as specific projects and within macroeconomic forecasts

Duke Energy's Term

"Measurement Uncertainty"

The challenge of accurately counting load once, without duplication

- Project-specific forecast
- Econometric forecast
- Risk of double-counting

Challenge 4: Geographic Concentration

Uneven Distribution

- New large loads are not distributed evenly across the grid
- Data centers cluster in areas with access to critical infrastructure:
  - Fiber optic networks
  - Water resources
  - Transmission capacity
- Creates pockets of intense grid stress requiring targeted infrastructure upgrades

Case Study: Virginia

Home to the world's largest data centre market

Concentrated in Loudoun County's "Data Center Alley"

Key Utilities Serving This Region:

- Dominion Energy
- NOVEC

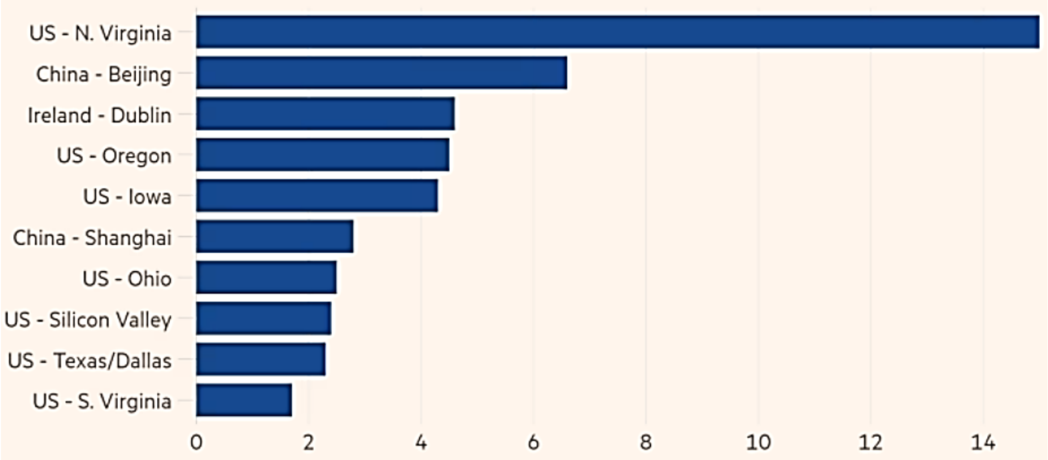
"By 2026, Northern Virginia will host approximately 56% of all the data center capacity in the United States."

— Industry forecast

NOVEC – Northern Virginia Electric Cooperative

Challenge 4: Geographic Concentration

- Facilities used by big tech companies like Amazon, Google and Microsoft
- The data centres used 7bn litres of water in 2023, compared to 4.2bn litres in 2019
- Overall, US data centres consumed more than 283bn litres of water in 2023, roughly equivalent to the amount London consumes in 4 months



Northern Virginia and the Greater Beijing area make up 22% of the total global hyperscale data centre capacity

FINANCIAL TIMES

JK COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI

Technology sector + Add to myFT

US tech groups' water consumption soars in 'data centre alley'

Usage has increased by almost two-thirds since 2019 in the world's largest concentration of computing infrastructure

Data centres in Loudoun county, Virginia. Water is used to cool computing equipment © AP

Data Centres in the UK

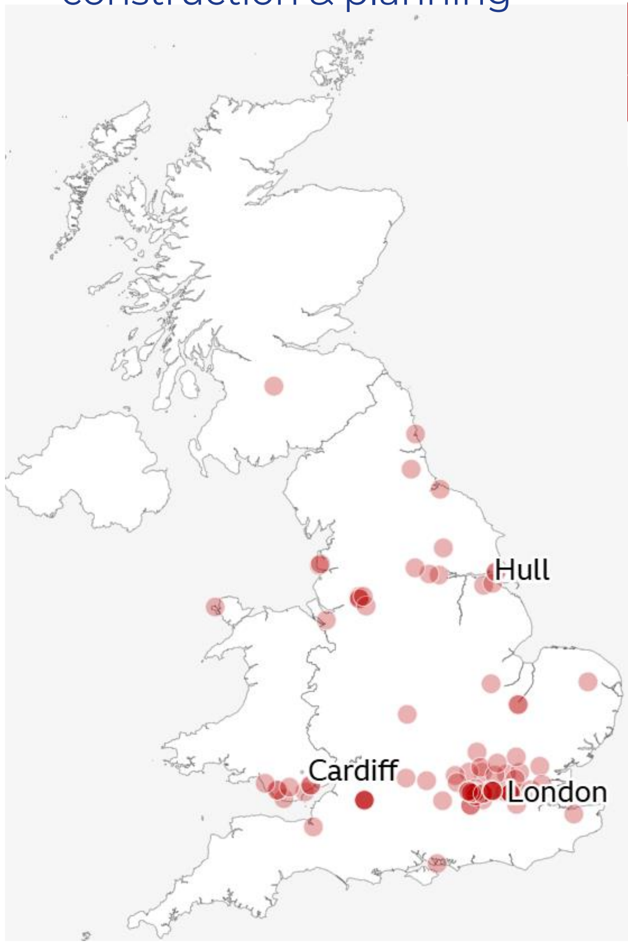
Currently there an estimated 477 data centres in the UK. This number will increase by 100 in the next 5 years

There are concerns about the huge amount of energy and water the new data centres will consume

NESO has projected that data centres could add up to 71 TWh of electricity demand in the next 25 years

Major locations are London, Wales, Greater Manchester, Hull, Blyth

Location of UK data centres – construction & planning




BBC Sign in Home News Sport Weather iPlayer

NEWS

Home | InDepth | Israel-Gaza war | War in Ukraine | Climate | UK | World | Business | Politics | Culture

Technology

Data centres to be expanded across UK as concerns mount

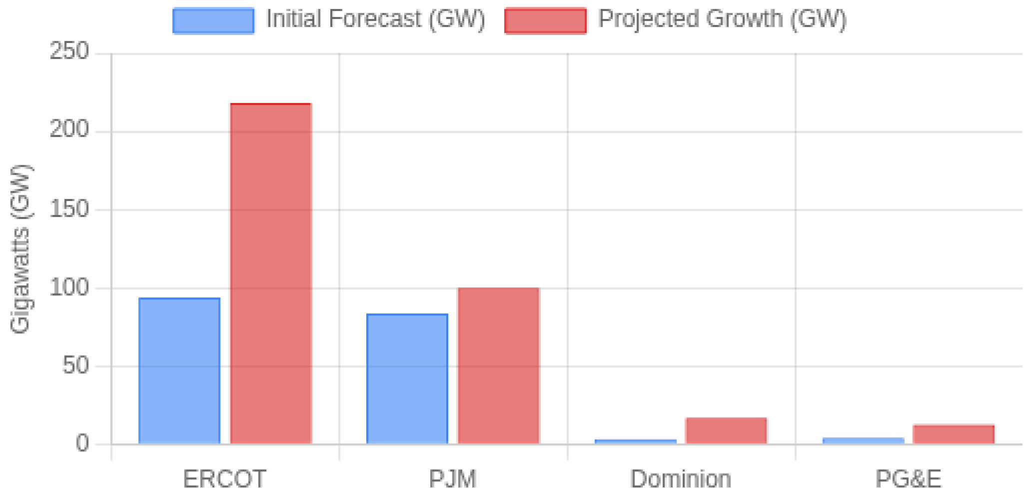
GETTY IMAGES

Data centres, like this one Google is building in Hertfordshire, are becoming a more familiar sight across the UK

Data Analysis: Scale of Projected Growth

Recent forecasts reveal a sharp break from historical trends

Projected Load Growth by Provider



Source: Utility forecasts and public filings, 2024-2025

ERCOT – Electric Reliability Council of Texas
PJM – Pennsylvania-New Jersey-Maryland Interconnection
PG&E – Pacific Gas and Electric Company

Key Growth Examples

ERCOT

Peak load forecast jumps from **94 GW** (2024) to **218 GW** (2030)

PJM

15-year growth rate: **2.4%**, 2030 forecast is **+16 GW** higher than p2024

Dominion

Data center forecast: **3.5 GW** (2024) to **17+ GW** (2045)

PG&E

Queue explosion: **4.4 GW** to **12.8 GW** in just 6 months

Table: Utility Practices—Forecast Adjustments

Comparing quantitative adjustments across different forecasting entities

METRIC	FORECASTER	QUANTITATIVE ADJUSTMENT	IMPACT
Capacity-to-Demand Ratio	ERCOT	Reduces new data centre load to **49.8%** of the requested amount in its "Adjusted Forecast"	Substantial
	CEC	Assumes peak load is **67%** of requested capacity, based on historical SVP data	Moderate
	APS	Estimates measured demand will ramp to **75%** of a data centre's capacity	Moderate
	PJM	Translates capacity requests to demand using a factor between **70-90%**	Variable
Schedule Delays	ERCOT	Applies a blanket **180-day delay** for all new projects	Universal
	APS	Models a **50% probability of a 12-month** start date delay for a typical project	Probabilistic

Table: Utility Practices—Forecast Adjustments

Comparing quantitative adjustments across different forecasting entities

METRIC	FORECASTER	QUANTITATIVE ADJUSTMENT	IMPACT
Project Viability	BPA	Includes projects in the "Load Forecast" only if they have a **>=70% probability** of completion	Threshold
	CEC/PG&E	Applies confidence levels based on application status (early inquiry: **10%** , engineering studies: **70%**)	Stage-based

PG&E – Pacific Gas and Electric Company
BPA – Bonneville Power Administration
CEC – California Energy Commission

Key Risk Factor

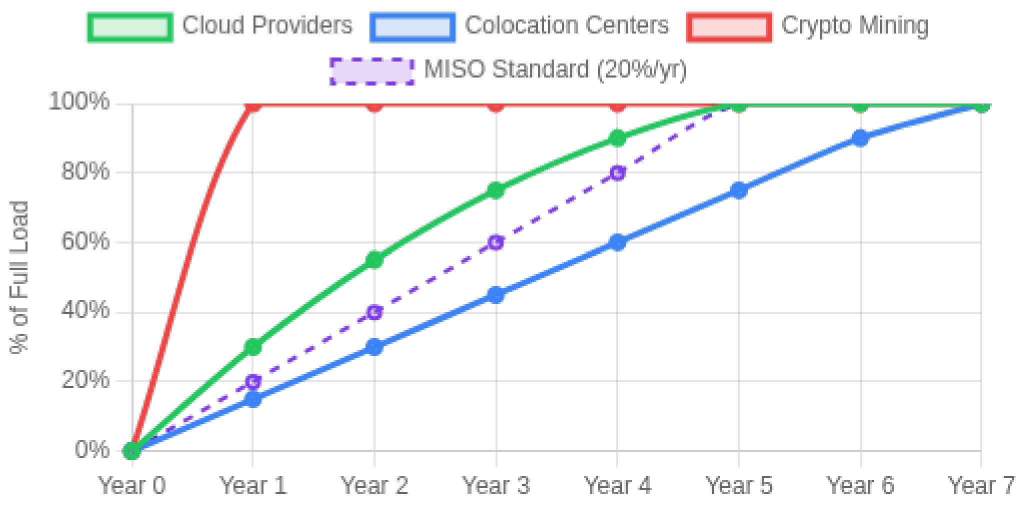
Raw, customer-provided data is **rarely used** without significant quantitative adjustments

Note: Utilities consistently reduce customer-provided data based on historical performance and expert judgment

Modeling the Ramp: How Fast Does Load Actually Arrive?

The speed at which facilities reach full load varies significantly by customer type

Ramp Rate Comparison by Customer Type



Percent of Full Load Capacity Achieved Over Time (Years)

MISO – Midcontinent Independent System Operator
CEC – California Energy Commission
APS – Arizona Public Service

Utility Ramp Rate Examples

APS

Limits a customer's proposed **3.3 MW/month** to a more realistic **2.0 MW/month** based on historical data

Dominion

Ramp speed by customer type:

Cloud Providers: **3-5 years**

Colocation Centres: **5-7 years**

Crypto Mining: **1 year**

MISO

Applies an expected ramp rate of approximately **20% annually** for new projects

CEC

Default ramp schedule assumes **149% year-over-year growth** for the first five years

Forecasting Techniques: Industry Approaches

The industry is evolving toward more sophisticated methodologies to address unprecedented load growth uncertainty.

Four distinct approaches have emerged:

Bottom-Up (Project-Specific)
Tracking individual large load projects with quantitative adjustments

Top-Down (Econometric)
Statistical models using historical relationships with macroeconomic drivers

Hybrid Approaches
Most common method combining baseline forecasts with specific project overlays

Stochastic/Probabilistic
Advanced approach explicitly modeling uncertainty through simulations

Detailed Comparison: Utility and RTO Methodologies

Comparative analysis of forecasting approaches across the industry

UTILITY / RTO	PRIMARY METHODOLOGY	KEY FEATURES & PROCESS
APS	**Bottom-Up Discounting**	Employs a **three-step discount:** ➢ Cap the ramp rate ➢ Apply probability for schedule delays ➢ Reduce total load to a percentage of nameplate capacity (75%)
NOVEC	**Granular Bottom-Up**	Uses a **rigorous, multi-stage contracting process** to screen projects. Forecasts built from individual building level up, using econometric models to predict utilisation rate.
ERCOT	**Hybrid with Mandated Inputs**	State law requires ERCOT to accept load forecasts provided in **"TSP Officer Attested Letters"**. ERCOT then creates its own **"Adjusted Forecast“** by applying risk adjustments like a 180-day delay and reducing data centre load to 49.8%.

NOVEC – Northern Virginia Electric Cooperative
ERCOT – Electric Reliability Council of Texas
APS – Arizona Public Service

Detailed Comparison: Utility and RTO Methodologies

Comparative analysis of forecasting approaches across the industry

UTILITY / RTO	PRIMARY METHODOLOGY	KEY FEATURES & PROCESS
PJM	**Hybrid: Top-Down with Adjustments**	Starts with a **multivariable regression model** and allows members to submit formal **"Large Load Adjustment“** requests. PJM screens these requests, translates capacity to demand (70-90% factor), and applies multi-year ramp schedule.
Duke Energy	**Hybrid: Macroeconomic Baseline**	Begins with bottom-up macroeconomic model and adds known large load projects as adjustments. Explicitly considers and tries to mitigate "Project Uncertainty" and "Measurement Uncertainty".
Santee Cooper	**Stochastic / Probabilistic**	Internal team evaluates each potential customer against a checklist. This feeds a **stochastic model** that runs 50,000 trials with probability distributions for key variables to produce a risk-adjusted forecast.
MISO	**Compendium Approach**	Uses a NAICS-based "compendium approach" for industrial forecast, gathering data from public announcements and third-party sources to form a consensus view, moderated with deployment adjustments and ramp rates.

Hybrid and probabilistic approaches are increasingly favored as they balance historical trends with emerging large load realities

NAICS – North American industry Classification System

The Future: Data Centre Load Flexibility

While currently viewed as inflexible, data centres have technical potential to provide grid services, though economic and contractual barriers exist.

Potential Sources of Flexibility

Compute Load

Shifting
Postponing non-time-sensitive batch processes (e.g., AI training) to off-peak hours

Scaling
Reducing processor power consumption using dynamic voltage and frequency scaling

Balance of Plant

Cooling Systems
Leveraging thermal inertia to temporarily reduce cooling load or pre-cool during low-price hours

On-site Power Assets

UPS/Batteries
Can provide fast frequency response and ancillary services

Standby Generators
Can reduce grid draw during peak events, subject to emissions limits

Key Barriers

Opportunity Cost
The lost revenue from interrupting computations is often the most significant factor

Service Level Agreements (SLAs)
Strict requirements for uptime and performance limit the ability to curtail load

Co-location Complexity
Serving multiple customers with different goals complicates coordinated flexibility actions

> While technical potential exists, unlocking data center flexibility will require new commercial structures, operational protocols, and regulatory frameworks

Key Takeaways (1/2)

1 A New Era of Load Growth is Here

Driven primarily by data centres and electrification, rapid load growth is creating significant challenges for grid planning after decades of predictable demand.

2 Traditional Forecasting is Obsolete

The unique characteristics of new large loads—especially their scale, speed, and uncertainty—mean that historical trend-based forecasting is no longer sufficient.

3 Uncertainty is the Core Challenge

Forecasters must contend with a pipeline of projects that may be speculative, have unrealistic schedules, or feature new technologies with no historical precedent.

4 Methodologies are Evolving

Utilities and ISOs are adopting more sophisticated deterministic, stochastic, and scenario-based approaches to better quantify and manage uncertainty. There is no single "best practice"; approaches must be tailored to local circumstances.

Key Takeaways (2/2)

5 Adjusting Customer Data is Essential

A critical step in modern forecasting is adjusting customer-provided information (capacity, ramp, schedule) based on historical realisation rates, professional judgment, and structured analysis.

6 Commercial Practices are Key to Risk Mitigation

New tariff structures, contracts with financial commitments (upfront payments, minimum bills), and stricter screening processes are vital to protect the system and all customers from the financial risks of overbuilding.

7 Collaboration is Critical

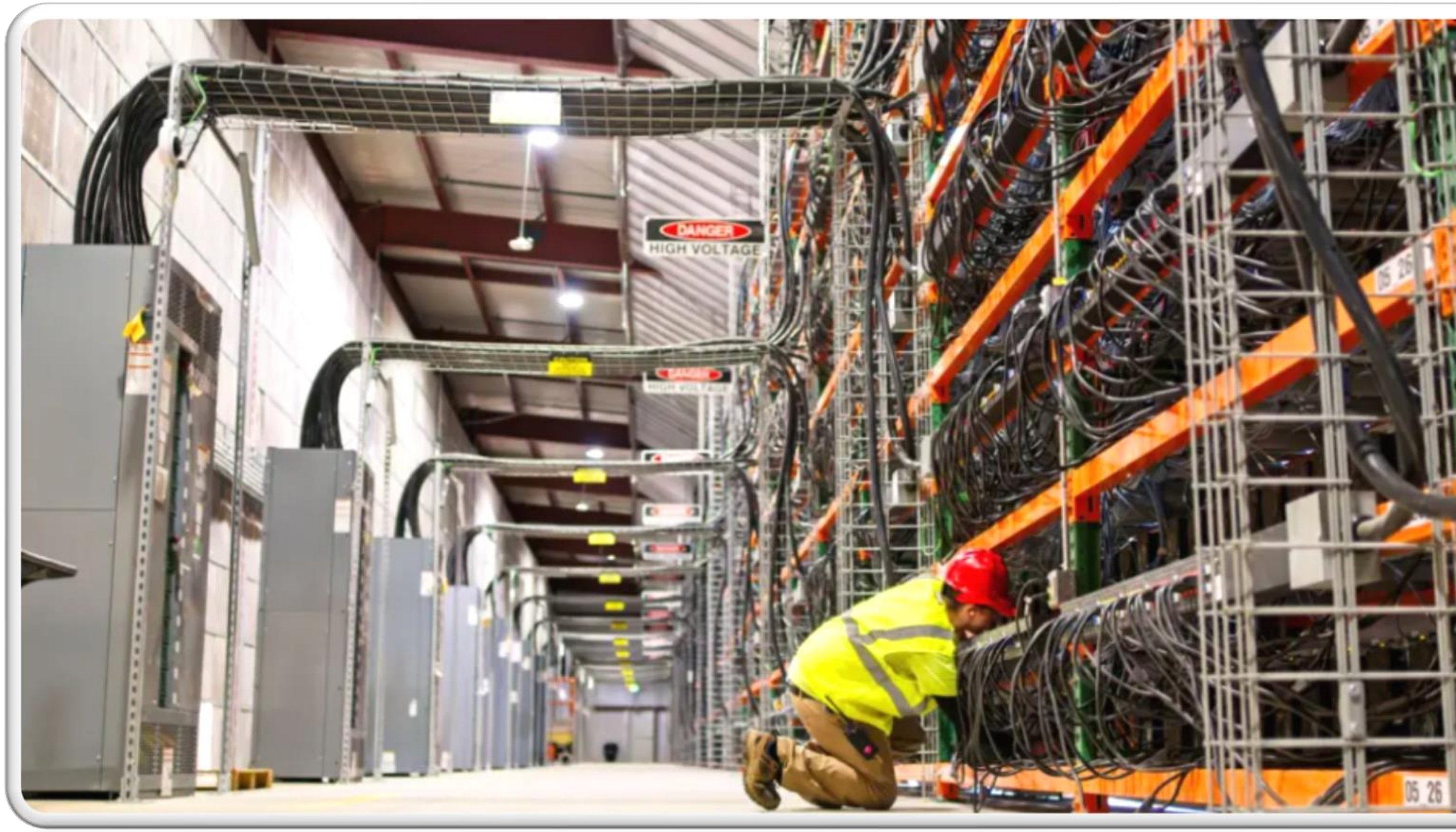
Improving forecast accuracy requires a cross-sectoral effort involving utilities, ISOs, regulators, and developers to share data and insights.

8 Flexibility is the Next Frontier

While challenging, unlocking the demand flexibility potential of data centres could provide significant benefits for grid reliability and efficiency.

Thank you

— **For your attention**



Diptargha Chakravorty