Bits and Mortar: The Microspatial Economics of Data Centers*

Rui Du† Xiaojin (Aaron) Sun‡ Siqi Zheng§

May 20, 2026

Abstract

Using a building-level panel of U.S. data center entries (2010–2024), we train a convolutional neural network to predict facility entry at 500-meter resolution. The model uncovers a structural bifurcation: hyperscale cloud providers prioritize power capacity, connectivity, and regulatory stability, whereas third-party colocation firms prioritize power access and land affordability. Counterfactual simulations through 2030 reveal that AI-driven demand intensifies constraints in established hubs. In contrast, on-site power generation and renewable grid expansion redistribute development toward previously unviable regions, demonstrating that the geographic equilibrium of data centers is highly sensitive to energy policy.

JEL Classification: R30, L86, Q40, O33, C45

Keywords: Data Center; Location Choice; Microspatial; Deep Learning

*The authors contributed equally to this work and are listed alphabetically. We gratefully acknowledge research support from the Real Estate Research Institute (RERI). For their insightful feedback, we thank our RERI mentors—Brent Ambrose (PSU), Mark Fitzgerald (Affinius Capital), and David Kelly (CBRE)—as well as Jacob Albers, Oliver Cowan, and Lexi Russel from Google's Data Center Portfolio Insights and Market Intelligence Team. We also appreciate the comments from Jorge De la Roca, Alvin Murphy, Amanda Ross, and Tingyu Zhou, and we thank Wenqi Liu and Peigen Mao for their excellent research assistance. All errors are our own.

†Department of Economics, Oklahoma State University, 347 Business Building, Stillwater, OK 74078, USA; Email: rui.du@okstate.edu.

‡Department of Economics, Oklahoma State University, 433 Business Building, Stillwater, OK 74078, USA; Email: xiaojin.sun@okstate.edu.

§Sustainable Urbanization Lab, Department of Urban Studies and Planning, and Center for Real Estate, Massachusetts Institute of Technology, Cambridge, MA 02139; Email: sqzheng@mit.edu.

1 Introduction

Data centers have emerged as the core physical capital of the digital economy, akin to the railroads of the industrial era and the fiber-optic networks of the Internet boom. Recent estimates suggest that planned U.S. data center capacity could expand by approximately 200 gigawatts over 2026–2032, implying a buildout on the order of $8 trillion that exceeds, as a share of GDP, the historical infrastructure cycles of railroads, electrification, and telecommunications (Van Nieuwerburgh, 2026). Although digitization facilitates ubiquitous connectivity, it has not eliminated the relevance of geography; rather, it has redefined the strategic role of spatial proximity. Distinct from retail anchors (Qian, Zhang and Zhang, 2024) or manufacturing plants (Greenstone, Hornbeck and Moretti, 2010), data centers exhibit extreme resource intensity and asset specificity, constituting a high-demand infrastructure asset class characterized by unique technical drivers, rapid technological obsolescence, and complex regulatory environments.[1]

Because data center location decisions are dictated by latent, highly localized factors, they create significant information asymmetries. Decoding these proprietary site-selection algorithms requires resolving both temporal dynamics and microspatial interactions—a task where existing empirical frameworks face inherent limitations. For example, while Fang and Greenstein (2025) document a strong preference for urban proximity among colocation providers and a cost-minimization strategy for hyperscalers, their static, county-level logit models obscure the essential micro-geographic forces. Similarly, although structural models (e.g., Elhai, 2026) provide valuable insights for quantifying selection forces and evaluating policy counterfactuals, their rigid parametric restrictions limit empirical fit and out-of-sample extrapolation.

Therefore, navigating the sector's extended development cycles and binding infrastructure constraints requires a flexible predictive framework. Such an approach must answer two central questions: what predicts the emergence of new digital hubs, and how do siting drivers differ between hyperscale cloud providers and third-party colocation operators? To address these questions, we bypass the limitations of traditional discrete choice models by developing a deep learning architecture that emulates real-world industry practices.

Building upon recent applications of computer vision in economics (Khachiyan et al., 2022; Pollmann, 2023; Qian, Zhang and Zhang, 2024), we represent the evolving economic and infras-

[1] Industry research describes data centers as a hybrid asset class that combines real estate and infrastructure characteristics, distinguishing them from traditional industrial or office assets (CBRE, 2026).

tructural landscape as multi-channel spatiotemporal grids and train a 2D Convolutional Neural Network (CNN) to decode these patterns. Our framework captures the temporal evolution of siting determinants through a strict predictive lag. By stacking multiple years of historical covariates to forecast subsequent market entry, the model learns from the momentum of localized developments, such as progressive grid upgrades or shifting demographic trends.

Moreover, the network evaluates spatial depth using a convolutional sliding window that acts as a learnable catchment radius. This mechanism continuously scans the periphery of each prospective site for critical physical resources, such as fiber-optic backbones and power substations, directly mirroring the forward-looking due diligence of developers. Through sequential convolutional filters and pooling layers, the architecture extracts nonlinear feature hierarchies and recovers complex spatiotemporal dependencies that conventional frameworks cannot observe.

To empirically ground our framework, we construct a geocoded longitudinal panel of 4,447 U.S. data centers managed by 807 owners between 2010 and 2024. Synthesizing proprietary industry data with satellite imagery, this dataset captures the sector's structural evolution wherein binding power, land, and cooling constraints began dictating placement. To evaluate the spatial logic underlying this expansion, we follow Fang and Greenstein (2025) by distinguishing hyperscale cloud providers from third-party colocation facilities—a bifurcation reflecting how firms balance internal operational linkages against external cluster benefits (Alcácer and Delgado, 2016). Within our sample, the 728 hyperscale facilities operate as a geographically sparse network, whereas the 3,719 third-party colocation sites exhibit heavy-tailed spatial clustering.

Our measurement strategy for the siting determinants explicitly reflects the extreme physical constraints of this unique real estate asset.[2] Motivated by these resource requirements, we construct a comprehensive, rasterized spatial panel at a 500-m resolution. This panel integrates physical infrastructure (grid topology, generation capacity, reliability, and fiber-optic connectivity), environmental factors (water stress, drought risk, and temperature), local economic dynamics (tax incentives, land values, and labor demographics), and a state-level legislative index reflecting regulatory frictions and clean energy mandates.

Validating out-of-sample through 2024, our framework demonstrates robust predictive accu-

[2]Modern data centers consume energy at rates 10 to 50 times higher than traditional office properties, accounting for approximately 4.4% of total U.S. electricity consumption (U.S. DOE, 2024; Shehabi et al., 2024). Furthermore, a single facility's cooling system can require up to five million gallons of water daily (Environmental and Energy Study Institute , EESI; Volzer, 2025), and hyperscale sites frequently exceed 100 acres, driving a 144% surge in average land transaction prices since 2022 (Cushman & Wakefield, 2025).

racy in isolating highly viable sites from the broader geographic landscape.[3] By translating these divergent constraints into continuous opportunity maps for the contiguous U.S., we delineate the feasible geographic frontier for future expansion. Ultimately, this approach pinpoints both realized investments and latent arbitrage opportunities: locations where optimal infrastructure fundamentals imply high suitability despite the absence of observed entry.

Furthermore, by interrogating the network's internal decision-making process, we recover the implicit hierarchy and microspatial economics of site-selection drivers.[4] Our results reveal a distinct structural bifurcation: bulk power capacity and fiber-optic connectivity strictly bound hyperscale cloud deployments, whereas third-party colocation facilities exhibit a pronounced sensitivity to local economic fundamentals—such as labor demographics and real estate costs—reflecting their need to minimize latency for proximate end users.

Finally, we conduct scenario-based counterfactual analyses through 2030, evaluating five shocks against a baseline status quo, including an AI boom, green grid expansion, climate stress, autarkic expansion, and regulatory headwinds. Under an AI boom—a capacity surge that loosens grid constraints while elevating power prices—soaring demand drives robust expansion into secondary markets. Meanwhile, an autarkic expansion scenario—where facilities achieve energy self-sufficiency through on-site generation—unchains developers from traditional technology hubs. This channels investment into remote geographies, triggering a profound spatial realignment that elevates previously unviable areas and erodes the dominance of established markets.

Furthermore, a green grid scenario redirects capital toward renewable zones, conferring a distinct competitive advantage on regions with abundant clean energy. In contrast, climate stress (which amplifies drought severity) and regulatory headwinds (which slash economic incentives and elevate local friction) severely penalize established hubs. These restrictive conditions depress suitability, exerting a moderate to severe drag on regional appeal and deterring future entry.

The physical placement of data centers dictates the allocation of hundreds of billions of dollars in capital, fundamentally reshaping regional economic trajectories and exposing the physical limits of modern grids. Yet, the spatial logic governing this digital foundation remains poorly understood. Our study demonstrates that for digital infrastructure, the strategic role of proximity has not been eliminated but has instead bifurcated. This confirms recent evidence that loca-

[3]The model achieves a mean Area Under the Receiver Operating Characteristic Curve (ROC-AUC) of 0.828 for third-party facilities and 0.781 for cloud deployments. It also generates a top-decile lift factor (Lift@10%) of 4.11x for third-party colocation operators and 3.63x for hyperscale cloud providers.

[4]Specifically, we extract the underlying feature importance by interrogating the network with the Integrated Gradients attribution method (Sundararajan, Taly and Yan, 2017).

tion choices and urban biases are highly conditional on a firm's operational model (Greenstein, Forman and Goldfarb, 2018; Wang, LaRiviere and Kannan, 2019). While third-party colocation facilities rely heavily on proximity to urban demand, hyperscale cloud providers pursue remote cost-minimization strategies. We contribute to a foundational literature examining the evolution of the internet and cloud services (Malecki, 2002; Zook, 2002; Byrne, Corrado and Sichel, 2018; Coyle, Nguyen et al., 2018; Greenstein, 2020, 2021). This scholarship has long debated whether digitization reduces information friction to enable a "death of distance" (Goldfarb and Tucker, 2019) or instead reinforces an "urban bias" that complements existing agglomeration (Sinai and Waldfogel, 2004; Blum and Goldfarb, 2006; Forman, Ghose and Goldfarb, 2009).

This spatial bifurcation highlights a significant departure from traditional theories of industry clustering.[5] While multiunit firms typically organize value chains across space to access specialized human capital and knowledge spillovers (Alcácer and Delgado, 2016), and research on IT adoption heavily focuses on the geography of digital labor (Bloom et al., 2014; Tambe, 2014; Jin and McElheran, 2017; DeStefano, Kneller and Timmis, 2023), data centers decouple physical operations from skilled workers. Instead, their spatial distribution is anchored by traditional physical infrastructure (Fernald, 1999; Leduc and Wilson, 2012; Donaldson, 2018) and institutional capital (Andonov, Kräussl and Rauh, 2021)—specifically, the capacity of the electricity grid. Because these facilities demand massive power loads, their expansion exposes severe transmission bottlenecks and allocative inefficiencies (Davis, Hausman and Rose, 2023; Hausman, 2025). These constraints are further exacerbated by backlogged interconnection queues (Gorman et al., 2025), while deploying energy-intensive operations within constrained grids generates negative spillovers through price shocks and resource crowding (Benetton, Compiani and Morse, 2023).

To untangle these competing spatial dynamics, researchers require methodologies capable of capturing nonlinear location decisions over time. We specifically advance the literature beyond the seminal work of Fang and Greenstein (2025), who provided a foundational analysis of data center geography using static, county-level logit models. In contrast, we develop a dynamic, predictive convolutional neural network (CNN) trained on a high-resolution, 15-year panel (2010–2024). This longitudinal approach maps the evolving micro-geographies of agglomeration, quantifying how infrastructure limitations act as binding frictions on regional growth. By simulating

[5]Parallel evidence from Canada reinforces this bifurcation: Carlo and Rolheiser (2026) document that the average announced Canadian facility is more than ten times the size of the currently active sites, with announced projects sited an average of 175 kilometres from the nearest downtown—compared to 19 kilometres for active facilities—indicating a similar structural shift toward hyperscale developments in remote, infrastructure-rich locations.

high-fidelity counterfactuals, our framework provides the predictive precision necessary to evaluate critical debates in real estate investment and public policy, such as the efficacy of targeted tax incentives and rural development subsidies (CBRE, 2013, 2015).

The remainder of this paper proceeds as follows. Section 2 reviews the institutional background and physical constraints governing data center location choices. Section 3 describes the construction of our building-level sample of data centers and high-resolution spatial panel. Section 4 presents the empirical strategy, including the CNN architecture, the choice-based training framework, and the transfer learning pipeline. Section 5 reports results on baseline suitability surfaces, structural determinants of siting, and counterfactual projections. Section 6 concludes.

2 Background: Data Center Location Choices

Data centers constitute the capital-intensive physical foundation of the digital economy, with the U.S. accounting for over 40% of the global market (McKinsey & Company, 2023; Jones Lang LaSalle, 2026). Contrary to the perception of digital industries as footloose, these facilities face strict physical constraints that historically drove spatial clustering in states such as Virginia and Texas (Electric Power Research Institute, 2024), generating self-reinforcing infrastructure synergies (Blair, 2017). However, recent literature highlights a strategic bifurcation based on operator models (Fang and Greenstein, 2025). While third-party colocation facilities maintain an urban bias to minimize latency for end-users, hyperscale cloud providers increasingly pursue cost minimization in lower-density areas to accommodate massive power and land requirements.

Foundational to both strategies are power density and telecommunications connectivity (Newmark, 2023). Hyperscale facilities operate with extreme energy intensity, requiring 40–60 watts of cooling for every 100 watts of computing (Hancock, 2009), making robust grid access a paramount criterion for site selection. Equally critical is proximity to high-capacity fiber-optic backbones, which ensures the low-latency data transmission required by modern digital workloads. Beyond these physical constraints, developers are highly sensitive to local taxation and state legislation. Because energy and equipment replacements dominate operating expenditures, fiscal policy acts as a decisive competitive advantage; consequently, targeted tax exemptions can substantially accelerate construction rates, particularly for large-scale facilities (Gargano and Giacoletti, 2025).

The extreme capital intensity of these requirements fundamentally alters the sector's financing structure. Driven by surging AI demand, developers mitigate financial risk by shifting from spec-

ulative to build-to-suit development, pre-committing over 80% of under-construction capacity to hyperscalers through long-term pre-leases (Zhang, 2023; CBRE, 2024). For institutional investors, these guaranteed offtake agreements are prerequisites to underwrite the compounding risks of long permitting timelines and congested grid interconnection queues (Gorman et al., 2025). Because facilities are pre-leased and custom-built for anchor tenants, this financing model decouples geographic placement from traditional urban agglomerations, redirecting development toward lower-density regions that satisfy the tenant's massive infrastructure requirements.

Despite the agglomeration benefits of established hubs, mounting sociopolitical and environmental limits increasingly dictate new location choices. The unprecedented scale of modern facilities generates severe negative externalities, including localized electricity price spikes, grid congestion, massive land consumption, and pervasive noise. Furthermore, while historically mitigating power demands (Masanet et al., 2011; Shehabi et al., 2018; Masanet et al., 2020), modern water-based cooling consumes up to 1.7 million gallons daily (Adams, 2013), threatening local resources (Markoff, 2016). The resulting political backlash in historical hubs forces developers toward remote markets with fewer constraints on grid capacity, water availability, and land access.[6]

3 Data and Measurement

Unlike previous studies that rely on county-level proxies, our inputs are preserved at the native spatial resolution of the underlying infrastructure if possible—ranging from census blocks for fiber availability to precise vector paths for power transmission lines. This granularity permits the model to isolate highly localized infrastructural synergies, such as specific land parcels where fiber optic backbones optimally intersect with high-voltage power nodes, which are smoothed out in aggregate statistics. Table 1 details this comprehensive feature space.

3.1 The U.S. Data Center Inventory

Data Source. We construct a high-resolution, building-level inventory of U.S. data centers by synthesizing proprietary data from primary industry aggregators (`datacenters.com`, `datacentermap.com`, and `baxtel.com`), current as of August 2025. Moving beyond traditional aggregate counts, our sample captures individual physical assets across the entire size spectrum, from

[6] Carlo and Rolheiser (2026) document a striking international parallel: as Canada's clean-grid provinces restrict new loads, Alberta captures 93% of planned capacity—averaging over 200 kilometers from major cities—despite hosting only 10% of active facilities and possessing an emissions intensity five times the national average. This confirms that regulatory friction frequently overrides energy-cost fundamentals in shaping digital infrastructure investments.

Table 1—Variable Definitions and Data Sources

Variable & Definition	Resolution	Source
Dependent Variables		
Cloud Provider Entry Indicator (=1).	500m Pixel	Self-Built
Third-Party Entry Indicator (=1)	500m Pixel	Self-Built
Physical Infrastructure		
Grid Topology: Dist. to power lines and substations.	500m Pixel	HIFLD
Generation Capacity: Dist. to + capacity of power plants.	500m Pixel	EIA
Grid Reliability: Power outage duration and frequency.	Utility	EIA
Fiber Availability: FTTP availability.	Block/Tract	FCC
Broadband: Max up/down speeds + provider density.	Tract	FCC
Environmental Constraints		
Water Stress: Baseline water risk metrics.	500m Pixel	WRI Aqueduct
Drought Severity: Palmer Drought Severity Index.	500m Pixel	PRISM
Regulatory & Legislative		
Data Center Legislative Index:	State	NCEL
- *Grid Reliability*: Power stability and outages.	State	–
- *Electricity Costs*: Energy pricing and cost allocation.	State	–
- *Reporting Requirements*: Energy and emissions disclosure.	State	–
- *Environmental Impacts*: Emissions and resource use.	State	–
- *Local Siting Control*: Zoning and permitting authority.	State	–
- *Clean-Energy Mandates*: Renewable energy requirements.	State	–
- *Tax Provisions*: Incentives and tax treatment.	State	–
Cost Structure		
Land Price Proxy: Mid-Tier Home Value Index (ZHVI).	County	Zillow
Real Estate Tax: Median real estate property taxes paid.	County	ACS
Construction Wages: Avg. weekly construction wage.	County	BLS
IT Labor Wages: Median hourly wage for IT occupations.	State	BLS
Electricity Price: Average industrial electricity price.	State	EIA
Sales Tax Rate: Combined state and local sales tax rate.	State	PDIT
Economic Incentives: Project Dev. & Investment Tax.	State	PDIT
Demand Shifters		
Market Size: Total population count.	Tract/County	ACS
Human Capital: % of population with a BA/BS or above.	Tract/County	ACS
IT Agglomeration: % total employment in IT sectors.	County	CBP

Notes: This table lists the spatial raster features used for model training, detailing their spatial resolution and data sources.

small edge nodes (<1k sqft) and mid-sized enterprise centers (1k–10k sqft) to major regional hubs (10k–100k sqft) and massive hyperscale structures (>100k sqft). The dataset provides a rich array of site-specific characteristics, including net technical floor space,[7] total power capacity (MW), rack power density (kW/rack), and grid connectivity strength.[8] It also tracks the specific business

[7]The net technical floor space is the raised-floor area equipped with power and cooling where servers actually live—as opposed to the office space or space for back-end infrastructure such as generators and chillers.

[8]The data distinguishes between medium-voltage (115kV+) and high-voltage (230kV+) transmission lines. We validate utility providers using Energy Information Administration (EIA) Form 861 owner identifiers.

models of each asset, ranging from physical hardware leasing to direct network interconnection.[9]

We classify the market into two segments based on tenancy structure and spatial strategy. The first group comprises 728 hyperscale facilities operated by major cloud providers (e.g., Amazon, Apple, Google/Alphabet, IBM, Meta, Microsoft, and Oracle). These single-tenant sites are highly capital-intensive (often exceeding $1 billion per site) and involve long-term investments (10–15 years). Because they serve global networks rather than local retail markets, they prioritize massive scalability and power efficiency over immediate proximity to end-users (Zhang, 2023).

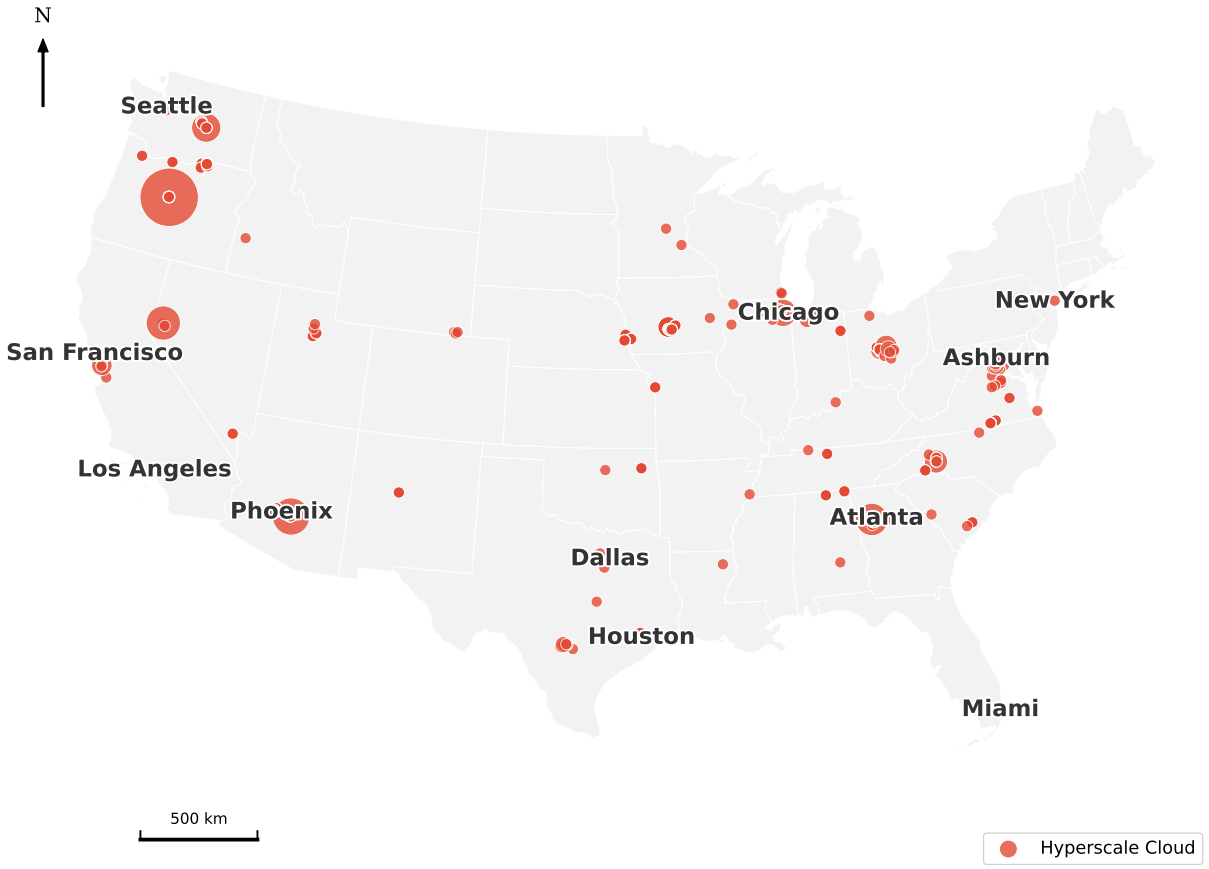
The second group includes 3,719 third-party wholesale and retail colocation data centers. These facilities lease infrastructure to multiple tenants on shorter contracts (1–5 years) and are highly sensitive to local infrastructure externalities and agglomeration economies. Following the taxonomy of Fang and Greenstein (2025), we stratify these sites into three zones reflecting their fundamental trade-offs between urban low-latency requirements and rural land/power availability: "Downtown" sites (< 5 km from the urban core; $N = 436$) optimize for latency reduction and interconnection; "Suburban" sites (5–30 km; $N = 1,829$) balance land costs with connectivity; and "Footloose" sites (> 30 km; $N = 1,454$) capitalize on power availability and land arbitrage.

Geographical Distribution. The spatial distribution of U.S. data centers reveals distinct deployment strategies governed by provider type. Hyperscale cloud providers (Figure 1) exhibit a pattern of high strategic concentration and massive physical scale, clustering within a select few high-density hubs (e.g., Northern Virginia, the San Francisco Bay Area). This clustering reflects the structural economics of hyperscale networks, which prioritize immense power capacity and minimal inter-node latency for machine-to-machine communication. Consequently, the cloud segment ($N = 728$) demonstrates pronounced spatial sparsity; as the truncated red distribution in Figure 2 shows, the vast majority of U.S. counties host no core cloud infrastructure. To reconcile this macro-level sparsity with rapid end-user response times, hyperscalers decouple compute from delivery, relying on diffuse third-party edge nodes for localized content distribution.

In contrast to this centralized approach, third-party colocation facilities (Figure 3) demonstrate a broadly distributed proximity model driven by diverse enterprise demand. While exhibiting heavy-tailed clustering in major hubs—meaning a few select counties host massive concentrations of infrastructure—these sites ($N = 3,719$) are significantly more prevalent across secondary mar-

[9]The dataset offers binary indicators for four distinct service types: (1) Bare Metal Servers (leasing dedicated physical hardware); (2) Retail Colocation (leasing rack space for customer-owned hardware); (3) Internet Exchange (IX) Points (physical locations where networks exchange traffic); and (4) Infrastructure-as-a-Service (IaaS) (virtualized computing resources).

Figure 1—Geographic Distribution of Hyperscale Cloud Data Centers in the U.S.

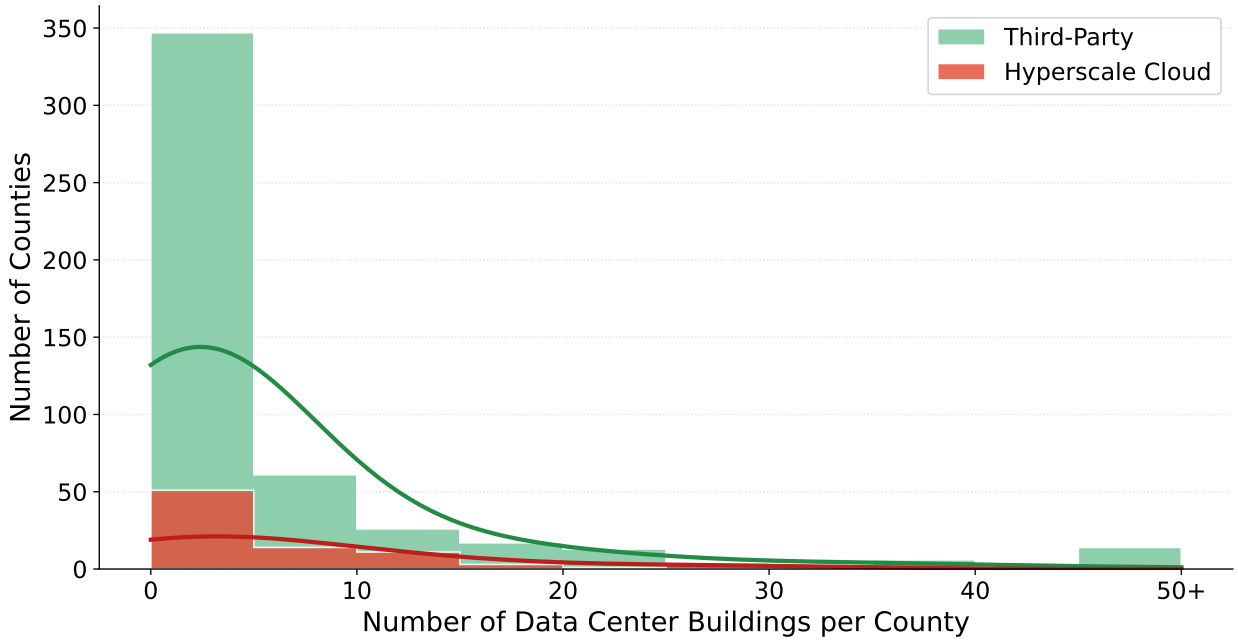


Notes: This figure maps the spatial distribution of hyperscale cloud data centers (e.g., sites owned and operated by Amazon, Apple, Google/Alphabet, IBM, Meta, Microsoft, and Oracle) in the U.S. Each marker represents a single building. Marker size is proportional to the building's gross square footage, reflecting physical asset scale at the building level.

kets. The green histogram in Figure 2 underscores this localized agglomeration, with a significant number of counties hosting 50 or more distinct buildings. Following the taxonomy of Fang and Greenstein (2025), our spatial stratification reveals three archetypes: dense "Downtown" facilities anchored within Central Business Districts (< 5 km from the urban core; $N = 436$), large-scale "Suburban" hubs (5–30 km; $N = 1,829$), and rural "Footloose" sites (> 30 km; $N = 1,454$). This heterogeneous footprint highlights the fundamental tension between the high costs of urban proximity and the physical constraints of data center scale.

Ultimately, this spatial stratification reveals a sharp divergence in optimization strategies. While hyperscale networks can isolate operations to prioritize macro-level power capacity, third-

Figure 2—Spatial Agglomeration versus Geographic Sparsity in U.S. Data Center Networks



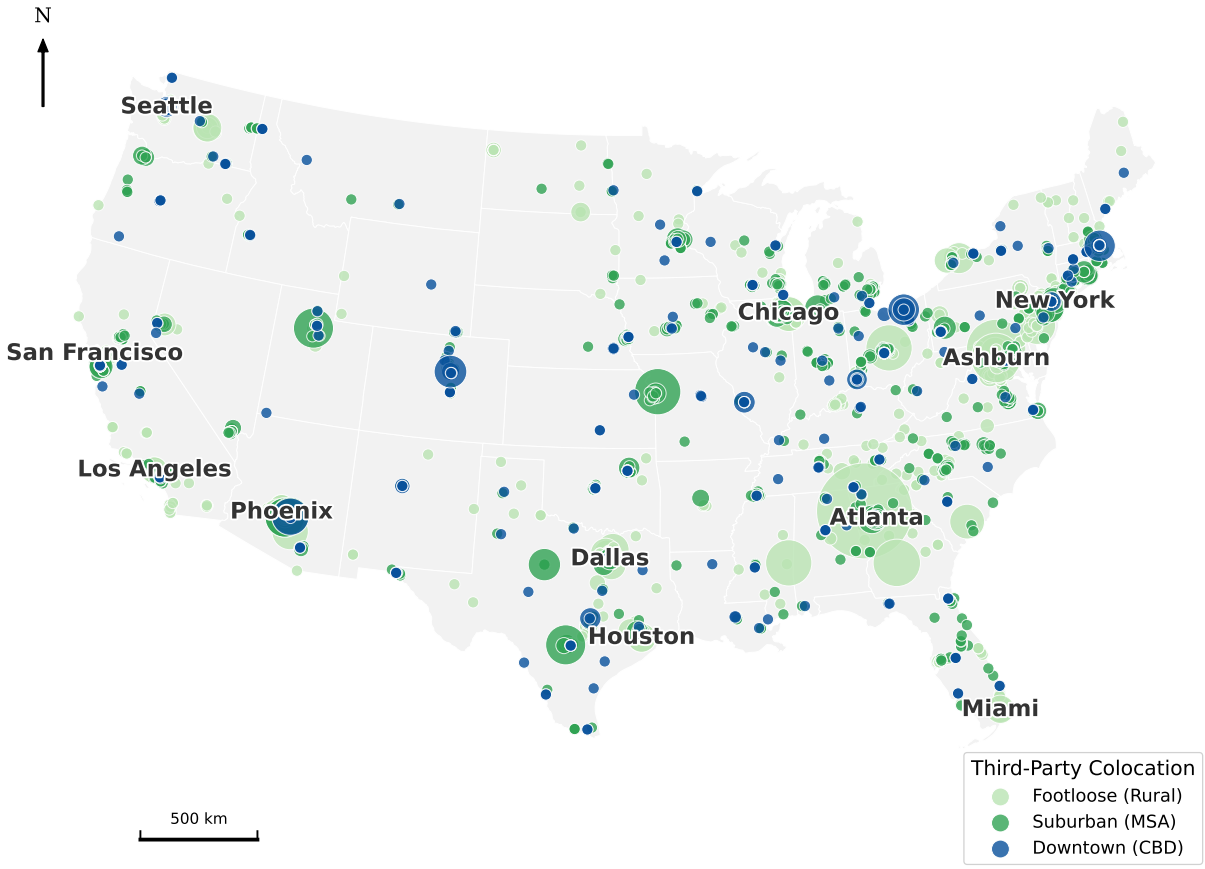
Notes: This figure presents histograms showing the concentration of data center buildings at the county level. The horizontal axis represents the number of distinct buildings located within a single county, while the vertical axis represents the number of counties exhibiting that specific building count. The red bars and curve display the distribution for cloud providers ($N = 728$), defined as single-tenant hyperscale facilities. The green bars and curve display the distribution for third-party data centers ($N = 3,719$), defined as colocation and wholesale facilities leasing space to multiple tenants. The distribution is clipped at 50 buildings to accommodate extreme right-skewness.

party operators rely heavily on physical proximity to fiber exchanges and carrier hotels.[10] This reliance creates localized network effects where location value scales with tenant density, imposing a complex objective function that balances regional infrastructure reliability, enterprise demand, and escalating urban land costs.

Industry Growth. To precisely measure site entry timing, we determine baseline operational years using industry databases and web archives. We systematically validate these dates and physical footprints against historical Landsat and Sentinel-2 satellite imagery (Appendix Figures A.1-A.3). We restrict our analysis to new facilities entering the market between 2010 and 2024. This period aligns with the widespread availability of high-resolution satellite data and isolates the industry's structural shift toward the hyperscale era—a period when location choices are dictated

[10]Carrier hotels are centralized telecommunication hubs where distinct networks physically converge.

Figure 3—Geographic Distribution of Third-Party Data Centers in the U.S.

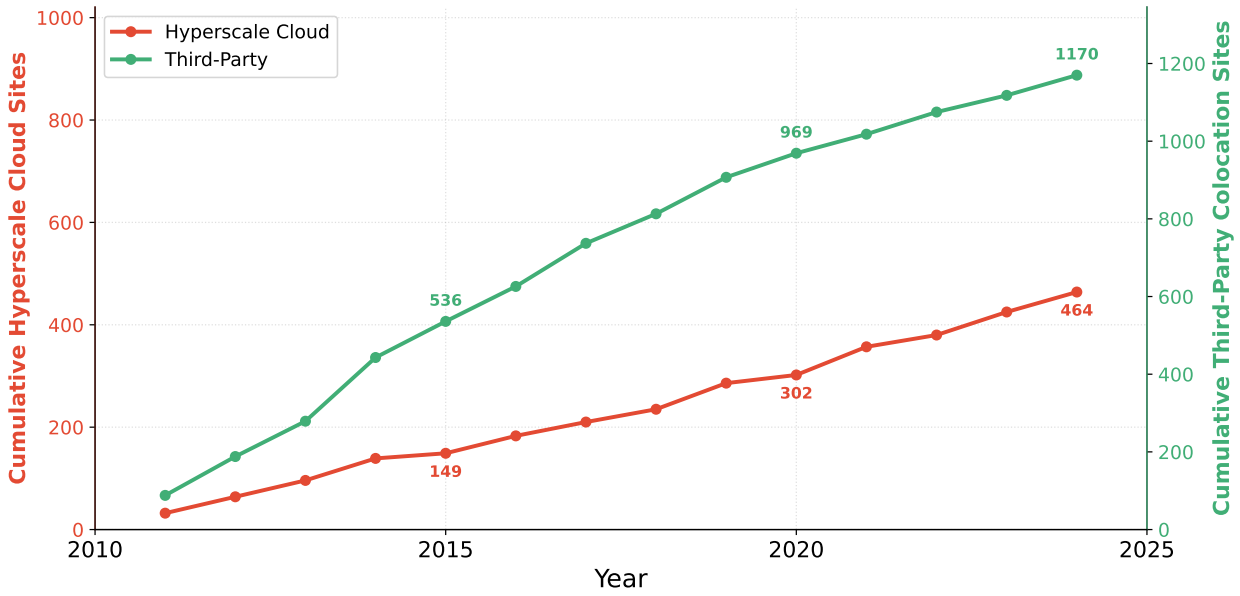


Notes: This figure maps the spatial distribution of third-party data centers (wholesale and retail colocation facilities) in the United States. Each marker represents a single building. Marker size is proportional to the building's gross square footage, reflecting physical asset scale at the building level. Third-party buildings are further stratified by spatial location into "Downtown" (within CBDs), "Suburban" (within MSAs but outside urban cores), and "Footloose" (rural or non-metropolitan areas).

by power, land, and cooling constraints rather than urban proximity.

Figure 4 illustrates the divergent expansion trajectories of these market segments. Hyperscale cloud providers have followed a measured growth path, reaching 464 sites by 2024, reflecting the massive physical scale and capital intensity of single-tenant facilities. In contrast, the third-party colocation segment exhibits rapid, sustained growth, reaching 1,170 sites. This volume gap highlights a fundamental industry dynamic: while hyperscalers dominate aggregate computing capacity and media attention, the broader physical architecture of the cloud relies heavily on a distributed network of multi-tenant facilities.

Figure 4—Divergent Growth Trajectories of Cloud and Third-Party Data Centers, 2011–2024



Notes: This figure displays the cumulative count of new data center sites established in the U.S. between 2011 and 2024. The red line (left vertical axis) tracks the entry of cloud providers, representing facilities owned or operated by hyperscalers such as Amazon, Google, and Microsoft. The green line (right vertical axis) tracks the entry of third-party data centers, including colocation and wholesale facilities operated by firms such as Equinix and Digital Realty. The horizontal axis indicates the year of initial site establishment, while the vertical axes represent the cumulative number of new operational sites by the end of each year.

3.2 Spatial Determinants of Location Choice

To characterize the local determinants of data center entry, we construct a high-resolution spatial panel spanning the period 2010–2024. We map our primary covariates at a 500-meter grid resolution. This granular architecture functions as a multi-layered digital map, allowing us to capture precise localized infrastructure constraints and directly compare realized data center locations against unchosen alternatives within regional power markets. Our spatial panel integrates five critical dimensions: power infrastructure, water resources, broadband connectivity, economic fundamentals, and regulatory environments. Detailed variable construction and data sourcing are provided in Appendix Section A.2.

First, we model power infrastructure by calculating proximity to high-voltage transmission lines and substations at the 500-meter pixel level using Homeland Infrastructure Foundation-Level Data (HIFLD). This is crucial because localized grid bottlenecks constrain development more heavily than broad market trends (Mamkhezri, Sun and Yang, 2025). We supplement this

with local generation capacity from EIA-860 (mapped to the 500-meter grid) and utility-level reliability metrics from EIA-861, i.e., the System Average Interruption Duration Index (SAIDI) and the System Average Interruption Frequency Index (SAIFI) .

Second, water resources and environmental constraints are quantified using baseline water risk from the World Resources Institute (WRI) Aqueduct 4.0 Water Risk Atlas and dynamic drought intensity via the Palmer Drought Severity Index (PDSI) from the PRISM climate model—both scaled to the 500-meter grid resolution.

Third, for broadband connectivity, we isolate terrestrial Fiber-to-the-Premises (FTTP) technology at the census-block level and maximum advertised speeds at the census-tract level using Federal Communications Commission (FCC) Form 477 data, establishing a highly specific proxy for network capacity.

Fourth, we capture agglomeration economies and labor market depth through tract- and county-level population, education, and IT employment data. We proxy fixed setup and variable operating costs using county-level Zillow Home Value Index (ZHVI) (Qu et al., 2025), county-level average weekly wage in the construction sector from the Bureau of Labor Statistics (BLS), county-level employment share in IT sectors from the BLS, and state-level EIA electricity prices.

Fifth, we evaluate pro-investment drivers separately through our tax incentive variables. Fiscal push-and-pull factors are modeled using state-level sales taxes, county-level property taxes (Giroud and Rauh, 2019), and state-level targeted tax incentives from the Panel Database on Incentives and Taxes (PDIT) (CBRE, 2013; Gargano and Giacoletti, 2025).

Finally, to capture regulatory frictions, we construct a regulatory index using dictionary-based content analysis (DCA) that categorizes text based on predefined keywords. We process 76 state data-center-related bills sourced from the National Caucus of Environmental Legislators (NCEL), classifying them into seven policy domains: grid reliability, electricity costs, reporting requirements, environmental impacts, local siting control, clean-energy mandates, and tax provisions. To precisely quantify regulatory pressure, we weight each bill by its legislative status (e.g., enacted vs. proposed) and enforcement intensity (e.g., binding mandates vs. exploratory commissions). These weighted values are aggregated and normalized to a 0-to-10 scale.

4 Empirical Strategy

4.1 Modeling Spatial Entry

Motivation: Emulating Developer Due Diligence. Traditional location choice models typically forecast entry by heavily weighting historical agglomeration—effectively assuming that future development will closely mirror established clusters. In contrast, data center developers underwrite risks and returns over 5- to 10-year horizons based on the evolving trajectory of a site's intrinsic capacity. We operationalize this forward-looking perspective using a multi-channel 2D Convolutional Neural Network (CNN) designed to replicate the evaluation process of a field surveyor assessing a prospective site. While our framework inherently relies on historical data, it does not extrapolate past entry. Instead, it processes a short temporal stack of exogenous fundamentals to capture the velocity and acceleration of local infrastructure.

Just as a surveyor walks a parcel and systematically inventories the surrounding utility ecosystem—checking proximity to substations, fiber access points, and water sources—our network scans each spatial neighborhood through overlapping convolutional kernels that detect localized resource co-occurrences. The broader catchment area that a surveyor evaluates when assessing regional connectivity corresponds to the network's effective receptive field—the cumulative geographic extent that influences the model's final prediction. By optimizing this mapping, the CNN model reverse-engineers the sophisticated underwriting criteria that surveyors and developers apply in practice, valuing locations on their latent structural potential rather than their historical path dependence.

Model Specification and Architecture. We model the probability of entry Y_{it}—a binary indicator equal to one if a new data center is established at location i in year t—as a nonlinear function of historical infrastructure endowments:

$$P(Y_{it} = 1) = \sigma(f(\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \mathbf{X}_{i,t-3}; \Theta))$$

where f represents the deep neural network mapping a three-year history of spatially varying covariates ($\mathbf{X}_{i,t-k}$) to the selection decision, and Θ denotes the learned parameters. To preclude look-ahead bias, we impose a rigorous temporal firewall, conditioning predictions exclusively on covariates observed during years $t-3$ through $t-1$. This stacked data structure (the input tensor) captures both the static level of infrastructure capacity and its growth momentum. We evaluate

each location within a 64×64 pixel window, providing a 32 km $\times$ 32 km geographic scope (spatial receptive field) that captures local labor and utility markets while masking existing facilities to ensure the model relies on structural fundamentals.

The model processes these data through a three-stage automated pattern-matcher (a convolutional backbone). In each stage, sliding filters (kernels) scan the maps to detect localized geographic signatures, such as the intersection of high-voltage lines and fiber backbones. A standardization step (batch normalization) and a threshold function (ReLU activation) stabilize learning and capture the nonlinear nature of infrastructure: a site either meets the discrete threshold for power access or it does not. Subsequent signal distillation (max pooling) retains only the strongest geographic markers, discarding redundant detail. This architecture mimics a developer's hierarchical screening: as the focus narrows from the full 32-km grid to a 4-km neighborhood, the complexity of tracked criteria expands from 32 to 128 distinct patterns.

In the final stage (the classifier head), the model flattens the resulting 8,192 spatial signals into a single vector. This vector is passed through a weighted scoring function—a fully connected layer with 256 units (neurons)—that functions as a standardized underwriting rubric. To ensure the model learns universal principles rather than memorizing training sites (overfitting), we implement a 50% stress-test (dropout), randomly deactivating half the scoring units during training to force the development of robust decision rules. The final output is a raw score (logit), which a standard logistic curve (sigmoid function) transforms into a calibrated suitability probability between zero and one. All model parameters are initialized using a variance-preserving scheme (Kaiming Normal initialization) to maintain stable numerical gradients.

4.2 Constructing the Spatial Panel

To construct the sample for network training, we first standardize all heterogeneous inputs onto a unified 500×500-meter raster grid using the Albers Equal Area Conic projection (EPSG:5070) to ensure the physical size of each cell remains geographically accurate (preserve area fidelity). This high-resolution continuous surface mitigates the modifiable areal unit problem (MAUP)—a spatial aggregation bias where statistical results artificially change depending on how boundary lines are drawn. By relying on a uniform grid rather than arbitrary administrative borders like counties, the model avoids averaging out highly localized data. This enables the detection of micro-geographic advantages that are often obscured within broad regional statistics.

Positive entry events ($Y_{it} = 1$) are identified using our geocoded inventory of 4,447 facili-

ties, with each building's entry year t dictating its temporal assignment on this master grid. A single observation is constructed as a spatially organized data snapshot for a specific year—a multi-channel image patch—measuring 64×64 pixels. At our 500-meter resolution, this patch represents a 32 km $\times$ 32 km catchment area (spatial receptive field). By presenting the model with this broad local neighborhood, the algorithm can evaluate the precise geographic intersection of necessary resources, mirroring the spatial due diligence performed by real estate developers. Unlike traditional cross-sectional analyses that collapse temporal variation into long-term averages, this approach preserves the longitudinal structure of entry events.

For a target entry year t, we construct the model input (the input tensor) by stacking three consecutive years of geographic maps (feature channels) spanning years $t-3$ through $t-1$. This retrospective window captures critical growth momentum: the network observes not only the most recent infrastructure levels ($X_{i,t-1}$), but also their rate of change (velocity, $X_{i,t-1} - X_{i,t-2}$), and their acceleration ($[X_{i,t-1} - X_{i,t-2}] - [X_{i,t-2} - X_{i,t-3}]$). This temporal depth allows the model to distinguish between sites with expanding capacity and those that have stagnated. We restrict this window to three years to avoid the decay of obsolete information and minimize early-period data loss (left-censoring of the panel).

These geographic maps are systematically layered in an order to reflect our core determinants. Each temporal slice begins with *Power Infrastructure*, comprising nine layers: grid topology (3 channels) and generation capacity (6 channels). Next, *Water Resources* are captured through 2 channels measuring water risk and drought severity, followed by *Connectivity*, documented via 6 channels of broadband deployment. The remaining 27 channels encompass the economic drivers of the market: *Labor & Demographics*, *Real Estate & Land Costs*, *Tax Incentives*, and *Regulation* (quantified via our state-level legislative index).

With a three-year retrospective horizon, the complete multi-dimensional data stack for a single location contains $44 \times 3 = 132$ layers. To ensure that features measured in disparate units—such as megawatts, bits per second, or local currency—are evaluated on a mathematically equivalent basis, we standardize each patch via channel-wise z-score normalization. This procedure rescales all variables to a common distribution, preventing features with larger raw magnitudes from dominating the network's internal scoring logic. Consequently, the model remains attuned to the subtle structural trade-offs that dictate site selection.

4.3 Network Training and Validation

Guardrails Against Data Leakage. Before training the model, we must address the spatial dependence inherent in phased construction. Treating sequential developments as independent location choices introduces severe data leakage—a flaw where the algorithm memorizes the shared geographic traits of co-located buildings during training, rather than learning generalizable siting rules. To account for this, we define a "campus" as the aggregation of all buildings owned by a single firm within a given county. This reflects a fundamental economic distinction: the initial site selection represents a discrete new investment decision (the extensive margin of entry), whereas subsequent buildings typically reflect capacity expansions within a pre-secured footprint.

To enforce this distinction, we partition the data into five geographically distinct testing subsets (Group K-Fold cross-validation) using campus identifiers as blocking groups. This procedure ensures that all facilities belonging to a single campus are assigned exclusively to either the training set or the testing set, never splitting across both. As a result, each training cycle utilizes four regions (80% of the data) to learn patterns and reserves the fifth region (20%) to validate accuracy on campuses it has never encountered. This rigorous spatial separation forces the neural network to look past familiar clusters and rely on latent structural determinants, ensuring that its predictive accuracy reflects genuine out-of-sample performance in entirely new markets.

Beyond sample partitioning, we must also prevent feature-level data leakage—inadvertently allowing the algorithm to base its predictions on the locations of existing facilities rather than underlying economic drivers. To achieve this, the model never observes realized entry locations. All spatial layers related to existing data center presence or proximity are structurally excluded from the multi-dimensional data matrix fed into the algorithm (the input tensor). This ensures that predicted probabilities derive solely from exogenous infrastructure and economic fundamentals. The resulting geographic grid (raster) constitutes a counterfactual baseline—a probability surface reflecting where the industry could theoretically expand based on intrinsic site characteristics alone, entirely independent of historical path dependence or established industry clusters.

Choice-Based Training and Hard-Negative Mining. Having established this spatial partitioning, Table 2 reports the class distribution across our training and validation splits, highlighting the key econometric challenge of modeling data center entry: extreme spatial sparsity.[11] Within our

[11] The training sample (1,286 third-party and 478 cloud sites) is substantially smaller than the full inventory of 4,447 facilities for three reasons. First, by restricting to new entry events from 2010 onward, we exclude approximately 2,800

geographic grid, locations hosting a third-party data center constitute only 1.01% of potential sites, while cloud deployments represent a mere 0.07%.[12] A standard yes-or-no classification algorithm would achieve a misleadingly high accuracy of 99.93% simply by predicting that development never occurs, completely failing to recover the latent profitability signals driving site selection.

Table 2—Class Distribution (Sample Counts)

Dataset Split	Positive Anchors (Total Choice Sets)	Hard Negative Counterfactuals	Available Negative Pool
Third-Party Providers			
Training (80%)	1,028	9,252	60,000
Validation (20%)	258	2,322	15,000
Total	**1,286**	**11,574**	**75,000**
Cloud Providers			
Training (80%)	382	3,438	60,000
Validation (20%)	96	864	15,000
Total	**478**	**4,302**	**75,000**

Notes: This table reports the distribution of positive empirical anchors, the constructed hard negative counterfactuals, and the available synthetic negative pool across the 5-fold spatial cross-validation splits. Because data center entry is a rare event, the number of realized sites (positive anchors) is remarkably small relative to the vast number of undeveloped land parcels (available negative pool). During model training, each positive anchor serves as the center of a single choice set. To dynamically construct these sets, the algorithm pairs each realized site with $K = 9$ hard negative counterfactuals—plausible alternative sites where no facility was built. These 9 hard negatives are randomly drawn from the available negative pool within a 50-mile radius of the anchor, strictly constrained to the same ISO/RTO power market. Therefore, every choice set evaluated by the model contains exactly 10 parcels (1 realized winner and 9 undeveloped alternatives). In our spatial framework, each of these "sites" or "parcels" corresponds to the central 500-meter grid cell of a given local catchment area.

To resolve this severe class imbalance, we formulate site selection as a conditional choice problem. Rather than evaluating parcels in isolation against an absolute standard of viability, we adopt a choice-based ranking formulation that mirrors the actual decision problem facing developers: weighing a menu of competing options within a regional market to identify the most advantageous site. As outlined in Table 2, we dynamically construct training choice sets comprising 10 locations: one actual data center site (the positive anchor) and nine hypothetical alternatives where a facility could have been built but was not (negative counterfactuals).

pre-existing facilities built before our sample period. Second, the model's three-year covariate lag ($t-3$ through $t-1$) requires complete historical data for each entry year, which may exclude a small number of early-period sites with insufficient covariate coverage. Third, the remote sensing validation (Appendix Figures A.1–A.3) flagged 63 records that could not be linked to a specific building due to unclear boundaries or absent identifiable features.

[12] In our spatial framework, each of these "sites" or "parcels" corresponds to the central 500-meter grid cell of a given local catchment area.

To ensure the network learns the subtle infrastructure differences that drive investment rather than relying on obvious macro-geographic disparities, we employ a targeted sampling strategy. The nine alternative sites are drawn from a 50-mile radius surrounding the realized site. This radius approximates a local commuting zone, holding macroeconomic factors constant while forcing the model to distinguish a chosen parcel from its immediate, undeveloped neighbors—a technique known as mining hard negatives. Crucially, we restrict all counterfactuals to the same Independent System Operator (ISO) or Regional Transmission Organization (RTO) power market and explicitly exclude undeveloped parcels that belong to the same existing data center campus.

Finally, to expand the effective sample size and ensure the model does not over-rely on specific spatial layouts (rotational invariance), we apply random 90-degree rotations and horizontal flips to the input maps during training. The network evaluates each choice set simultaneously, ranking the realized investment site against its local counterfactuals. We optimize this discrete choice using a cross-entropy loss function—a deep learning metric functionally equivalent to maximizing a parcel's latent utility (overall suitability for development) in a standard conditional logit model. Consequently, the network is forced to weigh the relative value of competing geographic features. This formulation directly optimizes the algorithm's ability to rank plausible alternatives within a competitive market, effectively mirroring a developer's site-selection due diligence.

Transfer Learning across Market Segments. While choice-based training resolves spatial sparsity, modeling hyperscale cloud entry presents an additional econometric challenge: extreme scarcity of positive observations. With only a few hundred realized hyperscale cloud sites across the country, standard deep learning models would likely memorize the specific idiosyncrasies of these few sites rather than learning generalizable rules of entry (overfitting). We resolve this "Small N" problem by adopting a sequential two-stage training pipeline with transfer learning.

In the first stage (the Teacher model), we train the full network on the substantially larger third-party colocation dataset to learn the general "physics" of infrastructure suitability. In the second stage (the Student model), we preserve the foundational geographic rules established by the Teacher. We achieve this by locking the network's core feature-extraction layers, a process computationally known as freezing the convolutional backbone. With this baseline spatial knowledge secured, we adjust only the model's final evaluation scorecard—fine-tuning the classifier head—using the sparse hyperscale cloud data.

Because the complex spatial rules are locked, the model effectively runs a standard regression

on the pre-extracted geographic features (a technique known as linear probing). This strategy relies on the core economic assumption that all data centers, regardless of their business model, depend on the same fundamental infrastructure constraints, such as access to high-voltage grids and robust fiber networks. This sequential approach allows the network to leverage universal site fundamentals while tailoring its final predictive weights to the distinct scale and risk preferences of the hyperscale market segment.

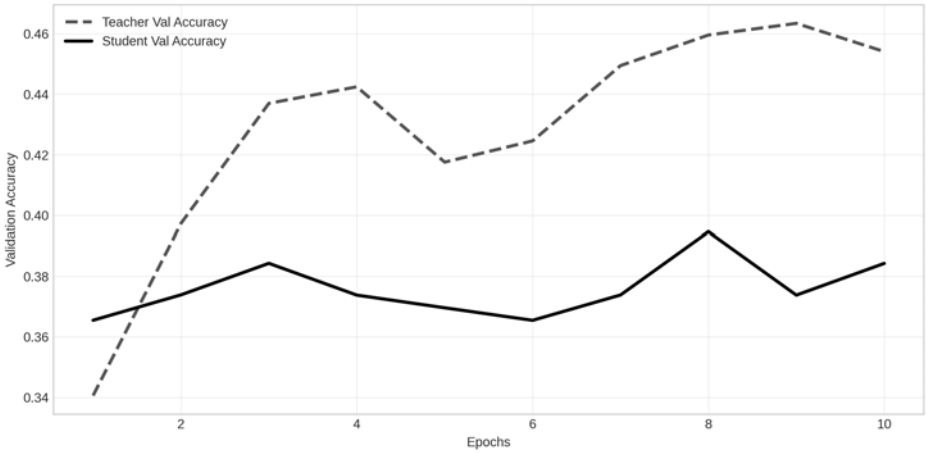
Predictive Performance. Panel (a) of Figure 5 tracks out-of-sample accuracy across successive passes through the training data (epochs). In our choice-based framework, this metric reports top-1 ranking accuracy—the frequency with which the model selects the realized data center as its first choice from a competitive set of 10 plausible parcels. Although top-1 accuracy settles near 46%, this represents a 4.6$\times$ improvement over the 10% base rate implied by random selection among ten candidates. The steady plateauing of the learning curves further confirms that the networks converge on generalizable structural rules rather than memorizing the geographic coordinates of specific training sites (overfitting).

Panel (b) of Figure 5 presents Receiver Operating Characteristic (ROC) curves that summarize performance across the pooled national dataset. Each curve plots the trade-off between correctly identified sites and false alarms as the model's decision threshold varies; the area beneath the curve (AUC) represents the probability of correctly ranking a true site above a random non-site across the entire country. The model demonstrates substantial predictive power, effectively distinguishing suitable infrastructure environments from the vast majority of unsuitable terrain.

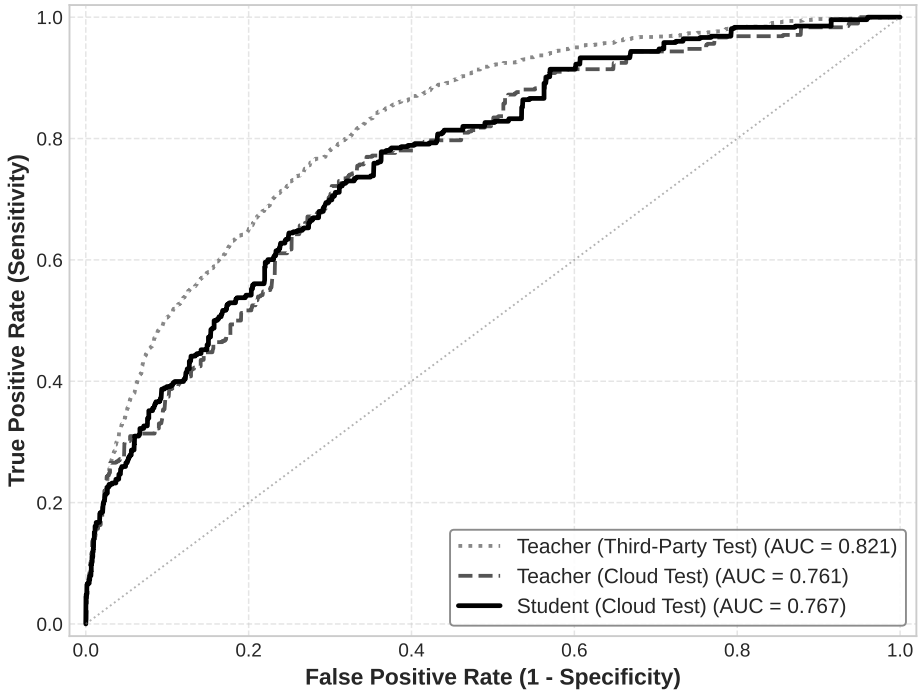
Because data center development is an exceptionally rare event—occurring in less than 1% of land parcels—standard accuracy metrics are misleadingly optimistic. For instance, an algorithm that blindly predicts "no development" for every parcel would artificially achieve 99% accuracy without identifying a single viable site. To address this extreme spatial sparsity, Table A.1 reports five rigorous evaluation metrics tailored to rare-event modeling: the Area Under the Receiver Operating Characteristic Curve (ROC AUC), the Area Under the Precision-Recall Curve (PR AUC), Lift in the top decile (Lift@10%), Recall in the top percentile (Recall@1%), and the Brier score.

Our primary evaluation focuses on the ROC AUC, which measures the model's overall ranking accuracy. In practical terms, this metric represents the probability that the algorithm correctly assigns a higher suitability score to an actual data center site than to a randomly chosen undeveloped parcel. Detailed definitions and economic interpretations of the remaining four metrics—

Figure 5—Predictive Performance and Training Dynamics



(a) Training Dynamics and Model Convergence



(b) Out-of-Sample ROC Curves

Notes: This figure evaluates the out-of-sample predictive performance of the spatial choice model. Panel (a) plots accuracy across training iterations (epochs) to confirm the algorithm extracts generalizable economic rules rather than memorizing specific geographic coordinates (overfitting). Accuracy here denotes the Top-1 ranking success rate—the frequency the model correctly identifies the realized data center from a choice set of 10 parcels (against a 10% random baseline). Panel (b) uses ROC curves to summarize global performance. The AUC represents the probability of correctly ranking an actual development site above a random undeveloped parcel across the national dataset.

which evaluate the reliability of highly confident predictions, screening efficiency, top-tier candidate identification, and model calibration—are provided in Appendix Section A.4.

Furthermore, to ensure the model learns generalizable economic rules rather than memorizing local idiosyncrasies, we evaluate these metrics across five non-overlapping geographic holdout regions. This protocol, computationally known as 5-fold spatial cross-validation, evaluates each fold independently before averaging the results. Consequently, the approach ensures that the metrics capture how effectively the model ranks competing parcels within the same local market, avoiding the trap of trivially distinguishing between obvious macro-regional differences.

Table A.1 additionally presents a cross-domain evaluation that directly tests our core hypothesis: that hyperscale cloud providers and third-party colocation operators, despite pursuing different business strategies, are constrained by the same underlying physical infrastructure. We formalize this shared dependence as a structural prior—the assumption that spatial representations learned from one market segment transfer to the other because both segments respond to the same grid, fiber, and water fundamentals.

Cross-Domain Evaluation. The Teacher model (trained and tested on third-party sites) establishes a strong baseline, achieving a mean cross-validated ROC AUC of 0.828, a Lift of $4.1\times$, and a Recall@1% of 0.068.[13] The Lift metric shows that, while random search has a near-zero likelihood of identifying a viable site, the model's top-decile predictions are more than four times as likely to contain a realized facility. Moreover, the Recall@1% metric indicates that the model isolates nearly 7% of all actual data centers into its exclusive top 1% of the recommended parcels. Together, these results suggest that the model has effectively learned the local fundamental constraints governing data center siting.

To test whether these rules generalize across market segments, we applied the Teacher model directly to the cloud sample without retraining (Teacher Structural). While the model maintained a mean ROC AUC of 0.766, its Lift dropped to $3.5\times$ and its Recall@1% fell to 0.048. This performance decay—visually apparent in the downward shift of the ROC curve in Panel (b) of Figure 5—confirms that, although third-party and hyperscale facilities share foundational infrastructure requirements (such as grid capacity and fiber topology), their distinct business models impose divergent spatial preferences that a single rigid set of scoring weights cannot fully capture.

[13]The cross-validated AUC reported in the table reflects within-market accuracy, which slightly exceeds the pooled national AUC (0.821) shown in Figure 5b. This distinction is economically meaningful: although absolute infrastructure scores vary between macro-regions, the model ranks competing parcels highly effectively within the same local market.

The Student model resolves this divergence through transfer learning—a technique that applies knowledge gained from a data-rich domain to a data-scarce one. Specifically, the model locks in the universal geographic rules already learned by the Teacher (freezing the convolutional backbone) and adjusts only its final evaluation scorecard to fit the sparse hyperscale data (fine-tuning the classifier head). This strategy recovers much of the lost performance, achieving a mean ROC AUC of 0.781, a Lift of $3.6\times$, and a Recall@1% of 0.042. The restored trajectory visible in Panel (b) of Figure 5 confirms that this approach enhances discrimination at the extremes without sacrificing the generalized infrastructure knowledge embedded in the shared architecture.

5 Results

5.1 Baseline Suitability Surfaces

To map national investment potential, we apply our trained algorithm across the contiguous U.S. By averaging predictions across five independently trained versions of the model (a five-fold ensemble), we generate highly reliable, continuous suitability surfaces for both hyperscale cloud and third-party data centers. We specifically target data center entry in 2024, relying exclusively on covariates observed from 2021 through 2023.

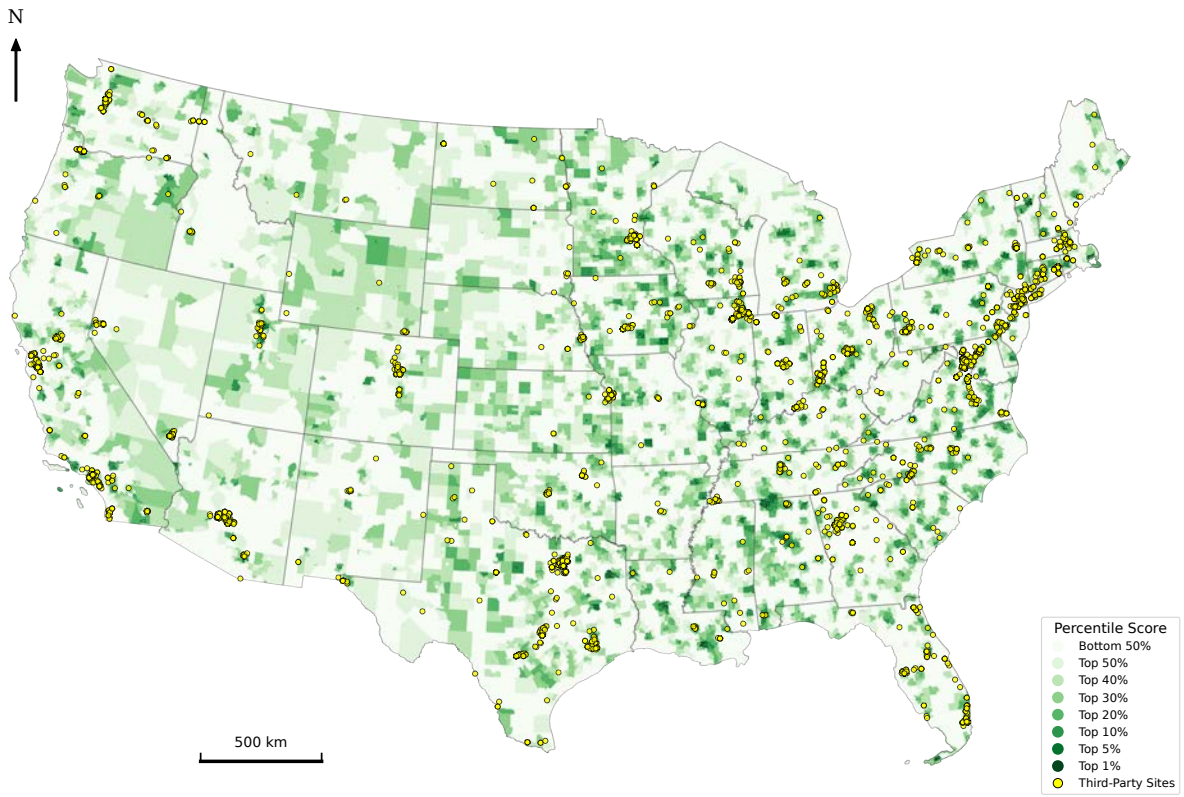
While predicting a future year would provide a pure forward-looking forecast, targeting 2024 serves two critical empirical purposes. First, it allows for rigorous out-of-sample validation; by predicting 2024, we can overlay realized entry events (ground truth) onto the predicted suitability surface to verify the model's accuracy. Second, it accounts for standard institutional reporting lags in critical infrastructure covariates (typically 6 to 12 months), such as broadband deployments and utility generation, ensuring the algorithm evaluates a complete and finalized data stack.

To translate these pixel-level predictions into policy-relevant administrative units, we compute zonal statistics—the mean predicted probability—within each census tract, which captures the full density of suitable infrastructure within each administrative unit and assigns meaningful scores to large rural tracts that would appear empty in a point-based analysis. To ensure geographic consistency, all spatial aggregations utilize constant 2010 cartographic boundaries.

National Suitability and Market Divergence. Figure 6 maps the predicted investment surface for third-party colocation facilities at the tract level, classified into national percentile bins. While the model correctly identifies peak suitability in established Tier 1 infrastructure hubs (such as

Northern Virginia, Dallas, and Silicon Valley), the overall spatial pattern is notably dispersed. Elevated suitability extends well into secondary markets—such as Denver, Minneapolis, and Nashville—and similar Tier 2 metropolitan areas. This geographic breadth reflects the colocation sector's strategic imperative to locate in proximity to end users across diverse regional economies, minimizing latency for local enterprise clients.

Figure 6—Predicted Suitability Surface: Third-Party Colocation Facilities

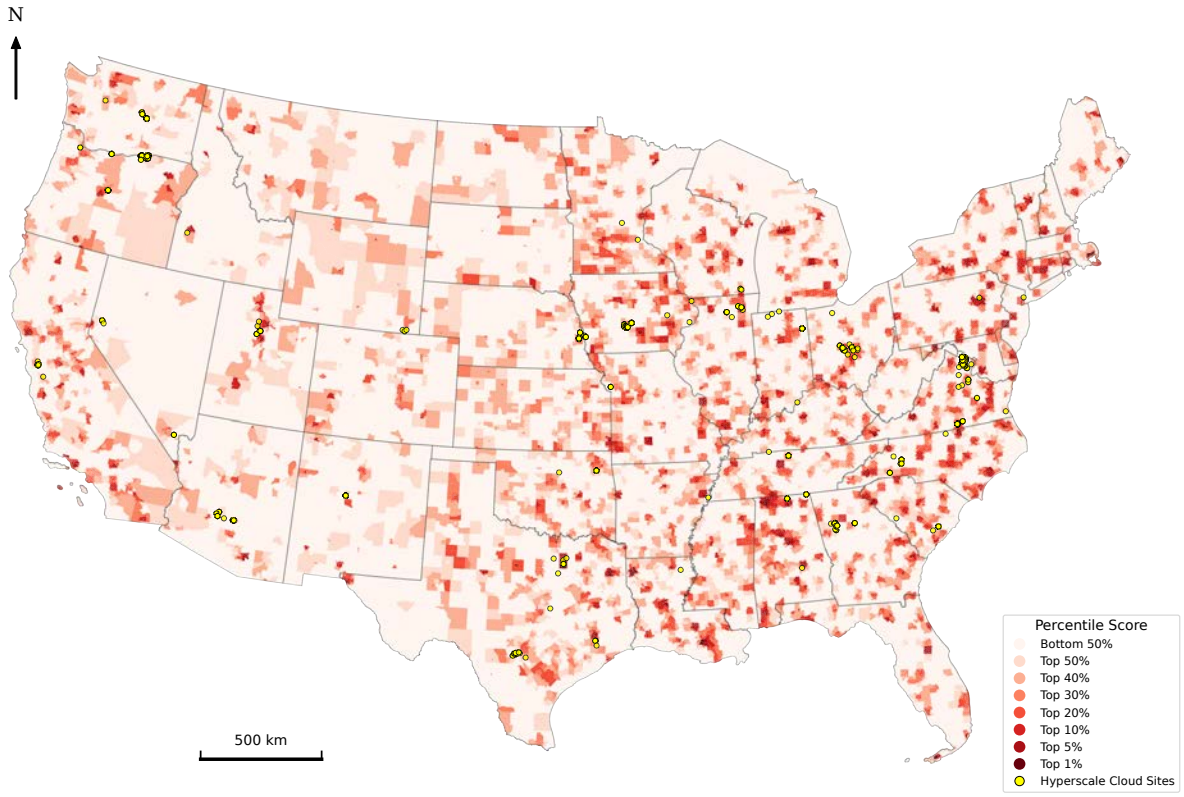


Notes: This map displays the baseline predicted infrastructure suitability for third-party colocation data centers, aggregated to 2010 census tracts. We specifically target data center entry in 2024, relying exclusively on covariates observed from 2021 through 2023. Values represent the mean pixel-level probability within each tract, derived from a five-fold CNN ensemble trained on 132 feature channels spanning power infrastructure, fiber connectivity, water resources, economic fundamentals, tax incentives, and regulatory stringency. Tracts are classified into percentile bins (50th through 99th); darker shading indicates higher predicted suitability. Yellow markers denote realized third-party data center sites.

By contrast, Figure 7 visualizes the investment potential for hyperscale cloud infrastructure at the tract level, effectively capturing a strategy characterized by sparse, high-density clustering. The CNN correctly identifies the industry's macro-level geography: beyond the well-established "Data Center Alley" in Northern Virginia and hydroelectric-rich regions in the Pacific Northwest

(e.g., Prineville, Oregon; Quincy, Washington), the algorithm's predictive surface highlights key interconnection hubs in Dallas and Chicago, alongside rapidly expanding footprints in Central Ohio and the Midwest.

Figure 7—Predicted Suitability Surface: Hyperscale Cloud Infrastructure



Notes: This map displays the baseline predicted infrastructure suitability for hyperscale cloud data centers, aggregated to 2010 census tracts. We specifically target data center entry in 2024, relying exclusively on covariates observed from 2021 through 2023. Values represent the mean pixel-level probability within each tract, derived from a five-fold CNN ensemble trained on 132 feature channels spanning power infrastructure, fiber connectivity, water resources, economic fundamentals, tax incentives, and regulatory stringency. Tracts are classified into percentile bins (50th through 99th); darker shading indicates higher predicted suitability. Yellow markers denote realized hyperscale cloud data center sites.

Notably, the predicted spatial pattern exhibits a distinct "halo effect" around major metropolitan areas; top-percentile counties frequently appear on the suburban outskirts rather than in dense urban cores, reflecting the hyperscale preference for massive land parcels adjacent to high-voltage power transmission. Furthermore, isolated rural hotspots successfully align with high-capacity renewable energy zones, while the neural network accurately identifies vast "empty zones" in the Rockies and Great Plains where foundational connectivity remains insufficient for deployment.

Latent Capacity and Spatial Arbitrage. Crucially, CNN successfully distinguishes between the centralized efficiency demanded by hyperscale cloud operators and the distributed access required by third-party colocation providers. Both maps identify "prime locations"—locations ranking in the top national percentiles for predicted suitability that currently lack realized facilities. For instance, the model highlights significant latent capacity in the suburban peripheries surrounding major hubs, such as Kaufman County near Dallas, Fayette County outside Atlanta, and portions of the Will County corridor near Chicago. These latent sites represent spatial arbitrage opportunities where existing infrastructure supply—such as high-voltage power grid capacity, fiber connectivity, and water resources—exceeds current market utilization. By quantifying this gap between structural infrastructure readiness and realized investment, the algorithm provides a leading indicator for future capital allocation in both the hyperscale and colocation segments.

Visualizing the Spatial Screening Logic. Figures A.4 and A.5 illustrate why a spatial CNN model outperforms traditional regression specifications for this problem. Each panel displays a single 32 km $\times$ 32 km catchment area as the model receives it: a stack of geographic layers (the input tensor) decomposed into three thematic rows—fiber and broadband connectivity, energy infrastructure, and physical constraints (water stress and drought)—across three representative sites selected from a 2,000-location scan.

The "High Suitability" column ($P \geq 0.92$) shows locations the model identifies as strong candidates for development. In both market segments, these sites exhibit rich, spatially heterogeneous texture across all three layers: dense fiber networks intersecting high-voltage grid corridors, surrounded by moderate physical constraints. A standard regression would summarize each of these layers as a single numeric covariate (e.g., average distance to the nearest substation), discarding the precise geometric layout that distinguishes a redundant, well-connected network from a single vulnerable transmission line. The CNN preserves this full spatial structure and learns to detect the co-occurrence of multiple infrastructure resources within the same local neighborhood.

The "Hard Negative" column ($P \leq 0.08$) demonstrates the model's capacity to identify binding constraints where the absence of a single critical resource renders other advantages irrelevant. For the hyperscale cloud example (Figure A.4), the hard negative site displays substantial energy capacity but nearly uniform, featureless connectivity layers, indicating an area beyond the reach of dense fiber networks. For the colocation example (Figure A.5), connectivity and energy are both present, but the physical constraints row reveals pronounced water stress—a disqualifying condi-

tion for facilities requiring large-scale cooling. A linear model would likely assign high predicted probability to both sites based on their energy endowments alone; the CNN correctly rejects them by detecting the spatial coincidence of a fatal deficiency alongside otherwise favorable conditions.

The "Rural Baseline" column ($P = 0.01$) serves as a control, depicting a location with uniformly weak infrastructure across all layers. Together, these three cases confirm that the model captures the strict, holistic screening logic that institutional site selectors apply in practice: a single missing prerequisite disqualifies an otherwise attractive parcel.

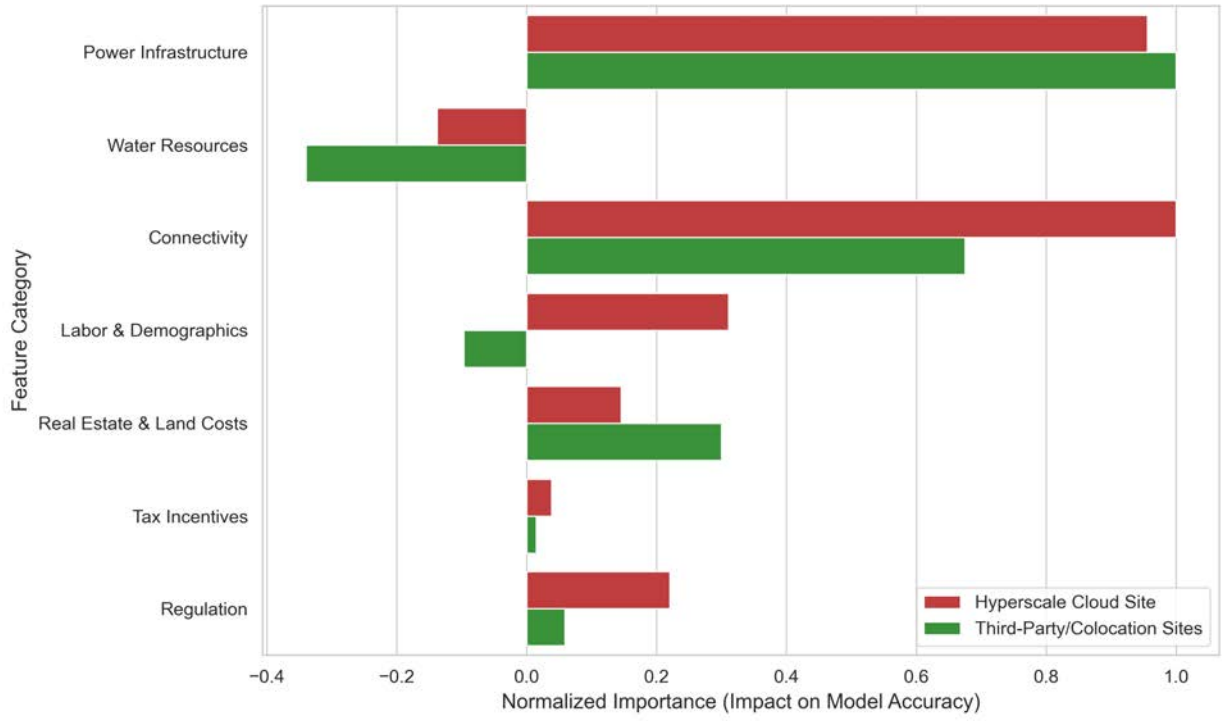
5.2 Structural Determinants of Data Center Siting

Grouped Permutation Importance. To identify the structural drivers of site selection, we employ two complementary interpretability methods. The first, grouped permutation importance, quantifies each feature category's marginal contribution to predictive accuracy by randomly shuffling its channel values across locations—destroying the spatial signal within that group—and measuring the resulting increase in prediction error. We apply this procedure to a representative fold of each model using a 2,000-site subsample, repeating each permutation five times to stabilize the estimates. The normalized results, reported separately for each market segment in Figure 8, reveal two distinct siting hierarchies that correspond to the divergent business models of hyperscale and colocation operators.

For hyperscale cloud providers, power infrastructure and network connectivity emerge as two dominant binding constraints. Labor and demographics, along with local regulation, serve as critical secondary drivers, substantially outweighing real estate and land costs. This reflects the operational complexity of modern hyperscale data centers: operators prioritize reliable grid access and ultra-high-bandwidth networks, while also relying on skilled regional labor pools and favorable regulatory environments to support massive, long-term deployments. In contrast, water resources are not a primary structural determinant; rather, they likely function as predictive noise, reflecting the industry's ability to substitute water-intensive evaporative cooling with air-cooled technologies in arid regions (Siddik, Shehabi and Marston, 2021).

In sharp contrast, third-party colocation providers exhibit a hierarchy strictly dominated by power access and physical real estate. Power infrastructure acts as the absolute primary predictor, exerting significantly more influence than network connectivity. Real estate and land costs rank third, indicating that colocation developers focus heavily on securing grid capacity alongside affordable parcels to construct leasable server shells. Unlike hyperscalers, colocation providers

Figure 8—Grouped Permutation Importance by Market Segment



Notes: This figure reports the grouped permutation importance of seven feature categories for hyperscale cloud providers (red) and third-party colocation operators (green). For each category, all constituent channels are simultaneously shuffled across locations in a 2,000-site subsample, and the resulting increase in binary cross-entropy loss is recorded. This procedure is repeated five times; the mean loss increase is then normalized by the maximum observed impact across all categories within each segment (scale -1.0 to 1.0). Positive values indicate that scrambling the category degrades accuracy (the category is a driver of entry); negative values indicate that scrambling improves accuracy (the category introduces noise or acts as a deterrent). Results are based on a single representative fold (fold 0) of the Teacher model (third-party) and Student model (cloud).

show no structural reliance on local labor markets—which instead present as predictive noise—likely because their enterprise tenants independently manage their own internal IT operations. Similarly, the negative importance of water resources suggests that, like the hyperscale segment, the colocation sector does not treat hydrological fundamentals as a decisive factor in site selection.

Site-Level Attribution. To move beyond aggregate national rankings and examine how these drivers operate at the level of individual sites, we apply a second method: Integrated Gradients (IG) attribution. This diagnostic technique decomposes a site's final suitability score to identify the specific contribution of each infrastructure feature. Specifically, IG evaluates the importance of a

characteristic by tracing the increase in probability as a site "improves" from a neutral baseline—a hypothetical location with no relevant infrastructure—to its actual observed state.

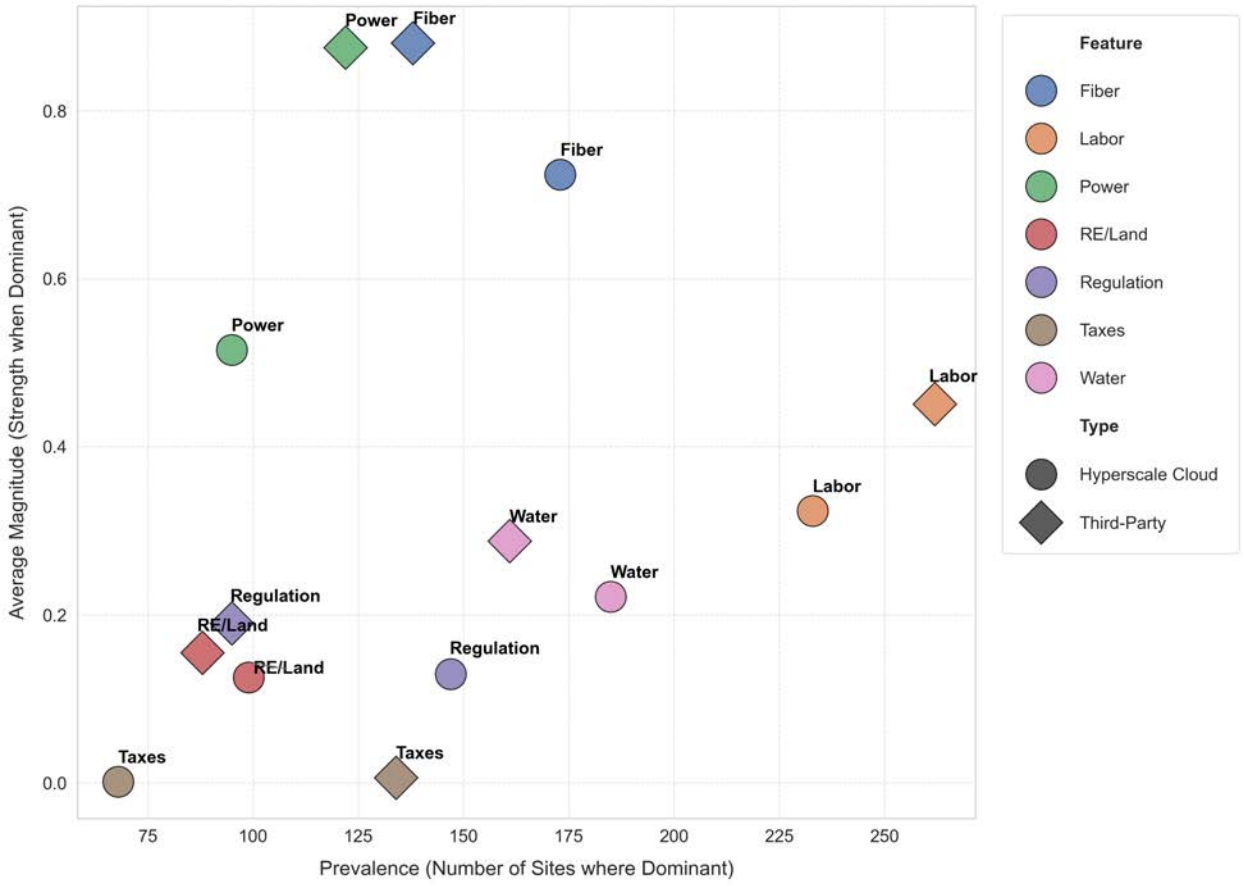
Crucially, this approach ensures that the model's final decision is fully transparent and entirely accounted for by the underlying data. This satisfies the axiom of completeness, meaning that the individual weights—or feature attributions—assigned to each infrastructure and economic factor sum exactly to the total predicted suitability score, leaving no part of the model's output unexplained. This property is particularly valuable for policymakers and investors, as it provides a rigorous "accounting" of site selection. By ensuring that every percentage point of the suitability score is tied back to a specific observable characteristic, the method removes the "black box" nature of deep learning. Instead, it offers a granular breakdown of how much credit should be assigned to the local structural determinant in driving the final investment signal.

To ensure the validity of these site-level results, we draw a 1,000-location subsample using a stratified random sampling approach. This procedure ensures that the subsample accurately mirrors the national distribution of both realized data center entries and highly suitable undeveloped parcels across diverse geographic regions. Using this representative subset, we aggregate the parcel-level attributions into our broader feature categories. We classify the category exerting the strongest positive influence on the model's decision as the dominant driver for that specific investment. Figure 9 maps the resulting investment landscape, plotting the prevalence of each category as a dominant driver (how frequently it ranks first) against its average magnitude when dominant (the intensity of its contribution to the score).

The results characterize a clear structural trade-off between infrastructure intensity and operational prevalence in data center siting. The findings identify Fiber and Power as the "high-magnitude" deal-makers: while these features dominate fewer total sites compared to labor, they exert the most intense influence on the model's suitability scores when they are present. Specifically, for third-party colocation facilities, Fiber and Power exhibit the highest average magnitudes, suggesting that these physical assets act as the primary structural anchors for investment.

In comparison, Labor & Demographics emerge as the most "prevalent" driver across both market segments, acting as the primary determinant for the largest number of sites. However, the magnitude of its impact is notably lower than that of critical infrastructure. This suggests that while broadband and energy capacity are the decisive factors that "make or break" a site, labor market characteristics serve as a ubiquitous baseline requirement that influences a broader geographic range of lower-intensity siting decisions.

Figure 9—Dominant Location Drivers: Prevalence versus Magnitude



Notes: This figure characterizes the prevalence and intensity of feature categories when they serve as the dominant (highest-scoring) driver for a given site. Attribution scores are computed via Integrated Gradients (20 approximation steps, mean-baseline from 100 random patches) on a 1,000-site subsample drawn from each market segment. A feature category is classified as dominant for a site if it yields the highest positive summed attribution score across all spatial pixels. The horizontal axis reports the number of sites for which each category is dominant; the vertical axis reports the average attribution magnitude across those sites. Third-party colocation sites are represented by diamonds and hyperscale cloud sites by circles.

Our findings identify an incentive paradox that adds a crucial spatial dimension to recent causal evidence. While state-level research suggests that tax incentives can double the volume of data center construction (Gargano and Giacoletti, 2025), our site-level analysis indicates that these fiscal interventions exhibit minimal importance compared to physical infrastructure. We resolve this discrepancy by recognizing that infrastructure acts as a fundamental prerequisite for site viability, while fiscal policy operates as a secondary factor that only influences decisions once these essential requirements are met. At the parcel level, access to high-capacity power and fiber

is mandatory; without them, a site remains unviable regardless of the tax regime. However, once these structural necessities are satisfied, incentives may serve as the decisive factor for choosing between competing jurisdictions.

Furthermore, the model reveals that hyperscale operators show a significantly higher structural reliance on favorable local regulation than colocation developers. This reflects the divergent risk profiles of the two segments: the multi-billion-dollar, long-horizon nature of hyperscale campuses necessitates a stable, predictable policy environment, whereas colocation operators prioritize speed-to-market and proximity to existing enterprise hubs.

Additional detail on these attribution results—including pixel-level heatmaps of local importance for four case study markets, the decomposition of positive versus negative factor contributions, and the distributional spread of driver strength conditional on dominance—is provided in Appendix Section A.5 (Figures A.6–A.8).

5.3 Counterfactual Projections of Data Center Growth

Simulation Framework. We extend the analysis to 2030 by modeling six counterfactual scenarios using a rolling-horizon recursive simulation—a dynamic forecasting technique where the model sequentially predicts site selections year by year. After each simulated year, the algorithm updates the underlying geographic data—the input tensor of available infrastructure—before making its predictions for the subsequent year. All scenarios share a common baseline structure. The *Status Quo* extrapolates historical trends at a 2% annual compound growth rate for economic and broadband covariates, while holding physical grid topology and generation capacity fixed at 2024 levels—a conservative assumption reflecting the multi-year permitting and construction timelines that characterize transmission infrastructure (U.S. Department of Energy, 2023). The five alternative scenarios then perturb specific covariate groups relative to this baseline to isolate the spatial consequences of distinct economic, environmental, and policy shocks.

The AI Boom Scenario. The spatial distribution of AI infrastructure is fundamentally shaped by the distinction between two core deep learning processes: model training and inference. Model training—the computationally intensive process of teaching an algorithm using vast datasets—can prioritize low-cost rural power because transmission latency (delay) is largely irrelevant (Knittel, Senga and Wang, 2025). In contrast, inference workloads—the application of these trained models to generate real-time responses for end users—demand ultra-low latency. The mass deployment of

real-time AI applications, therefore, necessitates a proliferation of "edge" facilities located directly within densely populated centers.

To operationalize this geographic shift, the *AI Boom* scenario isolates metropolitan areas to simulate a push-pull economic dynamic driven by explosive computational demand. First, we model a 50% surge in urban grid capacity to reflect grid infrastructure upgrades. Second, to capture load strain, we impose a 20% electricity price spike (Electric Power Research Institute, 2024; International Energy Agency, 2025). This acts as a *de facto* congestion tax that raises marginal operating costs, aligning with empirical congestion premiums and capacity auction spikes (Mamkhezri, Sun and Yang, 2025; Kearney and McLaughlin, 2025).

Crucially, the algorithm resolves this economic tension without hard-coded rules. Instead, the CNN evaluates the updated multi-layered geographic map (the input tensor). Because the model internalized the interdependent screening logic of real estate developers during training, it organically weighs the trade-off between superior physical grid access and sharply inflated operating costs. The algorithm mathematically adjusts the suitability scores, allowing the data to reveal whether the strategic necessity of low-latency urban infrastructure justifies the financial penalty of the congestion tax.

The Climate Stress Scenario. The *Climate Stress* scenario simulates a compound environmental shock along three channels: a sharp decline in drought severity (PDSI), a 40% increase in baseline water stress (WRI), and a 20% degradation in grid reliability (SAIDI). Data centers require continuous access to cooling water and uninterruptible power, making drought-prone and grid-vulnerable regions less viable as climate conditions deteriorate. Together, these perturbations model a future in which climate risk acts as a binding constraint on facility siting, potentially redirecting investment away from historically attractive but increasingly resource-stressed markets.

The Green Grid Scenario. The *Green Grid* scenario models a targeted 5% annual expansion of grid capacity within renewable-rich zones, limited to wind and solar to capture the high geographic scalability that distinguishes these technologies from constrained sources like hydroelectricity. This design mirrors the procurement strategies of hyperscale operators, whose corporate power purchase agreements increasingly fund new renewable developments rather than draw on existing grid supply (International Energy Agency, 2025)—making the spatial diffusion of wind and solar a leading indicator for future site selection.

Autarkic Expansion. The *Autarkic Expansion* scenario assumes that data center operators bypass commercial utility constraints entirely through on-site generation—such as small modular reactors or dedicated natural gas plants (International Energy Agency, 2025; U.S. Department of Energy, 2025). We model this by setting all grid distance covariates (distance to transmission lines, substations, and high-voltage nodes) to zero nationwide, eliminating grid proximity as a factor in site selection. This represents the industry's energy self-sufficiency: if operators can generate power on-site at competitive cost, proximity to the existing grid ceases to be a prerequisite for development, and location choice is governed entirely by non-energy fundamentals such as land cost, connectivity, and labor supply.

Regulatory Headwinds. The *Regulatory Headwinds* scenario models the opposite dynamic: increasing political resistance to data center development. We simulate this through a 50% increase in regulatory friction (applied to zoning and legislative stringency covariates) and a 50% reduction in tax incentive attractiveness. This parameterization captures the emerging pattern of local moratoriums, tightened permitting requirements, and scaled-back fiscal incentives observed in historically dense development corridors such as Northern Virginia, where community opposition to noise, water consumption, and grid strain has intensified.[14] The Regulatory Headwinds scenario can therefore be interpreted as quantifying the spatial reallocation that such legislative shocks would produce within the U.S. context.

Aggregate Trajectories. Figure 10 plots the projected national footprint under all six scenarios, aggregating both hyperscale cloud and third-party colocation segments. For each simulated year, the model scores every undeveloped 500-meter grid cell with a suitability probability; a scenario-specific demand parameter then determines how many new facilities enter by selecting the highest-scoring cells until annual demand is filled.[15] The left vertical axis reports imputed electrical capacity in gigawatts (GW); the right vertical axis displays the corresponding physical footprint in thousands of acres. We convert footprint to capacity using a fleet-average density of 0.44 MW per acre, which reflects the blended deployment intensity of hyperscale and colocation

[14] Beyond the U.S., this dynamic is also visible internationally: Carlo and Rolheiser (2026) document that Quebec's Bill 69 now requires ministerial authorization for any new load above 5 MW, Ontario's Bill 40 imposes connection requirements on covered data center facilities, and British Columbia has imposed a hard 400 MW capacity cap on data center and AI projects.

[15] Annual demand is drawn from a normal distribution centered on a scenario-specific mean—600 cells for AI Boom, 450 for Autarkic Expansion, 300 for Status Quo and Green Grid, 200 for Regulatory Headwinds, and 150 for Climate Stress—with a standard deviation of 50 cells. These baselines approximate observed annual entry rates for the Status Quo and are scaled proportionally for alternative scenarios.

facilities and aligns the Status Quo trajectory with independent national projections of approximately 130 GW of U.S. data center capacity by 2030 (Electric Power Research Institute, 2024; International Energy Agency, 2025).[16]

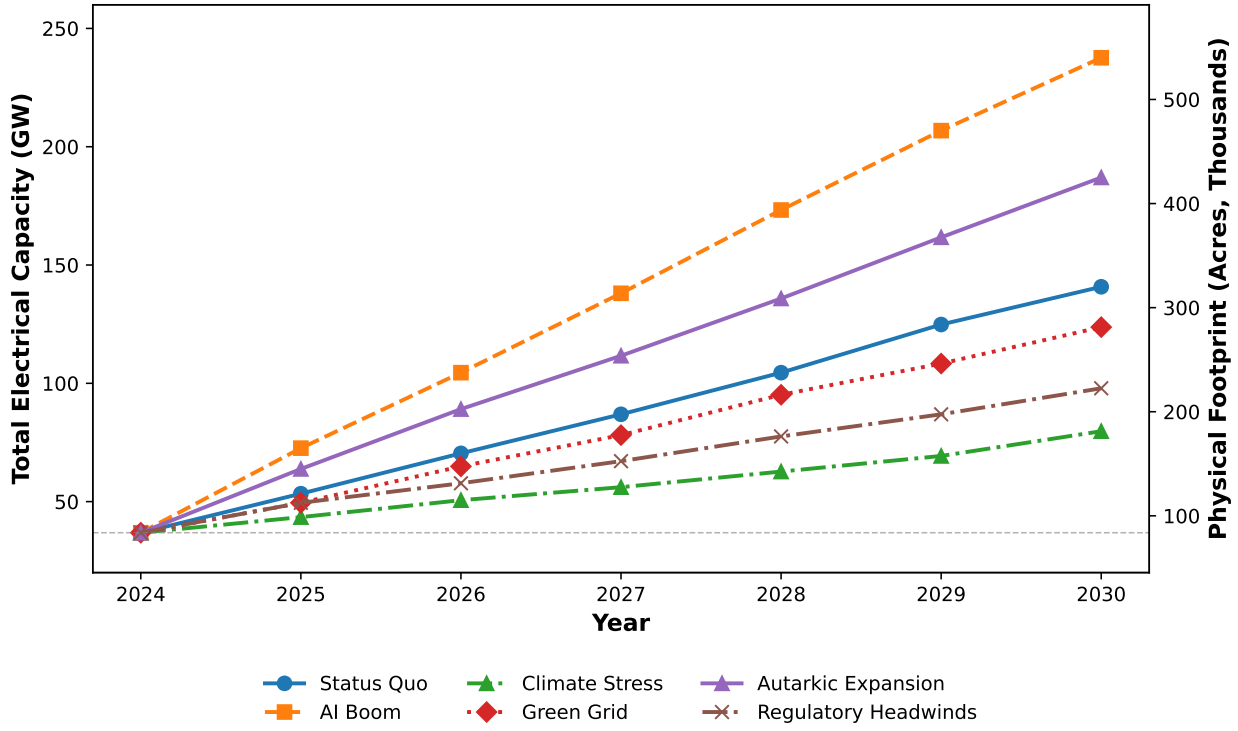
The trajectories diverge sharply from a common 2024 baseline, ranging from approximately 80 GW under *Climate Stress* to roughly 240 GW under *AI Boom* by 2030—a nearly threefold gap that underscores the sensitivity of infrastructure investment to the prevailing policy and market environment. The *AI Boom* trajectory dominates all others, producing the highest cumulative footprint by 2030. *Autarkic Expansion* ranks second at ∼190 GW, nearly matching the AI Boom despite operating without any explicit demand stimulus. This near-equivalence is itself a substantive finding: when on-site generation removes grid distance as a binding factor, latent suitability is unlocked at a scale comparable to a 50% expansion of urban grid capacity, suggesting that access to grid infrastructure is the single largest constraint on aggregate deployment.

In contrast, the *Green Grid* trajectory (∼125 GW) trails the unconstrained *Status Quo*. By restricting development to renewable-energy zones, operators abandon highly suitable carbon-intensive markets, creating a spatial bottleneck. Although infrastructure within these green corridors expands by 5% annually, localized gains cannot offset the broader geographic exclusions, forcing a strict trade-off between capacity expansion and sustainability. At the lower end, *Regulatory Headwinds* (∼100 GW) and *Climate Stress* (∼80 GW) produce the most constrained trajectories—consistent with regimes in which tightened permitting, reduced incentives, water scarcity, and grid unreliability simultaneously raise development costs and erode returns.

Regional Shifts and Spatial Reallocation. Figure 11 decomposes these aggregate trajectories into state-level winners and losers relative to the *Status Quo* by 2030, while Tables A.2 and A.3 report the comprehensive state-by-state projections. The *AI Boom* scenario generates a rising-tide effect with no state-level losers. Texas emerges as the dominant beneficiary, gaining roughly 8 GW of projected new capacity, followed by Montana and Nevada. This concentration reflects Texas's distinctive combination of deregulated energy markets, abundant land, and rapidly expanding grid infrastructure. When the urban capacity shock is applied, the model's spatial screening logic dynamically penalizes congested metropolitan areas and reallocates investment toward unconstrained, infrastructure-rich regions.

[16] Shehabi et al. (2024) report aggregate U.S. data center electricity demand of 325–580 TWh by 2028, corresponding to roughly 74–132 GW of average power demand at a 50% capacity utilization rate. Our 0.44 MW-per-acre density factor is calibrated to align the Status Quo trajectory with the midpoint of this range while preserving the relative magnitudes

Figure 10—Projected Aggregate Data Center Capacity and Footprint Under Alternative Scenarios

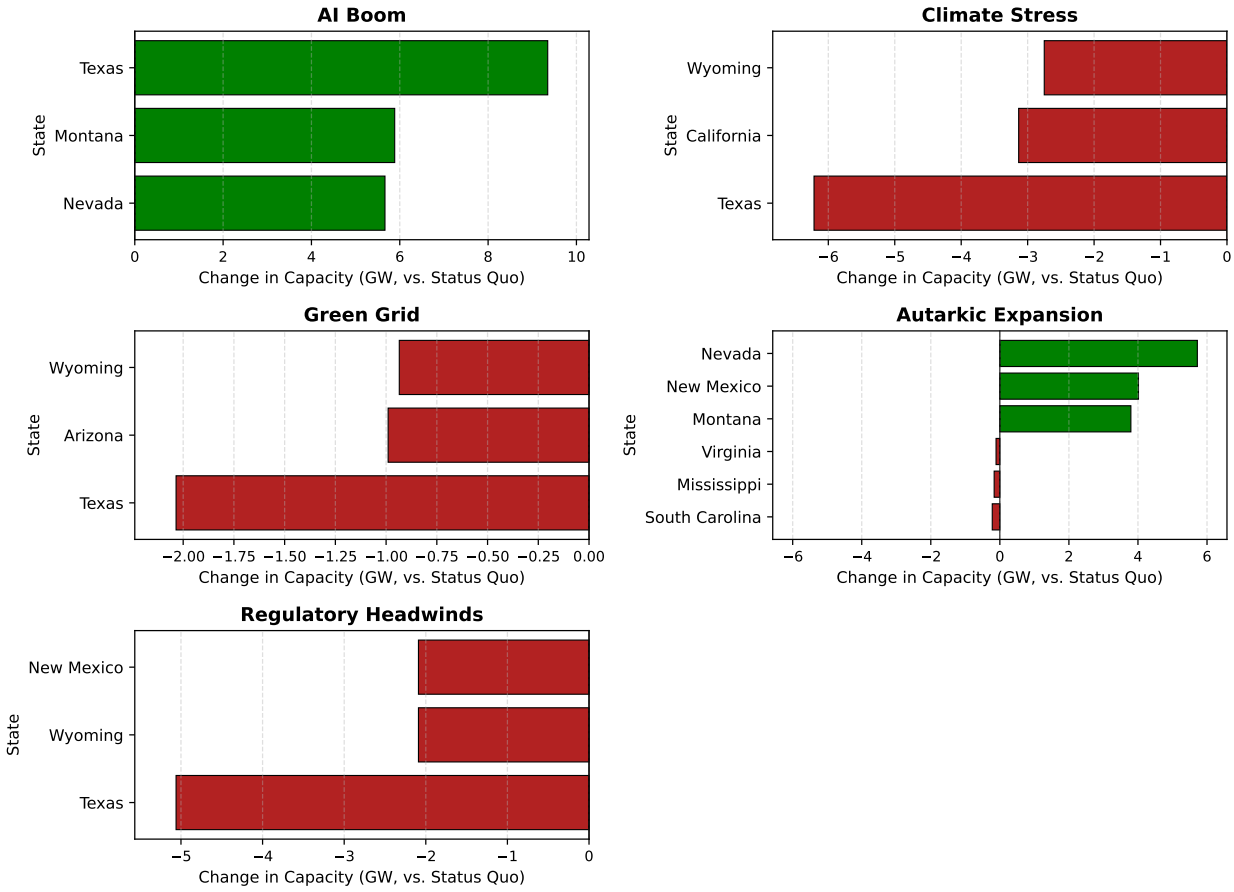


Notes: This figure plots the projected total electrical capacity and physical footprint of the U.S. data center industry from 2024 to 2030 under six counterfactual scenarios, aggregating both hyperscale cloud and third-party colocation segments. Projections are generated by a rolling-horizon recursive simulation in which the trained CNN sequentially predicts facility entry for each year, updating the covariate tensor after each step. The left vertical axis reports total imputed electrical capacity in gigawatts (GW); the right vertical axis displays the corresponding physical footprint in thousands of acres. Capacity is imputed from the simulated spatial footprint at a fleet-average density of approximately 0.44 MW per acre, which blends hyperscale and colocation deployment intensities. The Status Quo trajectory aligns with independent industry projections of approximately 140 GW of U.S. data center capacity by 2030. All scenarios share the same 2024 empirical baseline; variations across trajectories reflect how scenario-specific shocks compound year over year as the forecast unfolds.

By contrast, the *Climate Stress*, *Green Grid*, and *Regulatory Headwinds* scenarios impose binding constraints that shrink the aggregate development pipeline, yielding no state-level winners relative to the baseline. Under *Climate Stress*, Texas sheds the most absolute capacity (∼6.2 GW) as simulated water scarcity and grid unreliability erode its competitive position, with California (∼3.1 GW) and Wyoming (∼2.8 GW) experiencing the next-largest losses—a pattern that reflects the compounding effects of grid vulnerability and environmental exposure.

across scenarios.

Figure 11—State-Level Reallocation of Projected Data Center Capacity by 2030



Notes: Each panel reports the change in projected 2030 electrical capacity (GW) for the three states with the largest absolute gains (green) and losses (red) relative to the *Status Quo* baseline, under each of five alternative scenarios. Capacity is imputed from the simulated spatial footprint at a fleet-average density of approximately 0.44 MW per acre, which reflects the blended deployment intensity of hyperscale and colocation facilities and aligns the Status Quo trajectory with independent national projections of approximately 130 GW of U.S. data center capacity by 2030 (Electric Power Research Institute, 2024; International Energy Agency, 2025). The panels illustrate how distinct macroeconomic, environmental, and regulatory shocks redistribute competitive advantage across states.

The *Green Grid* scenario follows a similar redistributive pattern, with Texas, Arizona, and Wyoming bearing the largest reductions (∼2.0, 1.0, and 0.9 GW, respectively). Strict environmental mandates create spatial bottlenecks that limit expansion even in states with established renewable infrastructure, as the model dynamically excludes high-suitability acreage outside designated green corridors.

The *Autarkic Expansion* case produces the most geographically dispersed reallocation. States

such as Nevada (∼5.7 GW), New Mexico (∼4.0 GW), and Montana (∼3.8 GW) gain substantially, whereas historically transmission-rich states such as Virginia, Mississippi, and South Carolina lose modest shares. This pattern is economically intuitive: when on-site generation renders grid proximity irrelevant, the model's feature attribution dynamically shifts weight away from transmission lines and toward cheaper land and favorable non-energy fundamentals.

Finally, the *Regulatory Headwinds* panel shows widespread capacity shedding, with Texas (∼5.1 GW), Wyoming (∼2.1 GW), and New Mexico (∼2.1 GW) bearing the largest absolute losses. This redistribution captures the flight of capital from states where local moratoriums and tightened zoning raise effective development costs, permanently pricing viable acreage out of the national pipeline. State-level projections appear in Appendix Tables A.2 and A.3. County-level divergence maps, depicting the geographic redistribution of suitability probabilities under each alternative scenario relative to the Status Quo, are reported in Appendix Figures A.9–A.13.

6 Conclusion

This study demonstrates that data centers are a capital-intensive industrial asset class strictly bound by physical constraints. By deploying a CNN over a high-resolution, 15-year spatial panel, we reverse-engineer the complex site-selection logic of real estate developers. This approach—which uses spatial filters to mathematically scan geographic maps much like a human surveyor evaluates overlapping utility constraints—reveals a profound structural divergence in the sector. While third-party colocation providers cluster in urban peripheries to balance connectivity with land availability, hyperscale operators function as industrial utilities, optimizing for bulk power capacity and fiber access at the grid's remote periphery.

Our findings offer direct strategic utility for institutional developers and policymakers. By identifying near-miss locations—parcels that the model scores as highly suitable but that remain undeveloped—investors can implement targeted land-banking strategies to secure undervalued parcels and utility rights before markets mature. For policymakers, these locations pinpoint infrastructure or regulatory bottlenecks where targeted interventions—such as expedited permitting or grid modernization—could catalyze regional growth in data center infrastructure. The model's interpretability tools (feature attribution) further diagnose which specific deficits—whether inadequate grid access, insufficient broadband, or unfavorable policy environments—are suppressing development in particular regions, providing an actionable roadmap for closing those gaps.

Finally, our counterfactual simulations highlight the instability of the current geographic equilibrium. While an AI-driven demand shock amplifies growth in established, infrastructure-rich hubs, alternative futures—such as the adoption of on-site power generation or tightened environmental and regulatory constraints—can drastically redistribute the national data center footprint. As these facilities become critical infrastructure for the modern economy, their physical geography will increasingly shape the distribution of economic activity and digital productivity. By applying deep learning to high-resolution spatial data, our paper provides a flexible empirical framework for anticipating where that infrastructure will locate and how policy can steer it. This approach supports the long-term alignment of technological expansion with infrastructure planning.

References

Adams, Andrew. 2013. "New Utah NSA Center Requires 1.7M Gallons of Water Daily to Operate." *KSL*.

Alcácer, Juan, and Mercedes Delgado. 2016. "Spatial organization of firms and location choices through the value chain." *Management Science*, 62(11): 3213–3234.

Andonov, Aleksandar, Roman Kräussl, and Joshua Rauh. 2021. "Institutional Investors and Infrastructure Investing." *The Review of Financial Studies*, 34(8): 3880–3934.

Benetton, Matteo, Giovanni Compiani, and Adair Morse. 2023. "When Cryptomining Comes to Town: High Electricity-use Spillovers to the Local Economy." National Bureau of Economic Research Working Paper 31312.

Blair, Loudon. 2017. "Finding Strength in Numbers: The Data Center Clustering Effect." *Data Center Knowledge*, Accessed: 2026-01-10.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen. 2014. "The distinct effects of information technology and communication technology on firm organization." *Management Science*, 60(12): 2859–2885.

Blum, Bernardo S., and Avi Goldfarb. 2006. "Does the Internet Defy the Law of Gravity?" *Journal of International Economics*, 70(2): 384–405.

Byrne, David, Carol Corrado, and Daniel E Sichel. 2018. "The Rise of Cloud Computing: Minding Your P's, Q's and K's." National Bureau of Economic Research.

Carlo, Alexander, and Lyndsey Rolheiser. 2026. "Data Centred: the Shifting Landscape of Canada's Digital Infrastructure." Schulich School of Business Real Assets Research Paper Series. forthcoming.

CBRE. 2013. "Impact of Taxes & Incentives on Data Center Locations." CBRE Research Report, New York.

CBRE. 2015. "Site Selection Strategies for Enterprise Data Centers." CBRE Research Report, New York.

CBRE. 2024. "North America Data Center Trends H1 2024." https://www.cbre.com/insights/reports/north-america-data-center-trends-h1-2024, Accessed April 17, 2026.

CBRE. 2026. "Data Center Research and Insights." https://www.cbre.com/insights/data-center, Accessed March 4, 2026.

Coyle, Diane, David Nguyen, et al. 2018. "Cloud Computing and National Accounting." *Economic Statistics Centre of Excellence Discussion Paper*, 2019.

Cushman & Wakefield. 2025. "U.S. Data Center Development Cost Guide 2025." Accessed October 20, 2025. https://cushwake.cld.bz/DataCenter-Development-Cost-Guide-2025/2-3/.

Davis, Lucas W., Catherine Hausman, and Nancy L. Rose. 2023. "Transmission Impossible? Prospects for Decarbonizing the US Grid." *Journal of Economic Perspectives*, 37(4): 155–180.

DeStefano, Timothy, Richard Kneller, and Jonathan Timmis. 2023. "Cloud Computing and Firm Growth." *Review of Economics and Statistics*, 1–47.

Donaldson, Dave. 2018. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *The American Economic Review*, 108(4-5): 899–934.

Electric Power Research Institute. 2024. "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption." Electric Power Research Institute White Paper.

Elhai, Coly. 2026. "Data Centers: Location Choice and Grid Externalities." Unpublished manuscript, Harvard University.

Environmental and Energy Study Institute (EESI). 2025. "Data Centers and Water Consumption." Environmental and Energy Study Institute, Washington, DC. Accessed on October 20, 2025.

Fang, Tommy Pan, and Shane Greenstein. 2025. "Where the Cloud Rests: The Economic Geography of Data Centers." *Strategy Science*.

Fernald, John G. 1999. "Roads to Prosperity? Assessing the Link Between Public Capital and Productivity." *The American Economic Review*, 89(3): 619–638.

Forman, Chris, Anindya Ghose, and Avi Goldfarb. 2009. "Competition between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live." *Management Science*, 55(1): 47–57.

Gargano, Antonio, and Marco Giacoletti. 2025. "Subsidizing the Cloud: US State Incentives to Data Centers." *Available at SSRN 5881105*.

Giroud, Xavier, and Joshua Rauh. 2019. "State Taxation and the Reallocation of Business Activity: Evidence from Establishment-Level Data." *Journal of Political Economy*, 127(3): 1262–1316.

Goldfarb, Avi, and Catherine Tucker. 2019. "Digital Economics." *Journal of Economic Literature*, 57(1): 3–43.

Gorman, Will, Julie Mulvaney Kemp, Joseph Rand, Joachim Seel, Ryan Wiser, Nick Manderlink, Fredrich Kahrl, Kevin Porter, and Will Cotton. 2025. "Grid Connection Barriers to Renewable Energy Deployment in the United States." *Joule*, 9: 101791.

Greenstein, Shane. 2020. "The Basic Economics of Internet Infrastructure." *Journal of Economic Perspectives*, 34(2): 192–214.

Greenstein, Shane. 2021. "Digital Infrastructure." *Economic Analysis and Infrastructure Investment*, 409.

Greenstein, Shane, Chris Forman, and Avi Goldfarb. 2018. "How Geography Shapes—and Is Shaped by—The Internet." In *The New Oxford Handbook of Economic Geography*. , ed. Gordon Clark, MaryAnn Feldman, Meric Gertler and Dariusz Wojcik, 269–285. Oxford University Press.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti. 2010. "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings." *Journal of Political Economy*, 118(3): 536–598.

Hancock, Simon. 2009. "Iceland Looks to Serve the World." *BBC News*, October 9, 2009.

Hausman, Catherine. 2025. "Power Flows: Transmission Lines, Allocative Efficiency, and Corporate Profits." *American Economic Review*, 115(8): 2574–2615.

International Energy Agency. 2025. "Energy and AI." International Energy Agency, Paris.

Jin, Wang, and Kristina McElheran. 2017. "Economies before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing." *Rotman School of Management Working Paper*, 3112901.

Jones Lang LaSalle. 2026. "2026 Global Data Center Outlook." Jones Lang LaSalle (JLL) Industry Report.

Kearney, Laila, and Tim McLaughlin. 2025. "Power Costs Soar in PJM Region as Data Center Demand Spikes." *Reuters*, Accessed on March 12, 2026. https://www.reuters.com/business/energy/power-costs-soar-pjm-region-data-center-demand-spikes-2025-08-07.

Khachiyan, Arman, Anthony Thomas, Huye Zhou, Gordon Hanson, Alex Cloninger, Tajana Rosing, and Amit K Khandelwal. 2022. "Using neural networks to predict microspatial economic growth." *American Economic Review: Insights*, 4(4): 491–506.

Knittel, Christopher R., Juan Ramon L. Senga, and Shen Wang. 2025. "Flexible Data Centers and the Grid: Lower Costs, Higher Emissions?" National Bureau of Economic Research Working Paper 34065.

Leduc, Sylvain, and Daniel Wilson. 2012. "Roads to Prosperity or Bridges to Nowhere? Theory and Evidence on the Impact of Public Infrastructure Investment." *NBER Macroeconomics Annual*, 27(1): 89–142.

Malecki, Edward J. 2002. "The economic geography of the Internet's infrastructure." *Economic Geography*, 78(4): 399–424.

Mamkhezri, Jamal, Xiaochen Sun, and Yuting Yang. 2025. "The Hidden Cost of the Cloud: Data Centers and Electricity Market Inefficiency." United States Association for Energy Economics Working Paper.

Markoff, John. 2016. "Microsoft Plumbs Ocean's Depths to Test Underwater Data Center." *The New York Times*.

Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020. "Recalibrating Global Data Center Energy-Use Estimates." *Science*, 367(6481): 984–986.

Masanet, Eric R., Richard E. Brown, Arman Shehabi, Jonathan G. Koomey, and Bruce Nordman. 2011. "Estimating the Energy Use and Efficiency Potential of U.S. Data Centers." *Proceedings of the IEEE*, 99(8): 1440–1453.

McKinsey & Company. 2023. "Investing in the Rising Data Center Economy." McKinsey & Company Report.

Newmark. 2023. "2023 U.S. Data Center Market Overview Market Clusters." Accessed on January 24, 2026. https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview-market-clusters.

Pollmann, Michael. 2023. "Causal Inference for Spatial Treatments." *arXiv preprint arXiv:2011.00373*.

Qian, Franklin, Qianyang Zhang, and Xiang Zhang. 2024. "Identifying Agglomeration Spillovers: Evidence from Grocery Store Openings." Working Paper.

Qu, Xiaorui, Qinan Lu, Minghao Li, and Wendong Zhang. 2025. "Broadband Internet Speed Upgrades and the Farmland Market: A Shift-Share Instrumental Variable Approach." *American Journal of Agricultural Economics*, Early View.

Shehabi, Arman, Sarah J. Smith, Alex Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. "2024 United States Data Center Energy Usage Report." Lawrence Berkeley National Laboratory Technical Report LBNL-2001637.

Shehabi, Arman, Sarah J. Smith, Eric Masanet, and Jonathan Koomey. 2018. "Data Center Growth in the United States: Decoupling the Demand for Services from Electricity Use." *Environmental Research Letters*, 13(12): 124030.

Siddik, M Abu Bakar, Arman Shehabi, and Landon Marston. 2021. "The environmental footprint of data centers in the United States." *Environmental Research Letters*, 16(6): 064017.

Sinai, Todd, and Joel Waldfogel. 2004. "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?" *Journal of Urban Economics*, 56(1): 1–24.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." 3319–3328, JMLR.org.

Tambe, Prasanna. 2014. "Big Data Investment, Skills, and Firm Value." *Management Science*, 60(6): 1452–1469.

U.S. Department of Energy. 2023. "National Transmission Needs Study." U.S. Department of Energy, Washington, DC.

U.S. Department of Energy. 2025. "Advantages and Challenges of Nuclear-Powered Data Centers." Office of Nuclear Energy, U.S. Department of Energy, Washington, DC.

U.S. DOE. 2024. "U.S. Data Center Energy Use Report." U.S. Department of Energy.

Van Nieuwerburgh, Stijn. 2026. "Financing the AI Buildout." SSRN Working Paper 6444363.

Volzer, Helena. 2025. "Data Centers Are Increasing in the Great Lakes. At What Cost?" *Alliance for the Great Lakes*, Accessed October 20, 2025. https://greatlakes.org/2025/03/data-centers-are-increasing-in-the-great-lakes-at-what-cost/.

Wang, Shuang, Jacob LaRiviere, and Aadharsh Kannan. 2019. "Spatial Competition and Missing Data: An Application to Cloud Computing." *Working Paper*, 1–23.

Zhang, Mary. 2023. "Types of Data Centers: Enterprise, Colocation, Hyperscale." *Dgtl Infra*.

Zook, Matthew A. 2002. "Grounded capital: venture financing and the geography of the Internet industry, 1994–2000." *Journal of Economic Geography*, 2(2): 151–177.

Appendix

A.1 Temporal Identification via Remote Sensing

As noted in Section 3, public data center registries lack construction dates for new data center entrants. Because our empirical strategy conditions predictions on a strict three-year lag ($t-3$ through $t-1$ predicting entry in year t), misassigning a facility's completion date by even a single year would contaminate the temporal firewall and undermine identification. To resolve this ambiguity, we developed the hybrid remote-sensing protocol formalized in Figure A.1.

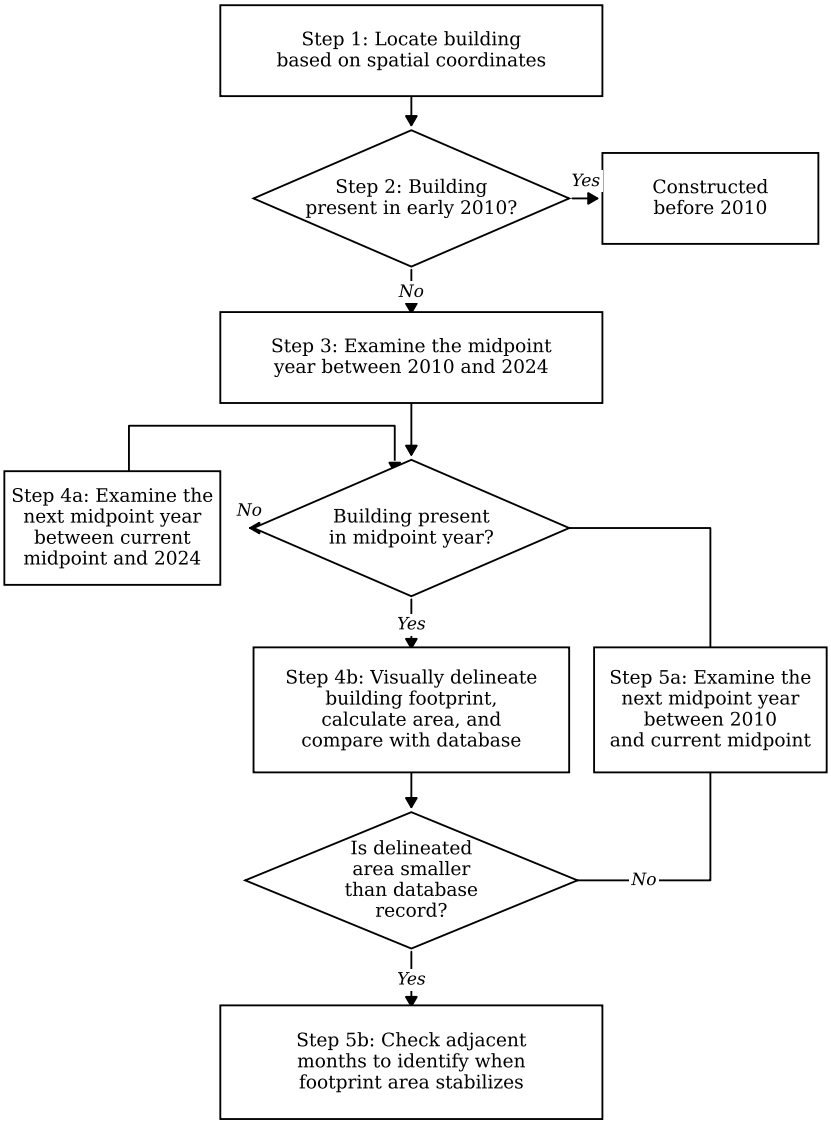
The approach relies on a binary iterative search algorithm applied to historical satellite imagery. We adopted a semi-automated protocol based on manual visual interpretation of Landsat (30-meter resolution, 2000–2018) and Sentinel-2 (10-meter resolution, 2018–2024) time series.[17] The workflow defines a search window between 2010 and the present, iteratively examining the midpoint year until the construction date is narrowed to the specific month of completion. This month-level precision ensures that no facility is misassigned across the three-year prediction horizon that anchors our CNN training pipeline. Records where construction occurred in multiple phases are coded based on the date the entire site was fully completed.

Figure A.2 demonstrates the first step of this workflow: establishing the extensive margin of entry. The left panel displays a baseline Landsat 5 image from 2010, while the right panel shows current high-resolution aerial imagery. By verifying that parcels currently occupied by data center campuses were undeveloped in 2010, we confirm these facilities were constructed after our baseline year—ensuring that the entry indicator ($Y_{it} = 1$) used to train the CNN captures genuine new investment rather than pre-existing infrastructure. As of August 2025, this process successfully resolved temporal data for 254 records that initially lacked construction date information, confirming 168 valid completions while flagging 63 records that could not be linked to a specific building due to unclear boundaries or absent identifiable features.

A further challenge in assembling the building-level panel is distinguishing data centers from observationally similar structures—particularly logistics warehouses—within the same industrial zones. This concern is especially acute given our reliance on geocoded building footprints to construct the entry indicator (Y_{it}) that serves as the dependent variable in the CNN. Figure A.3

[17] Automated spectral indices such as the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Built-up Index (NDBI) are unreliable at the building scale due to interference from cloud cover, shadows, and atmospheric noise.

Figure A.1—Remote Sensing of Data Center Entry Timing



Notes: This figure illustrates the hybrid remote-sensing approach used to determine data center entry timing (month and year of construction completion) for new construction facilities missing temporal records. The process employs a binary iterative search across historical satellite imagery (Landsat 5/7/8 for 2000–2018 and Sentinel-2 for 2018–2025) to narrow the search window efficiently. To address multi-phase expansion, the workflow includes a validation step comparing the visually delineated footprint against the aggregate gross square footage recorded in industry databases. The final entry date is defined as the month in which the building footprint stabilizes and matches the reported physical asset size.

illustrates the validation mechanism we employ to mitigate this source of measurement error. We manually delineate the building footprint visible in satellite imagery and calculate its aggre-

Figure A.2—Detecting New Data Center Construction

Notes: This figure demonstrates the validation process for establishing the extensive margin of data center entry. The left panel presents a Landsat 5 satellite image (30-meter resolution) captured in 2010, serving as the baseline observation for the start of the sample period. The right panel displays recent high-resolution aerial imagery from Google Maps (August 2025), with red markers indicating the current physical footprint of a data center campus. The comparison reveals that the specific land parcels currently occupied by the facility (circled in red) were undeveloped in 2010 (characterized by the spectral signature of soil and vegetation rather than impervious surfaces), thereby confirming the facility as a new entrant constructed after 2010. This baseline verification is the first step in the temporal identification workflow before applying the binary iterative search for the exact completion month.

gate surface area. This satellite-derived area is then cross-referenced against the gross square footage of data centers reported in industry records. A match within $\pm 15\%$ confirms the facility's identity, a tolerance that accounts for measurement discrepancies between satellite-derived roof area and administratively reported interior floor space (e.g., roof overhangs, covered loading docks, and multi-story structures where gross square footage exceeds the visible footprint). Facilities where visual footprints produce significant mismatches with administrative data, or where boundaries cannot be clearly defined (e.g., multiple buildings of similar size in proximity), are explicitly flagged and excluded, preserving the integrity of the entry variable used for model training and the campus-level blocking groups used for spatial cross-validation.

A.2 Construction of Spatial Determinants of Location Choice

This section provides detailed variable construction procedures for the five covariate categories summarized in Section 3. All covariates are mapped onto the unified 500-meter raster grid described in Section 4, spanning the 2010–2024 panel period. Table 1 provides a complete listing of

Figure A.3—Validation of Facility Identification via Physical Footprint Cross-Referencing

Notes: This figure illustrates the validation mechanism used to verify the accuracy of the remote sensing identification strategy. The image displays a high-resolution satellite view of a data center campus, with the final building footprints manually delineated (highlighted). To confirm the facility's identity and rule out false positives, the aggregate surface area of these delineated polygons is calculated and cross-referenced against the gross square footage reported in the sample. This procedure ensures that the structures identified via satellite correspond precisely to the specific facility in our dataset, effectively distinguishing data centers from similar logistics warehouses or manufacturing plants within the same industrial zones.

variables, resolutions, and sources.

Power Infrastructure. Because localized grid bottlenecks constrain data center development more heavily than broad regional electricity market conditions (Mamkhezri, Sun and Yang, 2025), we model power access at the highest feasible spatial resolution. We calculate the Euclidean distance from the centroid of each 500-meter grid cell to the nearest high-voltage transmission line, substation, and high-voltage node using geospatial data from the Homeland Infrastructure Foundation-Level Data (HIFLD). These three distance layers constitute the grid topology covari-

ates referenced in the counterfactual scenarios (Section 5.3).

To quantify local generation capacity, we integrate plant-level data from the EIA-860 survey, constructing a six-band panel of plant locations, nameplate capacities, and fuel types for the 2010–2024 period. Grid reliability is measured using utility-level outage metrics—specifically the System Average Interruption Duration Index (SAIDI) and the System Average Interruption Frequency Index (SAIFI)—derived from EIA-861 reliability reports. SAIDI also serves as the grid reliability channel perturbed in the Climate Stress counterfactual scenario. To capture broader electricity market conditions, we also incorporate state-level EIA-861 panel data on electricity sales, revenues, industrial pricing, and customer bases.

Broadband and Fiber Connectivity. Rather than relying on general broadband availability proxies, we measure the specific network infrastructure required for enterprise-grade data center operations. We construct a granular panel of terrestrial broadband deployment using Federal Communications Commission (FCC) Fixed Broadband Deployment data (Form 477), which is available from 2014 through 2021. For observations falling outside this window, we carry forward the nearest available year: the 2014 values are applied to all prior years in the panel, and the 2021 values are held constant through 2024.[18]

We exclude satellite connections and isolate Fiber-to-the-Premises (FTTP) technology at the census-block level to identify areas capable of the millisecond latency that data center interconnection requires, establishing a precise proxy for local fiber density. We supplement these physical infrastructure indicators with maximum advertised upload and download speeds, provider density, and a Simpson diversity index measuring the degree of competition among local broadband providers at the tract level. These connectivity channels capture both the physical presence of high-capacity networks and the competitive depth of the local telecommunications market.

Water Resources and Constraints. Data centers require continuous access to large volumes of cooling water, making hydrological conditions a binding physical constraint on site viability. We quantify baseline water stress using the World Resources Institute (WRI) Aqueduct 4.0 Water Risk Atlas. We complement this static measure with a dynamic assessment of moisture conditions

[18]FCC Form 477 data are available from 2014 through 2021. For years prior to 2014, we apply the earliest available (2014) values, reflecting the fact that broadband infrastructure is largely cumulative—fiber deployed by 2014 was overwhelmingly present in prior years. For years after 2021, we hold the 2021 values constant; FCC Form 477 was discontinued and replaced by the Broadband Data Collection (BDC) system, which uses a different reporting methodology that is not directly comparable.

using the annual mean Palmer Drought Severity Index (PDSI) derived from the PRISM climate model. Both variables are scaled to the 500-meter grid and span the full 2010–2024 panel.

Regulatory and Legislative Environment. To capture the institutional frictions that increasingly constrain data center siting—particularly in historically dense development corridors such as Northern Virginia—we construct a state-level legislative index using dictionary-based content analysis (DCA). Our corpus comprises 76 data-center-related bills introduced in 2025 across 27 states, sourced from the National Caucus of Environmental Legislators (NCEL). We categorize this legislation into seven policy domains: grid reliability, electricity costs and rates, reporting and planning requirements, environmental impacts, siting and local control, clean-energy mandates, and tax provisions. Bills are assigned to these categories based on keyword frequencies; for instance, terms like "resource adequacy" signal grid reliability, while "cost recovery" flags electricity rates. To quantify regulatory pressure, we weight each bill by its enforcement intensity (prioritizing binding prohibitions over exploratory study commissions) and legislative status (weighting enacted laws more heavily than introduced bills). The resulting scores are aggregated and normalized to a 0–10 scale. This index strictly isolates regulatory friction; pro-investment fiscal drivers are modeled separately through the tax incentive variables described below.

Cost Structure, Agglomeration, and Demand Shifters. We model the economic forces governing site selection as a balance between the pull of local demand and agglomeration and the push of capital and operating costs. We proxy land acquisition costs using the county-level Zillow Home Value Index (ZHVI) for mid-tier homes, which also reflects the infrastructure premium capitalized into parcels near data-center-ready sites (Qu et al., 2025). Construction labor costs are measured using the county-level average weekly wage in the construction sector from the Bureau of Labor Statistics (BLS). Variable operating costs are measured using state-level industrial electricity prices from EIA-861 and state-level hourly wages for computer-related occupations from the BLS.

The fiscal environment operates through both push and pull channels. Building on evidence that state taxation drives the spatial reallocation of business activity (Giroud and Rauh, 2019), we account for baseline tax burdens via state-level sales tax rates and county-level median property taxes. Targeted tax incentives function as pull factors, and because such subsidies can effectively boost hyperscale construction in qualifying areas (CBRE, 2013; Gargano and Giacoletti, 2025), we integrate state-level project development and investment tax incentives from the Panel Database

on Incentives and Taxes (PDIT).

Demand shifters and agglomeration economies are captured through tract- and county-level total population counts, the share of the population holding a bachelor's degree or above, and the county-level share of IT employment from the County Business Patterns (CBP). The dual tract/county resolution captures both localized neighborhood effects—such as zoning opposition tied to residential density—and broader regional labor market dynamics.

A.3 Visualizing the Model's Input Data

Figures A.4 and A.5 display the raw geographic data that the CNN evaluates when scoring a prospective site. Here, we provide a detailed reading of the spatial patterns visible in each panel and explain how they map to the covariate construction described in Appendix Section A.2.

Each figure presents three 32 km $\times$ 32 km catchment areas drawn from a random subsample of 2,000 locations in the model's inference dataset. From this subsample, we select three archetypes, each displayed as a column in the figure: the highest-probability realized site (High Suitability, left column), the undeveloped site with the richest infrastructure endowment yet low predicted probability (Hard Negative, center column), and the undeveloped site with the weakest infrastructure endowment (Rural Baseline, right column).

The rows decompose the 132-channel input tensor into three thematic groups: fiber and broadband connectivity (Row 1, 6 channels per year $\times$ 3 years = 18 channels), energy infrastructure comprising grid topology and generation capacity (Row 2, 9 channels per year $\times$ 3 years = 27 channels), and physical constraints capturing water stress and drought severity (Row 3, 2 channels per year $\times$ 3 years = 6 channels). Each channel is standardized by subtracting its spatial mean and dividing by its spatial standard deviation within the catchment area, converting raw values (megawatts, kilometers, index scores) to z-scores on a common unitless scale. Extreme values are winsorized at the 2nd and 98th percentiles to suppress artifacts at the edges of the raster window.

For the connectivity and energy rows, lighter colors indicate higher density, greater capacity, or closer proximity to infrastructure, while darker colors indicate absence. For the physical constraints row, blue tones indicate low water stress and the absence of drought, red tones indicate elevated water risk, and the neutral white midpoint reflects areas with no meaningful spatial variation in hydrological conditions.

Hyperscale Cloud Sites (Figure A.4). The high-suitability site ($P = 0.97$) exhibits the visual signature of a major infrastructure corridor. The connectivity row displays rich spatial heterogeneity—bright yellow concentrations interspersed with dark voids—indicating dense fiber nodes surrounded by areas beyond network reach. A standard regression would summarize this pattern as a single tract-level average, discarding the precise spatial layout; the CNN instead preserves the full geographic structure and learns that multiple fiber routes converging within the same neighborhood signal the redundancy required for enterprise-grade connectivity.

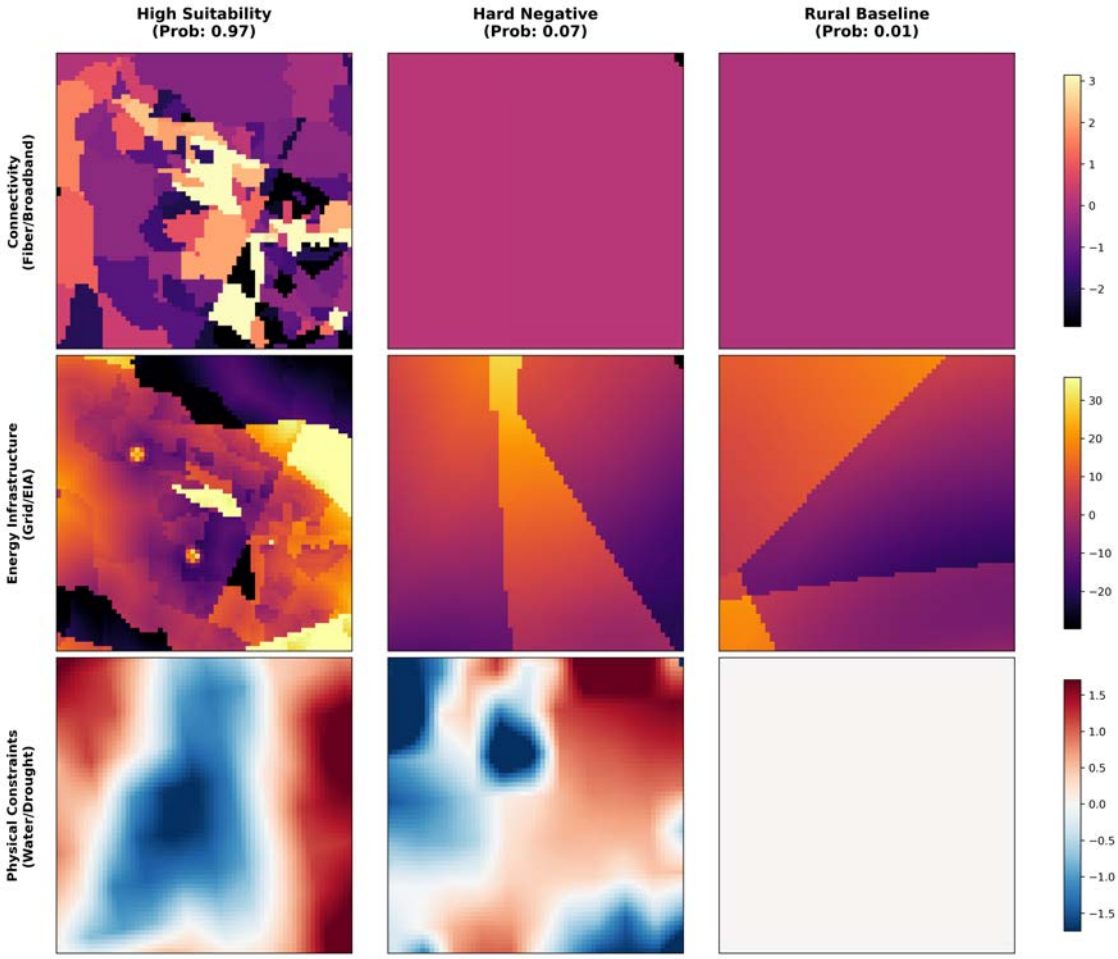
The energy row reveals a similarly complex structure: bright yellow features trace high-voltage transmission lines intersecting at substations, forming a grid junction that provides the power redundancy required for hyperscale operations. The physical constraints row shows neutral tones throughout, indicating neither elevated water stress nor severe drought risk—a stable hydrological environment well suited to large-scale cooling.

The hard negative ($P = 0.07$) illustrates the model's capacity to detect a binding deficiency that disqualifies an otherwise promising site. The energy row displays substantial generation capacity (warm orange tones across much of the catchment), which would register as favorable in any univariate screen. However, the connectivity row is nearly uniform dark purple—featureless and lacking the bright aggregation nodes visible in the high-suitability case. This indicates a location beyond the reach of dense fiber networks, a fatal deficiency for hyperscale cloud facilities that require ultra-low-latency interconnection between distributed server clusters. The CNN integrates these conflicting signals and correctly downgrades the site despite its energy endowment. The physical constraints row shows moderate conditions that neither help nor harm, confirming that fiber absence alone drives the rejection.

The rural baseline ($P = 0.01$) serves as a control. The connectivity and energy rows are uniformly dark, reflecting the near-complete absence of digital and power infrastructure. The physical constraints row appears as a flat, neutral field (light gray), indicating that the PDSI and WRI covariates report no meaningful variation in this remote area. This confirms that the model assigns near-zero suitability only when foundational infrastructure is absent across all dimensions, not when a single covariate happens to be low.

Third-Party Colocation Sites (Figure A.5). The high-suitability site ($P = 0.92$) displays a qualitatively different spatial signature from its hyperscale counterpart, reflecting the distinct siting logic of multi-tenant colocation facilities. The connectivity row shows dense, spatially concen-

Figure A.4—Visualizing the Spatial Determinants of Hyperscale Cloud Site Selection



Notes: This figure displays the raw geographic data (input tensor) evaluated by the CNN when scoring three representative hyperscale cloud locations. Columns represent location archetypes: High Suitability (left, $P = 0.97$), Hard Negative (center, $P = 0.07$), and Rural Baseline (right, $P = 0.01$). Rows decompose the 132-channel input into three thematic groups: fiber and broadband connectivity (Row 1), energy infrastructure including grid topology and generation capacity (Row 2), and physical constraints including water stress and drought severity (Row 3). Each panel covers a 32 km $\times$ 32 km catchment area at 500-meter resolution. Values on the color scales represent normalized standard deviations (z-scores), winsorized at the 2nd and 98th percentiles. For the connectivity and energy rows, lighter colors indicate higher density, greater capacity, or closer proximity to infrastructure, while darker colors indicate absence. For the physical constraints row, blue tones indicate low water stress and the absence of drought, red tones indicate elevated water risk, and the neutral white midpoint reflects areas with no meaningful spatial variation in hydrological conditions. The three sites were selected from a random 2,000-location scan to represent the extremes of the model's scoring distribution.

trated fiber coverage (bright yellow clusters) consistent with proximity to a metropolitan telecommunications hub. The energy row exhibits a complex, high-intensity pattern with multiple bright

features indicating nearby substations and transmission corridors. Critically, the physical constraints row reveals a blue-dominant pattern in the water stress layer, indicating low baseline water risk—a favorable condition for facilities relying on evaporative cooling systems.

The hard negative ($P = 0.08$) demonstrates a different failure mode. Unlike the hyperscale case, both connectivity and energy layers display meaningful infrastructure presence. The disqualifying factor is visible in the physical constraints row, which lacks the blue tones (low water stress) seen in the high-suitability site and instead displays pronounced red tones indicating elevated water risk—a binding constraint for facilities relying on water-intensive cooling. The CNN correctly rejects the site despite adequate connectivity and energy, illustrating how the binding constraint can differ across market segments. The rural baseline ($P = 0.01$) mirrors the hyperscale control: uniformly dark connectivity and energy layers with a featureless physical constraints field, confirming that the model assigns near-zero suitability when foundational infrastructure is absent across all dimensions.

A.4 Evaluation Metrics and Cross-Domain Validation

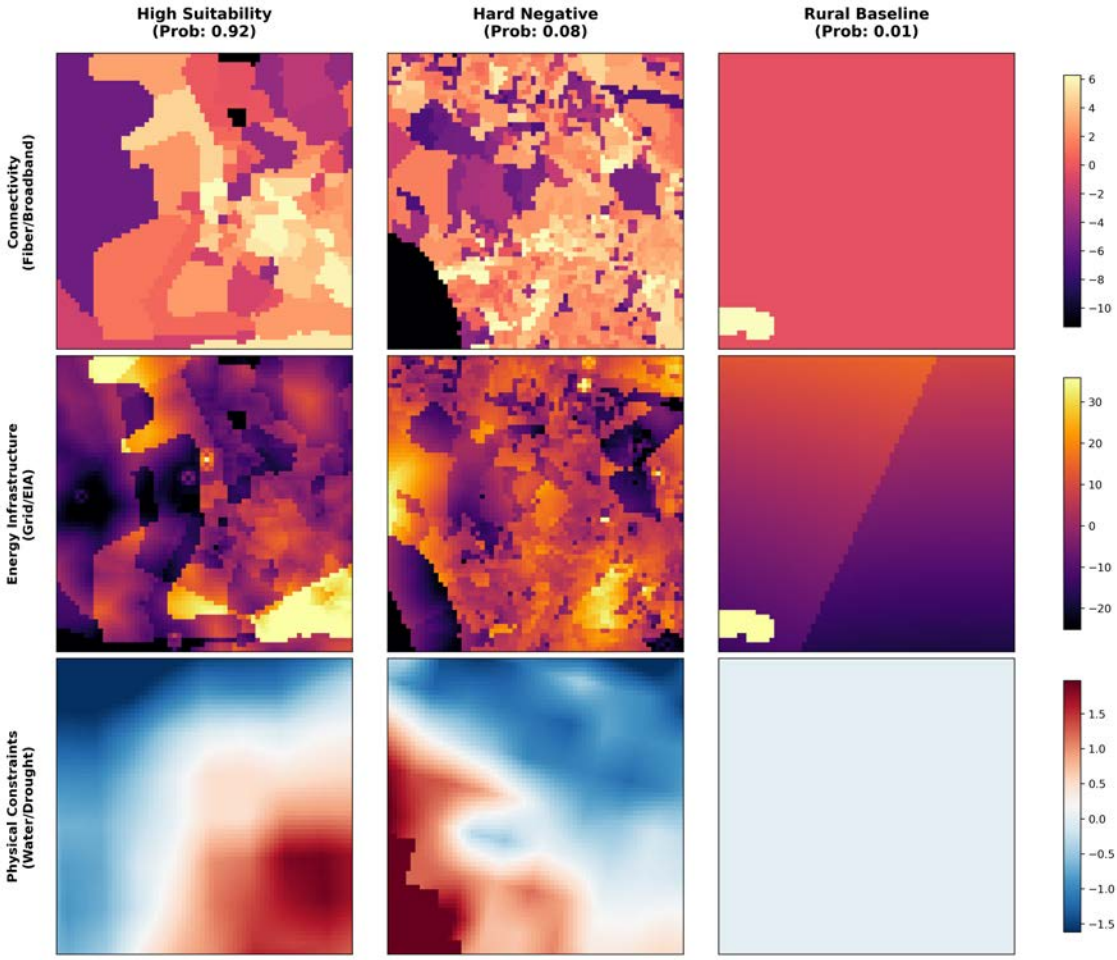
This section provides formal definitions and economic interpretations for the five evaluation metrics reported in Table A.1. Standard accuracy metrics are misleading when our outcome of interest—data center entry—occurs in fewer than 1% of land parcels. An algorithm that blindly predicts no development everywhere would achieve 99% accuracy while failing to identify a single viable site. The metrics below are specifically designed for this rare-event setting, and each maps onto a distinct stage of the real estate investment process.

ROC AUC: Overall Ranking Accuracy. The Area Under the Receiver Operating Characteristic Curve (ROC AUC) measures the model's ability to rank sites correctly across the full scoring distribution. Formally, for a set of n^+ realized sites and n^- undeveloped parcels, the ROC AUC equals the probability that a randomly drawn realized site receives a higher predicted score than a randomly drawn undeveloped parcel:

$$\text{ROC AUC} = \frac{1}{n^+ \cdot n^-} \sum_{i \in \mathcal{S}^+} \sum_{j \in \mathcal{S}^-} \mathbf{1}[\hat{p}_i > \hat{p}_j]$$

where $\hat{p}_i$ denotes the model's predicted suitability for location i. A value of 0.5 indicates random ranking (no discriminatory power); a value of 1.0 indicates perfect separation. In practical terms,

Figure A.5—Visualizing the Spatial Determinants of Third-Party Colocation Site Selection



Notes: This figure displays the raw geographic data (input tensor) evaluated by the CNN when scoring three representative third-party colocation locations. Columns represent location archetypes: High Suitability (left, $P = 0.92$), Hard Negative (center, $P = 0.08$), and Rural Baseline (right, $P = 0.01$). Rows decompose the 132-channel input into three thematic groups: fiber and broadband connectivity (Row 1), energy infrastructure including grid topology and generation capacity (Row 2), and physical constraints including water stress and drought severity (Row 3). Each panel covers a 32 km $\times$ 32 km catchment area at 500-meter resolution. Values on the color scales represent normalized standard deviations (z-scores), winsorized at the 2nd and 98th percentiles. For the connectivity and energy rows, lighter colors indicate higher density, greater capacity, or closer proximity to infrastructure, while darker colors indicate absence. For the physical constraints row, blue tones indicate low water stress and the absence of drought, red tones indicate elevated water risk, and the neutral white midpoint reflects areas with no meaningful spatial variation in hydrological conditions. The three sites were selected from a random 2,000-location scan to represent the extremes of the model's scoring distribution.

an ROC AUC of 0.828 (Teacher Baseline) means that if an investor randomly selects one realized site and one undeveloped parcel, the model assigns the higher score to the correct location 82.8%

Table A.1—Cross-Domain Predictive Performance (Structural Prior Evaluation)

Model Evaluated	ROC AUC	PR AUC	Lift (Top 10%)	Recall (Top 1%)	Brier Score
Teacher (Baseline)	0.828	0.405	4.114	0.068	0.103
	(0.014)	(0.043)	(0.390)	(0.010)	(0.019)
Teacher (Structural)	0.766	0.331	3.465	0.048	0.103
	(0.087)	(0.144)	(2.088)	(0.024)	(0.016)
Student (Fine-Tuned)	0.781	0.341	3.633	0.042	0.115
	(0.064)	(0.156)	(1.885)	(0.029)	(0.013)

Notes: This table reports out-of-sample predictive performance across five spatial cross-validation folds, with means on the main line and standard deviations in parentheses. All folds use campus-level blocking—grouping all buildings owned by the same firm within a county—to prevent data leakage from phased construction at the same site. The *Teacher (Baseline)* is trained and tested on third-party colocation sites, establishing the strongest benchmark. The *Teacher (Structural)* applies the same third-party-trained model to hyperscale cloud sites without retraining, testing whether siting rules generalize across market segments. The *Student (Fine-Tuned)* preserves the Teacher's geographic pattern recognition but retrains only the final scoring weights on cloud data, adapting the model to the distinct preferences of hyperscale operators. ROC AUC reports the probability that a randomly drawn realized site receives a higher score than a randomly drawn undeveloped parcel. PR AUC evaluates how often the model's most confident positive recommendations correspond to actual development, penalizing false positives in this rare-event setting. Lift@10% reports how many times more likely parcels in the model's top decile are to contain a realized site compared to random selection. Recall@1% reports the share of all realized sites captured within the model's most selective top percentile of recommendations. The Brier score measures the average squared gap between predicted probabilities and observed outcomes (lower indicates better-calibrated estimates).

of the time. This metric captures global ranking quality across all possible decision thresholds, making it the primary measure of whether the model has learned to distinguish infrastructure-rich environments from unsuitable terrain.

PR AUC: Reliability of High-Confidence Predictions. The Area Under the Precision-Recall Curve (PR AUC) evaluates how trustworthy the model's most confident recommendations are. Unlike ROC AUC, which weights true negatives equally with true positives, PR AUC focuses exclusively on the quality of positive predictions—the parcels the model flags as suitable. Precision measures the fraction of flagged sites that are actually realized facilities; recall measures the fraction of all realized facilities that the model successfully flags. The PR AUC summarizes this trade-off across all scoring thresholds:

$$\text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)}, \quad \text{Recall}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}$$

where TP, FP, and FN denote true positives, false positives, and false negatives at threshold t. In rare-event settings, the baseline PR AUC equals the prevalence rate (approximately 0.01), so even modest values represent substantial improvement over random selection. For institutional investors, PR AUC answers a critical underwriting question: when the model identifies a parcel as highly suitable, how often does that recommendation correspond to a location where development actually occurred? A higher PR AUC means fewer wasted site visits and due diligence expenditures on false leads.

Lift@10%: Screening Efficiency. The top-decile lift factor (Lift@10%) quantifies how much more efficiently the model identifies viable sites compared to random search. It is computed as the ratio of the hit rate within the model's top 10% of scored parcels to the overall base rate of development:

$$\text{Lift@10\%} = \frac{\text{Precision in top 10\% of predictions}}{\text{Overall prevalence of entry}}$$

A Lift of 4.1$\times$ (Teacher Baseline) means that parcels in the model's top 10% of recommendations are 4.1 times more likely to contain a realized data center than parcels selected at random from the full national landscape. This metric directly mirrors the initial screening stage of commercial real estate due diligence, where developers narrow thousands of candidate parcels to a manageable shortlist. For a portfolio manager evaluating hundreds of potential acquisitions, a Lift of 4.1$\times$ translates into a fourfold reduction in the number of parcels requiring costly on-the-ground evaluation.

Recall@1%: Top-Tier Candidate Identification. Recall in the top percentile measures the fraction of all realized data center sites that fall within the model's top 1% of nationally scored parcels:

$$\text{Recall@1\%} = \frac{|\{i \in \mathcal{S}^+ : \text{rank}(\hat{p}_i) \leq 0.01 \cdot N\}|}{|\mathcal{S}^+|}$$

where N is the total number of scored parcels. This metric captures the model's ability to concentrate realized development into an extremely selective shortlist. A Recall@1% of 0.068 (Teacher Baseline) indicates that nearly 7% of all actual data center sites are captured within the model's most exclusive top 1% of recommendations—a 7$\times$ concentration relative to what random selection would achieve. For developers operating under strict capital constraints who can only evaluate a handful of locations, this metric determines whether the model's highest-confidence picks contain

a meaningful share of the market's actual investment targets.

Brier Score: Probability Calibration. The Brier score measures the mean squared deviation between the model's predicted probabilities and observed binary outcomes:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$$

where $y_i \in \{0, 1\}$ is the realized entry indicator. Unlike the ranking-based metrics above, the Brier score evaluates whether the model's numerical probability estimates are well calibrated—that is, whether a predicted probability of 0.30 corresponds to a location where entry occurs roughly 30% of the time. Lower values indicate better calibration (0.0 is perfect). This property matters for downstream applications that require reliable probability inputs rather than mere ordinal rankings, such as the counterfactual simulations, where the algorithm selects a fixed number of highest-scoring undeveloped cells each year to simulate new facility entry. A poorly calibrated model might rank sites correctly but produce probability estimates that are systematically too high or too low, distorting the annual selection of new entry sites in the counterfactual simulations and biasing the simulated footprint.

Comparative Performance. All five metrics are evaluated using 5-fold spatial cross-validation, where campus-level blocking ensures that all buildings owned by the same firm within a given county are assigned to the same fold—preventing the model from memorizing the geographic signature of phased developments and forcing generalization to entirely new campuses. Each fold serves as the holdout exactly once.

Table A.1 reports the mean and standard deviation (S.D.) across the five holdout folds. The S.D. captures geographic heterogeneity in model performance: folds containing major infrastructure hubs (e.g., Northern Virginia) tend to produce higher within-fold AUC than folds dominated by secondary markets, reflecting the greater density of discriminating infrastructure features in established corridors. By averaging across all five folds, the reported metrics represent the model's expected performance in an arbitrary new regional market, avoiding upward bias from evaluating exclusively in data-rich environments.

The Teacher (Baseline), trained and tested exclusively on third-party colocation sites, establishes the strongest benchmark among the three models. Its ROC AUC of 0.828 (S.D. 0.014) in-

dicates that the model correctly ranks a realized site above a random undeveloped parcel nearly 83% of the time, with minimal variation across geographic folds. The PR AUC of 0.405 confirms that the model's most confident positive predictions are reliable despite the extreme class imbalance (baseline prevalence of approximately 1%). The Lift of 4.1× and Recall@1% of 0.068 together demonstrate strong screening efficiency: the top decile of recommendations is four times more efficient than random search, and nearly 7% of all realized sites are concentrated in the model's most selective top percentile. The Brier score of 0.103 indicates well-calibrated probability estimates.

Applying the Teacher directly to the hyperscale cloud sample without retraining (Teacher Structural) tests whether third-party siting rules generalize across market segments. Performance declines across all ranking metrics: ROC AUC falls to 0.766, PR AUC to 0.331, Lift to 3.5×, and Recall@1% to 0.048. Notably, the S.D. also increases substantially (e.g., ROC AUC S.D. rises from 0.014 to 0.087), indicating that the transferred rules perform unevenly across regions—working well in markets where third-party and cloud siting logic overlap (e.g., fiber-rich corridors) but poorly in regions where hyperscale operators pursue distinct strategies (e.g., remote power-rich locations). The Brier score remains unchanged at 0.103, suggesting that the probability scale is preserved even as ranking accuracy deteriorates. This pattern confirms that the two segments share foundational infrastructure requirements but diverge in how they weigh competing factors.

The Student (Fine-Tuned) model partially closes the gap between the Teacher Structural and the Teacher Baseline by locking the Teacher's geographic feature extraction layers and retraining only the final scoring weights on cloud data. ROC AUC rises to 0.781, PR AUC to 0.341, and Lift to 3.6×. The fold-to-fold variability also decreases relative to the Teacher Structural (ROC AUC S.D. falls from 0.087 to 0.064), indicating more consistent performance across diverse geographic regions. The Recall@1% of 0.042 remains slightly below the Teacher Structural's 0.048, reflecting a trade-off: fine-tuning sharpens the model's ability to discriminate among plausible hyperscale sites (higher AUC) but redistributes probability mass away from the extreme tail, slightly reducing concentration in the top percentile. The Brier score increases modestly to 0.115, a consequence of the reduced training sample size for the cloud segment (478 sites versus 1,286 for third-party), which limits the precision of probability calibration.

Taken together, these results validate the transfer learning strategy. A model trained on third-party data (Teacher Structural) retains meaningful predictive power for cloud sites—confirming the shared infrastructure prior—while the Student's recovery of ranking accuracy through fine-tuning of only the final scoring layer suggests that the two segments differ primarily in how they

weigh a common set of geographic features rather than in which features they consider. For practitioners, these metrics suggest that the model's top-decile recommendations provide a reliable shortlist for site prospecting in both market segments, with the caveat that probability estimates for the cloud segment carry somewhat greater uncertainty due to the smaller training sample.

A.5 Local Feature Attribution and Driver Heterogeneity

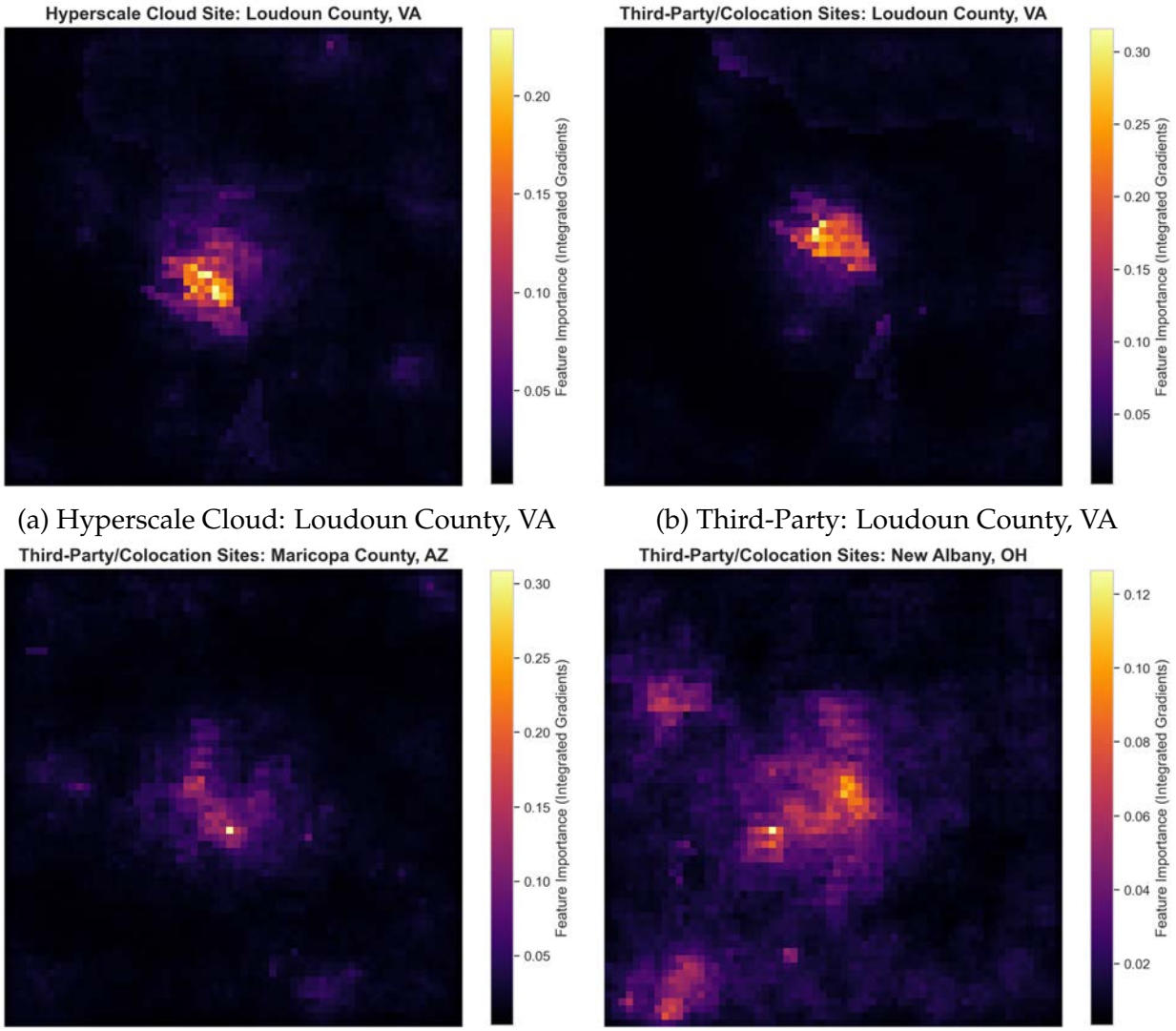
Here, we provide three complementary views of the Integrated Gradients (IG) attribution results that reveal finer spatial and distributional detail: local heatmaps of attribution intensity (Figure A.6), aggregate positive and negative factor contributions (Figure A.7), and the distributional spread of driver strength conditional on dominance (Figure A.8).

Spatial Heatmaps of Local Attribution. Figure A.6 maps the pixel-level IG attribution scores for four case study locations, revealing the highly localized and spatially heterogeneous nature of the factors driving site selection. Each heatmap covers the same 32 km $\times$ 32 km catchment area that the CNN evaluates, with brighter pixels indicating locations where the underlying infrastructure covariates exert the strongest positive influence on the model's predicted suitability.

Panels (a) and (b) present the same geographic area—Loudoun County, Virginia, the largest data center market in the world—evaluated separately by the hyperscale cloud model and the third-party colocation model. Both panels show tightly concentrated bright spots rather than diffuse regional gradients, confirming that the model's siting decisions are driven by precise micro-geographic features (such as proximity to specific substations or fiber junctions) rather than broad county-level characteristics. Notably, the two models pinpoint a remarkably similar core despite serving distinct market segments. This spatial overlap suggests that in a mature, infrastructure-dense hub like Loudoun County, the physical scarcity of viable power and fiber corridors forces extreme spatial agglomeration regardless of the operator's business model—a finding consistent with the shared structural prior documented in the cross-domain evaluation.

Panels (c) and (d) extend the analysis to two additional third-party markets. Maricopa County, Arizona (panel c) displays a single dominant bright cluster, suggesting that the model identifies one highly localized infrastructure nexus within a broader metropolitan area. New Albany, Ohio (panel d) shows a comparatively diffuse and lower-intensity pattern, with attribution scores an order of magnitude smaller than those observed in Loudoun or Maricopa. This contrast reflects the relative infrastructure maturity of these markets: established hubs generate sharp, high-

Figure A.6—Spatial Heatmaps of Local Feature Importance in Data Center Siting



(a) Hyperscale Cloud: Loudoun County, VA

(b) Third-Party: Loudoun County, VA

(c) Third-Party: Maricopa County, AZ

(d) Third-Party: New Albany, OH

Notes: This figure maps the pixel-level contribution of infrastructure covariates to the model's suitability score within a 32 km $\times$ 32 km catchment area. Brighter pixels indicate stronger positive influence on predicted suitability; dark pixels indicate minimal contribution. Panels (a) and (b) compare hyperscale cloud and third-party attribution within Loudoun County, Virginia. Panels (c) and (d) contrast third-party attribution in Maricopa County, Arizona, and New Albany, Ohio. Attribution is computed via Integrated Gradients, which measures each feature's contribution to the suitability score by gradually transforming a neutral baseline (the average infrastructure profile of 100 randomly sampled national locations) into the site's actual observed profile over 50 incremental steps.

magnitude attribution signals, while emerging markets produce weaker and more spatially distributed signals. For policymakers, these heatmaps pinpoint the specific sub-county locations

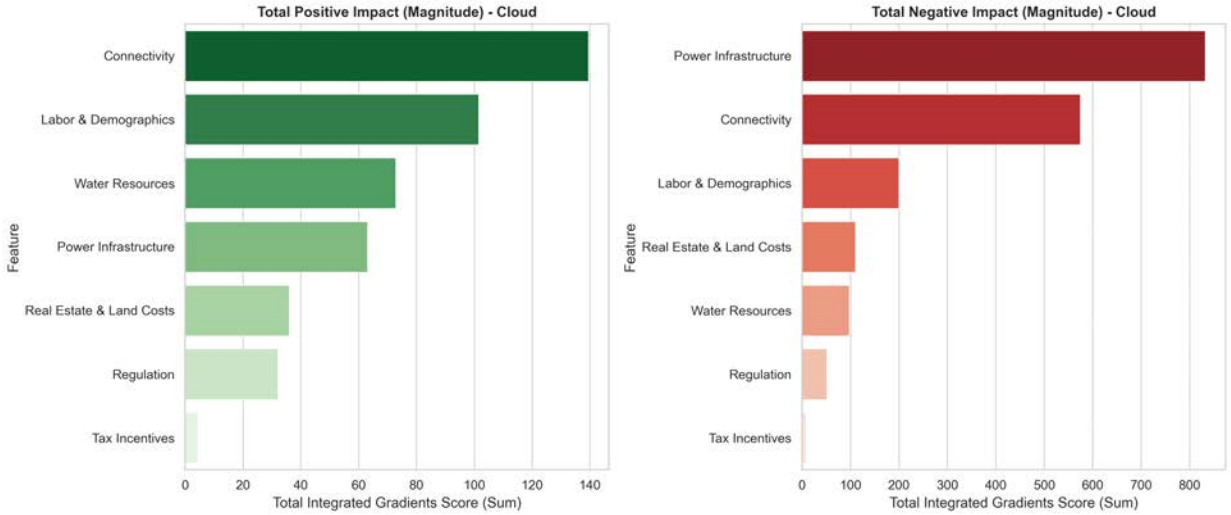
where targeted infrastructure investments—such as substation upgrades or fiber extensions—would most effectively raise local suitability scores.

Aggregate Positive and Negative Factor Attribution. Figure A.7 decomposes the total IG attribution into positive contributions (features that pull the model toward predicting entry) and negative contributions (features that push the model away), aggregated across a 1,000-site random sample for each market segment. This decomposition complements the permutation importance analysis reported in the main text, but answers a fundamentally different question. Permutation importance asks: if we scramble a feature, how much worse does the model perform across all locations? It identifies which features the model depends on most to make accurate predictions nationwide. IG attribution instead asks: at a specific site, how much does each feature raise or lower that site's individual suitability score? It reveals the directional contribution of each factor at the parcel level. A feature can rank low in permutation importance (because it rarely varies enough to affect national accuracy) yet generate large IG attribution at specific sites where it happens to be unusually strong or weak.

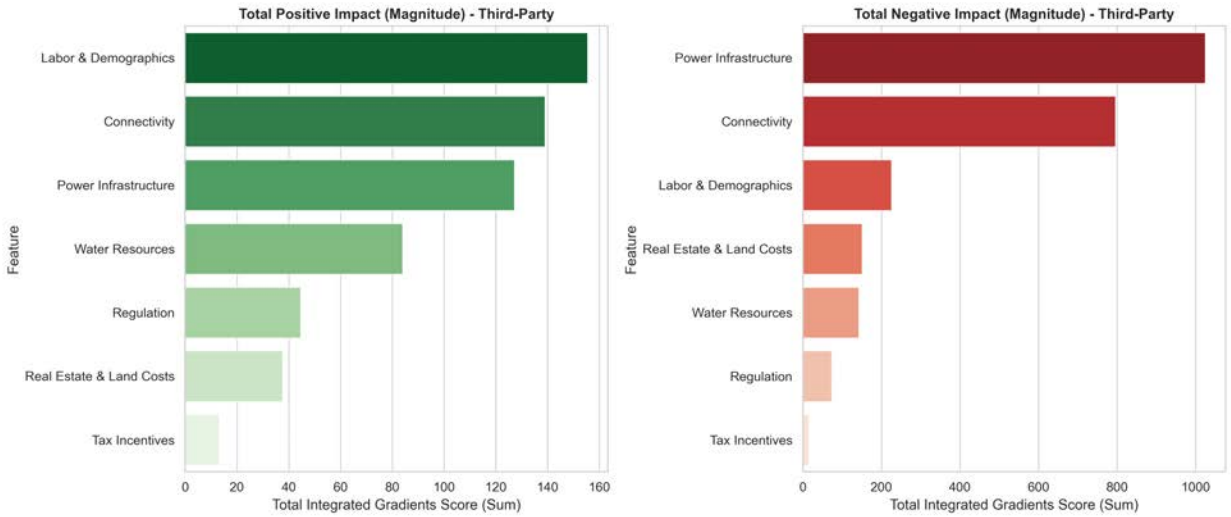
For cloud providers (panel a), connectivity dominates the positive attribution, followed by labor and demographics, water resources, and power infrastructure. This ranking differs from the permutation importance hierarchy, where power and connectivity jointly dominated. The distinction is economically meaningful: power infrastructure is the feature whose removal most damages the model's national accuracy (high permutation importance), but at locations where hyperscale facilities actually locate, it is connectivity that most actively elevates the suitability score (high positive IG attribution). Power access functions primarily as a pass/fail screen—its presence is necessary but does not strongly differentiate among sites that already meet the threshold. On the negative side, power infrastructure generates the largest barrier effect by a wide margin, followed by connectivity and labor. This asymmetry confirms the screening logic: inadequate power access produces the most intense negative signal, effectively disqualifying a site regardless of other advantages, while abundant connectivity produces the strongest positive pull.

For third-party colocation facilities (panel b), labor and demographics generate the largest positive attribution, followed closely by connectivity and power infrastructure. This ordering reflects the colocation sector's proximity-driven model: sites near large, educated populations receive the strongest positive boost because these locations serve local enterprise demand. The negative side mirrors the hyperscale pattern, with power infrastructure again producing the dominant barrier

Figure A.7—Total Factor Attribution in Local Data Center Siting Decisions



(a) Hyperscale Cloud Providers



(b) Third-Party Data Centers

Notes: This figure decomposes the Integrated Gradients attribution into aggregate positive contributions (green bars, left) and aggregate negative contributions (red bars, right) for each feature category, summed across a random sample of 1,000 sites. Positive attribution indicates features that actively pull the model toward predicting entry at a given location; negative attribution indicates features that push the model away from predicting entry. The horizontal axis reports the total summed IG score across all sites in the sample. Panel (a) reports results for hyperscale cloud providers; panel (b) reports results for third-party colocation facilities. The asymmetry between positive and negative magnitudes within a given feature category reveals whether that factor primarily functions as an attractor (large positive, small negative) or as a screening barrier (large negative, small positive).

effect. The gap between positive and negative magnitudes is more compressed for third-party facilities than for hyperscale operators, indicating that colocation siting decisions involve a more balanced weighting of competing factors rather than the sharp threshold screening that characterizes hyperscale deployment.

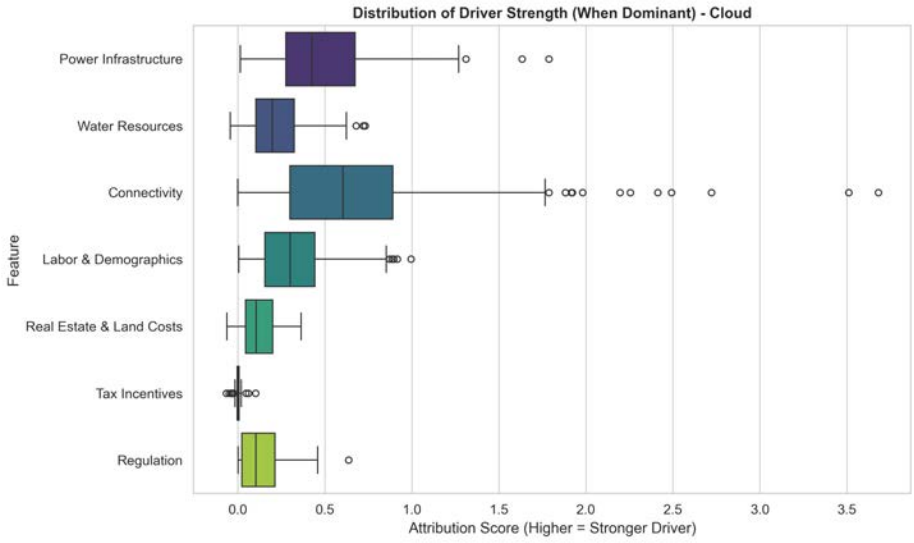
Tax incentives and regulation both contribute minimal positive or negative attribution in both segments. For tax incentives, this is consistent with the incentive paradox: without adequate power and fiber, no tax break can make a site viable, so incentives influence the final choice among sites that have already passed the infrastructure screen rather than driving the initial suitability score. Regulation similarly registers as a weak direct contributor to individual site scores, likely because state-level legislative conditions vary too slowly across space to differentiate neighboring parcels within the same 50-mile choice set.

Distribution of Driver Strength. Figure A.8 presents box plots of the IG attribution score for each feature category, conditional on that category being the dominant (highest-scoring positive) driver at a given site. This conditional distribution reveals how intensely each factor binds when it matters most, and how variable that binding strength is across locations.

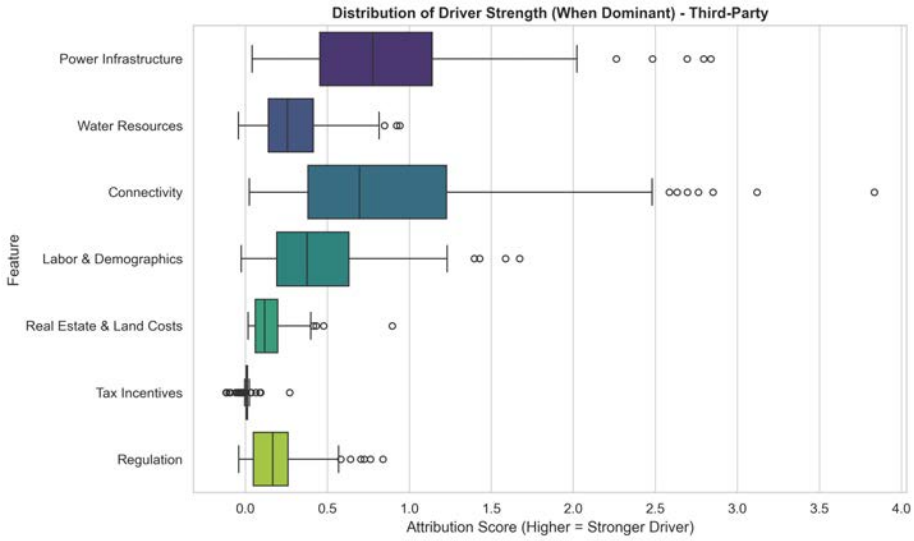
For hyperscale cloud providers (panel a), power infrastructure displays the highest median attribution score when dominant, with a wide interquartile range and several extreme outliers. This indicates that when power access is the decisive factor at a hyperscale site, its contribution to the suitability score is both large and highly variable, reflecting the enormous differences in grid capacity across locations. Connectivity shows a similar pattern with high outliers but a somewhat lower median. By contrast, labor and demographics, real estate, tax incentives, and regulation all cluster near the lower end of the distribution with narrow interquartile ranges, confirming that these factors serve as baseline requirements with relatively uniform contributions rather than as sharp differentiators. Water resources shows moderate median strength with limited variance.

For third-party colocation facilities (panel b), the distributional picture is notably different. Power infrastructure retains the highest median, but connectivity now shows substantially higher outliers than in the hyperscale case. This suggests that at specific colocation sites, fiber access can exert an exceptionally strong pull, likely reflecting locations near major carrier hotels or internet exchange points where interconnection density creates powerful agglomeration effects. Labor and demographics show a broader interquartile range for third-party sites than for hyperscale, consistent with the colocation sector's greater sensitivity to local market depth. Real estate and land

Figure A.8—Distribution of Primary Location Drivers



(a) Hyperscale Cloud Providers



(b) Third-Party Data Centers

Notes: For each of 1,000 randomly sampled locations, we classify the feature category with the highest positive Integrated Gradients attribution as the site's dominant driver. The box plots display the distribution of attribution scores for each category across the subset of sites where it ranks first, showing the median, interquartile range, and outliers. A wide spread indicates that the factor's influence varies substantially across locations; extreme outliers identify sites where a single infrastructure dimension dominates the suitability score. Panel (a) reports results for hyperscale cloud providers; panel (b) for third-party colocation facilities.

costs also display meaningful variance, confirming that land affordability serves as a binding constraint at a non-trivial subset of colocation sites, particularly in high-cost urban markets where the tension between proximity and affordability is most acute. Tax incentives and regulation remain compressed near zero in both segments.

Together, these three figures provide a granular view of the siting logic that complements the aggregate rankings. The heatmaps demonstrate that infrastructure advantages are spatially precise rather than regionally diffuse. The positive/negative decomposition reveals the asymmetric role of power infrastructure as the dominant barrier and connectivity as the dominant attractor. The box plots confirm that the binding strength of any given factor varies substantially across locations, underscoring the value of a flexible nonlinear model that can adapt its scoring weights to the local infrastructure environment rather than applying a single fixed coefficient nationwide.

A.6 Projections and Spatial Divergence Maps

This appendix provides the complete state-level projections (Tables A.2 and A.3) and county-level spatial divergence maps (Figures A.9–A.13) that underlie the summary exhibits presented in the main text.

State-Level Capacity Projections. Tables A.2 and A.3 report the imputed electrical capacity (in gigawatts, GW hereafter) for each contiguous state and the District of Columbia under all six scenarios. The baseline column reflects the empirical 2024 footprint; the six projection columns report the cumulative 2030 capacity after six years of rolling-horizon simulation. Capacity is converted from spatial footprint at a fleet-average density of 0.44 MW per acre. Several patterns merit attention beyond the top-three rankings discussed in the main text.

First, the magnitude of projected expansion varies enormously across scenarios. Under the AI Boom, Texas reaches over 23 GW by 2030—roughly six times its 2024 baseline of 3.9 GW—while under Regulatory Headwinds the same state reaches only 8.7 GW, less than half the AI Boom projection. This sensitivity to the policy environment is not unique to Texas: Montana, for instance, ranges from 2.2 GW under Climate Stress to 10.5 GW under AI Boom, a nearly fivefold difference driven by the scenario assumptions applied to a common baseline.

Second, certain states exhibit remarkable stability across scenarios, suggesting that their competitive position is robust to the specific shocks we simulate. Virginia, for example, ranges from 3.6 GW (Regulatory Headwinds) to 5.1 GW (AI Boom)—a comparatively narrow band relative to

Table A.2—Projected State-Level Data Center Capacity (Part 1)

State	Base 2024	Projected Capacity by 2030 (GW)					
		Status Quo	AI Boom	Climate Stress	Green Grid	Autarkic Expansion	Regulatory Headwinds
Alabama	0.35	2.34	4.13	1.58	2.07	2.53	1.74
Arizona	1.30	4.76	8.51	2.58	3.78	6.55	3.04
Arkansas	0.08	2.01	3.59	0.95	1.69	2.58	1.39
California	3.62	9.32	13.26	6.20	8.73	11.50	8.10
Colorado	0.95	4.35	7.83	2.26	3.81	6.12	2.77
Connecticut	0.38	0.79	0.82	0.54	0.73	0.71	0.60
Delaware	0.14	0.46	0.52	0.22	0.41	0.43	0.27
DC	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Florida	1.74	3.51	5.76	2.53	3.23	3.59	2.96
Georgia	1.22	2.77	4.19	1.55	2.31	2.96	1.98
Idaho	0.14	2.04	4.30	0.73	1.77	4.62	1.01
Illinois	1.77	3.42	5.38	2.47	3.15	3.94	3.02
Indiana	0.46	1.77	2.66	1.11	1.66	2.26	1.06
Iowa	0.63	3.23	4.68	2.01	2.88	3.53	2.07
Kansas	0.24	3.67	5.79	1.28	3.37	3.86	1.69
Kentucky	0.38	1.69	2.64	0.87	1.52	2.15	1.36
Louisiana	0.19	2.15	2.99	1.22	1.96	2.39	1.77
Maine	0.08	0.82	1.93	0.24	0.65	1.49	0.46
Maryland	0.52	0.90	1.01	0.71	0.82	0.92	0.73
Massachusetts	0.82	1.11	1.41	0.92	1.01	1.03	0.92
Michigan	0.79	3.26	6.52	1.71	2.75	5.27	2.15
Minnesota	0.60	3.97	6.31	1.82	3.13	4.00	2.77
Mississippi	0.22	1.93	3.18	1.14	1.74	1.79	1.44
Missouri	0.60	2.69	4.21	1.52	2.45	3.32	1.96
Montana	0.11	4.62	10.49	2.20	3.91	8.43	2.80

Notes: This table reports the imputed state-level data center electrical capacity in gigawatts (GW) for the contiguous United States and the District of Columbia. The baseline column reflects the empirical capacity observed in 2024, computed by summing the spatial footprint of all realized data center pixels and converting to electrical capacity at a fleet-average density of 0.44 MW per acre. Subsequent columns project the cumulative capacity by 2030 under six counterfactual scenarios generated by the rolling-horizon recursive simulation described in Section 5. All scenarios share the same 2024 empirical baseline; differences reflect the cumulative effect of scenario-specific covariate perturbations compounded over six simulation years.

its 2.3 GW baseline. This stability reflects Virginia's uniquely mature infrastructure ecosystem: its dense fiber networks, established power corridors, and deep enterprise demand provide a diversified foundation that no single shock can easily dislodge. By contrast, states whose competitiveness depends heavily on a single advantage—such as Wyoming's low land costs or Nevada's favorable tax regime—display much wider ranges, as the scenarios selectively amplify or neutralize their primary asset.

Third, the District of Columbia remains fixed at its 0.05 GW baseline across all scenarios, re-

Table A.3—Projected State-Level Data Center Capacity (Part 2)

State	Base 2024	Projected Capacity by 2030 (GW)					
		Status Quo	AI Boom	Climate Stress	Green Grid	Autarkic Expansion	Regulatory Headwinds
Nebraska	0.41	2.61	5.27	1.09	2.09	4.19	1.60
Nevada	0.57	3.67	9.32	1.69	2.88	9.35	2.09
New Hampshire	0.14	0.38	0.52	0.24	0.35	0.46	0.27
New Jersey	1.06	1.22	1.33	1.06	1.17	1.25	1.17
New Mexico	0.27	4.21	8.89	1.88	3.53	8.24	2.15
New York	1.77	3.62	4.92	2.61	3.34	4.08	2.91
North Carolina	1.11	3.15	4.49	2.36	2.94	3.67	2.69
North Dakota	0.22	3.02	5.25	0.95	2.20	3.07	1.33
Ohio	1.66	2.94	4.24	2.36	2.85	3.29	2.55
Oklahoma	0.38	2.66	4.27	1.30	2.17	2.88	1.88
Oregon	0.82	4.02	6.82	2.26	3.81	6.22	2.85
Pennsylvania	0.98	2.80	3.56	1.88	2.58	3.07	2.20
Rhode Island	0.08	0.22	0.24	0.14	0.16	0.22	0.14
South Carolina	0.27	1.36	2.23	0.76	1.22	1.14	1.03
South Dakota	0.11	3.13	5.68	1.11	2.47	4.46	1.63
Tennessee	0.90	2.26	3.62	1.71	2.01	2.88	1.71
Texas	3.86	13.78	23.21	7.58	11.77	17.10	8.73
Utah	0.52	2.69	6.20	1.47	2.31	5.44	1.85
Vermont	0.08	0.35	0.49	0.22	0.35	0.41	0.24
Virginia	2.28	3.97	5.11	3.18	3.83	3.86	3.56
Washington	1.41	3.29	6.01	2.09	2.96	3.97	2.45
West Virginia	0.14	1.01	1.69	0.33	0.82	1.17	0.60
Wisconsin	0.63	2.91	4.40	1.74	2.69	3.78	2.23
Wyoming	0.14	4.32	7.91	1.60	3.40	6.25	2.20

Notes: Continuation of Table A.2. See that table for methodology and variable definitions.

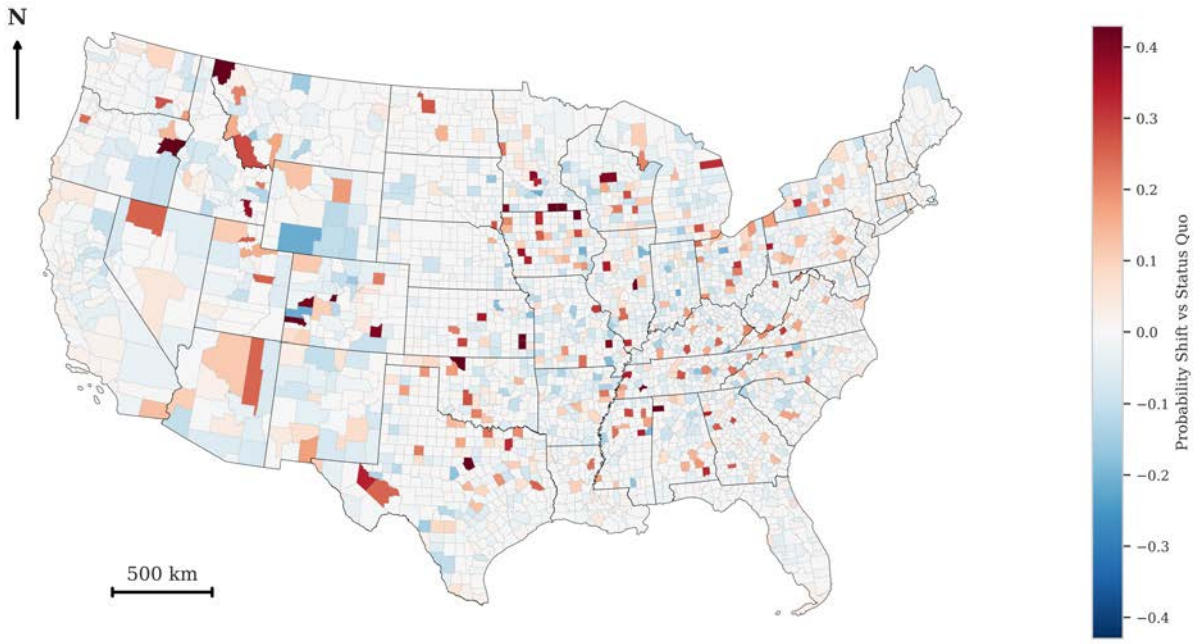
flecting the physical impossibility of greenfield data center development in a fully urbanized jurisdiction with no available land for large-scale facilities.

County-Level Spatial Divergence. Figures A.9 through A.13 map the county-level change in the model's predicted suitability probability under each alternative scenario relative to the Status Quo baseline. These maps are constructed by running the full CNN ensemble under both the Status Quo and each alternative scenario for the 2030 projection year, then computing the difference in the combined (cloud plus third-party) probability surface at each pixel. These pixel-level differences are aggregated to county means via zonal statistics and displayed on a diverging red-to-blue color scale, where red indicates counties whose suitability increases relative to the Status Quo and blue indicates counties whose suitability declines.

The AI Boom divergence map (Figure A.9) displays the most intense and geographically concentrated shifts, with probability changes ranging from -0.4 to $+0.4$. The strongest positive sig-

nals (deep red) appear in scattered counties across the Mountain West, Great Plains, and parts of the Southeast—regions with available land and expanding grid capacity that benefit most from the simulated urban capacity surge. Notably, many established coastal hubs show neutral or mildly negative shifts, consistent with the push-pull dynamic: improved urban grid access is offset by the 20% electricity price increase, producing ambiguous net effects in already-dense markets.

Figure A.9—Spatial Divergence in Data Center Site Selection: AI Boom Scenario

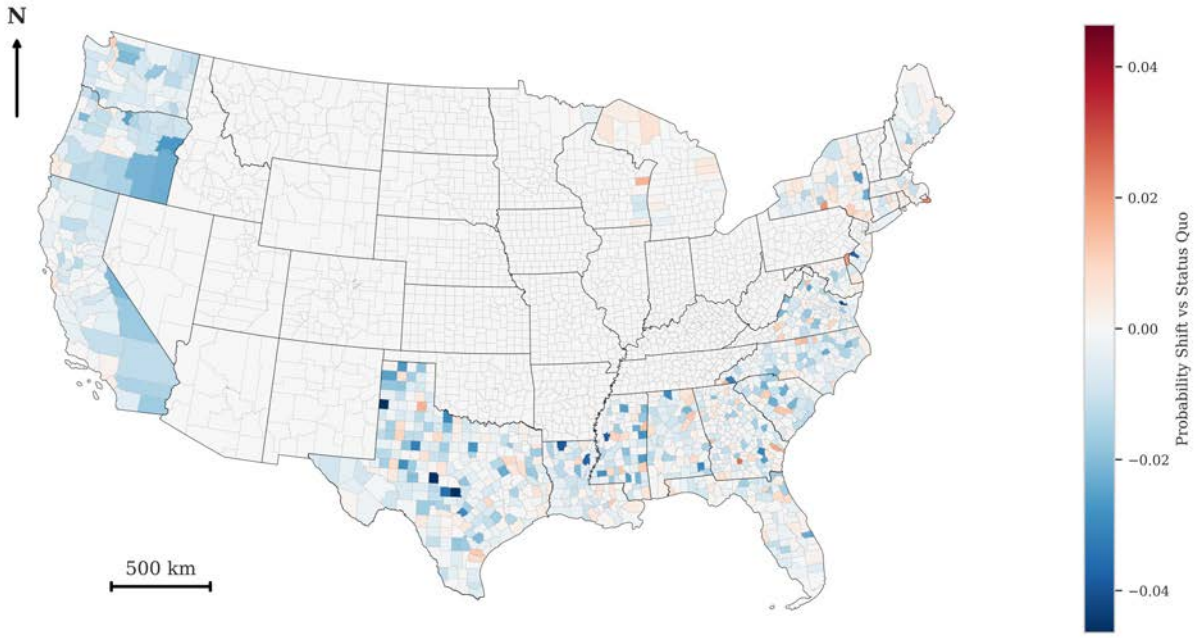


Notes: This map displays the county-level change in predicted data center suitability under the AI Boom scenario relative to the Status Quo baseline for 2030. Values represent the mean difference in combined (cloud plus third-party) suitability probabilities within each county, computed by running the full CNN ensemble under both scenarios and subtracting the Status Quo probability surface from the AI Boom surface. Red shading indicates counties where suitability increases (gaining competitive advantage from the simulated urban grid capacity surge and electricity price increase); blue shading indicates counties where suitability declines. The color scale ranges from -0.4 to $+0.4$. County boundaries are drawn from the 2010 Census TIGER/Line shapefiles; state boundaries are overlaid in black.

The Climate Stress map (Figure A.10) shows a qualitatively different pattern: the divergence magnitudes are an order of magnitude smaller (± 0.04), and the spatial distribution is dominated by widespread blue shading across the South and Southwest. This broad suitability decline reflects the compound effect of deteriorating drought conditions, elevated water stress, and degraded grid reliability in regions already operating near environmental limits. The few red counties—

concentrated in the Pacific Northwest and northern Great Plains—gain a relative advantage from their cooler climates and lower baseline water risk.

Figure A.10—Spatial Divergence in Data Center Site Selection: Climate Stress Scenario

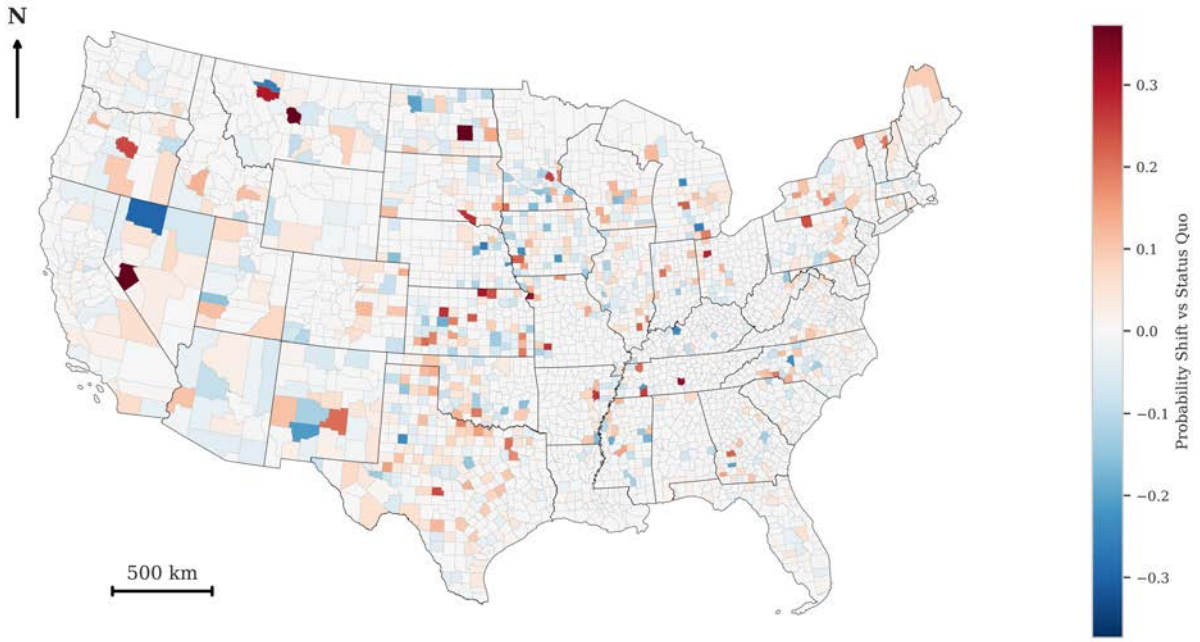


Notes: This map displays the county-level change in predicted data center suitability under the Climate Stress scenario relative to the Status Quo baseline for 2030. Red counties gain a relative advantage from lower environmental vulnerability (cooler climate, lower baseline water stress); blue counties experience reduced suitability due to the simulated compound shock of declining drought severity (PDSI), 40% elevated water stress (WRI), and 20% degraded grid reliability (SAIDI). The color scale ranges from -0.04 to $+0.04$—an order of magnitude smaller than the AI Boom divergence—reflecting the more gradual and diffuse nature of climate-driven suitability shifts.

The Green Grid map (Figure A.11) displays moderate divergence (± 0.3) with a distinctive spatial pattern: red counties cluster along wind corridors in the Great Plains, solar-rich zones in the Desert Southwest, and renewable-heavy pockets in the upper Midwest. Blue counties appear in the Southeast and parts of the Mid-Atlantic where renewable generation capacity is limited. This pattern directly reflects the scenario's design, which expands grid capacity exclusively within pre-defined renewable-rich zones while leaving carbon-intensive markets unchanged.

The Autarkic Expansion map (Figure A.12) produces the widest probability range (± 0.6) and the most geographically dispersed reallocation. By setting all grid distance covariates to zero, this scenario removes the single largest source of spatial differentiation in the model's scoring

Figure A.11—Spatial Divergence in Data Center Site Selection: Green Grid Scenario

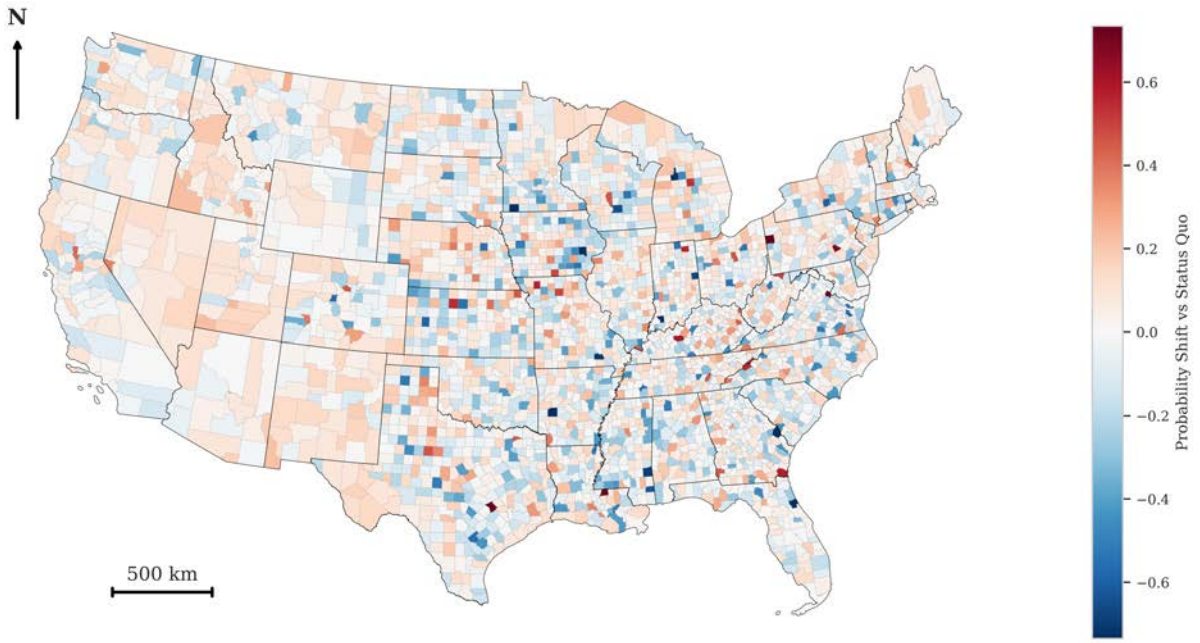


Notes: This map displays the county-level change in predicted data center suitability under the Green Grid scenario relative to the Status Quo baseline for 2030. Red counties benefit from the 5% annual expansion of grid capacity within pre-defined renewable-rich zones (wind and solar); blue counties lose relative suitability because their grid infrastructure remains unchanged while renewable-zone competitors improve. The spatial pattern of red clusters along Great Plains wind corridors and Desert Southwest solar zones directly reflects the geographic targeting of the scenario's grid expansion.

logic. The resulting map reveals which locations gain or lose when grid proximity is no longer a factor. Red counties appear throughout the rural interior—particularly in the Mountain West and northern Great Plains—while blue shading concentrates in historically grid-advantaged corridors such as Northern Virginia and parts of the Southeast.

The Regulatory Headwinds map (Figure A.13) shows moderate divergence (± 0.15) with a geographically diffuse pattern. The scenario increases zoning friction and reduces tax incentives uniformly across states, but the model's response is spatially heterogeneous because the baseline regulatory and fiscal environment varies by state. Red shading appears in states with currently permissive regulatory environments (parts of the Mountain West and Southwest), while blue shading concentrates in the Midwest and parts of Texas—regions where the simulated regulatory tightening interacts with already-marginal economic conditions to push suitability below

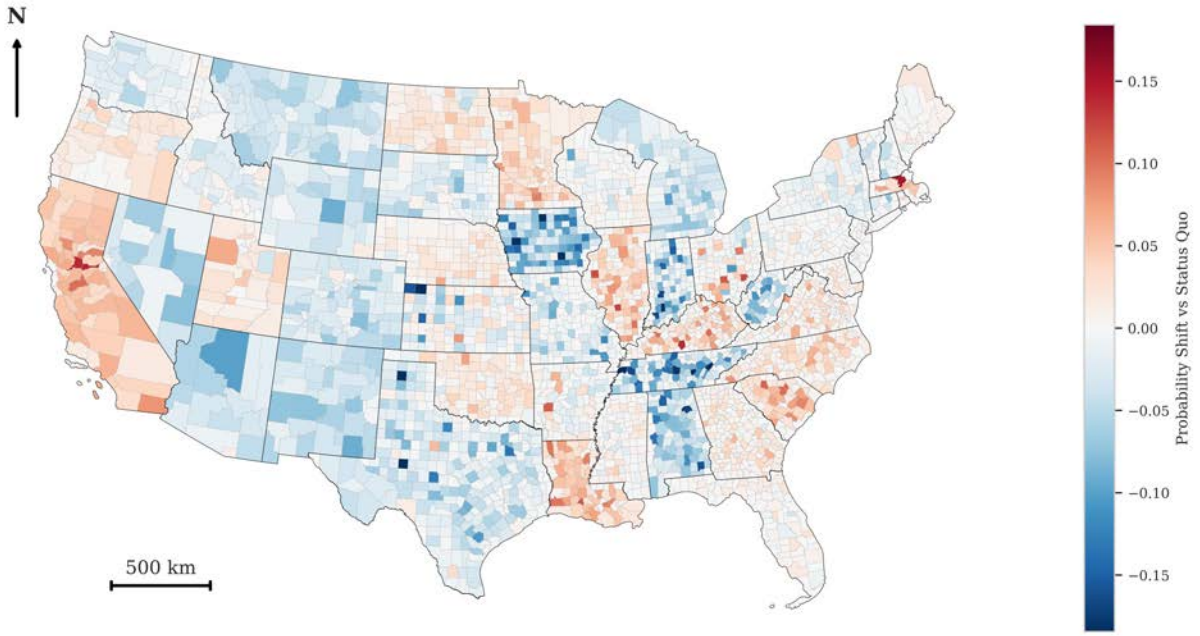
Figure A.12—Spatial Divergence in Data Center Site Selection:
Autarkic Expansion Scenario



Notes: This map displays the county-level change in predicted data center suitability under the Autarkic Expansion scenario relative to the Status Quo baseline for 2030. All grid distance covariates (distance to transmission lines, substations, and high-voltage nodes) are set to zero nationwide, simulating universal energy self-sufficiency through on-site generation. Red counties gain suitability as grid proximity ceases to differentiate locations, revealing the non-energy fundamentals (land cost, connectivity, labor) that favor previously remote areas. Blue counties—concentrated in historically grid-advantaged corridors—lose their spatial premium. The color scale (± 0.6) is the widest among the five scenarios, reflecting the large magnitude of removing the model's single most important spatial differentiator.

the entry threshold. The relatively modest magnitude of the probability shifts, compared to the AI Boom or Autarkic Expansion scenarios, is consistent with the permutation importance results, which ranked regulation and tax incentives below power infrastructure and connectivity as determinants of site selection.

Figure A.13—Spatial Divergence in Data Center Site Selection:
Regulatory Headwinds Scenario



Notes: This map displays the county-level change in predicted data center suitability under the Regulatory Headwinds scenario relative to the Status Quo baseline for 2030. The scenario applies a 50% increase in zoning and legislative stringency covariates and a 50% reduction in tax incentive attractiveness. Red counties gain a relative advantage in jurisdictions with currently permissive regulatory environments; blue counties experience reduced suitability where the simulated regulatory tightening compounds existing economic marginality. The moderate magnitude of the divergence (± 0.15) is consistent with the permutation importance results, which ranked regulation and tax incentives below physical infrastructure as determinants of site selection.