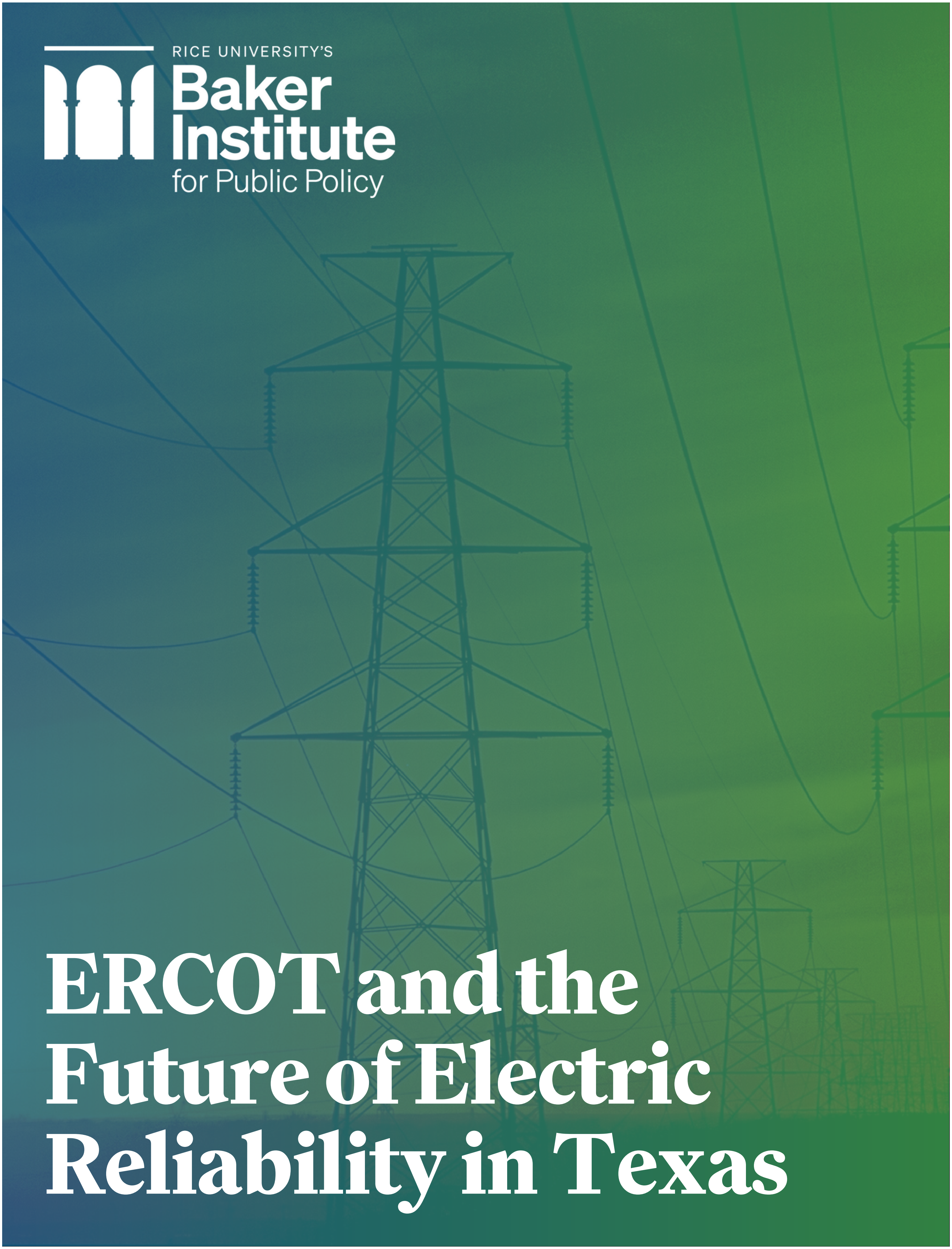


RICE UNIVERSITY'S
Baker Institute
for Public Policy

ERCOT and the Future of Electric Reliability in Texas

ERCOT and the Future of Electric Reliability in Texas

Peter R. Hartley, Kenneth B. Medlock, III, and Shih Yu (Elsie) Hung

The authors thank Blackrock for providing funding to support this research. The research was conducted independently without input from Blackrock and has been reviewed by two anonymous referees. All findings and conclusions are the responsibilities of the authors.

ERCOT and the Future of Electric Reliability in Texas

Peter R. Hartley, Kenneth B. Medlock, III, and Shih Yu (Elsie) Hung

Executive Summary

Electricity reliability and resource adequacy in the Electric Reliability Council of Texas (ERCOT) have been top legislative, regulatory, and commercial priorities in Texas for the past few years. This stems from strong electricity demand (load) growth due to an expanding population and robust economic activity. Moreover, load is likely to continue to expand for the foreseeable future. Texas' population is expected to continue to grow, and the economy is on solid footing, with added vectors for demand growth from electric vehicle (EV) penetration, cryptocurrency mining, carbon capture and storage (CCS) and hydrogen market expansion, and general trends toward increased electrification. There has even been a push toward electrification of oil and gas operations, which has manifested in very strong regional load growth in high oil and gas producing regions.

While the evolution of demand in ERCOT is, by itself, enough to raise concerns about the future of resource adequacy, there has also been inadequate investment in dispatchable sources of generation in ERCOT — leading to greater overall system instability. In fact, ERCOT loads have been increasingly in excess of dispatchable generation capacity in both frequency and quantity since 2018. In 2023 alone, load exceeded dispatchable generation capacity for over 1,000 hours. This raises the risk of resource inadequacy and has coincided with an increase in ERCOT conservation notices and energy emergency alerts.

There is also a general geographic inconsistency between load growth and generation capacity investment across ERCOT. This is not likely to self-correct, given that the drivers of load growth are centered around new industrial activity and population growth in high-load areas in the Texas Triangle — the urban megaregion consisting of the Houston, Dallas-Fort Worth, Austin, and San Antonio metropolitan areas.

Analysis of locational marginal prices (LMPs) reveals that the market is already signaling that there are binding constraints related to patterns of generation, capacity investment, load growth, and transmission. Proper accounting and subsequent internalization of these signals is important for informing the optimal configuration of generation capacity, transmission, and storage with respect to load.

Reliability can be enhanced with proper "insurance," and ERCOT has a portfolio of options available. But policy will ultimately influence which options can be profitably exercised. Among the options identified herein are:

- investment in dispatchable forms of generation that can be called upon when intermittent resources are not available while load is high,
- investment in storage capacity in utility areas and/or alongside industrial consumers to facilitate a reduction of purchases from the grid during periods of high demand,
- investment in production area storage capacity alongside wind and solar generation to allow a “smoothing” of sales from intermittent resources and promote a more efficient use of transmission capacity,
- expansion of transmission capacity to alleviate existing constraints, fully recognizing that the frequency and severity of constraints matter for the economic feasibility of the transmission capacity investment, and
- siting future generation capacity closer to load centers to avoid grid-level bottlenecks.

No market structure can be void of risk because there will always be unexpected incidents and low-probability events that can compromise any system. But allowing structural risks to reliability that can be avoided at a reasonable cost is unacceptable. Therefore, appropriate market design and sufficient regulatory oversight are critical. This opens the door for policy discussions that include, but are not limited to, implementing market structures and/or incentives that ensure sufficient backup capacity and imposing adequate penalties for underperformance by generators under specific obligations. In the end, resource adequacy and reliability are in the best interests of producers and consumers alike, as they establish a platform for long-term growth. Identifying opportunities to provide reliability is paramount, and ERCOT has a substantial portfolio of options.

Framing the Issues

In Texas, electricity reliability has been a point of increasing concern, rising to the top of public discourse in the wake of Winter Storm Uri in February 2021. In previous research, we performed a detailed examination of various factors that were blamed for the extended power outage on the ERCOT electricity grid during Uri.[1] We found that no single factor was fully responsible as all forms of generation capacity experienced failures, and coordination failures in identifying and addressing risks along fuel supply chains were also at fault.

One issue we identified was the need to carefully analyze reserve margins as intermittent generation capacity expands. Texas is No. 1 in the nation in terms of *existing* wind capacity, as well as No. 1 in terms of *planned* capacity additions for wind and solar power. However, such aggressive growth of intermittent resources can compromise reliability when it is accompanied by little-to-no addition of dispatchable forms of generation, especially if system load continues to grow.

In ERCOT, load has grown substantially over the past 20 years, driven by factors such as strong population growth, economic expansion, and increasing electrification of oil and gas operations. In fact, total electricity consumption in Texas has doubled since 2000, with commercial sector demand increasing at the highest clip — by 66%.

In 2022, the residential, commercial, and industrial sectors in Texas accounted for 36%, 34%, and 30%, respectively, of total load. The state's electrification of home heating (61.5%) is already significantly higher than the national average (39.8%), and recent developments — for example, increased cryptocurrency mining, the adoption of EVs, and an overall drive to expand electrification of energy use more generally — serve to place high and growing demands on ERCOT. At the same time, burgeoning activities in CCS and the "green" hydrogen industry will increase the demand for electricity, with the speed and scale of those developments determining the extent to which they will impact the overall market.[2] How ERCOT will respond to emergent and expanding concerns about resource adequacy and reliability is a core question.

We begin with a discussion of the evolution of the Texas electrical grid, culminating in the formation of ERCOT and the creation of a competitive wholesale and retail marketplace. Then, we discuss ERCOT generation resources and the evolution of generation and load. This allows a deeper dive into current concerns about resource adequacy and reliability, with some key insights afforded from statistical analysis of locational marginal prices (LMPs). We conclude with a summary of the findings and some recommendations for further study.

From Then Until Now: The Evolution of ERCOT

What Is ERCOT?

ERCOT is one of the three interconnected systems of the North American Electric Reliability Corporation (formerly the National Electricity Reliability Council), known as NERC, and one of six independent system operators and regional transmission organizations within NERC that work to ensure grid reliability. As such, it manages the grid and electricity flows to more than 26 million customers, which is roughly 90% of Texas's load, 75% of whom are competitive-choice customers. As an independent system operator, ERCOT does not own any generation resources and does not participate as a buyer or seller in the market.

ERCOT is regulated by the Public Utility Commission of Texas (PUCT) and the Texas Legislature. The PUCT was authorized by the Public Utility Regulatory Act of 1975 (PURA) and was given regulatory oversight over all market participants. So, the PUCT has oversight over ERCOT protocols, operating guidelines, and other rules, and it regulates transmission and distribution lines (rates and siting) and retail rates for entities that have opted out of the competitive ERCOT market.[3] It also participates in transmission planning with regions outside of ERCOT.

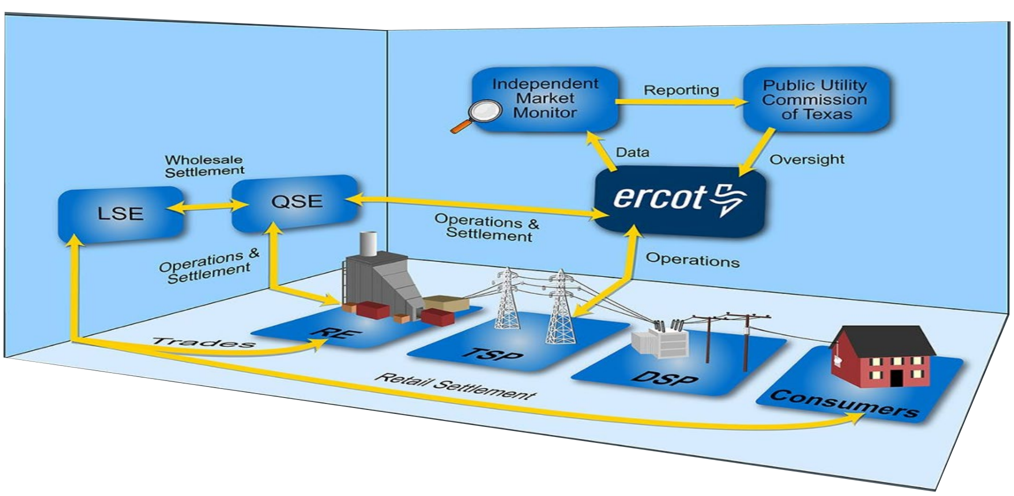
As the Texas Legislature determined during electricity market restructuring in 1999, ERCOT's role is to:

- maintain system reliability,
- facilitate competitive wholesale and retail electricity markets, and
- ensure open access to transmission.

Stakeholders and members of ERCOT include consumers, cooperatives, independent generators, independent power marketers, retail electricity providers, investor-owned electric utilities (transmission and distribution providers), and municipally owned electric utilities.[4] ERCOT market participants interact in real-time and day-ahead wholesale markets, energy and ancillary services market, and the retail market.[5]

The ERCOT physical market is depicted in Figure 1. There are over 1,800 market participants. Qualified scheduling entities (QSEs) are resource entities (REs) that can submit offers to sell and/or bids to buy energy in the day-ahead market and real-time markets, offer or procure ancillary services needed to serve their load, and financially settle with ERCOT. Load serving entities (LSEs) buy wholesale power to provide electric service to individual and wholesale customers. Transmission/distribution service providers (TDSPs) own/operate equipment and facilities to transmit and/or distribute electricity and are required to provide nondiscriminatory access to the grid. Other participants include power marketers and aggregators. ERCOT serves as an intermediary between retail customers and wholesale market participants. Wholesale market participants face real-time price volatility that reflects supply-demand balance, while retail customers do not. However, the costs associated with wholesale market volatility are ultimately passed to retail customers in rate offerings.

Figure 1 — ERCOT Physical Market and Electricity Flows



Source: ERCOT.

ERCOT's Evolution: Independent Oversight, Reform, and Competition

The evolution of ERCOT helps explain its structure and why it has minimal interconnection with the rest of the continental United States.[6] As indicated in Figure 2, the path to the creation of ERCOT began in the first half of the 20th century, when the Public Utility Holding Company Act of 1935 (PUHCA) and Federal Power Act of 1920 (FPA) ended the existence of major multistate holding companies in the United States. The PUHCA gave the U.S. Securities and Exchange Commission regulatory authority over electric utility holding companies. The FPA gave authority over interstate electricity transactions to the Federal Power Commission (FPC), which decades later became the Federal Energy Regulatory Commission (FERC).

Prior to these acts, many Texas power companies operated as part of multistate holding companies that were, in some cases, interconnected and that controlled a significant portion of the electricity assets in the state as well as across the country. To capture economies of scale and increase reliability, these companies developed and expanded transmission infrastructure to connect assets across power markets. However, to avoid federal oversight after the enactment of the PUHCA and FPA, some companies ceased interstate trade.

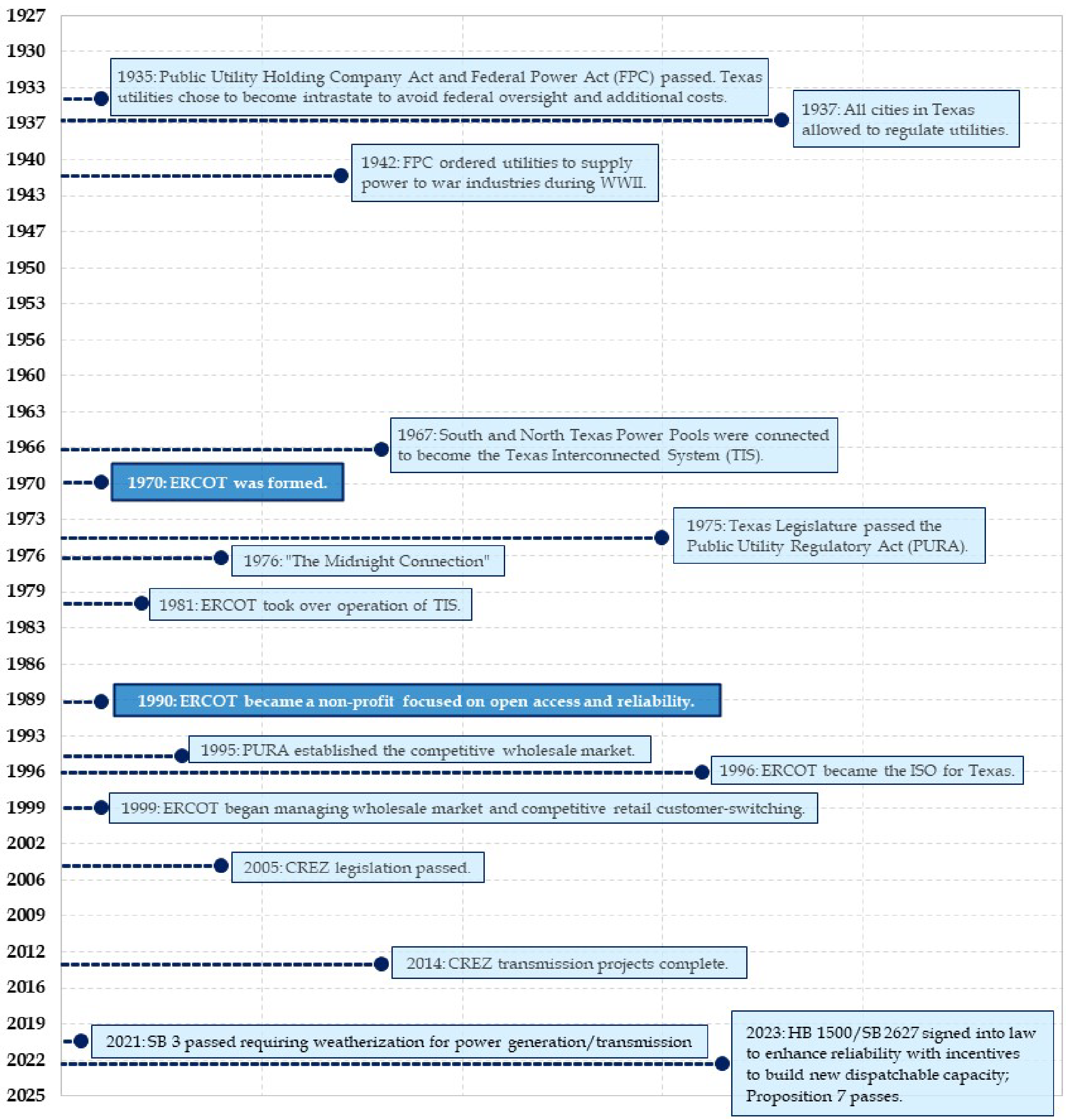
In 1942, during World War II, interstate trade resumed as the FPC relieved utilities of federal oversight on the condition that they supply power to the war industries. However, after the war, companies in Texas abandoned interstate interconnections and chose to operate only within the state. This was strong evidence of the preference of utilities in Texas to operate independently of federal oversight.

In 1967, Texas power companies augmented legacy architecture developed to supply power to defense industries during World War II to connect the South and North Texas Power Pools. This was the first incarnation of the "Texas grid." Shortly thereafter, in 1970, state utilities established ERCOT as a private entity to ensure that system reliability guidelines promulgated under NERC, an initiative that emerged as a national policy priority following the 1965 Northeast Blackout.[7] In 1975, the Texas Legislature passed the Public Utilities Regulatory Act (PURA), establishing state-level oversight of electricity rates and conditions for market entry.

In 1976, West Texas Utilities system operators secretly connected the Texas Interconnected System to the interstate network, an incident known as the "Midnight Connection." This caused companies such as Houston Light & Power and Texas Electric Service Companies to disconnect from the grid to avoid being entered into federal jurisdiction. This again underscores a strong preference among Texas utilities to avoid federal oversight while remaining interconnected with other utilities in the state. This preference, as revealed on multiple occasions across several decades, set the stage for the current "island" state of

ERCOT, which in turn affects resource requirements for meeting present loads. To this day, the state remains largely disconnected from the rest of the U.S., except for some direct current (DC) ties that allow power to flow in and out of Texas and some "switchable" generators that can supply power either to ERCOT or to a neighboring regional transmission organization.

Figure 2 — Timeline of the Evolution of the Texas Grid



Note: Data for Figure 2 is pulled from multiple sources and compiled by the authors.

In 1981, ERCOT assumed operation of the Texas Interconnected System as an independent nonprofit entity. Subsequent regulatory interventions have shaped the role and function of ERCOT. In 1995, the Texas Legislature passed a PURA amendment that established a competitive wholesale market in Texas. In 1996, the PUCT assigned ERCOT as a nonprofit independent system operator to ensure open access to the transmission grid. Texas Senate Bill (SB) 7, passed in 1999 and subsequently signed into law, mandated ERCOT to develop a market architecture that promoted retail competition. After the PUCT approved the protocols for a wholesale market structure established by ERCOT in 2001, the current competitive market in Texas emerged.

Along with federal subsidies, a renewable portfolio standard adopted as part of SB 7 promoted growth in wind generation capacity. However, the best wind resources in Texas are not located near load centers; transmission constraints emerged and signaled a need for more transmission to connect wind in West Texas to load in the Texas Triangle. So, in 2005, the Texas Legislature passed SB 20 to establish the Competitive Renewable Energy Zone (CREZ) initiative.[8] The CREZ initiative authorized PUCT and ERCOT to develop transmission that would deliver wind energy from West and Northwest Texas to load centers in the east and central parts of the state. By 2014, all CREZ transmission lines were completed at a cost of about $7 billion, totaling 3,600 miles of open access 345-kilovolt (kV) lines with a designed transmission capacity of 18.5 gigawatts (GW).[9] This further catalyzed expansion of wind capacity in West Texas, where prolific wind resources are located.

Although the competitive market in ERCOT was not without its detractors, its benefits to consumers were notable.[10] ERCOT remained largely unchanged until the fallout from Winter Storm Uri in February 2021, which exposed fragilities in the Texas energy system.[11] Some of the problems were avoidable, but they triggered a massive wave of deliberation in the Texas Legislature around competitive market structure, interconnecting ERCOT with neighboring states, implementing a capacity market, mandating weatherization of electricity infrastructure, and reevaluating the performance of ERCOT as a grid operator. Subsequent action in the Texas Legislature was largely focused on (i) requiring generation and transmission to improve performance in extreme weather (SB 3), (ii) restructuring the board of ERCOT (SB 2), and (iii) evaluating potential PUCT wholesale market design reforms.

The PUCT split its reforms into two phases. Phase I includes various market design elements, including lowering the operating reserves demand curve (ORDC) cap from $9,000 to $5,000 per megawatt-hour (MWh) and increasing the minimum reserves deemed necessary to avoid system-wide failure to 3,000 megawatts (MW).[12] The newly created "firm fuel supply service," which targets gas generators and on-site dual fuel resources, is intended to enforce "firm fuel" availability for 48 hours for increased reliability. Another initiative involves greater administrative procurement of ancillary services — including emergency response, fast frequency response, contingency reserves, non-spinning reserves, and voltage support — to incentivize the availability of capacity in the day-ahead market. Phase II focuses on ensuring ERCOT has sufficient operating capacity during extreme events resulting from weather, high demand, and/or low output from renewables. This phase also introduced a "load-side reliability mechanism" in which multiple proposals were developed and evaluated.[13]

Activity around ERCOT market reform has since continued. In January 2023, the PUCT voted to adopt a Performance Credit Mechanism (PCM) that incentivizes increased availability of dispatchable generation such as coal, natural gas, and nuclear power.[14] ERCOT was directed to evaluate bridging options until the PCM is fully implemented. In June 2023, the ERCOT Contingency Reserve Service (ECRS) was adopted to enhance reliability and mitigate real-time grid stress. ECRS is a daily procured ancillary service that complements other ancillary service products.[15] Proposition 7, which passed on Nov. 7, 2023, with almost 65% of the popular vote, amended the Texas Constitution to create the Texas Energy Fund, to be administered by the PUCT. The Texas Energy Fund provides low-interest rate loans to finance the construction of new and upgraded dispatchable generation resources. SB 2627, which is the implementing legislation, was passed during the 2023 legislative session.

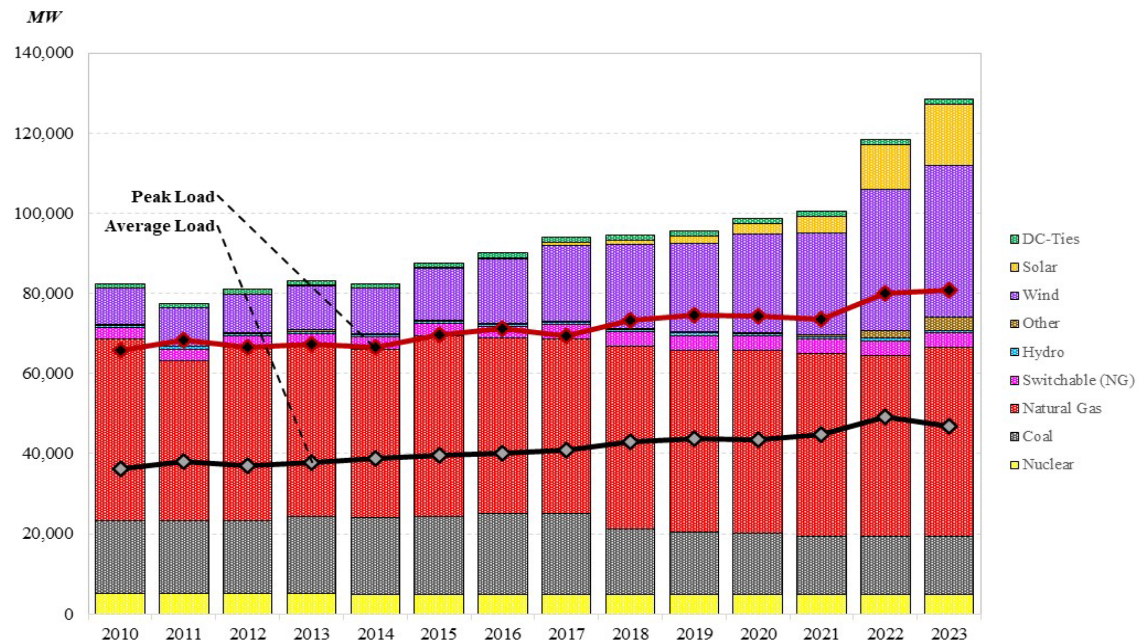
ERCOT Generation Resources

The ERCOT generation portfolio is substantial, totaling over 127 GW in nameplate capacity. As of summer 2023, ERCOT nameplate capacity consisted of wind (37.7 GW), solar (15.5 GW), nuclear (5.0 GW), coal (14.4 GW), natural gas (47.1 GW), and hydro/biomass/batteries (4.0 GW), plus DC ties to neighboring regions (1.2 GW) and natural gas capacity that is switchable (3.7 GW) into and out of ERCOT. Figure 3 details the evolution of nameplate capacity by resource type from 2010-2023.

While impressive, there are valid concerns about whether the ERCOT generation fleet is sufficient. To begin, new natural gas plants have merely compensated for the retirements of older coal facilities. Wind and solar, however, have each seen massive growth. Since 2000, wind capacity has expanded from 160 MW of installed capacity to over 37,000 MW — an average annual growth rate of more than 26%. Over the same period, solar capacity has increased from 15 MW to almost 16,000 MW — an annual average growth rate of over 70%.

While *average* generation from wind and solar have also increased in line with capacity, averages are meaningless for grid stability in electric power markets. Grid stability requires adequate resources to match the temporal fluctuations of load. If resources cannot deliver power on demand, the grid will be exposed to an increasing risk of failure, at which point the need to call on various emergency measures increases. Because wind and solar are not dispatchable, they are not traditional "on-demand" resources, and grid stability can be compromised if they are not available when load exceeds the sum of dispatchable resources. Thus, while overall *nameplate* capacity has expanded considerably with the addition of wind and solar, *dispatchable* capacity has not. Given the demand growth in ERCOT, system load has exceeded dispatchable capacity with increasing frequency since 2018.

Figure 3 — Generation Capacity by Energy Type Plus Peak and Average Load, 2010–23



Source: ERCOT. Data compiled by authors.

ERCOT's Growing Load

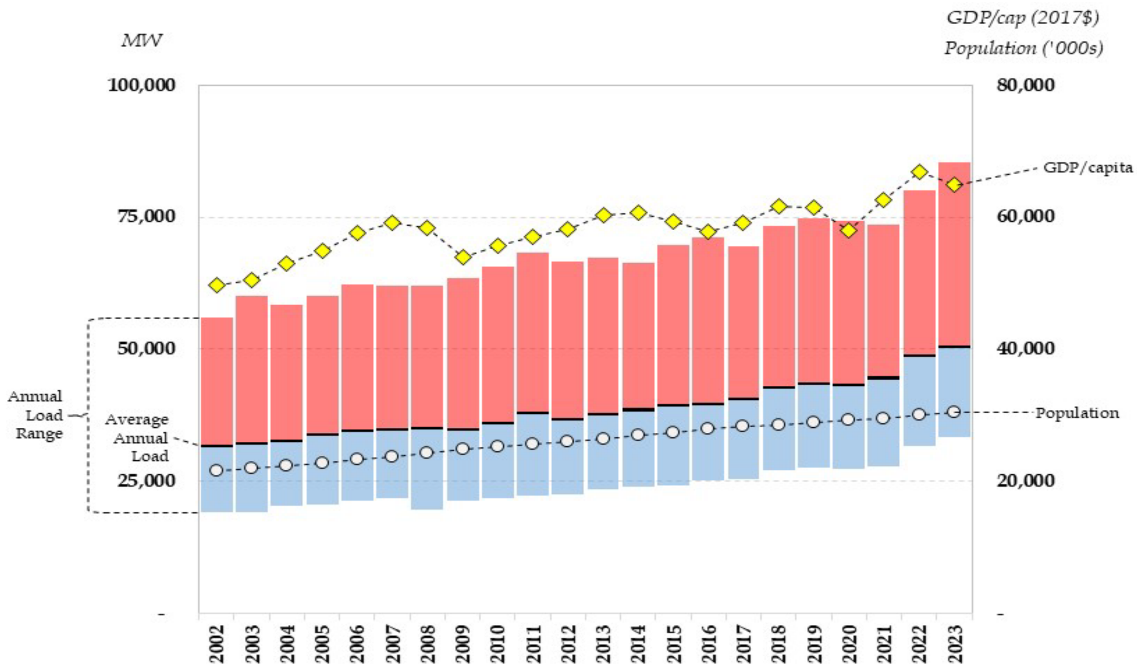
ERCOT load has steadily increased over the past two decades in lockstep with population growth and increases in gross domestic product (GDP) in Texas (see Figure 4). In fact, regressing average annual load ($LOAD$) on state GDP (GDP), state population (POP), and the average annual retail electricity price (P) over the period 2002–22, all transformed to natural logs, reveals

$$\ln LOAD_t = \underset{(2.769)}{1.448} + \underset{(0.343)}{0.648} \ln POP_t + \underset{(0.151)}{0.356} \ln GDP_t - \underset{(0.099)}{0.027} \ln P_t$$

$$\text{Durbin-Watson, d} = 1.701;\ R^2 = 0.938$$

with standard errors in parentheses under the parameter estimates.

Figure 4 — ERCOT Load, Texas GDP per Capita, and Texas Population, 2002–23



Sources: ERCOT, the Federal Reserve database (FRED), and the U.S. Census Bureau.

The parameter estimates can be directly interpreted as elasticities, given the variables are in natural log form, and the parameter estimates on population and GDP are statistically significant with t-stats of 1.89 and 2.36, respectively. The parameter estimate on price is small and not statistically significant, but there may be some simultaneity issues with annual data.[16] Overall, the model indicates a 1% increase in population will yield a 0.65% increase in average annual load, while a 1% increase in GDP will yield a 0.36% increase in average annual load, all else equal.

Since 2002, the average annual load has increased by 2.2% per year, amounting to an increase of about 60%, or almost 19 GW. Over the same period, peak load increased by almost 30 GW. This highlights a major point of emerging stress for the grid, because flexible, dispatchable resources are needed most acutely in peak periods. Resource adequacy is not about average loads; it is about peak loads. Continued economic and population growth in ERCOT will continue to raise the demand for flexible, dispatchable resources.

Existing projections, based on ERCOT's Capacity, Demand, and Reserves reports, indicate that most future growth in generation capacity is expected to continue to come from wind and solar resources. Battery capacity is also expected to grow substantially but is starting from a very low base. Batteries also have limited ability to store energy compared to the energy stored in fuels used by thermal generators.

The outcomes of recent policy changes, such as Proposition 7, remain to be seen, but ERCOT represents a large electricity market that is growing rapidly. The evolution of system capacity has struggled to keep pace, setting the stage for some very tangible concerns related to system stability and grid reliability. It also opens the door for opportunities to profitably grow dispatchable resource capacity.

ERCOT System-Wide Concerns: A Question of Resource Adequacy

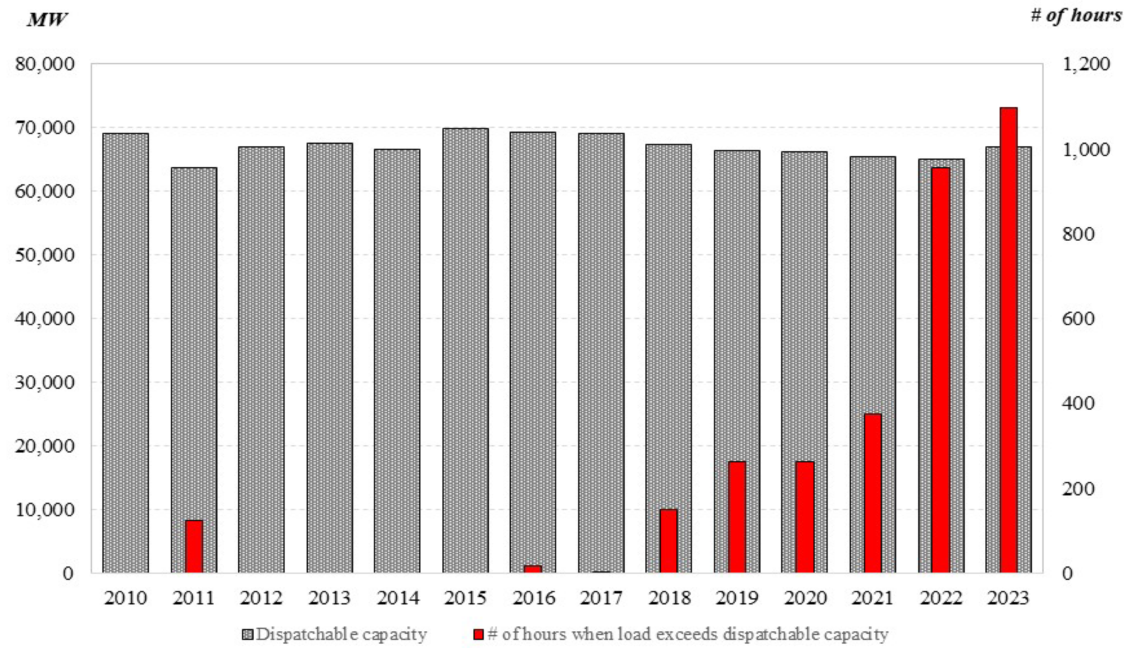
The expected expansion of wind and solar generation in ERCOT owes to Texas' resource endowment (land, wind, and sun), the ease of doing business in the state, and federal incentives that have been bolstered by recent legislation, such as the Inflation Reduction Act of 2022 (IRA). If load growth in ERCOT continues to be strong, as is projected, especially during peak periods, resource inadequacy could become more acute. Formerly arcane discussions about the social value of reliability or the value of lost load will become much more salient. Increased use of demand management tools, while important, may raise other controversies. More generally, reliability will need to be "priced" into the market through regulatory interventions and/or market rulemaking to ensure sufficient investment in dispatchable capacity.

The largest source of flexible dispatchable generation capacity in ERCOT — combined-cycle natural gas generation — has not grown in recent years. Meanwhile, the fleet of single-cycle gas turbines has diminished as units have been retired. Low wholesale electricity prices, combined with increased ramping costs and reduced operating hours as non-dispatchable generation, when producing, displaces higher marginal cost plants, have negatively impacted profitability assessments of new dispatchable capacity. The resulting system lacks sufficient dispatchable generation capacity to serve as a backstop for intermittent resources.

Backstop reliability service is analogous to an insurance policy. Like insurance, it will be provided only with an adequate risk premium. Higher average prices will be required to justify long-term investment in dispatchable resources as the risks associated with their revenue and cost streams increase. Box 1 highlights this insurance concept.

As shown in Figure 5, the number of hours when loads exceed dispatchable capacity has increased, indicating the annual data in Figure 3 does not simply reflect peak demands on abnormal days. This generally corresponds with the frequency of Energy Emergency Alerts (EEAs) and Voluntary Conservation Notices issued by ERCOT. It also highlights the growing risk of continued load growth in ERCOT.

Figure 5 — Hours of Load in Excess of Dispatchable Capacity, 2010–23

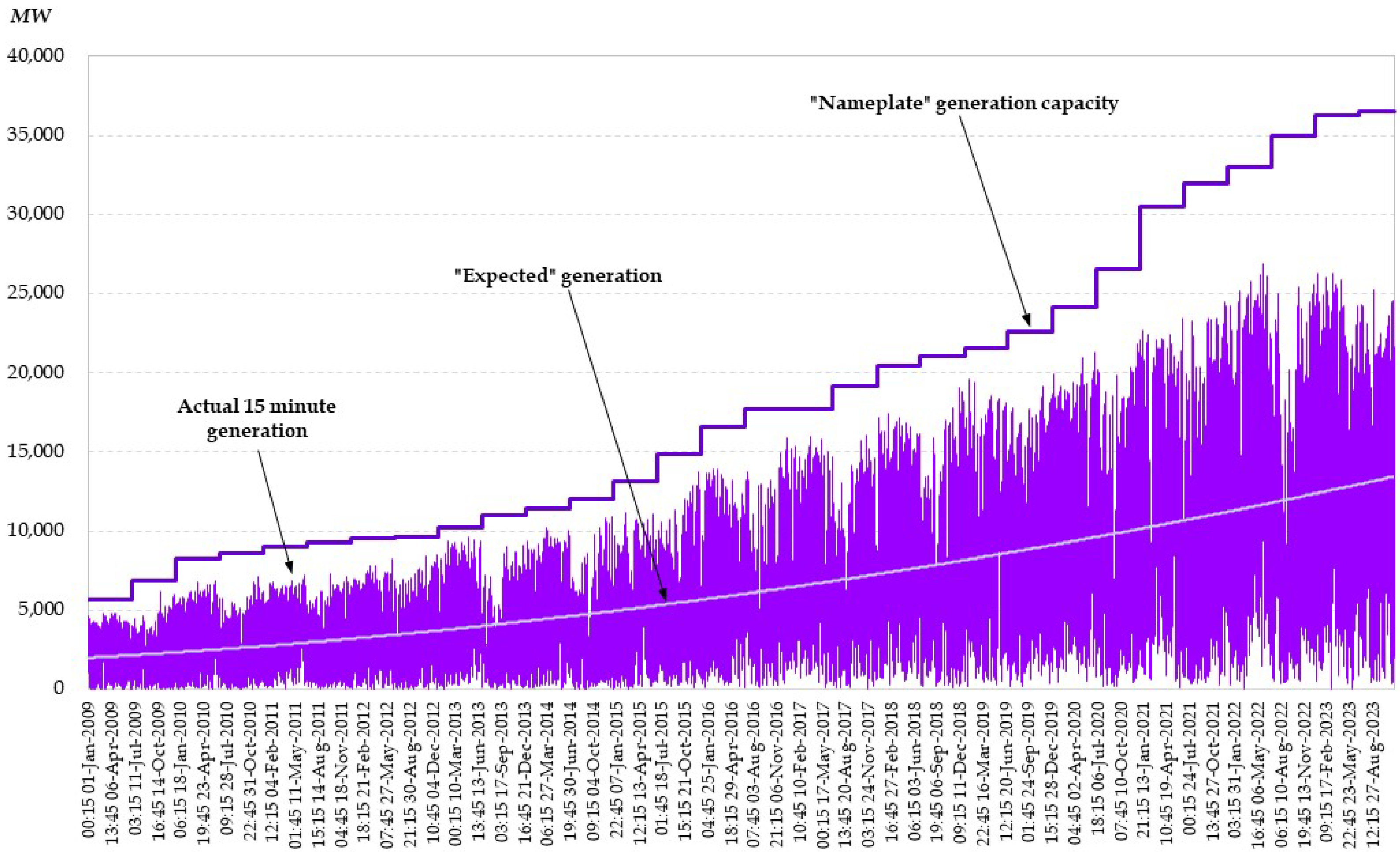


Source: ERCOT. Data compiled by authors.

An emerging, ominous trend for system reliability is the retirement of dispatchable capacity. Over 6,500 MW of coal and natural gas were retired from 2018 to 22, most of which was sited close to demand centers. Today, over 36% of Texas' operational thermal capacity is over 40 years old, including more than 65% of existing coal capacity and 30% of existing natural gas capacity. Expected retirements and new interconnections translate to a supply-demand gap of roughly 30 GW to be filled by 2027, excluding future load growth. And even though system-wide capacity has increased and is expected to continue to increase, the growth has all been from wind and solar installations.

Certainly, wind and solar often deliver well above their rated capacities, especially on days that are windier or sunnier than usual, but they also frequently deliver well below their rated capacities (Figures 6 and 7). Thus, the risk profile of the generation fleet is growing. Given government incentives for renewable energy, regulators and system operators must address the increasing risk of inadequate operational generation resources. Moreover, as indicated in Figure 8, relying on continued expansion of wind and solar on the grounds that they are complementary is not a panacea.

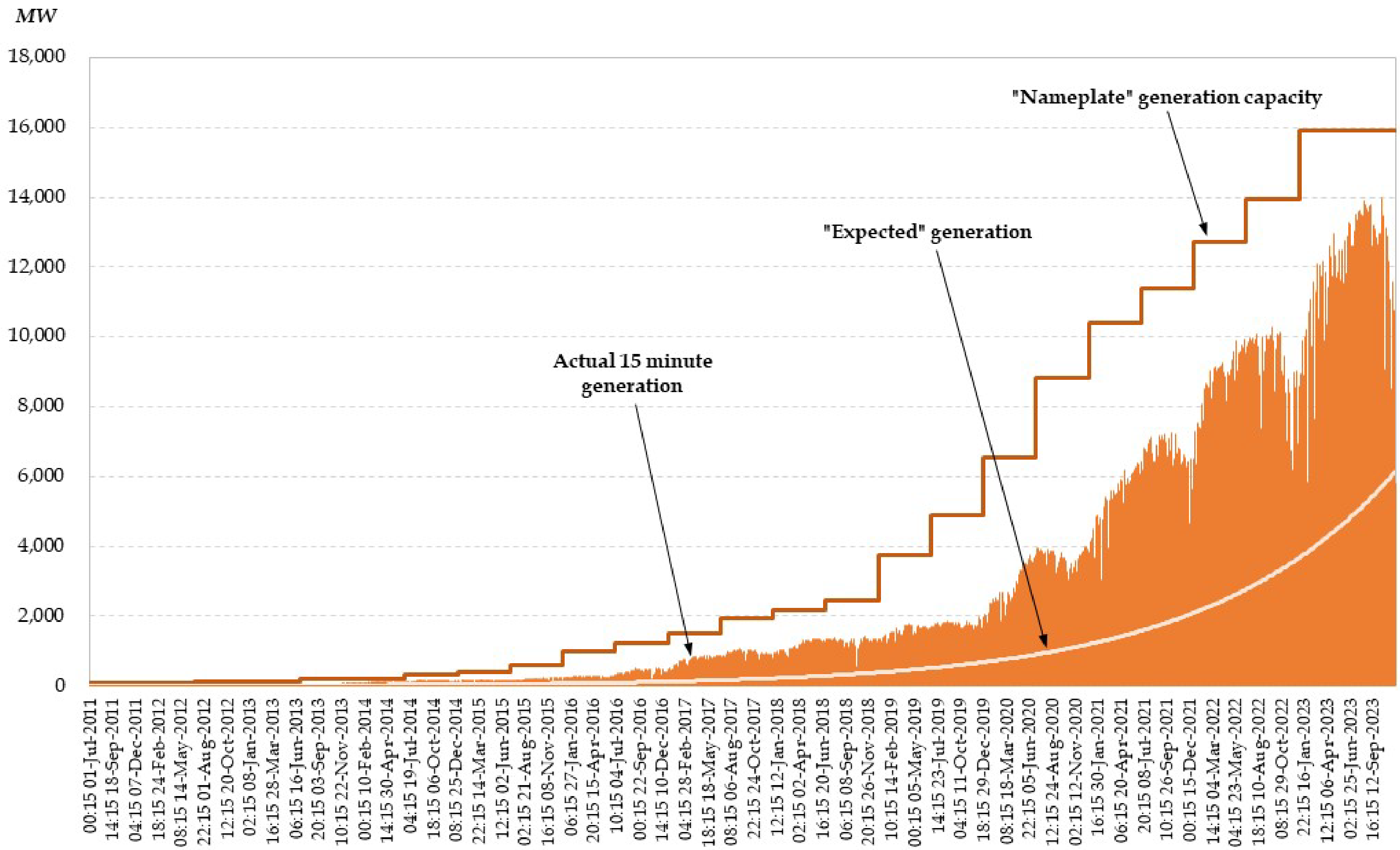
Figure 6 — Wind Generation in ERCOT, 15-Minute Intervals, Jan. 1, 2009–Nov. 30, 2023



Source: ERCOT. Data compiled by authors.

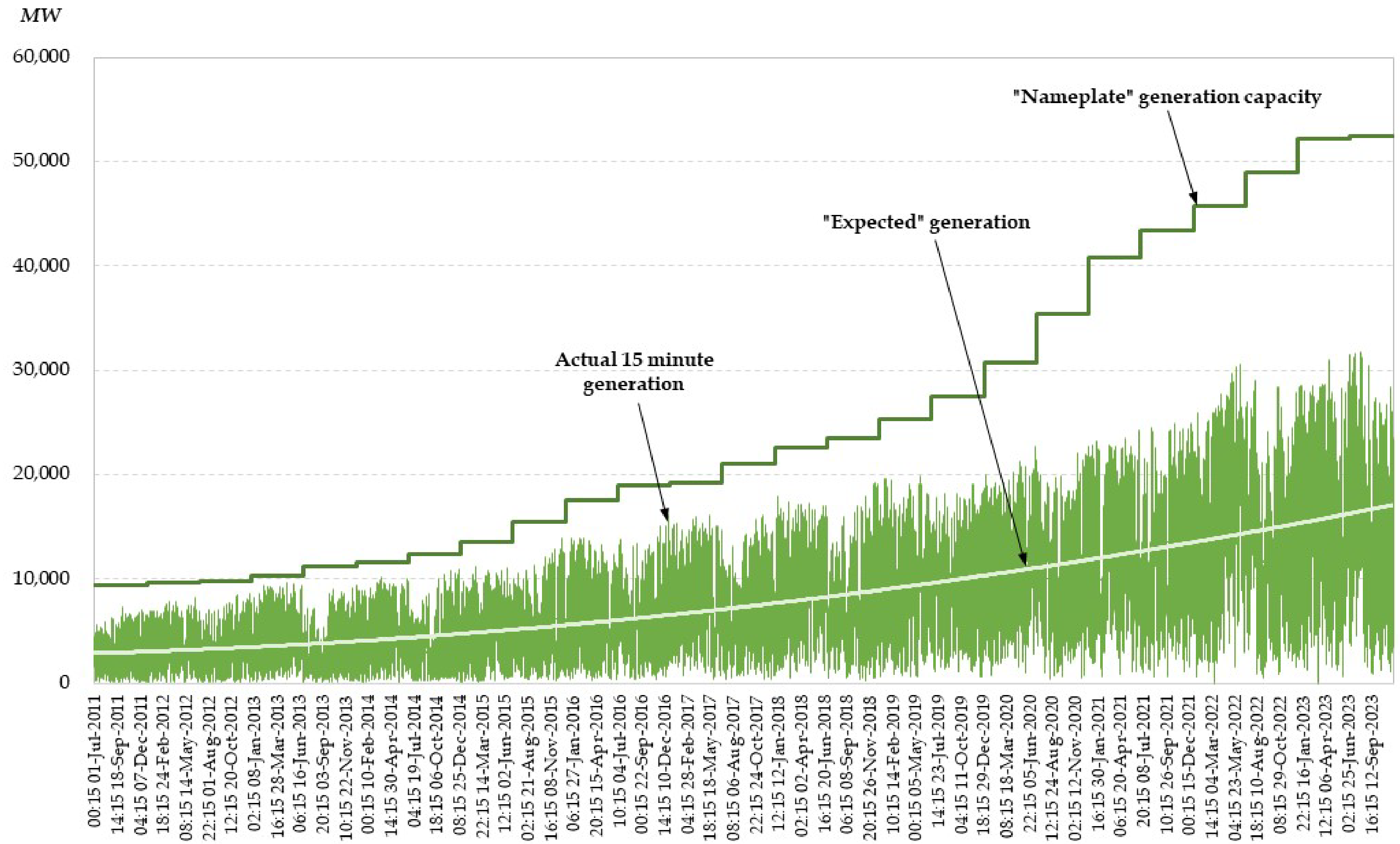
Note: “Expected” generation is included for illustrative purposes only. It is the regression results of the monthly averages over the entire sample and is consistent with an “average” wind generation outcome. “Nameplate” capacity data is not available in 15-minute increments, so it is depicted in six-month windows as the installed capacity at the end of the six-month period.

Figure 7 — Solar Generation in ERCOT, 15-Minute Intervals, July 1, 2011–Nov. 30, 2023



Source: ERCOT. Data compiled by authors.
Note: “Expected” generation is included for illustrative purposes only. It is the regression results of the monthly averages over the entire sample and is consistent with an “average” solar generation outcome. “Nameplate” capacity data is not available in 15-minute increments, so it is depicted in six-month windows as the installed capacity at the end of the six-month period.

Figure 8 — Wind *Plus* Solar Generation in ERCOT, 15-Minute Intervals, July 1, 2011–Nov. 30, 2023



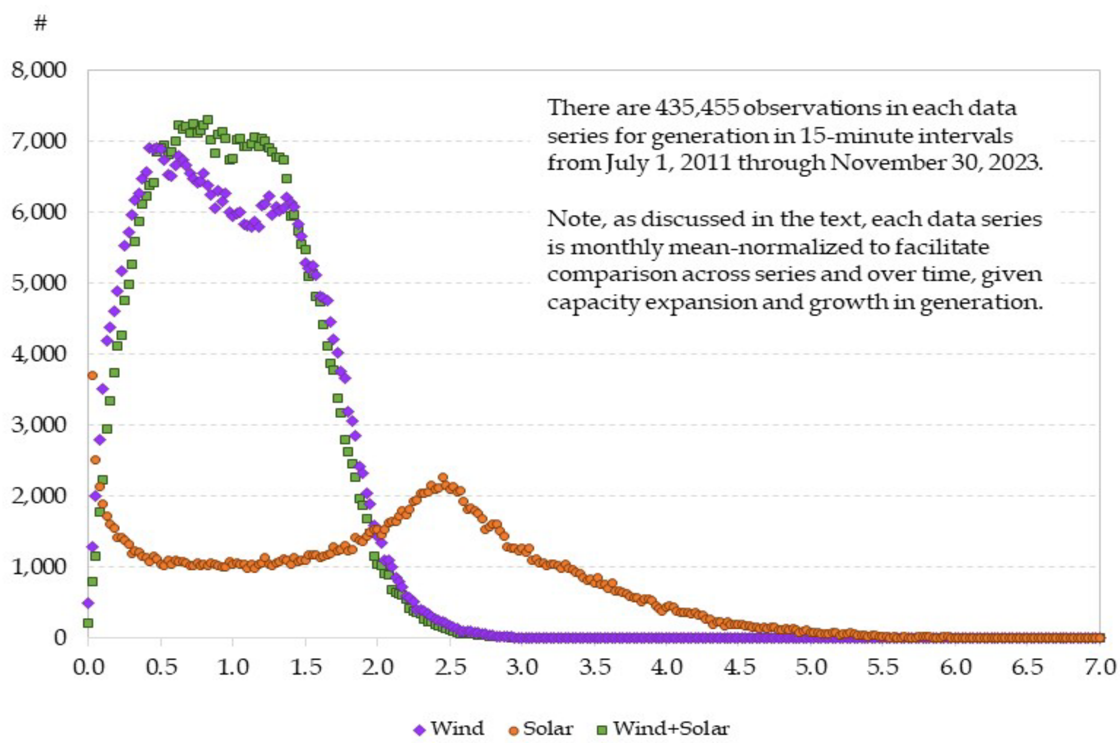
Source: ERCOT. Data compiled by authors.

Note: "Expected" generation is included for illustrative purposes only. It is the regression results of the monthly averages over the entire sample and is consistent with an "average" wind *plus* solar generation outcome. "Nameplate" capacity data is not available in 15-minute increments, so it is depicted in six-month windows as the installed capacity at the end of the six-month period.

Importantly, the variation indicated in Figures 6-8 is independent of load at any given time because it is dictated by whether the wind is blowing, or the sun is shining. Variation in every other generation source is a function of dispatch, which is the call on those generation sources given the availability of other sources and their relative cost, given system load. Thus, variation in dispatchable generation is *endogenous*, or dependent on other variables in the electricity system, while variation in non-dispatchable generation is *exogenous*, or independent of other variables in the electricity system. Reliability requires sufficient endogeneity of system resources to optimize responsiveness.

Figure 9 is constructed to reveal the mean-normalized distribution of generation for wind, solar, and wind *plus* solar. Mean-normalization divides the generation in each 15-minute interval by the monthly average generation within that same month. This normalizes the data against growth over time, as well as monthly seasonality in generation patterns, thus allowing a comparison independent of time.

Figure 9 — Mean-Normalized Distributions of Wind, Solar, and Wind Plus Solar



Source: ERCOT. Analysis done by authors.

By construction, Figure 9 yields a series mean that is always equal to 1, while the standard deviations will reflect how much generation varies relative to its average for each month. For instance, the average generation from wind in July 2011 was 2,634 MW, with a maximum of 5,578 MW and minimum of 63 MW. In November 2023, the average was 10,851 MW, with a maximum of 24,597 MW and minimum of 361 MW. Hence, the mean-normalized values for average, maximum, and minimum in July 2011 were 1.000, 2.117, and 0.024, respectively, while for November 2023 the same values were 1.000, 2.267, and 0.033, respectively.[17]

As Figure 9 reveals, the distribution of the mean-normalized data for wind plus solar is narrower than either wind or solar alone, indicating there is a reduction in variance around the monthly average when the two data series are combined. Adding solar to wind increases the number of observations that are closer to the mean, but the differences are minor. The standard deviation of the distribution of wind *plus* solar (= 0.51) is only marginally smaller than the standard deviation of wind alone (= 0.55).

Moreover, a standard deviation that is roughly half the mean in a mean-normalized data series indicates substantial variation in the observed generation over time. This has implications for resource requirements to adequately ensure reliability. For instance, in November 2023, if system planning is based on an expected generation that is consistent with the indicated value in Figure 9, a fully insured system would need about 15 GW of dispatchable generation capacity available on standby in the event the combined wind plus solar generation fell to its minimum value that month. Note, this is independent of load, as it assumes that total capacity, based on expected performance, is sufficient to meet load in any given month. If this is not the case, then the *overall* capacity requirement would be higher.

Because avoiding (i) catastrophic system failure is critical, (ii) power quality (frequency and voltage stability) is a prerequisite for commercial and industrial users, and (iii) assured supply is vital for the welfare of residential consumers, the recent emphasis in proposed long-term market design reforms from the PUCT has been on resource adequacy.[18] Proposals over the past couple of years have included

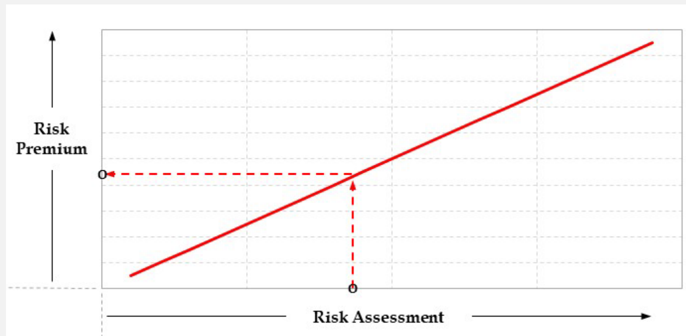
- higher penalties for nonperformance by ancillary service market participants,
- requiring market participants (from generators to local utilities) to secure backup capacity if intermittent generators underperform,
- calls for on-site fuel storage for dispatchable generation capacity, and
- arguments for instituting a capacity market in ERCOT.

Note that in every case, the fundamental argument is rooted in concerns about system reliability and creating mechanisms to ensure adequate operational (dispatchable) capacity is available to meet system load requirements. Effectively, the effort has been focused on establishing an appropriate insurance premium to warrant resource adequacy (see Box 1 for a brief discussion of this concept).

Box 1 — A Role for Risk Premiums and Insurance in ERCOT

A risk premium is commonly defined as the difference between the expected return on an asset with quantifiable risk and the lower expected return on the asset that would leave it equally attractive if it were risk-free. An asset's risk premium in a given investment reflects the additional compensation over that of a risk-free investment that an investor needs to tolerate the risk.

Risk premiums are often discussed as potential tools for correcting market distortions and providing price signals to market participants about a particular course of action, either through investment or operations, in various situations. This incentivizes risk-averse actors to internalize the costs of their actions. As indicated in the figure below, as the risk associated with a particular course of action rises, the risk premium also rises. This is at the core of how insurers price insurance policies on different activities, with riskier activities requiring higher premiums. To adequately warrant against undesirable outcomes, risk must be priced into the activity.



In the Texas electricity market, risk premiums should capture the value of reliability to market participants. Prices needed to ensure resource adequacy and other services for reliability would incorporate these premiums. For instance, to the extent that wind variability introduces a reliability risk, an appropriate risk premium to account for this, along with support backup capability (which could be additional natural gas capacity, battery installations, demand-side management programs, etc.), is needed. This would compensate investors for providing the necessary capacity additions.

In its role as regulatory overseer of ERCOT, the Texas Legislature has increased its focus in the past two sessions on legislation intended to ensure reliability. SB 3 (2021) required the PUCT to reform ERCOT market design to enhance reliability. In December 2021, the PUCT released its "Approval of Blueprint for Wholesale Electric Market Design and Directives to ERCOT." Implementation of Phase I of the blueprint began in January 2022 and was focused on providing "enhancements to wholesale market mechanisms" to improve reliability.

The primary goal of electricity market design is to leverage the incentives and price signals markets provide to minimize costs and ensure adequate operational capacity and flexibility to serve customers reliably. Of course, all actions taken at the state level must account for federal-level market interventions. Some recent relevant actions at the federal and state levels that will impact electricity market structure and reliability, in addition to state legislation previously discussed, include:

- *U.S. Infrastructure Investment and Jobs Act of 2021* (IIJA). Includes a National Electric Vehicle Infrastructure Formula Program and Discretionary Grant Program for Charging and Fueling Infrastructure.
- *U.S. Inflation Reduction Act of 2022* (IRA). Extends the business investment tax credit through 2025, creates a clean electricity investment tax credit for energy storage projects, and creates tax incentives to produce “green” hydrogen.
- *Texas SB 2627*. Establishes a Texas Energy Fund to provide up to $7.2 billion in funding through loans and grants to incentivize up to 10 GW of new or upgraded dispatchable generation.
- *Texas HB 1500*. Places several guardrails on the PUCT’s capability to establish and implement a PCM, which, starting in 2027, requires interconnected facilities to maintain output during peak load hours, instructs ERCOT to create Dispatchable Reliability Reserve Service, and establishes a standard allowance to incentivize new generation to site closer to load centers.

The IIJA could impact electricity demand, for example, by incentivizing greater electrification in the transportation sector. This means an even greater increase in load, which, without sufficient investment in generation capacity, will increase the risk of grid instability. The IRA also provides incentives that could increase electricity demand, such as for hydrogen and carbon capture, but it also extends and enhances incentives to expand renewable generation capacity as well as electricity storage capacity. While more renewable generation capacity could exacerbate emerging risks, increased energy storage could alleviate risks. Only time will tell which will dominate over the next few years.

Both SB 2627 and HB 1500, combined with Proposition 7, could alleviate emerging risks, but their effectiveness will depend on the pace of expansion of dispatchable resources stimulated by these state-level regulatory interventions. Regional siting decisions and transmission constraints in ERCOT that ultimately dictate the price received for generation may limit the profitability of these capacity expansions, even with the new incentives.

More generally, there is a powerful argument that electricity price should reflect the value of reliability in addition to short-run supply costs. This could motivate market designs that place greater value on dispatchability. Direct subsidies to dispatchable resources, such as loan guarantees, tax credits, or other incentives, are one approach. But such subsidies are difficult to set at the appropriate level or to adjust as conditions change, and they often succumb to special interest lobbying. An alternative is to create indirect subsidies to dispatchable resources that manifest through, for example, regulatory requirements that non-dispatchable resources must warrant firm supply to participate in the day-ahead or ancillary services markets. This might provide more flexibility in response to market conditions. In either case, the objective would be to internalize the value of reliability and resource adequacy.

ERCOT Subregional Concerns and LMP Signaling

In addition to the system-wide issues facing ERCOT, some subregional issues also demand attention. Many of these are directly connected to transmission. The devil is in the details.

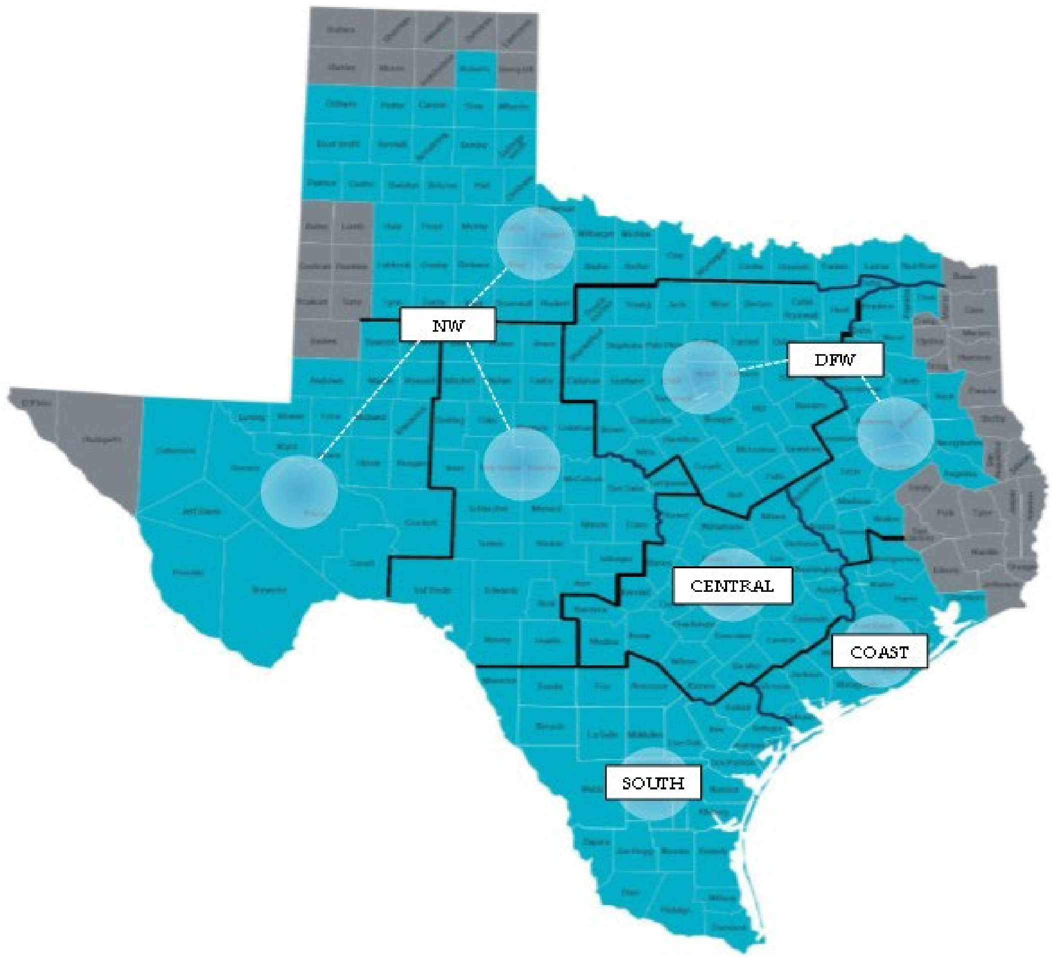
To begin, the top nine most populated counties, totaling 16.8 million people, represent 56% of the state population.[19] Accordingly, this is where load is highest. Dispatchable capacity, such as coal, natural gas, and nuclear, accounts for roughly 90% of the installed capacity within these counties. The remaining 10% of capacity consists mostly of solar. Wind is generally sited in more remote locations, where wind resources are better and land prices lower, given the required land footprint. This translates into significant distances between demand centers and renewable resources. Add in the substantial growth in wind capacity in recent years, and it is not surprising that ERCOT faces a well-recognized transmission challenge.

A 2022 ERCOT report indicated transmission constraints for two route types: (1) exports away from renewable resource centers and (2) imports into demand centers.[20] Another ERCOT study identified the need for new transfer pathways and long-distance transfer technologies beyond the typical 345-kV circuit line to improve export capability.[21] Transmission constraints for imports can be addressed by adding generation closer to demand centers. ERCOT's latest estimate of capital costs for new units found that the overnight cost for all natural gas technologies is lower than any other technologies except for two-hour battery storage.[22]

These considerations raise two interrelated questions critical for the future of regional stability on the ERCOT grid: (1) How should resources be sited and developed to avoid transmission constraints, and (2) what transmission upgrades or enhancements are needed to alleviate constraints, including how they should be financed, given transmission investments will affect resource siting choices? A first step toward understanding this is to identify where constraints are manifest.

It is useful to disaggregate ERCOT into regions, as depicted in Figure 10, with the Texas Triangle previously referred to comprising most of the "DFW," "Central," and "Coast" regions in Figure 10. There are some key differences across market areas, so defined. As can be seen in Figure 11, the Texas Triangle has historically had the largest loads, although load in the "NW" region has recently overtaken load in the Central region. Load growth in NW corresponds to rapid increases in oil and gas production in the Permian Basin coupled with efforts to electrify upstream operations, a trend driven at least in part by environmental considerations.

Figure 10 — ERCOT Territory and Regional Characterization

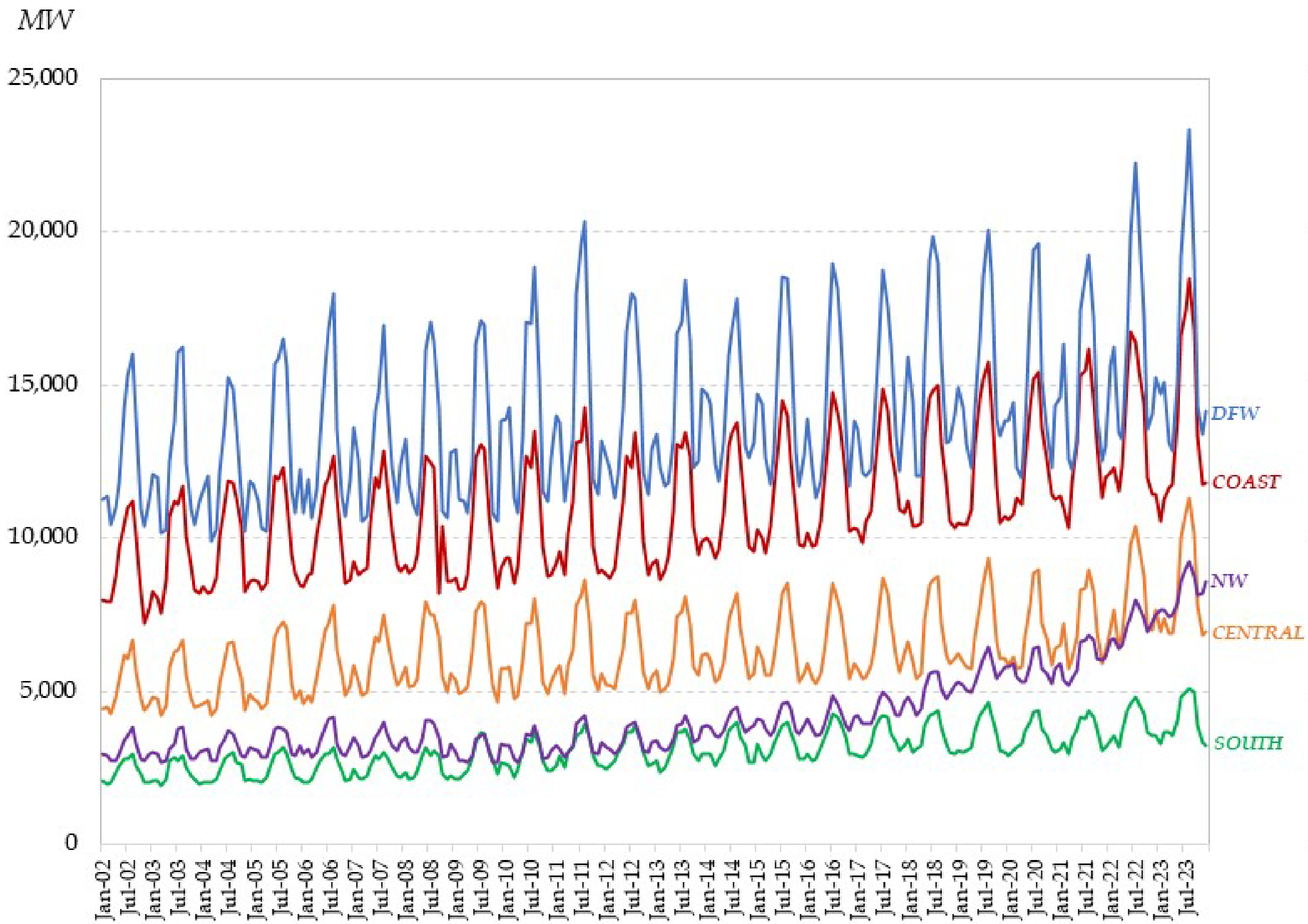


Note: Figure 10 is based on ERCOT weather zones (see https://www.ercot.com/news/mediakit/maps). For the analysis done herein, the North, West, and Far West zones are grouped into the "NW" region. The North Central and East zones are grouped into the "DFW" region, the South Central zone comprises "Central," and the South and Coast zones are unchanged.

Using the data for monthly averages of hourly loads in Figure 11, load has increased more than 266% from 2002–23 in NW. Load growth is second highest in the South region, where the increase was more than 70%. Central is not far behind, with a growth of 62%. In DFW, the highest load region, growth was slightly above 32%, while in Coast it was almost 53%.

The embedded table in Figure 11 also reveals that regional peak loads and the hourly standard deviations of load have increased by a greater amount than regional average loads. The need to respond to greater peaks alongside increased variation at the regional level exacerbates the need for greater flexibility in grid resources at the system-wide level resulting from the growth of intermittent versus dispatchable resources.

Figure 11 — Load by Region, Monthly Averages of Hourly Loads, Jan. 1, 2002–Dec. 31, 2023



		MW
DFW	Avg. 2002	12,337
	StdDev, 2002	*3,321*
	Peak 2002	23,069
	Avg. 2023	16,330
	StdDev, 2023	*4,710*
	Peak 2023	31,585
COAST	Avg. 2002	9,020
	StdDev, 2002	*1,898*
	Peak 2002	14,574
	Avg. 2023	13,798
	StdDev, 2023	*3,410*
	Peak 2023	23,963
NW	Avg. 2002	3,072
	StdDev, 2002	*617*
	Peak 2002	5,325
	Avg. 2023	8,232
	StdDev, 2023	*789*
	Peak 2023	10,674
CENTRAL	Avg. 2002	5,146
	StdDev, 2002	*1,308*
	Peak 2002	9,544
	Avg. 2023	8,338
	StdDev, 2023	*2,254*
	Peak 2023	15,174
SOUTH	Avg. 2002	2,375
	StdDev, 2002	*535*
	Peak 2002	3,984
	Avg. 2023	4,048
	StdDev, 2023	*977*
	Peak 2023	6,608

Source: Hourly data retrieved from ERCOT.

Note: Monthly averages of hourly data are graphed in Figure 12. The embedded table indicates the annual average, standard deviation, and peak load for 2002 and 2023 to illustrate the evolution of load.

Examination of LMPs across ERCOT helps to highlight the regional challenges faced within Texas. It should be noted, however, that LMPs, as currently calculated by ERCOT, differ geographically only because of transmission constraints. Normal marginal transmission losses are instead allocated to all users regardless of location. This tends to mask the full impact of transmission bottlenecks and siting decisions. System-wide efficiency is adversely affected because incentives for decision-makers to take those losses into account are lost.

In 2018, the PUCT requested that ERCOT assess the expected benefits of including marginal transmission losses into the LMPs.[23] There were several interesting outcomes of the subsequent analysis that centered on the marginal revenue implications for different plants sited in different locations. Notably, incorporating marginal losses into LMPs was found to increase unit revenue shortfalls, which is the difference between daily energy revenue and the operating costs of units committed to maintain local grid reliability. ERCOT indicated that this could "result in an increase in units being committed out of merit order so as to maintain local reliability and an increase in unit make-whole payments." They further noted that an increase in unit start-up costs when LMPs included marginal transmission losses may be a possible explanation. This may be either because of "an increase in the number of unit starts and/or an increase in the number of starts of units with higher start-up costs."

ERCOT also noted that it is difficult to translate the changes in production costs and generator revenues into price changes for consumers. The main reason is that the change in LMPs would alter the price structures. Under the current system, average transmission losses are recovered by fixed charges for service. Under a marginal transmission loss system, the costs of the losses would become part of the variable charge per kWh.

Although LMPs currently calculated by ERCOT do not faithfully reflect differences in costs of supply across locations, they nevertheless reveal information about transmission constraints. Since transmission is typically built to cope with peak loads, binding transmission constraints should be an unlikely occurrence when loads are far below peak. Nevertheless, line outages and other equipment problems or unusual distributions of demand or supply across the network could lead to some constraints. When load approaches system capacity, transmission constraints are likely to become ubiquitous.

Binding Transmission Constraints

We examined the distribution of LMPs across the approximately 16,700 electrical buses on the network in 44,681 successive observations (around 152 days of mostly five-minute intervals).[24] Given our expectation that binding transmission constraints should be unlikely unless loads are near peak, we were surprised to find that they were present in almost 78% of the 44,681 five-minute intervals. Admittedly, the sample period covered the most recent 2023 summer, but that binding constraints were present for such a large fraction of the time indicates serious transmission capacity problems in the ERCOT network. This is reinforced by the finding that binding transmission constraints occur throughout the full range of loads, including some of the lowest ones in our sample. Evidently, there have been difficulties expanding transmission capacity to keep up with load growth in the network, as well as the expansion of supply sources that are distant from the main load centers, wind in particular. In fact, a strong tendency for wind generation to be higher in five-minute intervals when there is at least one binding transmission constraint is the most noticeable difference between the subsets with and without such constraints.

Given how ERCOT calculates LMPs, the cross-sectional standard deviation in LMPs will be zero if no transmission constraints are binding ($\chi = 0$) and positive otherwise ($\chi = 1$). Statistical analysis of the resulting binary variable, χ, reveals that an increase in load that can be met entirely by thermal generation has a slight positive effect on the probability of encountering at least one binding transmission constraint. If the increase in load is met instead by an equal increase in solar generation, the likelihood that transmission capacity will be constrained is noticeably greater. The effect is largest if the increased load is met entirely by wind generation.

Consistent with the purpose of higher physical response capability, additional physical responsive capability in the absence of any change in load tends to substantially reduce the likelihood of a binding transmission constraint. Over-forecasting load tends to lessen the likelihood of at least one binding transmission constraint. This may be because an excessive forecast leads to more generating capacity being in a state of readiness. The opposite error of under-forecasting load more strongly increases the chance of encountering a binding transmission constraint. A possible explanation is that low load forecasts lead to the "less than ideal" generators being active on the network. Consistent with the rapid growth in NW load identified earlier, forecasts of higher fractions of load coming from the West zone tend to increase the chance of encountering a binding transmission constraint. In other words, when demand rises at the end or supply rises at the beginning of an isolated or not-well-connected line, there is a propensity for transmission capacity to become constrained.

The market-clearing price of power in each five-minute interval, also called the "shadow price of power balance" by ERCOT, is a component of all the LMPs across the ERCOT system. It should depend on where the load intersects the supply stack. Since high loads also strain the transmission system, we found, unsurprisingly, that the price is on average 85% higher in intervals when at least one transmission constraint is binding.

We also found that an increase in load fully met by thermal or wind generation raises the shadow price regardless of whether any transmission constraints bind. If the increase in load is satisfied entirely by solar, however, the shadow price increases on average only when at least one transmission constraint is binding.

We also found that increasing wind or solar, while backing out an equivalent amount of thermal generation, reduces the shadow price of power balance, regardless of whether any transmission constraints bind. Such an outcome would be consistent with the so-called "merit order effect." The addition of zero (or, with subsidies, negative) marginal cost generation pushes the supply stack to the right and tends to lower equilibrium power prices. However, the estimated differences in the effects of marginal changes are all no larger than the estimated standard errors. Evidently, the merit order effect is quite weak in the current ERCOT system.

Having more physical responsive capability (RC) available tends to reduce the shadow price of power balance regardless of whether any transmission constraints are binding. However, the effect is more than four times stronger when at least one transmission constraint is binding. This is consistent with the observation that having dispatchable generation resources close to load can substitute for inadequate transmission capacity.

When at least one transmission constraint is binding, the components of the LMPs apart from the shadow price of power balance give us a measure of the severity of the binding constraints. Suppose a transmission link flowing power from Location A to Location B is operating at capacity. The LMPs at both A and B will be affected. The LMP at A will be reduced as a result of the constraint, while the LMP at B will be increased. These price changes send signals to potential investors in new generating plants and/or new loads (such as cryptocurrency mining) about where to best site new facilities, as well as where to build new or upgraded transmission capacity between A and B.

The prices also send a signal, especially if the constraints bind only periodically, about where to site new storage facilities. The resulting lower prices at A disincentivize locating new generation, but incentivize locating new load sources, at A. The opposite is true of B. Similarly, if a private firm could build transmission capacity without needing to take account of its impact on the rest of the system, the arbitrage opportunity reflected in the price differential between A and B is the potential revenue source that could pay for the cost of the transmission investment.[25] If the constraint is only present periodically, storage can be sited at B to take advantage of lower power prices when the constraint is nonbinding, and higher prices when the constraint is binding — i.e., buy low, sell high.

The extent to which a binding transmission link between A and B affects the LMPs at A and B depends on what ERCOT calls "shift factors." These reflect how increased net demand at the receiving Location B or increased net supply at the source, Location A, changes flow across that line from A to B. Intuitively, a location that is better connected to the rest of the network will have a lower shift factor since it has many alternative locations to either send power to, or obtain power from, in addition to the single location at the other end of the constrained line.[26]

Impact on LMPs

Statistical analysis of the shadow prices of binding transmission constraints suggest, first, that binding transmission constraints tend to have greater negative than positive impacts on LMPs. The implication is that the shift factors on locations flowing power into constrained lines (called "source" locations) are larger than the shift factors on locations where power is flowing out of the same constrained lines (called "sink" locations). In turn, that implies that source locations tend to be less well-connected to the rest of the network than sink ones, which signals that the most severely congested lines tend to connect isolated, remote generators to locations that are much better connected to the rest of the network.

A second observation is that magnitudes of the effects of exogenous factors on the LMPs increase greatly once load gets beyond around 72 GW (which is about 13 GW short of the 2023 peak load for ERCOT). During those peak periods, there are many very large positive and negative LMPs, but the negative ones are generally about double the magnitudes of the positive ones. An increased system load exacerbates the effects of binding transmission constraints, especially for links connecting isolated, remote generators to the network, such as far-flung wind farms. Providing further incentives to increase the capacity of such generators would mean that they would add even less output to the network than their nameplate capacity and typical wind speeds in their location would imply.

A more detailed statistical analysis was undertaken to relate the costs of binding transmission constraints to generation sources and physical responsive capability. Multiple sources of evidence revealed substantial differences between effects on the *number* of constrained links versus the *severity* of the constraints. The evidence suggested that the source ends of constrained links tend to be connected to solar or wind generation, but that wind tends to be much more seriously bottlenecked than solar.

In summary, consistent with the observation of regional demand growth and regional capacity expansion across ERCOT, we find that LMP behavior reflects the discrepancy between the location of generation capacity and load. In particular, when transmission constraints emerge, upstream resources that are bottlenecked tend to see significant discounts, while downstream load centers that are constrained tend to see somewhat lower, but nevertheless substantial, premiums. Interestingly, the severely bottlenecked locations tend to be concentrated in South Texas, close to the Mexican border. More moderately constrained locations are in the Houston and Dallas metro areas. The constrained sinks tend to be concentrated in the central part of the state — in San Antonio, Austin, and surrounding regions. In sum, this indicates a multipronged approach may be needed. Expanded transmission capacity from the South region may be warranted, but encouraging new generation or storage resources sited in the ERCOT regions that tend to see the highest loads — DFW, Central and Coast, in particular — may also be worthwhile.

Final Remarks and Recommendations

Electricity reliability and resource adequacy in ERCOT have been at the top of legislative, regulatory, and commercial agendas in Texas since Winter Storm Uri in February 2021. But Uri was the stress test that exposed issues that needed attention even in the absence of such extreme events. Texas has seen tremendous growth in wind and solar generation capacity over the past 20 years, and it is now No. 1 in the nation in terms of *existing* wind capacity and No. 1 in terms of *planned* capacity additions for wind and solar. Such aggressive growth of intermittent resources, while motivated by environmental goals, will compromise reliability if there is little-to-no concurrent addition of dispatchable forms of generation. Continued growth in system load and changes in its geographic distribution only exacerbate matters.

In Texas, electrification of home heating is already significantly higher than the national average (61.5% versus 39.8%), and aspirations to electrify everything — including the continued electrification of oil and gas operations, cryptocurrency mining, adoption of EVs, expansion of CCS, and growth of the "green" hydrogen industry — will likely drive substantial additional load growth. As such, the issue of resource adequacy to ensure reliability must be tackled now, rather than later.

An understanding of the evolution of ERCOT to now highlights why significant expansion of transmission into and out of the region is unlikely. The desire to avoid federal oversight of intrastate networks predates the creation of ERCOT by decades. While this leaves gains from trade with neighboring regions off the table, any commercial incentive to capture those gains has not been sufficient to alter course, and the future of reliability in ERCOT is likely to remain a function of what happens in ERCOT.

The generation mix in ERCOT has changed considerably since 2000, and it has been tilted heavily toward wind and solar capacity, supported by robust subsidies and mandates. This has produced significant increases in generation from each resource, and since both resources are intermittent, it has also increased variability and unpredictability in overall supply. Therein lies the challenge. Grid stability is a function of resource adequacy that matches the temporal frequency of load. If resources cannot be made available on demand, the risk of grid failure increases. This means grid stability can only be warranted if there is sufficient dispatchable capacity available on standby. This cannot be baseload generation that is already dedicated. Holding such capacity in reserve, however, is expensive because backing it down when renewable output is available increases the levelized cost of the backup supply.

The outcomes of recent policy changes remain to be seen. But actions such as Texas SB 2627 and HB 1500, combined with Proposition 7, have the potential to alleviate emerging risks. That stated, there remain financial risks for capacity expansion — even with the new incentives — that are connected to regional siting decisions and transmission constraints in ERCOT. Since profitability must be sufficiently high to justify allocating capital to capacity investments, transmission constraints can encumber investment, particularly when the new capacity will only run when wind and solar are not available. As such, there is a strong case that the electricity price determined in the wholesale market needs to reflect the value of reliability.

ERCOT is also undergoing some significant changes on a regional basis. For example, regional peak loads have increased by a greater amount than regional average loads, and the variation in hourly load has also increased. This is occurring independently of growth in dispatchable resources, and it highlights an even greater need for flexibility in the generation mix as well as appropriate signals to have it optimally located to respond to higher peaks and larger variation.

The combination of greater variability in both load and generation creates stresses on the grid that manifest in LMPs. Higher generation from wind, and to a lesser extent solar, resources tends to exacerbate transmission constraints at almost all times, but the consequences for LMPs are especially extreme during peaks. LMP behavior also reflects the discrepancy between the location of generation capacity and load. Resources that are upstream of transmission constraints tend to see significant discounts, while load that is downstream of transmission constraints tends to see significant premiums. The most severely bottlenecked locations tend to be concentrated in South Texas, while more moderately constrained locations are in the Houston and Dallas metro areas. Weaker constraints are evident near San Antonio, Austin, and surrounding regions. As a result, targeted investments in transmission and proper siting of future dispatchable generation and storage capacity could all play a significant role in ensuring reliability in ERCOT.

In summary, opportunities to overcome system-wide concerns as well as subregional issues abound. There are significant opportunities to relieve current stresses on the grid caused by historical inadequate investment in dispatchable capacity and resource siting that has been geographically inconsistent with system load. As such, reliability can be enhanced with an adequate “insurance” policy, which can take many forms.

- Investing in dispatchable forms of generation that can be called on when intermittent resources are not available while load is high.

- Investing in storage capacity where downstream constraints exist. Specifically, locating storage capacity in utility areas and/or alongside industrial consumers can present opportunities to reduce purchases from the grid during high-demand periods, thereby lowering pressure on LMPs and allowing power to be redistributed.
- Siting longer-term storage capacity upstream of constraints, especially where congestion occurs due to above-normal wind and solar generation. This effectively acts like production area storage in natural gas markets, allowing a "smoothing" of sales from intermittent resources and promoting a more efficient use of transmission capacity.
- Expanding transmission capacity to alleviate existing constraints. However, the frequency and severity of those constraints matter for the economic desirability of the transmission capacity investment. Capacity that is only needed at times of stress and is heavily underutilized at all other times can be wasteful of scarce investment capital.
- Locating future dispatchable generation capacity closer to load centers to avoid bottlenecks and reduce transmission losses, thereby aiding in its profitability.
- Rescinding policies or decisions that tend to, as an unintended consequence, magnify reliability problems.

Given the abundance of options for alleviating concerns about resource adequacy and grid stability, it is likely that some combination of all of them would yield the lowest cost outcome. In the end, policy and regulation will play a key role in dictating which opportunities will be most attractive and executable.

Increased operational, dispatchable, redundant capacity on the electric grid is a necessary condition for continued growth in intermittent generation resources, but what form that capacity takes — batteries, natural gas, geothermal, etc. — is an open question that investor returns will ultimately determine. Given the abundance of opportunities that exist in Texas, it may be that a portfolio approach (including dispatchable generation capacity, demand response, conservation, efficiency in end use, distributed generation, etc.) that seeks the lowest cost outcomes is the best approach for expanding reliability in the long run.

To date in ERCOT, natural gas has been the workhorse of dispatchable redundancy. But longer-term innovations and investment in sources such as small modular nuclear and micronuclear capabilities, grid-scale battery capacity, and battery capacity strategically located in utility service areas could change this. It also should be recognized, however, that continued expansion of renewables makes some of these options more expensive by reducing the capacity factor of the backup capacity and increasing ramping costs. To achieve the most flexible path forward, policies that seek to enhance reliability should be agnostic to technology options.

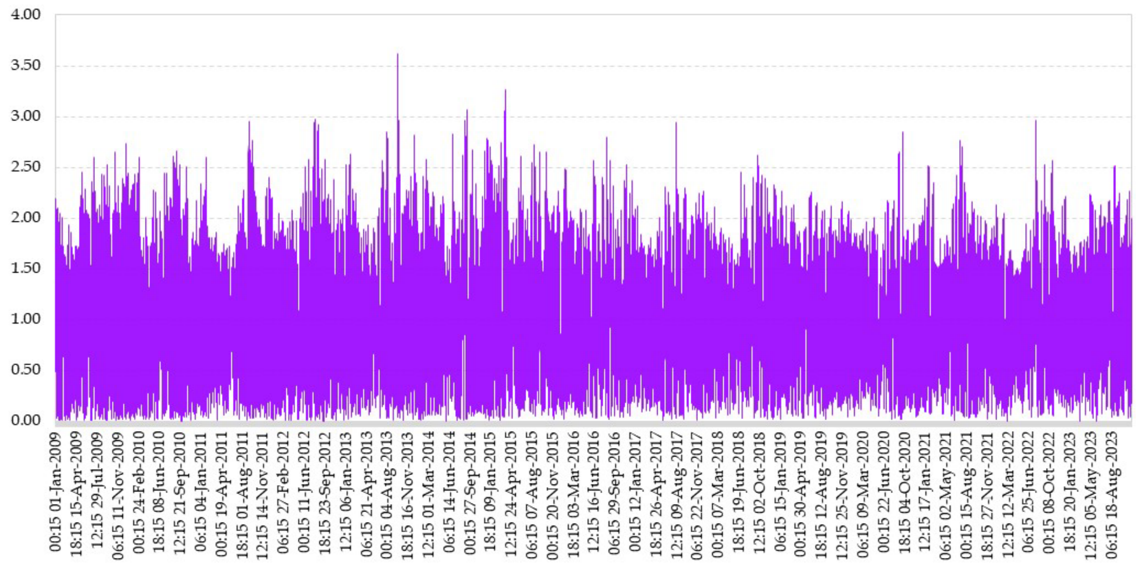
Currently, the value of reliability is not adequately priced into the electric power market. Furthermore, while environmental concerns are cited as the motivation for subsidizing and mandating renewable generation, a fundamental principle of environmental economics is that mandating particular technologies to reduce a negative externality is rarely the most efficient approach. It is far better to internalize the externality via technology-neutral emission taxes and let market actors decide as much as possible the best way to respond.

The current policy setting is even more problematic in that it attempts to address the environmental externality while ignoring the reliability externality. This leads to incentives to address one issue, but not the other. In turn, ERCOT has seen significant expansion of intermittent renewable generation capacity, but not the needed balancing service that dispatchable generation provides. Reliability problems are the unintended consequence. System reliability is essential to enabling a functional electricity market and the enormous value it provides in a modern economy and society. Pricing reforms to achieve reliability must be a priority.

No market structure is absent risks; there will always be unexpected incidents and low-probability events that can compromise any system. But allowing structural risks to reliability is unacceptable. Therefore, appropriate market design and sufficient regulatory oversight is critical. This can involve market structures that ensure sufficient backup capacity, or adequate penalties for underperformance by generators under specific obligations. In the end, resource adequacy and reliability are in the best interests of all market participants, producers and consumers alike, as they establish a platform for long-term growth. Identifying investment opportunities to provide reliability is paramount, and ERCOT has a substantial portfolio of options.

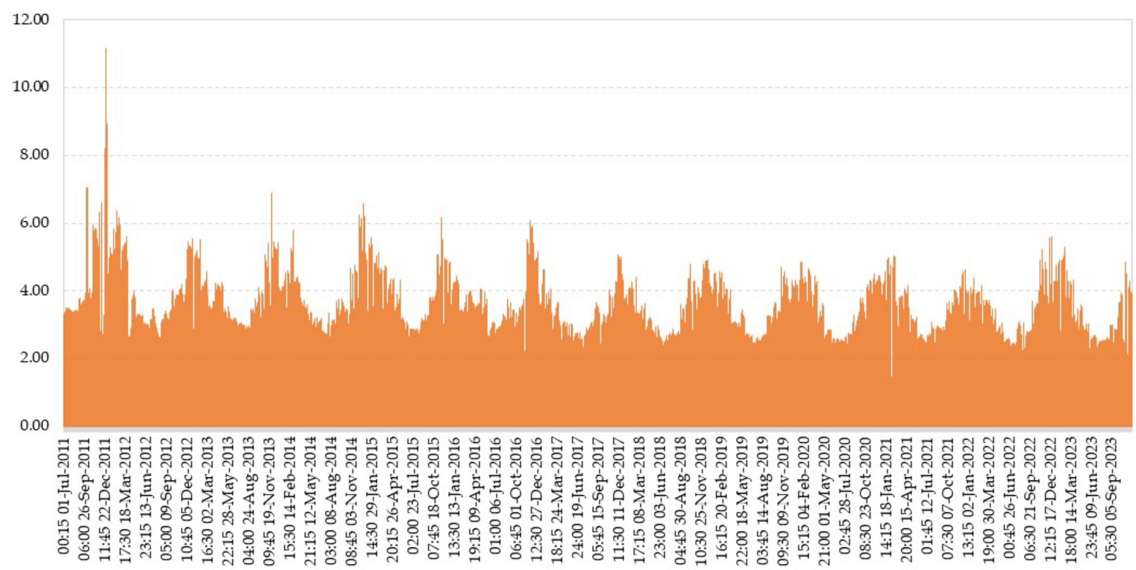
Appendix A. Mean-Normalized Wind, Solar, and Wind plus Solar Generation

Figure A1 — Mean-Normalized Wind Generation



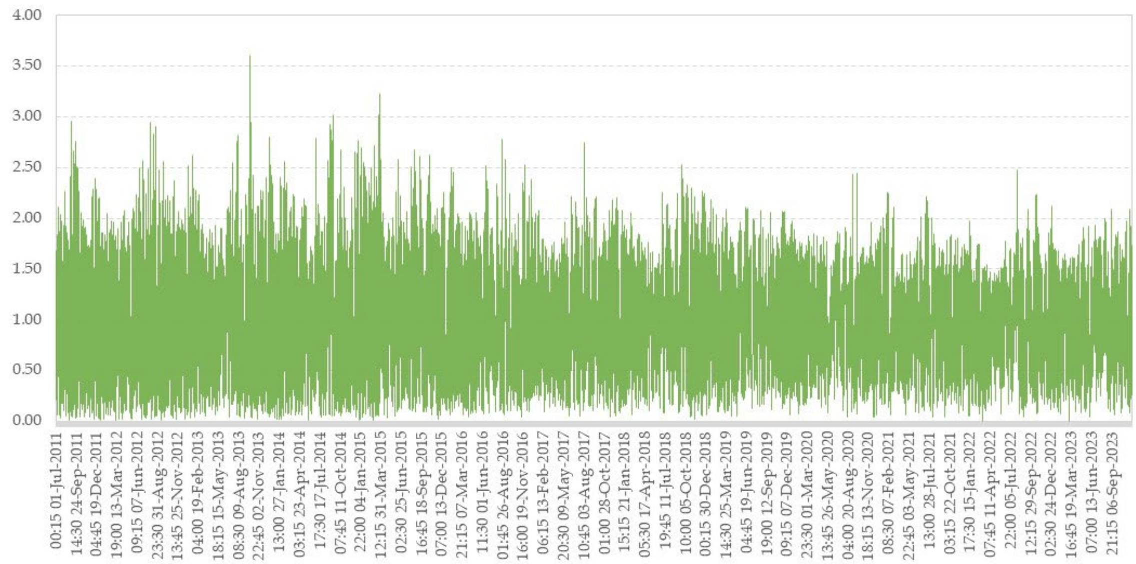
Source: Authors.

Figure A2 — Mean-Normalized Solar Generation



Source: Authors.

Figure A3 — Mean-Normalized Wind *Plus* Solar Generation



Source: Authors.

Appendix B. LMP Analysis

ERCOT calculates what it designates as locational marginal prices (LMPs) every five minutes and at occasional times in between the five-minute intervals when load is close to capacity and prices are at their upper bounds. At the most disaggregated level, they calculate LMPs for thousands of electrical buses on the network. These are then aggregated into hundreds of "settlement point" prices. For determining actual wholesale market payments to generators from distributors and other buyers in the wholesale market, the settlement point prices are aggregated temporally into 15-minute intervals. Our goal is to assess the economic role played by these LMPs. To do so, we also want to understand how the prices respond to variables affecting electricity supply and demand.

The LMPs calculated by ERCOT are "designated" as locational marginal prices, since they are not marginal prices as an economist would understand the term. The marginal price at a given location should reflect the marginal cost to the system of meeting a marginal increase in demand at that location, or the marginal value to the system of a marginal increase in supply at that location. In general, marginal changes at one location impact all other locations on the network as generation levels, loads, and flows adjust in response. From the perspective of the transmission system, a set of genuine LMPs should reflect marginal transmission losses. They should also reflect the shadow value of transmission constraints.

First, consider the case where there are no transmission capacity constraints. Suppose A and B are two electrical buses on the network with power flowing from A to B. The LMP at A (the source) should equal the LMP at B (the sink) less the marginal transmission loss between A and B. An increase in supply or decrease in demand at B, or a decrease in supply or increase in demand at A, would reduce the need to transmit power from A to B. This would in turn reduce the transmission loss. If power flows from A to B most of the time, the pricing differential would incentivize locational decisions that would tend to reduce transmission losses over time.

By contrast, the LMPs calculated by ERCOT give a single price at all electrical buses on the network when there are no transmission capacity constraints. In effect, the transmission losses throughout the network are "socialized" and allocated to all users regardless of location. Incentives for decision-makers to take those losses into account are thereby lost.

When each transmission line has a maximum capacity, the optimized load flow problem (used to solve for the cost-minimizing generation levels and transmission flows needed to satisfy a given magnitude and geographic distribution of loads) needs to include these capacity constraints with accompanying Lagrange multipliers.

If a particular transmission constraint is not binding, the corresponding multiplier is zero. If the constraint binds, however, the multiplier becomes positive. The LMP of increasing net demand at the sink should increase by the value of the multiplier on that link times a "shift factor." The latter reflects how increased net demand at the sink changes flow across the constrained link.[27] Conversely, the LMP of increasing net supply at the source should decrease by the value of the multiplier on the constrained link times a different shift factor. Intuitively, a bus that is better connected to the rest of the network will have a lower shift factor, since it has many alternative buses to either send power to or obtain power from than the single bus at the other end of the constrained line.

In principle, adjusting market prices to account for congestion incentivizes wholesale market participants to change their behavior in a way that helps ERCOT manage network congestion. In addition, positive Lagrange multipliers signal the value of *expanding the capacity* of the transmission link. When capacity is scarce, its value increases. For it to be worthwhile to incur the up-front costs of increasing the capacity of a transmission link, the cost of doing so needs to be less than the expected discounted value of the Lagrange multipliers (or value of the transmission link) over the life of the investment.

The LMPs currently calculated by ERCOT are based only on the Lagrange multipliers on transmission links that are constraining flows. They do not account for marginal transmission losses. Moreover, ERCOT caps the transmission constraint multipliers at $4500/MW for 345-kilovolt (kV) lines, $3500/MW for 138-kV lines, and $2800/MW for 69-kV lines. The current protocol also caps the shadow price of power balance at $5000/MW. These price caps imply that the short-run scheduling of generation to meet the pattern of loads will not yield minimum cost supply when any of the caps are binding. In the longer run, price caps also give incorrect incentives for the siting of generation, major loads, and transmission line upgrades.

In 2018, the Public Utilities Commission of Texas (PUCT) asked ERCOT to assess the expected benefits of including marginal transmission losses in LMPs. ERCOT used the Uplan Network Power Model to assess cost-minimizing dispatch to meet a specified level and pattern of load on the system while respecting various physical and operational constraints on system operation. They use the same approach to assess the need for transmission upgrades. The analysis allowed for three different natural gas prices based on the Energy Information Agency's 2018 Annual Energy Outlook: a low price of $2.55 per million British thermal units (MMBTU), a medium price of $3.55/MMBTU, and a high price of $3.96/MMBTU.

Including marginal transmission losses in LMPs changed production costs (which were defined to be the sum of fuel costs, variable operations, and maintenance costs), and unit start-up costs. In the low and medium gas price cases, production costs fell after including marginal transmission losses in LMPs. The primary reason was that the resulting shift in the location of active generators reduced the amount of energy required to serve a given load. In turn, the reduced fuel use was mainly the result of reduced transmission losses. ERCOT suggested that the slight rise in production costs in the high gas price case resulted from increased generation from coal-fired generators that are more remote from major loads than the gas-fired units they displaced. However, even in the high gas price case, total annual generation was lower when LMPs included marginal transmission losses.

ERCOT also found that annual generator revenues from energy sales declined when LMPs included marginal transmission losses. The reductions increased with the gas price. However, the revenue changes were not uniform across the four zones in ERCOT. The largest increases were in the Houston zone, followed by the South zone. The West zone would incur moderate losses under all three gas price scenarios, while the declines in North zone generator revenues from energy sales would be larger in magnitude than the gains in the Houston zone. The model output showed no change in wind generation (curtailments did not change), while thermal units lost revenue due to price changes *and* output reductions.

ERCOT noted that it is difficult to translate the changes in production costs and generator revenues into price changes for consumers. The main reason is that the change in LMP prices would alter the structure of prices. Under the current system, average transmission losses are recovered by fixed charges for service. Under a marginal transmission loss system, the costs of the losses would become part of the variable charge per kilowatt-hour (kWh).

Understanding When Transmission Constraints Are More Likely to Bind

To examine some of the factors influencing the LMPs, we obtained 44,681 successive observations (around 152 days of mostly five-minute intervals) of LMPs for the approximately16,700 electrical buses on the network from July 6, 2023, to Dec. 5, 2023.[28] Binding transmission capacity constraints were absent in about 22% of these five-minute intervals. Given how ERCOT calculates LMPs, all the LMPs in those intervals are identical, and the cross-sectional standard deviation in LMPs, σ, is zero. To investigate the *causes* of transmission constraints, we defined an indicator variable, $\chi = 0$ for $\sigma = 0$ and $\chi = 1$ for $\sigma > 0$, and estimated models with χ as the dependent variable. To understand factors affecting the *severity* of constraints, we examined models of the cross-sectional distribution of LMPs in the 78% of five-minute intervals where $\sigma > 0$.

As potential explanatory variables, we obtained load, forecast load, physical response capability, and wind and solar generation data for each five-minute interval in the 152-day sample period. To limit the possibility of reverse causation from the calculated prices to the explanatory variables, we ensured that all data measured in the explanatory variables predated the observed LMPs. Regarding load, physical response capability, and wind and solar generation, we used data from the five-minute interval closest to, but preceding, the time when the LMPs were measured.

Since transmission links are built to cope with peak loads, binding transmission constraints should be unlikely when loads are far below peak. Nevertheless, line outages and other equipment problems, or unusual distributions of demand or supply across the network, could strain some links. Transmission constraints are likely to become much more prevalent when load approaches system capacity. Furthermore, the shadow price associated with each binding constraint will increase with load.

While it is common to assume that load is exogenous to real-time price, some ERCOT customers have demand-side management contracts that allow their demand to be curtailed at short notice. So far as we know, these curtailments do not respond directly to price. However, circumstances where demand curtailments are typically utilized, such as when the operating reserve margins cross certain thresholds, are likely to be correlated with price. Furthermore, some large industrial customers may be able to moderate demand in response to real-time settlement point prices. The latter are averages of about 20 bus LMPs — albeit ones that are likely geographically concentrated and therefore highly positively correlated.[29] It would seem implausible, however, that load during a five-minute interval could respond to LMPs calculated after the *end* of that interval. Furthermore, customers regularly experiencing transmission constraints can hedge the expected costs using so-called congestion revenue rights (CRR). These are either obligations or options tied to a transmission link. In the case of an obligation, the holder receives hourly revenue equal to the MWh amount of CRR owned times the difference between the sink and source average hourly LMPs on the link. In the case of an option, the holder receives the maximum of the obligation amount and zero. CRR ownership should make load more sensitive to the average than to the short-run fluctuations in LMPs.

Forecast loads might also affect transmission flows by increasing the amount or location of generating capacity scheduled to be available when forecast loads are high. The way we use the load forecasts requires more explanation. The forecasts pertain to loads expected over an entire hour, but for each hour, two forecasts are made approximately an hour apart: a preliminary forecast and then a forecast that is updated later. In addition, the forecasts include separate allocations of load to four zones — Houston, South (excluding Houston), North, and West. They also include the times that the forecasts were published.

It may be easiest to explain the procedure used to construct expected loads with an example. Suppose we have a set of LMPs calculated at 11:18:10 am. We then find the forecast for the 11 am-to-noon hour of that day that is published closest to, but still precedes, 11:18:10 am. We also find the forecast for the noon-to-1 pm hour that is published closest to, but still precedes, 11:18:10 am. We then calculate a weight, $\omega = (18*60 + 10)/3600$, representing the fraction of seconds within the hourly interval that has elapsed between 11 am and when the LMPs were calculated. The expected load corresponding to those LMPs and the expected fractions of load occurring in each zone are then calculated as w times the chosen 11 am-to-noon hour forecast plus $(1-\omega)$ times the chosen noon-to-1 pm hour forecast.

Even though the forecasts we use are published at least 30 or 90 minutes before the LMPs are calculated, the correlation between load and expected load is quite high. A regression of load against expected load yields a coefficient on the latter of 1.00041 with a standard error of 0.00052 and an R^2 of 0.9880. Thus, we do not include both the actual and expected load in the regressions. However, the *difference* between the forecast and the actual load could measure stress on the system. The effects of the actual exceeding the forecast load could differ from the effects of the forecast exceeding the actual. Hence, we defined two variables (both greater than or equal to zero) measuring the size of the forecast errors in each case:

$$FE_L = \max\{load - expected\ load, 0\} \text{ and } FE_H = \max\{expected\ load - load, 0\}.$$

The expected zonal loads included in the analysis, calculated using the same weight w, are the proportions Hou = Houston load/expected load, Nth = north zone load/expected load, and Sth = south zone load/expected load (the fraction of load in the West zone is omitted).

The physical responsive capability, denoted *RC*, should allow ERCOT to limit the negative effects of emergencies or unexpectedly large loads. ERCOT defines physical responsive capability as "the generation and load resources that are available to respond quickly to system events in case of sudden changes, such as an unexpected outage at a large generating unit." ERCOT may attempt to schedule more responsive capability as the system approaches capacity and more constraints bind. However, the time series for *RC* is relatively smooth compared to load and especially compared to the average of the LMPs. Most importantly, it is difficult to see how *RC* for a given five-minute interval would be endogenous to the LMPs calculated beyond the *end* of that interval.

The amounts of wind (*wind*) and solar (*solar*) generation could influence the likelihood of binding transmission constraints for at least two reasons. On the one hand, low output from these uncontrollable sources could impose more stress by requiring more thermal generation to compensate. These marginal thermal resources are likely to have higher marginal costs but may also be sited in different locations compared to inframarginal dispatchable capacity. We also include total thermal generation (*thermal*) as an explanatory variable to capture this effect.[30]

Changes in wind and solar generation would also affect the locations of active generators. While many conventional plants have been located relatively close to major loads, wind and solar plants are located where the natural resources are more readily available and thus tend to be much more widely dispersed. Moreover, the existing transmission system was developed largely to accommodate the locations of conventional plants relative to major loads. The different locations of wind and solar plants compared to conventional plants may have required the construction of new, or upgrade of existing, transmission lines. Lags in these responses could cause more transmission constraints to bind when wind and solar output is relatively high.

Most of the time, wind and solar generation depends on weather conditions and is exogenous to what is happening in the electricity market, including the market price. However, binding transmission constraints could require output from these generators to be curtailed. Nevertheless, curtailment decisions during a particular five-minute interval are unlikely to be affected by LMPs calculated beyond the *end* of that interval.

Table B1 presents the ranges, means, and standard deviations of the explanatory variables. Table B2 gives the results for the logit model with c as the dependent variable. Robust estimates of the standard errors were used to allow for heteroskedasticity and potential serial correlation in the errors. The odds ratios in Table B2 measure the effect of a "1 unit" change in the relevant explanatory variable on $p/(1-p)$, where p is the probability that $\sigma > 0$.

Table B1 — Moments of Explanatory Variables Included in the Logit

	Minimum	Maximum	Mean	Standard Deviation
load (in GW)	34.7320	85.6120	56.4221	13.1832
thermal (in GW)	14.9200	68.3970	40.7786	12.1222
wind (in GW)	0.3587	25.3709	11.0630	5.7422
solar (in GW)	0	13.7884	4.1816	5.0630
RC (in GW)	2.1367	13.4731	7.4259	1.4949
FE_L (in GW)	0	7.3960	0.6255	0.8674
FE_H (in GW)	0	6.0220	0.5029	0.8401
Hou	0.2039	0.3084	0.2578	0.0111
Nth	0.2642	0.4106	0.3346	0.0244
Sth	0.2258	0.3315	0.2620	0.0118

Source: Authors.

We can use several criteria to examine the adequacy of the logit model. If we define $D = 1$ when the predicted probability $p \geq 0.5$ and $D = 0$ when the predicted $p < 0.5$, then the correspondence between D and χ is as given in Table B3. Overall, 77.81% of outcomes were correctly classified. The percentages in parentheses in each cell are (column percentage, and row percentage). These indicate that the logit model has substantial difficulty correctly predicting when $\sigma = 0$.

Table B2 — Logit for χ as Dependent Variable

Variable	Estimated Coefficient	Robust Std Err	p-value H_0: coef = 0	Implied Odds Ratio
load	0.4324	0.0315	0.000	1.5410
thermal	-0.3536	0.0325	0.000	0.7021
wind	-0.2046	0.0316	0.000	0.8149
solar	-0.2860	0.0318	0.000	0.7512
RC	-0.2295	0.0106	0.000	0.7949
FE_L	0.1881	0.0221	0.000	1.2069
FE_H	-0.0231	0.0172	0.180	0.9772
Hou	-58.1238	2.1638	0.000	5.72E-26
Nth	-51.0310	2.1139	0.000	6.88E-23
Sth	-21.7309	2.2784	0.000	3.65E-10
constant	34.3241	1.4152	0.000	8.07E+14

Source: Authors.

Note: H_0: All coefficients are zero, $\chi^2_{10} = 5524.29$, pseudo-$R^2 = 0.1231$.

There is other evidence that the logit model is inadequate. Estimating a new logit model with the linear predicted value from Table B1, $\hat{X}$, and its square, $\hat{X}^2$, as explanatory variables yield an estimated coefficient on $\hat{X}$ of 0.3856, with a standard error of 0.0346. This implies that the original model has explanatory power. However, the estimated coefficient on $\hat{X}^2$ of 0.2228 has a standard error of 0.0151. This suggests that the model is not correctly specified. Adding squared and cross-product terms in variables ameliorated, but did not correct, the problem.

Table B3 — Predicted and Actual Instances when $\sigma > 0$

	$\chi = 1$	$\chi = 0$	Totals
$D = 1$	33,351 (95.84%, 79.75%)	8,468 (85.68%, 20.25%)	41,819
$D = 0$	1,447 (4.16%, 50.56%)	1,415 (14.32%, 49.44%)	2,862
	34,798	9,883	44,681

Source: Authors.

Estimating a probit model instead of a logit slightly reduced the statistical significance of the coefficient on $\hat{X}^2$, but the hypothesis that the coefficient is zero was still rejected at a level better than 1%. In the probit model, the probability that $\chi = 1$ is

$$\Pr\left(\chi_t = 1\right) = \Phi\left(\sum_{j=1}^{11} \beta_j x_{jt}\right),$$

where Φ is the standard normal cumulative distribution function. It can be generalized to allow the distribution to have non-constant variance. Assuming that the standard deviation at t is $\exp\left(\sum_{k=1}^{m} \gamma_k z_{kt}\right)$, the heteroskedastic probit model implies

$$\Pr\left(\chi_t = 1\right) = \Phi\left(\sum_{j=1}^{11} \beta_j x_{jt} \Big/ \exp\left(\sum_{k=1}^{m} \gamma_k z_{kt}\right)\right).$$

Table B4 gives maximum likelihood estimates of the heteroskedastic probit model, allowing the variance to depend on *load*, *thermal*, *RC*, *wind*, *solar* and the forecast errors. To test the specification, we again estimate a probit model with the linear predicted value from Table B4, $\hat{X}$, and its square, $\hat{X}^2$, as explanatory variables, using the same model for the variance. The estimated coefficient on $\hat{X}$ is now 1.1437, with a standard error of 0.1419. The estimated coefficient on $\hat{X}^2$ becomes -0.5629, with a standard error of 0.3451, which is not statistically significantly different from zero at the 10% level. In addition, the estimated coefficients in the variance equation apart from those on *RC* and FE_L change by less than 1%.[31] We conclude that the model in Table B4 is the preferred specification.

Table B4 — Heteroskedastic Probit for χ as Dependent Variable

Variable	Estimated Coefficient	Robust Std Err	*p*-value H_0: coef = 0
load	-0.0035	0.0043	0.416
thermal	0.0095	0.0054	0.077
wind	0.0195	0.0067	0.004
solar	0.0124	0.0048	0.010
RC	-0.0200	0.0026	0.000
FE_L	0.0121	0.0054	0.026
FE_H	-0.0028	0.0013	0.031
Hou	-4.7628	0.9910	0.000
Nth	-4.6277	0.9989	0.000
Sth	-2.6240	0.5951	0.000
constant	3.2650	0.6742	0.000
$\ln\sigma$ coefficients:			
load	-0.4188	0.0387	0.000
thermal	0.4045	0.0399	0.000
wind	0.3512	0.0390	0.000
solar	0.3604	0.0373	0.000
RC	-0.0208	0.0203	0.306
FE_L	-0.0559	0.0476	0.241
FE_H	-0.1622	0.0187	0.000

Source: Authors.

Note: H_0: All coefficients are zero; mean model, $\chi^2_{10} = 117.44$, and variance model, $\chi^2_7 = 874.60$.

The estimated coefficients on *load* should be evaluated in concert with those on *thermal*, *wind*, and *solar*. As previously mentioned, while thermal plus wind plus solar generation does not equal total load, the difference is usually less than 5% of ERCOT load. Hence, changes in *load* are largely matched by changes in some combination of *thermal*, *wind*, and *solar*.

The estimated coefficients imply that an increase in load that can be met entirely by thermal generation, holding reserve capacity and the other variables fixed, has a slight positive effect on the probability of encountering at least one binding transmission constraint. By contrast, an increase in load that is met by an equal increase in solar generation has a larger impact on the likelihood that transmission capacity will be constrained. The effect is largest if the increased load is met entirely by wind generation.

Consistent with the purpose of higher physical response capability, additional physical responsive capability (*RC*) in the absence of any change in *load* tends to substantially reduce the likelihood of a binding transmission constraint. Another implication of the relatively large coefficient on *RC*, however, is that an increase in load that is met by a reduction in *RC*, while *thermal*, *wind*, and *solar* are held constant, has an even larger impact on the likelihood of encountering binding transmission constraints than when increased wind generation supplies the higher load.

Over-forecasting load ($FE_H > 0$) tends to lessen the likelihood of at least one binding transmission constraint, while under-forecasting it ($FE_L > 0$) more strongly increases the chance of encountering a binding transmission constraint. A possible explanation is that incorrect forecasts lead to the "less than ideal" generators being active on the network.

Forecasts of higher fractions of load in the Houston, North, and South zones tend to reduce the likelihood of binding transmission constraints. Forecasting a higher fraction of load in the West zone (the omitted category) thus tends to increase it.

The model in Table B4 also has implications for the variability of the underlying latent variable determining whether transmission constraints bind. An increase in load matched by an increase in thermal, wind, or solar generation tends to reduce variance, as does an increase in physical response capability and any load forecast error. From the formula for $\Pr(\chi = 1)$ in the heteroskedastic model, any such reductions in the variance tend to increase the probability that at least one transmission constraint will bind.

Determinants of the Shadow Price of Power Balance

When transmission constraints are not binding, all the LMPs equal the market-clearing power price in each interval, which is called the "shadow price of power balance" by ERCOT and denoted λ. In practice, λ is the Lagrange multiplier on the power balance constraint resulting from the solution to a security-constrained economic dispatch (SCED) algorithm. The inputs to the algorithm are the prevailing level and pattern of load on the system and the supply schedules submitted by producers. The SCED solves for the least-cost dispatch of resources to meet the loads while respecting various physical and operational constraints. The most important physical constraints for our purposes are the configurations and capacities of the various transmission links. Satisfying the constraints may require units to be committed out of merit order as determined by the submitted supply schedules.

When at least one transmission constraint binds, the value of λ in each period is still a component of all the LMPs. However, an additional shadow price is generated for each capacity-constrained transmission link. As mentioned in the introduction of this report, these additional prices affect the LMPs at both ends of a constrained link. The shadow price times a shift factor is subtracted from the bus flowing power into the constrained link, while the shadow price times a (usually different) shift factor is added to the LMP at the bus receiving power flowing out of the constrained link. It follows that the added components reflecting transmission constraints will generally differ across links. In addition, the LMPs at all buses on links that are not congested equal λ. We thus can identify λ_t with the modal LMP at t.

Table B5 gives some summary statistics for the distributions of λ from periods when $\chi = 0$ versus periods when $\chi = 1$. Negative values of λ can occur only when a generator is willing to pay to be allowed to supply more power. This can happen when renewable generators receive a production subsidy and are willing to pay up to the amount of the subsidy to avoid curtailment. It can also happen, for example, with nuclear plants, when ramping costs are high and the generator is willing to pay to avoid changing output. In these cases, the lowest horizontal segments in the submitted supply schedules are at negative prices. While that is a necessary condition for a negative value of λ, it is not sufficient. To obtain $\lambda < 0$, a generator submitting a supply schedule with a negative price must be the marginal generator actively contributing output to the network. The first column of Table B5 shows that, in our sample, such situations only occur when one or more transmission constraints are simultaneously binding.

Table B5 — Summary Statistics for the Distribution of λ as a Function of χ

Statistic	$\chi = 0$ ($N = 9{,}883$)	$\chi = 1$ ($N = 34{,}798$)
Minimum	0	-16.65
1%	10.51	0
5%	14.60	12.32
10%	16.23	15.77
25%	18.87	20.29
50%	22.48	25.32
75%	27.27	37.69
90%	43.01	75.53
95%	75.00	164.63
99%	676.66	2378.51
Maximum	5000	5202.27
Mean	53.33	98.61
Std Dev	294.55	453.44
Skewness	14.47	8.70
Kurtosis	226.24	84.25

Source: Authors.

More generally, the distribution of λ in periods when $\chi = 1$ is less positively skewed and has a greater proportion of values of $\lambda < 16$. Nevertheless, there are also more periods with $\lambda > 20$ when $\chi = 1$, and the mean and median values of λ are higher in the sample with $\chi = 1$. The standard deviation of λ also is higher when $\chi = 1$. Thus, periods with at least one binding transmission constraint ($\chi = 1$) tend also to be periods where load presses against the available generation capacity, so λ is high.

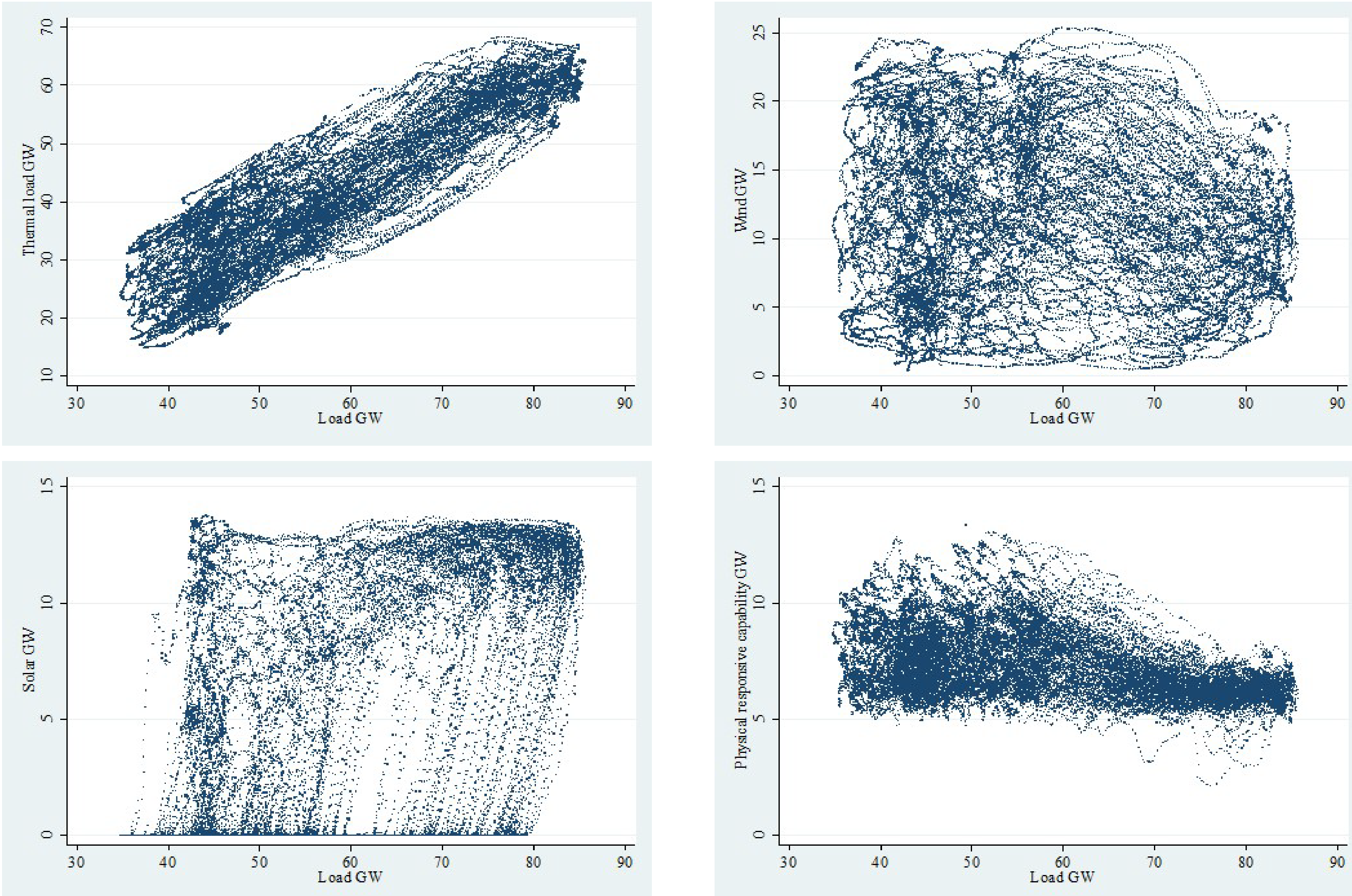
We turned to a regression model to understand the key determinants of λ. As already noted, λ is the solution from the SCED algorithm for the Lagrange multiplier on the power balance constraint. In a least-cost dispatch without any physical and operational constraints, λ should match the bid to supply incremental power from the marginal generator, that is, the active supplier where the load intersects the supply stack. However, the various constraints will modify this relationship. Thus, λ is likely to be a nonlinear function of the exogenous variables affecting supply and demand for power. From Table B5, the skewness (the third moment about the mean divided by the standard deviation cubed) and kurtosis (the fourth moment about the mean divided by the variance squared) measures of the distributions of λ imply that they are positively skewed, have very "fat" tails, and are far from normal.

With these considerations in mind, we used a local-linear nonparametric kernel regression to relate λ to the first five explanatory variables in Table B1. We reduced the number of potential explanatory variables because kernel regression is very computationally intensive. In addition, the first five variables relate to the current demand and supply factors impinging on the system. The remaining variables in Table B1 relate primarily to the forecasts of load, which should be less relevant to determining the current solution to the SCED algorithm. Their main impact would be indirect in that they would influence the generating resources scheduled to be operational prior to the period in question.

Figure B1 shows scatter plots of *thermal*, *wind*, *solar*, and *RC* against *load* for the 34,798 intervals when $\chi = 1$. These illustrate a tendency for *wind* to be lower when *load* is higher, while the opposite is true for *solar*. A decline in *RC* with *load* is also evident. The latter suggests that generating plants that provide physical responsive capability when *load* is lower are called on for normal supply at times when *load* is higher.

The scatter plots also reveal many "relatively straight or wiggly lines." These are subsets of successive intervals where the variables appear to be closely related. This may reflect a tendency for changes in weather that affect *load* (especially ambient temperatures) to be correlated with changes that affect output from *wind* (wind velocity) or *solar* (cloudiness or time of day) generators. Thermal generation would then also be correlated with *load* as dispatchable generators need to ramp up and down to accommodate the fluctuations in renewable output.

Figure B1 — Scatter Plots of the Kernel Regression Explanatory Variables



Source: Authors.

Table B6 gives moments of the explanatory variables as a function of χ. The upper half relates to intervals where there are no binding transmission constraints; the bottom half relates to intervals where there is instead at least one such constraint.

Table B6 — Moments of Explanatory Variables as a Function of χ

Statistic	*load*	*thermal*	*wind*	*solar*	*RC*
$\chi = 0$					
Minimum	34.861	19.6408	0.3587	0	4.2583
1%	36.929	22.1411	0.7049	0	4.9716
5%	39.892	26.1429	1.9265	0.0001	5.4867
10%	42.227	28.4092	2.7873	0.0001	5.7802
25%	45.909	33.3759	4.7938	0.0002	6.4048
50%	51.96	39.1043	8.1391	0.1875	7.4780
75%	58.888	46.1316	12.6731	8.8741	8.7139
90%	68.272	53.5996	15.7671	11.9558	9.867
95%	74.16	59.688	16.972	12.6455	10.4321
99%	78.331	65.2288	19.2071	13.2354	11.4115
Maximum	84.276	68.0718	21.8032	13.6531	13.4731
Mean	53.4423	40.2236	8.7887	4.0609	7.6560
Std Dev	9.9183	9.6909	4.8620	4.8790	1.5586
Skewness	0.7186	0.5340	0.3231	0.6464	0.5071
Kurtosis	2.9994	2.9611	2.0608	1.7492	2.8111
$\chi = 1$					
Minimum	34.732	14.92	0.3641	0	2.1367
1%	36.33	17.6663	1.2605	0	4.9791
5%	38.73	21.6439	2.7024	0	5.5330
10%	41.373	24.6661	4.1051	0	5.7998
25%	45.162	31.352	6.8765	0.0002	6.2596
50%	54.7675	39.0644	11.48	0.1330	7.0071
75%	69.106	51.6169	16.2798	9.8181	8.3009
90%	78.431	59.6295	19.8772	12.2597	9.4341
95%	81.719	61.9797	21.3166	12.8002	10.0893
99%	84.259	65.4311	23.3493	13.3056	11.6566
Maximum	85.612	68.397	25.3709	13.7884	13.3637
Mean	57.2684	40.9362	11.7089	4.2159	7.3606
Std Dev	13.8558	12.7241	5.8085	5.1136	1.4698
Skewness	0.4014	0.1923	0 .1283	0.6296	0.7977
Kurtosis	1.9240	2.0464	2.0555	1.6548	3.4618

Source: Authors.

The moments in Table B6 show that intervals where at least one transmission constraint is binding tend to be associated with more extreme high values of *load*, although the lower values of *load* are distributed similarly in the two subsets. In the case of *thermal*, the distributions across the two subsets are more similar for high values than low ones. In other words, whether there are binding transmission constraints, the dispatchable thermal generators can be called on to supply similarly high amounts — but thermal generation does not get as low when at least one transmission constraint is binding. By contrast, the whole distribution of *wind* is shifted to the right in intervals when χ = 1. The variability of *wind* generation across periods is also higher. The distribution of *solar* is the most similar across the two subsets of any of the variables in Table B6. Physical responsive capability differs most in the lower values. With some binding transmission constraints, the physically responsive generators are evidently called on at least to some extent to help mitigate the effects of the constraints. There are also many intervals, however, when *RC* remains large, despite the presence of transmission constraints. That implies that the *RC* is likely not always well located to alleviate the constraints.

Considering the results from Tables B5 and B6 showing that the distributions of λ and the exogenous variables differ depending on the value of χ, we estimated separate regression models for the two regions. Table B7 presents the results for the 9,883 intervals when χ = 0, while Table B8 does the same for the 34,798 intervals when χ = 1.

As with the other kernel regressions reported later in the paper, the bootstrap standard errors were obtained using 500 replications.[32] The bandwidth was chosen by cross-validation to minimize the integrated mean squared error of the prediction. An Epanechnikov kernel was used. All estimations were done using the STATA routine npregress.

Table B7 — Kernel Regression Model for λ when χ = 0

Variable	Estimated Coefficient	Bootstrap Std Err	*p*-value H_0: coef = 0
Mean			
λ	59.0015	3.5434	0.000
Effects (Averages of Derivatives)			
load	35.3141	6.7866	0.000
thermal	-33.0012	6.9210	0.000
wind	-34.8864	6.9530	0.000
solar	-36.6088	7.1541	0.000
RC	-21.4800	2.2135	0.000
R^2 = 0.6949			

Source: Authors.

The coefficients labeled "Mean" in Tables B7 and B8 are the average of the predicted means from the local linear regressions. They are close to the respective unconditional means of λ in Table B5. The coefficients labeled "Effects" in Tables B7 and B8 are averages of the derivatives of λ with respect to each variable. Although the means of λ are different in the two subsamples, the effects of marginal changes in each of the explanatory variables, apart from *RC*, are similar.

Table B8 — Kernel Regression Model for λ when $\chi = 1$

Variable	Estimated Coefficient	Bootstrap Std Err	p-value H_0: coef = 0
Mean			
λ	107.8794	2.6053	0.000
Effects (Averages of Derivatives)			
load	37.5752	3.3891	0.000
thermal	-33.2994	3.4681	0.000
wind	-36.4283	3.2704	0.000
solar	-34.1892	3.4708	0.000
RC	-88.0728	3.4483	0.000
$R^2 = 0.7008$			

Source: Authors.

As with the logit and probit models for χ, the coefficients on *load* are opposite in sign to those on *thermal*, *wind*, and *solar*. They imply that an increase in load fully met by thermal or wind generation raises the shadow price, λ, regardless of whether any transmission constraints bind. If increased load is satisfied entirely by solar, however, λ increases on average when $\chi = 1$ but slightly decreases on average when $\chi = 0$, albeit not statistically significantly different from zero.

The estimates also imply that increasing *wind* or *solar* while backing out an equivalent amount of thermal generation reduces λ, regardless of whether any transmission constraints bind. Such an outcome would be consistent with the so-called merit order effect. The addition of zero (or, with subsidies, negative) marginal cost generation pushes the supply stack to the right and tends to lower equilibrium prices. However, the estimated differences in the effects of marginal changes are all no larger than the estimated standard errors.

Having more physical responsive capability available (higher *RC*) tends to reduce λ regardless of the value of χ. However, the effect is more than four times stronger when at least one transmission constraint is binding.

Determinants of ERCOT LMPs when Some Transmission Constraints Bind

We next focus on the effects the same exogenous variables have on the transmission constraint shadow prices in the 34,798 intervals when $\chi = 1$. The LMP at bus i in period t can be written as $\text{LMP}_{it} = \lambda_t + \mu_{it}$. Subtracting λ_t from LMP_{it} yields μ_{it}. In this section, we examine the effect of *load*, *thermal*, *wind*, *solar*, and *RC* on the μ_{it}.

Recall that a constrained transmission link flowing power from bus i to bus j in period t leads to a negative μ_{it} and positive μ_{jt} that depend on the same Lagrange multiplier. However, these positive and negative values will only have the same absolute values if the shift factor for bus *i* equals that for bus *j*. In turn, the shift factor for a bus will depend on how well connected that bus is to the rest of the network. A better-connected bus will have a smaller shift factor.

Since we can have many thousands of non-zero μ_{it} values in any period, we reduced the dimensionality of the problem by investigating the behavior of some moments of the cross-sectional distributions of the μ_{it} across the 34,798 periods. Although many moments could be chosen, this discussion will focus on the mean and the median, the standard deviation, and the 5th and 95th percentiles of each distribution.

Table B9 presents summary statistics for the distribution of each moment across 34,798 intervals. Since a negative μ_{it} is coupled with a positive μ_{jt} on a constrained link between *i* and *j*, it is not surprising that the mean and median values tend to cluster near zero. The median (across periods) of the cross-sectional medians (in each period) is exactly zero, while even the 25th and 90th percentiles of the cross-sectional medians are less than 1. The median, 25th and 75th percentiles, and mean of the cross-sectional means are also close to zero.

Table B9 — Summary Statistics for Moments of the Cross-Sectional Distributions of μ_{it}

Statistic	Mean	Median	Std Dev	5%	95%
Minimum	-519.32	-471.35	0.00036	-4190.36	-11.07
1%	-69.20	-38.12	0.16	-1050.44	0
5%	-7.88	-11.44	0.72	-78.78	0
10%	-4.82	-5.98	1.60	-41.08	0
25%	-1.41	-0.61	4.32	-18.74	0.4
50%	-0.20	0	9.41	-5.82	5.22
75%	0.39	0.03	27.39	-1.03	22.87
90%	3.28	0.25	86.13	-0.01	67.00
95%	6.37	1.55	131.15	0	119.64
99%	18.86	67.49	559.70	0	397.36
Maximum	225.95	909.08	2117.25	194	2194.87
Mean	-1.55	-0.16	35.79	-41.44	27.98
Std Dev	13.42	22.36	94.10	230.63	78.69
Skewness	-9.09	7.29	7.53	-10.45	8.16
Kurtosis	159.63	176.03	77.91	126.71	116.17

Source: Authors.

This conclusion is reinforced by the negative skewness in the distribution of the means and the fact that the minimum and 1 percentile values for the means are lower than the corresponding statistics for the medians. Finally, the extreme values in the 5% negative tails of the cross-sectional distributions are much further from zero than the corresponding extreme values in the 95% positive tails of the distributions. There also is superficially conflicting evidence. While 99% of the medians are below 67.49, the possibility that the median can be as high as 909.08 shows that large positive μ_{it} are possible. Table B10 presents some statistics for the cross-sectional distributions with the 20 largest medians out of 34,798.

Table B10 — Other Statistics for Cross-Sectional Distributions with the 20 Largest Medians

Median	Minimum	Mean	Maximum
303.71	-4652.74	-126.98	4828.98
303.90	-4243.95	-113.31	2248.43
307.21	-5541.89	-90.19	5311.54
308.56	-5817.35	-119.28	1278.16
310.75	-5900.93	-120.88	1278.88
314.37	-5471.97	-87.21	5497.27
314.53	-5471.81	-87.24	5497.43
316.70	-4717.85	-106.65	2973.56
317.74	-4619.82	-133.81	2761.20
321.14	-6477.20	-123.42	1145.90
324.09	-5173.40	-129.72	2753.83
325.39	-4633.41	-134.74	2769.17
333.78	-5121.19	-134.82	2780.23
346.66	-6277.57	-134.67	2164.51
350.37	-5098.74	-142.50	4374.38
355.00	-5205.02	-140.91	2302.95
363.61	-7495.00	-130.22	4830.95
370.92	-6485.63	-148.81	4161.44
388.80	-5141.16	-139.13	2897.77
909.08	-6473.70	-519.32	2199.52

Source: Authors.

Comparing Tables B9 and B10, it is evident that these distributions are also among those with the most *negative* means. In fact, they are in the lowest half of 1% of distributions ranked by mean. Positive *medians* imply these distributions have more positive than negative μ_{it}, but the negative means imply that the negative values tend to be larger in magnitude. This is attested to by the fact that the absolute value of the minimum is, on average, almost 2,300 above the maximum. For the three cases where the maximum is larger, it is barely so.

The tendency for negative μ_{it} to be larger in magnitude than positive μ_{jt} implies that the shift factors on the buses *i* where $\mu_{it} < 0$ are larger than the shift factors on the buses *j* where $\mu_{jt} > 0$. In turn, that implies that buses *i* flowing power into the network on constrained links tend to be less well-connected to the rest of the network than buses *j* on the other ends of those constrained links. The implication is that the most congested lines tend to connect isolated, remote generators to buses that are much better connected to the rest of the network.Various statistics in Table B9 suggest that transmission constraints tend to have larger negative than positive impacts on LMPs. The fact that the average mean is negative while the median of the medians is zero suggests that negative μ_{it} tends to be larger in magnitude than positive μ_{it}.

Table B11 shows how the cross-sectional distributions μ_{it} vary across subsets of *load*.[33] The variable N equals the number of periods making up each average. The moments in columns 2–4 of Table 11 are similar to each other, but differ more noticeably from those in the first column and substantially from those in column 6. The moments in column 5 appear to be a mixture of those in columns 2–4 and those in column 6. Using the approach usually taken in the economic analysis of electricity supply, we could associate loads up to around 40 GW as base loads, the 40-70 GW range as intermediate loads, and above 70 GW as peak loads.

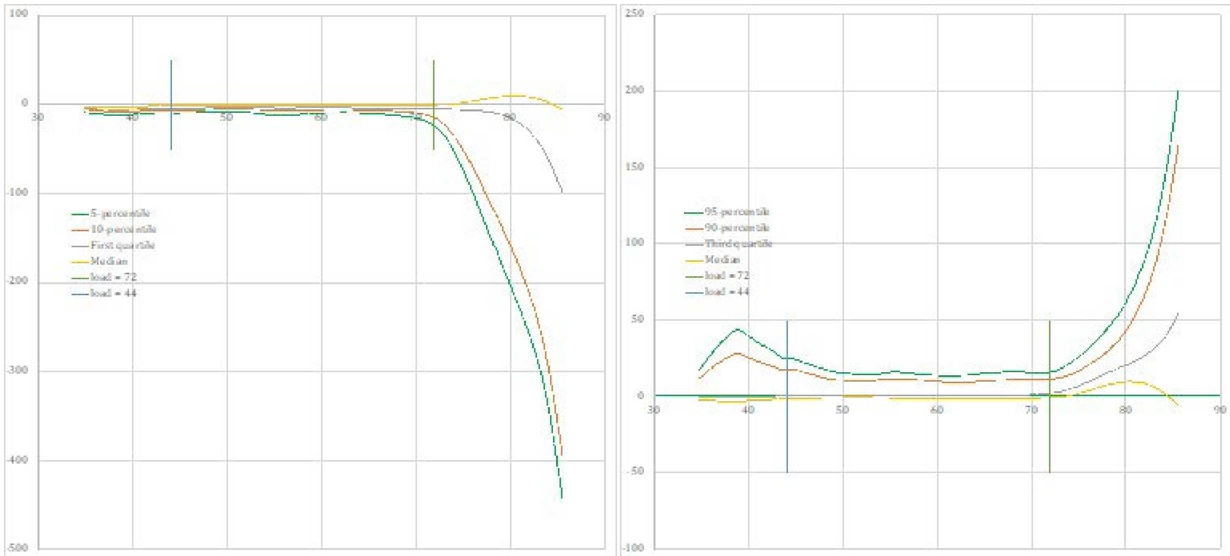
Table B11 — Means of Moments of Cross-Sectional Distributions of Ξ_{it} by Subsets of Load

	< 40 GW	40-50 GW	50-60 GW	60-70 GW	70-80 GW	>80 GW
Statistic	N = 2619	N = 11094	N = 7752	N = 5096	N = 5547	N = 2690
Minimum	-243.62	-150.75	-176.64	-155.23	-346.95	-678.14
1%	-30.47	-23.25	-41.63	-25.54	-118.71	-315.68
5%	-12.06	-9.41	-10.84	-10.28	-81.29	-267.17
10%	-8.39	-6.78	-6.53	-6.55	-57.31	-223.07
25%	-6.19	-4.56	-3.63	-4.13	-5.97	-33.12
50%	-3.28	-1.60	-1.09	-1.56	2.88	7.85
75%	-0.59	0.35	0.83	0.62	8.03	28.24
90%	26.12	16.50	10.74	10.11	18.02	71.52
95%	40.55	24.26	14.78	14.83	27.22	95.56
99%	57.43	36.49	32.04	38.40	82.60	185.44
Maximum	726.24	404.76	324.45	379.01	804.73	1926.96
Mean	1.330	0.415	-0.333	-0.084	-3.404	-14.976
Std Dev	31.125	18.532	17.520	18.666	55.520	155.963
Skewness	4.245	-1.646	1.566	2.107	2.189	9.102

Source: Authors.

A graphical version of the results in Table 11 can be obtained by estimating Lowess smoothing regressions of the 34,798 values of selected quantiles of the cross-sectional distributions of the μ_{it} on the load. The procedure derives smoothed values as the predicted values of the dependent variable from a set of weighted regression equations using moving windows of the data.[34] Figure B2 graphs these for the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 95^{th} percentiles. The vertical lines at 44 GW and 72 GW are candidate load thresholds for separating the intermediate from the base and peak loads. Clusters of points at loads below about 44 GW and above about 72 GW are evident in several of the plots in Figure B1.

Figure B2 — Lowess Smoothed Relationship between Quantiles of μ_{it} Distributions and Load



Source: Authors.

To further investigate determinants of high and low μ_{it} that result from transmission constraints, we estimated regression equations for the moments, presented in Table B9, of the cross-sectional distributions. Splitting the sample into periods where *load* < 44, 44 ⩽ *load* < 72, and *load* ⩾ 72, we found strong evidence that the relationship of the μ_{it} distributions to the exogenous variables apart from *load* also differed across these different regions. For example, simple OLS regressions estimated with robust (Huber-White sandwich) standard errors gave statistically significantly different coefficients on the exogenous variables across the three regions. It makes sense that the effect of a reduction in *wind* output, for example, is likely to depend on the set of generators already supplying the system and the marginal costs of generators that could be ramped up to compensate. However, it is more consistent with the underlying economic determinants of the outcomes to allow the relationships to change continuously as *load* (and the other explanatory variables) changes rather than separating the sample into three discrete categories based on *load*. To allow for flexible functional forms, we again estimated nonparametric kernel regressions with bootstrap standard errors using 500 replications. Bandwidth and kernel choices and the potential explanatory variables remained the same as for the models for λ.

Table B12 presents estimates of how the cross-sectional mean value of μ_{it} in each period t relates to *load*, *thermal*, *wind*, *solar*, and physical response capability *RC*. Table B13 gives corresponding estimates for the medians. The results in Tables B12 and B13 reinforce the conclusions from the analysis of moments as a function of *load* only. The conditional mean (across periods) of the cross-sectional mean values of μ in each period is more negative than the conditional mean of the medians. In fact, the latter is not statistically significantly different from zero. This again implies that binding transmission constraints have a greater impact on buses near supply source connections than on buses downstream of constraints.

Table B12 — Nonparametric Kernel Regression for the Cross-Sectional Mean of μ

Variable	Estimated Coefficient	Bootstrap Std Err	*p*-value H_0: coef = 0
Mean			
$E(\mu)$	-1.7749	0.0834	0.000
Effects (Averages of Derivatives)			
load	-1.7843	0.1293	0.000
thermal	1.7038	0.1302	0.000
wind	1.4210	0.1252	0.000
solar	1.5244	0.1322	0.000
RC	1.4550	0.0806	0.000
$R^2 = 0.5320$			

Source: Authors.

Table B13 — Nonparametric Kernel Regression for the Cross-Sectional Median of μ

Variable	Estimated Coefficient	Bootstrap Std Err	*p*-value H_0: coef = 0
Mean			
$\{\mu \mid \Pr(\mu_i \leq \mu) = 0.5\}$	-0.0596	0.1263	0.637
Effects (Averages of Derivatives)			
load	1.9831	0.2432	0.000
thermal	-1.8001	0.2459	0.000
wind	-1.7445	0.2348	0.000
solar	-2.0803	0.2508	0.000
RC	-1.3189	0.1957	0.000
$R^2 = 0.4858$			

Source: Authors.

An increase in *load* matched by an increase in *thermal* or *wind* generation (while leaving χ = 1) tends to reduce the mean and increase the median of μ. The magnitudes of both effects are greater for *wind*. Thus, even though the *magnitudes* of the negative μ values tend to increase, the *number* of negative versus positive μ values tends to fall. Where μ_i becomes more negative, some binding transmission constraints are becoming more severe. But an increase in the median suggests a greater number of more moderately constrained lines.

The results are different for *solar*. The estimated coefficient on *solar* in Table B12 implies its effect on the mean of μ lies between the effects of *wind* and thermal generation. On the other hand, the coefficient on *solar* in Table B13 is more negative than the amount by which the coefficient on *load* is positive. Hence, an increase in *load* that is matched by an increase in *solar* tends to reduce *both* the mean and the median values of μ. But since the mean of μ falls by more in the case of *wind*, then more *wind* generation is likely to produce large negative μ values, while more *solar* is more likely to increase the number of small negative μ values. These results suggest that when constraints are binding, the source end of the link tends to be connected to *solar* or *wind* generation. The differences in the behaviors of the means versus medians suggest that *wind* tends to be more seriously bottlenecked than *solar*.

The coefficients on *RC* in Tables B12 and B13 reveal that an increase in physical responsive capability holding *load* and generation fixed tends to increase the cross-sectional mean and reduce the median of μ. This is consistent with higher *RC* allowing for fewer binding transmission constraints. However, the estimate in Table 8 implies that an increase in *RC* holding *load* and generation fixed, while maintaining χ = 1, also tends to substantially reduce λ. This would, by itself, tend to reduce the costs associated with the transmission constraints.

Table B14 shows the results of a kernel regression with the cross-sectional standard deviations as the dependent variable. An increase in *load* that is matched by an increase in thermal generation tends to raise the standard deviation of μ. The effect is again greater if the increased *load* is instead matched by increased *wind* output. Table B14 shows, however, that an increase in *load* that is matched by an increase in *solar* generation tends to increase $\sigma(\mu)$ by about the same amount as when thermal generation supplies the increase in *load*. An increase in *RC* holding *load* and generation fixed strongly reduces $\sigma(\mu)$.

Table B14 — Nonparametric Kernel Regression for Cross-Sectional Standard Deviation of μ

Variable	Estimated Coefficient	Bootstrap Std Err	*p*-value H_0: coef = 0
Mean			
$\sigma(\mu)$	37.2857	0.5633	0.000
Effects (Averages of Derivatives)			
load	18.4331	1.0116	0.000
thermal	-17.4601	1.0197	0.000
wind	-16.8259	0.9733	0.000
solar	-17.8710	1.0359	0.000
RC	-16.2710	0.6840	0.000
R^2 = 0.6073			

Source: Authors.

Finally, Tables B15 and B16 present the kernel regression results for the lower and upper tails of the cross-sectional distributions of μ. From Table B9, and as illustrated in Figure B2, the former tends to be negative and the latter positive throughout the range of loads.

Table B15 — Nonparametric Kernel Regression for the Lower 5% Tail of μ

Variable	Estimated Coefficient	Bootstrap Std Err	*p*-value H_0: coef = 0
Mean			
$\{\mu \vert \Pr(\mu_i \leq \mu)=0.05\}$	-45.1171	1.5037	0.000
Effects (Averages of Derivatives)			
load	-40.6847	4.8519	0.000
thermal	35.5341	6.8012	0.000
wind	30.4997	8.2293	0.000
solar	32.4812	7.8371	0.000
RC	36.7822	3.1556	0.000
R^2 = 0.6151			

Source: Authors.

Table B16 — Nonparametric Kernel Regression for the Upper 5% Tail of μ

Variable	Estimated Coefficient	Bootstrap Std Err	*p*-value H_0: coef = 0
Mean			
$\{\mu \vert \Pr(\mu_i \leq \mu)=0.95\}$	29.6193	0.4889	0.000
Effects (Averages of Derivatives)			
load	8.1059	0.8499	0.000
thermal	-7.8840	0.8614	0.000
wind	-8.0265	0.8261	0.000
solar	-9.5005	0.8632	0.000
RC	-13.4043	0.6411	0.000
R^2 = 0.4073			

Source: Authors.

An increase in load that is matched by an increase in *thermal* generation exacerbates existing binding transmission constraints. The magnitudes of both extremely negative and extremely positive values of μ tend to increase.

An increase in *load* that is instead matched by an increase in *wind* has a substantially larger impact on the negative tails, but a smaller impact than thermal generation on the positive extreme values of μ. This result again emphasizes that increased *wind* strains the capacity of lines leading from the wind generation sites to better connected regions that are not as strongly impacted by the line constraint.

Increased *solar* has an impact on negative extreme values of μ that is intermediate between *wind* and thermal generation. However, an increase in *load* that is matched by an increase in *solar* tends to *reduce* large positive extreme values of μ.

An increase in physical responsive capability while holding *load* and generation fixed strongly reduces the magnitudes of extremely negative and extremely positive values of μ.

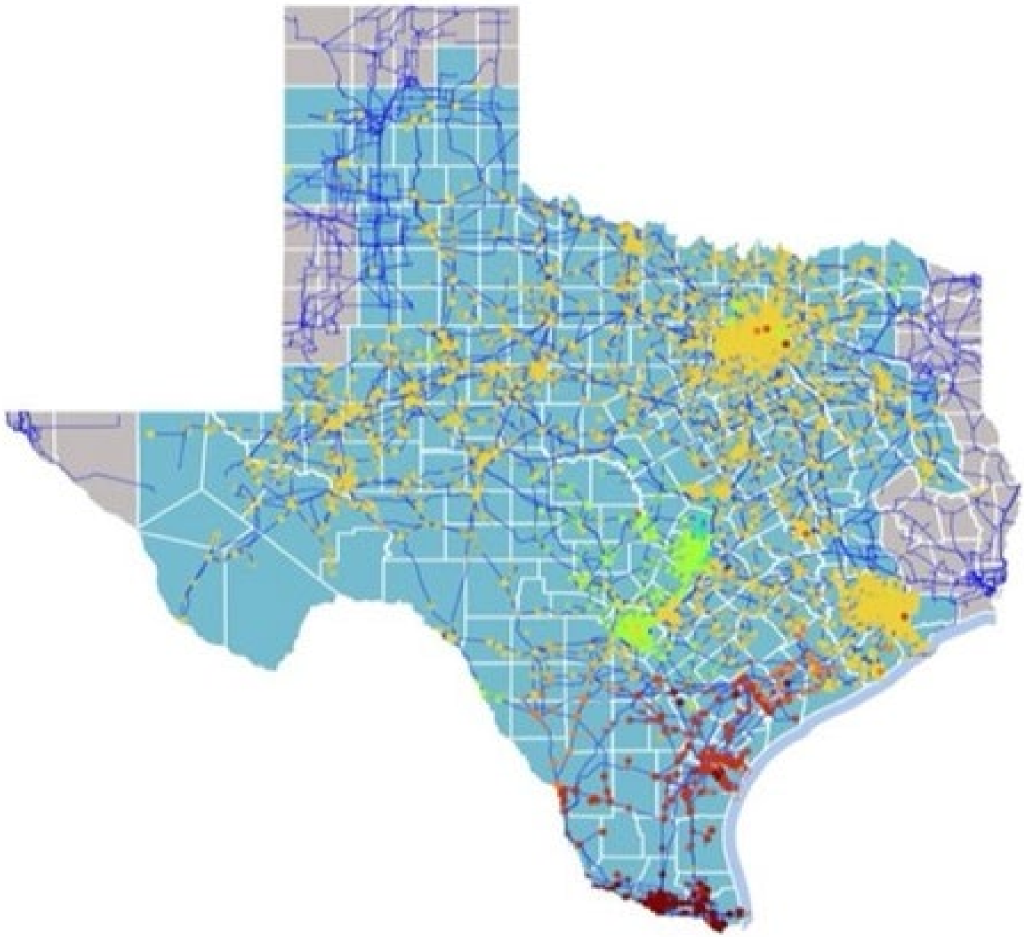
Locations of Binding Transmission Constraints

Thus far, we have examined the cross-sectional distribution of LMPs for each time period. Another view can be obtained by averaging the LMPs across time periods for each electrical bus on the network. As noted earlier, the number of buses in the complete data set changes over time as ERCOT adds new generation and demand centers or new transmission lines or as old facilities are retired. For this analysis, we focused on a subset of 1,440 periods (10 days) where peak loads on most days were near the highest within the full sample period, and the number of buses (16,677 in total) did not change.

The mean of all 1,440×16,677 LMPs in the data set is $320.70. A bus with an average LMP across the 1,440 time periods that is below $320.70 will be more likely than average to be a source on one or more constrained transmission links. We call this a “bottlenecked” bus. One with an average above $320.70 will be more likely than average to be a sink on one or more constrained transmission links. Even if a bus is a sink on one constrained link, however, it may not experience above-average LMPs if it is well connected to the network via other unconstrained links. In that case, its shift factor will tend to be low. We thus refer to buses with higher-than-average LMPs as “weakly connected sinks.”

Unfortunately, ERCOT does not disclose the location of individual buses on the network. However, Velocity Suite has a data set of location coordinates for a subset of the buses. Of the 16,677 buses in our restricted data set, around 11,000 could be matched to entries in the Velocity Suite data set. Figure B3 maps these matched buses by their average LMP. In this figure, buses with an average LMP close to $320.70 are colored yellow. Buses are more bottlenecked as yellow shades to orange and then dark red. Buses that are instead more weakly connected sinks are shaded from yellow to green and then to blue and gray. The severely bottlenecked buses tend to be concentrated in South Texas and especially close to the Mexican border. More moderately constrained buses (orange) are in the Houston and Dallas metro areas and in West Texas. The somewhat weakly connected sinks (greens) tend to be concentrated in the San Antonio and Austin metro areas and surrounding regions.

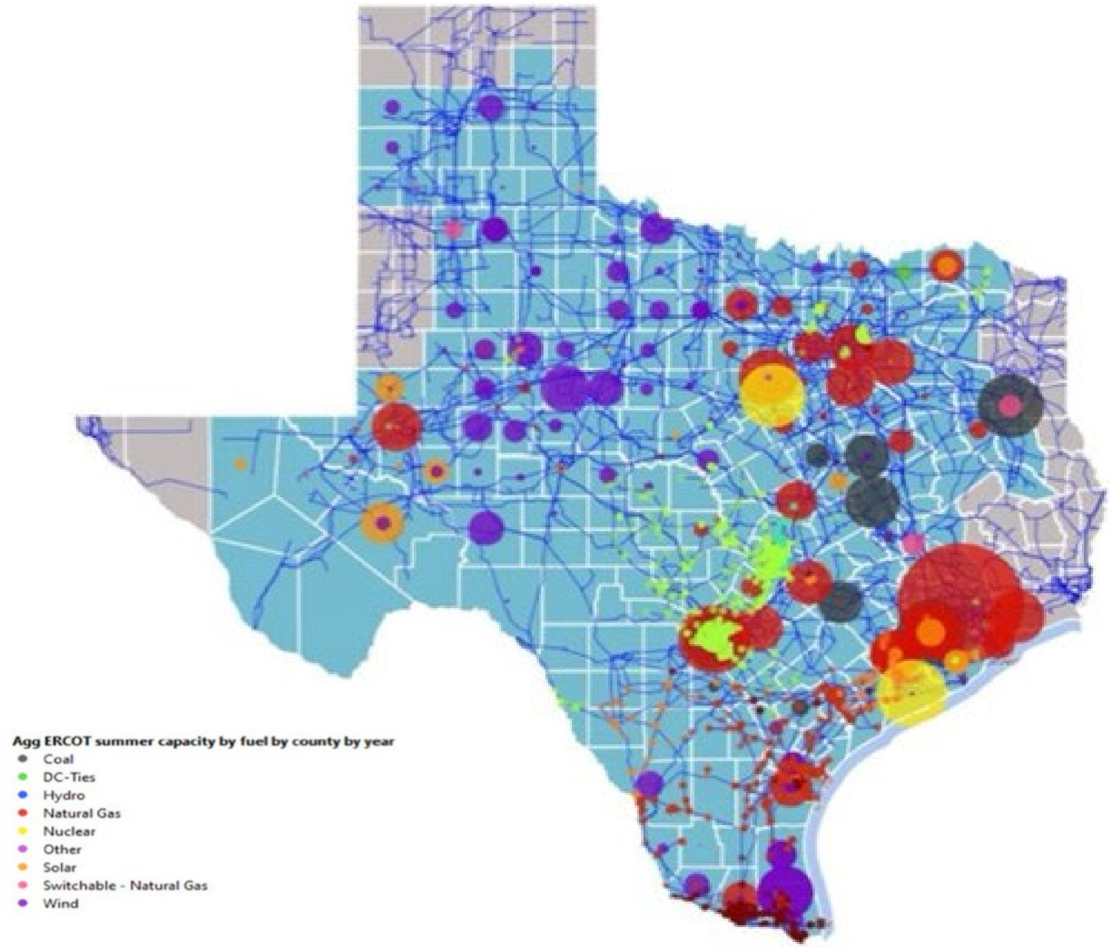
Figure B3 — Bus Locations Colored by Average Locational Marginal Price



Source: Authors.

Figure B4 shows the locations of different types of generation units in summer 2023. This suggests that the bottlenecked buses in South Texas may be due in part to the rapid expansion of wind generation in that coastal area without a corresponding build-out in new transmission capacity. The opposite problem in the San Antonio and Austin metro areas may be the result of rapid growth in demand in that part of the state that has not been matched by sufficient increased generating or transmission capacity.

Figure B4 — ERCOT Capacity by Fuel by County, Summer 2023



Source: Authors.

Concluding Comments

The component of ERCOT LMPs that is the result of binding transmission constraints can add substantially to the average level of wholesale prices. Nevertheless, these components subtract from some prices while adding to others. In most periods, the overall impact across all buses on the network is close to zero.

Buses that are frequently on the supply end of constrained links will tend to have very low or even negative prices most of the time. Buses on the demand end of the same links will tend to have prices exceeding the average price across all buses most of the time. In principle, these price differentials should signal where it is desirable to expand network capacity. Simultaneously, they should signal to producers where it is less desirable to expand generation output or increase electricity demand.

Our analysis of ERCOT data covering around 152 days in the summer and fall of 2023 found evidence that wind generators tend to be the most severely affected by inadequate transmission capacity, while utility-scale solar plants are also affected, but to a lesser extent. On the demand side, there was also evidence that the West zone is more affected by transmission constraints.

LMPs currently calculated in ERCOT reflect only transmission constraints. They do not reflect marginal transmission losses. To send efficient signals about the locations of new generating capacity, the location of large new industrial consumers of electricity, or the best places to upgrade transmission lines, the prices ought to be changed to reflect marginal transmission losses in addition to capacity constraints. At the request of the PUCT, ERCOT undertook a study to ascertain the impact of including marginal transmission losses in LMPs. They found evidence that this would incentivize a more efficient use of generation resources and improve the operation of the system in the short run. ERCOT did not assess, however, how such a change could affect long-run investment incentives. Taking those effects into account would substantially increase the benefits of moving to truly cost-reflective LMPs.

Notes

[1] Peter Hartley, Kenneth B. Medlock III, and Shih Yu (Elsie) Hung, "The Texas Deep Freeze of February 2021: What Happened and Lessons Learned?," *Economics of Energy and Environmental Policy* 12, no. 2 (2023): 5–29. An earlier, longer analysis is available here: Peter Hartley, Kenneth B. Medlock III, and Shih Yu (Elsie) Hung, "ERCOT Froze in February 2021. What Happened? Why Did It Happen? Can It Happen Again?" (Houston: Rice University's Baker Institute for Public Policy, February 2, 2022), https://www.bakerinstitute.org/research/ercot-froze-february-2021-what-happened-why-did-it-happen-can-it-happen-again.

[2] While not addressed in detail in this study, these structural impacts are compounded when there is an inability to track, for example, new sources of load, such as electric vehicles (EVs). Currently, there is no mechanism to require tracking, reporting, or disclosure of electricity use by EVs, which leaves a critical gap in potential demand response capacity. This issue is relevant to state and federal energy market designs. Any lack of data transparency must be addressed to assess future load realistically.

[3] Public Utility Regulatory Act of 1975, Title II, Texas Utilities Code (as amended September 1, 2019), https://www.puc.texas.gov/agency/rulesnlaws/statutes/pura19.pdf; Julie A. Cohn, "Connecting Past and Future: A History of Texas' Isolated Power Grid" (Houston: Rice University's Baker Institute for Public Policy, December 1, 2022), https://doi.org/10.25613/dpmy-r389.

[4] "ERCOT Organization Backgrounder," Electric Reliability Council of Texas (ERCOT), last modified January 24, 2024, https://www.ercot.com/news/mediakit/backgrounder.

[5] Bilateral contracts fall outside of ERCOT-administered markets.

[6] Much of this section is based on other Baker Institute analysis, specifically Cohn, "Connecting Past and Future."

[7] "The Great Northeast Blackout," HISTORY, A&E Television Networks, last updated November 6, 2020, https://www.history.com/this-day-in-history/the-great-northeast-blackout.

[8] Olivera Jankovska and Julie A. Cohn, "Texas CREZ Lines: How Stakeholders Shape Major Energy Infrastructure Projects" (Houston: Rice University's Baker Institute for Public Policy, November 17, 2020), https://doi.org/10.25613/261m-4215.

[9] Warren Lasher, "The Competitive Renewable Energy Zones Process" (Austin: ERCOT, August 11, 2014), https://www.energy.gov/sites/prod/files/2014/08/f18/c_lasher_qer_santafe_presentation.pdf.
[10] Peter R. Hartley, Kenneth B. Medlock III, and Olivera Jankovska, "Electricity Reform and Retail Pricing in Texas," *Energy Economics* 80 (2019): 1–11.
[11] Hartley, Medlock, and Hung, "ERCOT Froze in February 2021."
[12] For a description of Phase I of the reforms, see Mark Watson, "Texas Regulators Order Power Market Reform Phase I, Analysis of Bigger Changes," *S&P Global Commodity Insights*, December 16, 2021, https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/natural-gas/121621-texas-regulators-order-power-market-reform-phase-i-analysis-of-bigger-changes.
[13] Kenneth B. Medlock III and Shih Yu (Elsie) Hung, "Resource Adequacy in ERCOT: How Long-Term Market Design Reforms Could Enhance Reliability" (Houston: Rice University's Baker Institute for Public Policy, December 12, 2022), https://www.bakerinstitute.org/research/resource-adequacy-ercot-how-long-term-market-design-reforms-could-enhance-reliability.
[14] Public Utility Commission of Texas (PUCT), "Public Utility Commission of Texas Unanimously Adopts Performance Credit Mechanism Reliability Service," news release, January 19, 2023, https://www.puc.texas.gov/agency/resources/pubs/news/2023/puct_adopts_performance_credit_mechanism_reliability_service.pdf.
[15] "Ancillary Services," ERCOT, last updated January 25, 2024, https://www.ercot.com/gridmktinfo/dashboards/ancillaryservices.
[16] Wholesale electricity prices respond to load in the short run to the point that they are often assumed endogenous and load exogenous. Since retail electricity prices generally take time to reflect wholesale prices, they are not very cost-reflective in the short run but will respond to load and costs in the longer run. Hartley, Medlock, and Jankovska, "Electricity Reform and Retail Pricing in Texas" found that monthly retail electricity prices in Texas closely track monthly wholesale prices, especially in markets that are more competitive.
[17] The mean normalized data series are plotted in the Appendix, in Figures A1–A3.
[18] Medlock and Hung, "Resource Adequacy in ERCOT."
[19] This excludes El Paso County, which ranks 10th in population, as it is outside of the ERCOT service area.
[20] "Report on Existing and Potential Electric System Constraints and Needs" (Austin: ERCOT, December 2022), https://www.ercot.com/files/docs/2022/12/22/2022_Report_on_Existing_and_Potential_Electric_System_Constraints_and_Needs.pdf.
[21] "Long-Term West Texas Export Study" (Austin: ERCOT, January 2022), https://www.ercot.com/files/docs/2022/01/14/Long-Term-West-Texas-Export-Study-Report.pdf.
[22] "2024 LTSA Update" (Austin: ERCOT, May 16, 2024), 15, https://www.ercot.com/files/docs/2023/05/12/2024_LTSA_update_20230516_v1.0.pdf.
[23] "Study of the System Benefits of Including Marginal Losses in Security-Constrained Economic Dispatch" (Austin: ERCOT, June 29, 2018), https://www.ercot.com/files/docs/2018/06/29/Study_of_the_Benefits_of_Marginal_Losses_FINAL.pdf.
[24] The number of buses changes as new supplies and demands are connected to the network, new transmission lines are built, and some old transmission lines or supply or demand points are eliminated. Infrequent calculations made when the load is close to capacity add some observations.
[25] In practice, changing the capacity of just one link in the network will, in general, affect flows on many other links in the network. This is known as the problem of "loop flows." This is why no existing market-based electricity supply system allows anyone to build or upgrade a transmission link without some sort of public process to assess the value of such a link relative to its costs, all impacts considered.
[26] A nonlinear model of the network is used to calculate these offline, separately from the linear model used to calculate locational marginal prices (LMPs). The shift factors used in ERCOT are available only to market participants.
[27] Since power flows must respect Kirchhoff's circuit laws, alterations to demand or supply at either end of a line will change flows not only on that line, but also on many lines in the network. In a nonlinear manner, it

will also affect other factors, such as voltage, reactive power, and frequencies. As a result, a model of the network is needed to calculate the marginal impact on flows from a small change in net demands at given locations. The shift factors used in ERCOT are available only to market participants.

[28] Infrequent calculations made when the load is close to capacity add some observations. Although some successive observations are thus made at much less than five minutes apart, for convenience we always refer to the time dimension of the observations as "five-minute intervals." We started with 44,863 intervals but dropped 182 because of missing physical response capability values; the number of buses changes as new supplies and demands are connected to the network, new transmission lines are built, and some old transmission lines or supply or demand points are eliminated.

[29] In the period our data covers, there were 859 settlement points compared to 16,677 electrical buses. As noted, wholesale market payments are based on prices that are also aggregated temporally into 15-minute intervals.

[30] Thermal plus wind plus solar generation does not equal total load, but the difference is usually less than 5% of ERCOT load. Since load is actual energy delivered in megawatt-hours (MWh), total supply will generally exceed load because of transmission losses. In addition, sources of supply apart from thermal, wind, and solar are hydro, supply from storage (consumption by these devices is included in load), "other" (mainly biomass or capacity from settlement-only distributed generators), and imports of electricity from neighboring systems.

[31] Perhaps this indicates that some nonlinearity in reserve capacity could improve the fit.

[32] Increasing the number of replications from 400 to 500 changed the standard error estimates by about the same amount as changing the seed value in the pseudo-random number generator used to choose the bootstrap samples.

[33] Recall that the load (and values of other explanatory variables in Table 2) applying to period t is measured over the five-minute interval closest to but preceding t.

[34] Denote the load in period t by x_t and the corresponding dependent variable by y_t, for $t = 1, 2, \ldots, 34{,}798$. The data is first ordered based on increasing values of x_t and then divided into a set of overlapping subsets, called "moving windows." In our application, the moving window for period t includes the closest 3,480 loads above and 3,480 loads below x_t and the corresponding values of y_t. These values are then used to estimate a (local) weighted linear regression. The weight applied at a given load in the moving window declines cubically with the distance between that load and x_t. The value plotted at x_t is the predicted value from this local regression.