Correlated generator failures and power system reliability

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Engineering and Public Policy

Sinnott J. Murphy

B.S., Biochemistry and Molecular Biology, University of California—Davis
M.S., Agricultural and Resource Economics, University of California—Davis
M.S., Transportation Technology and Policy, University of California—Davis

Carnegie Mellon University
Pittsburgh, PA

May, 2019

Acknowledgments

My doctoral studies represent an immense period of growth, and my success would not have been possible without the help and support of numerous individuals.

I thank my advisor and committee chair, Jay Apt, for his attentive guidance, responsiveness, and reminders of the finite number of heartbeats one can expect to enjoy during their lifetime. Through him I received the opportunity to contribute to the advancement of power system reliability and I am proud of what we accomplished together.

I am deeply indebted to Fallaw Sowell for teaching me how to conduct empirical statistical research. It's one thing to take a statistics class; it's quite another to use the tools of statistics to answer complex, unstructured research questions. Developing that skill set was the primary reason I came back to graduate school and it would not have been possible without him.

I am humbled to have Granger Morgan on my committee. Getting to speak with him about the program over a beer back in February 2014 put EPP firmly at the top of my list, and I have never looked back. It's been both an honor and a pleasure getting to work with him.

If Jay, Fallaw, and Granger make up the lion's share of what I am today, Zico Kolter is the model of what I want to become. His expertise in statistics and machine learning led to significant improvements in the work. I'm looking forward to finally having time to watch recordings of his computational methods for the smart grid course this summer.

I have been fortunate to take courses from many outstanding faculty while at CMU. In addition to all four of my committee members falling into this category, I would like to acknowledge Tom Mitchell, Larry Wasserman, and especially Cosma Shalizi.

I have been blessed with wonderful colleagues in EPP who have motivated me to achieve more than I thought possible, raised my spirits when I hit stumbling blocks, and made this very long marathon bearable (even enjoyable at times). Thanks in particular to Brian Sergi for being my next-desk neighbor five years running; Jeremy Keen for teaching me Python on the fly in a course he didn't even take for credit; Michael Craig for letting me beat him in Mario Kart; and most of all

to Chukky Udeani, my brother-in-arms and the handsomer half of the undisputed EPP dream team.

Special thanks go to Luke Lavin. His efforts to enhance the capabilities of the RECAP model greatly increased the value of the final chapter of this thesis.

Thanks to Doug Sicker, Jerry Cohon, Scott Matthews, Vicki Finney, Adam Loucks, and Debbie Kuntz for their unwavering support and belief in me. It has meant a lot over the years.

From academic life before CMU, thanks to Bryan Jenkins, Chris Knittel, and Dave Rapson for being wonderful mentors and letter-writers.

Thanks to my parents, Jim and Marianne, for their boundless love and encouragement; this accomplishment is as much theirs as it is mine. Thanks to my brother Steve, my go-to statistics resource, unofficial fifth committee member, and the reason I thought I might be able to hack a Ph.D. in the first place.

Speaking of family, thanks to Yen, Kevin and Jifen, my proxy Pittsburgh family. I knew I'd get on just fine here once I discovered Fuku Tea.

Thanks to the Carnegie Mellon Electricity Industry Center and the North American Electric Reliability Corporation for funding my doctoral studies; John Moura and Tom Falin for project sponsorship and data access; Patricio Rocha-Garrido, Francis J. Bell III, Matthew Varghese and many others at PJM and NERC for answering numerous questions; and Ron Fluegge for teaching me everything there is to know about GADS.

Abstract

This thesis contributes new knowledge toward understanding the relationship between capacity procurement and power system reliability through rigorous analysis of generator-level availability data.

In Chapter 2 I analyze four years of data (2012-2015) from the Generating Availability Data System (GADS) database maintained by the North American Electric Reliability Corporation (NERC) to evaluate key assumptions made by power system planners when determining capacity requirements. Using block subsampling and binomial modeling, I demonstrate that large unavailable capacity events have occurred with much greater frequency than should be expected if current-practice assumptions hold.

In Chapter 3 I propose a nonhomogeneous Markov model to explain the observed correlated failures. I use logistic regression to fit a simple model specification that allows generator transition probabilities to depend on ambient temperatures and system load. I fit the model using 23 years of GADS data for the PJM Interconnection (PJM), the largest system operator by generation capacity in North America. Temperature and load are each statistically significant for two-thirds of generators. Temperature dependencies are observed in all generator types, but are most pronounced for diesel and natural gas generators at low temperatures and nuclear generators at high temperatures. The nonhomogeneous Markov model predicts system-level unavailable capacity substantially better than the homogeneous Markov model used currently by industry.

In Chapter 4, joint work with Luke Lavin, I quantify the reliability risks implied by temperature dependence in PJM's generator fleet. We modify an open-source resource adequacy modeling tool to allow generator availability to depend on temperature. We then parameterize the tool for PJM's system using temperature-dependent forced outage rates developed in Chapter 3. We find that temperature dependence substantially increases capacity requirements to achieve the target level of reliability, though PJM procures still more than our model finds is required. Given the seasonality in temperatures and loads, we also demonstrate that average annual capacity requirements could be

significantly reduced were PJM to set separate monthly targets, rather than a single annual target. Finally, we explore the resource adequacy implications of various future generator resource and climate change scenarios for PJM.

Contents

List of Tables

List of Figures

Chapter 1 Introduction and motivation

Having sufficient generation capacity is crucial for maintaining adequate reliability in electric power systems. Not only must electric supply and demand be balanced at all times, but generation reserves are required to guard against unanticipated generator outages and deviations from forecasted loads. All else equal, greater procurement of generation capacity is desirable because it reduces the probability of loss-of-load events. However, procuring capacity is costly and thus electric power system planners strive to strike a reasonable balance between the costs and benefits of reserves.

The key question is how a given level of capacity procurement translates to the probability of not being able to serve load—particularly the probability of not being able to serve the forecast peak demand for the system. With this knowledge, a system planner could immediately determine the level of capacity that satisfies the agreed-upon risk of generation shortage. System planners have long assumed that generators fail independently of one another and with constant probability throughout the year. This has worked well historically. However, recent events, such as the Polar Vortex of January 2014, have threatened grid reliability in the eastern United States despite capacity procurement levels far above those deemed sufficient under these assumptions.

In this thesis, I contribute new knowledge toward understanding the relationship between capacity procurement and power system reliability. In Chapter 2 I conduct what is to my knowledge the first systematic evaluation of the two key assumptions just mentioned—that generators fail independently of one another and that generators are equally likely to fail at all times of the year—for several large power systems in North America. To do this, I analyze four years of data from the Generating Availability Data System (GADS), a proprietary database of generator-level availability data for approximately 8,000 generators maintained by the North American Electric Reliability Corporation (NERC). I devise a novel method of translating the event records reported in GADS into hourly time series of unavailable capacity for each reporting generator. I then use multiple statistical methods to demonstrate that historical generator failure patterns are inconsistent with these assumptions. The key takeaway from this analysis is that large generator outages, where many generators are

simultaneously unavailable, occur much more frequently than system planners expect them to.

Given these results, in Chapter 3 I then propose a model to explain the observed correlated failures. Instead of assuming that generator failure probabilities are constant, and thus that generators fail unconditionally independently of one another, I relax this assumption by allowing environmental variables to moderate generator failure probabilities, making generator failures now conditionally independent of one another. I use logistic regression to fit a simple model specification that allows failure probabilities to depend on ambient temperatures and on system load (electricity demand), but many additional or alternative covariates could be included. I fit the model on the 1,845 generators that report to the GADS database maintained by the PJM Interconnection (PJM), the largest system operator by generation capacity in North America. My results demonstrate that PJM's generator fleet is substantially less available at both very cold and very hot temperatures. Given that both temperature extremes correspond to high demand for electricity in PJM, this result suggests that part of the challenge of the Polar Vortex of 2014 was due to the limitations of current reliability planning models.

In Chapter 4, joint work with Luke Lavin, I quantify the reliability risks implied by the temperature dependence of PJM's fleet. We modify an open-source resource adequacy modeling tool to allow generator availability to depend on temperature. We then parameterize the tool for PJM's system. We find that temperature dependence substantially increases capacity requirements to achieve the target level of reliability, though PJM procures still more than our model finds is required. This suggests substantial benefit could be achieved by helping system planners to quantify capacity requirements necessary to achieve a desired reliability target. Given the seasonality in temperatures and loads, we also demonstrate that average annual capacity requirements could be significantly reduced were PJM to set separate monthly targets, rather than a single annual target.

Chapter 2 An empirical evaluation of resource adequacy modeling assumptions*

Abstract

To keep the electric power system reliable, grid operators procure reserve generation capacity to protect against generator failures and significant deviation from the load forecast. Current methods for determining reserve requirements use historical generator availability data (recorded as failure events) to compute the fraction of the time each unit in the power system was unavailable unexpectedly. These values are then combined using analytical or simulation methods to yield a distribution of available capacity. From this distribution, the reserve capacity needed to maintain a particular reliability target may be determined. Such an approach implicitly assumes that generator failures occur independently of one another and that generator availability is not seasonal. To test these assumptions, we process the more than two million event records reported to the Generating Availability Data System (GADS) database maintained by the North American Electric Reliability Corporation (NERC) between January 1, 2012 and December 31, 2015. This allows us to construct complete availability histories for each of the approximately 8,000 generating units reporting to GADS during this period. Using these time series, we find strong evidence of correlated failures in most regions, even when removing Hurricane Sandy and the exceptionally cold month of January 2014 from the data. We find that correlated failures occur in all seasons. We do not find evidence of seasonality but note that seasonal structure may emerge with more data. In addition we determine the distribution of unscheduled unavailable capacity, unscheduled derating magnitudes, event durations, event arrival probabilities, and mean time between failure (MTBF) and mean time to recovery (MTTR) values. In each case, we report fit parameters to facilitate use by practitioners. We find statistically significant differences in mean time between failure for small and large units for three unit types when aggregating over regions. Finally we present time series of unavailable capacity from unscheduled, maintenance, and scheduled events. These may be used in conjunction with load data to directly study resource adequacy risks without assuming independent failures or constant availability. Our findings suggest that power system resource planners should consider correlated outages as they identify reliability and reserve capacity requirements.

* Published as: Murphy, S, Apt, J., Moura, J., & Sowell, F. (2018). Resource adequacy risks to the bulk power system in North America. Applied Energy, 212, 1360-1376. https://doi.org/10.1016/j.apenergy.2017.12.097

2.1 Introduction

Extended low temperatures in much of the United States (U.S.) and Canada in January of 2014 resulted in significant losses of electricity generation capacity. In the control area of PJM Interconnection LLC (PJM), a large regional transmission organization (RTO) in the eastern U.S., more than 20% of total capacity was unavailable during the peak of the polar vortex event [1].[1] To avoid blackouts, PJM had to enact emergency measures, including making public appeals for conservation, calling on demand-response resources, reducing system voltage, and scheduling shared reserves with neighboring systems.

Resource adequacy modeling (RAM) is the process of determining how much capacity is needed to achieve a given reliability standard.[2] Probabilistic methods have been used to determine required reserve generation from power plant outage data for more than 80 years [6, 7]. Significant advances were made in the immediate postwar period and led to the creation of a joint program of the Edison Electric Institute and the American Institute of Electrical Engineers on the application of probability methods [8–11].

Current industry practice proceeds as follows. First, historical availability data are used to calculate an "availability statistic" for each generating unit. The predominant availability statistic in use in the U.S. is the equivalent forced outage rate of demand (EFORd) which seeks to estimate the conditional probability of a unit being unavailable when needed by the power system [12]. Second, the availability statistics for each generating unit in the power system are used to determine the distribution of available capacity for the system through analytical or simulation methods. Finally, the resulting distribution is compared to a forecast of system load to determine capacity requirements [13].[3]

The availability statistic approach to RAM distills multiple years of availability history for each

[1] The Pennsylvania-New Jersey Interconnection was a power pool formed in 1927. It was renamed the Pennsylvania-New Jersey-Maryland (PJM) Interconnection in 1956 when Maryland-based utilities joined. Its current footprint includes all or parts of 13 U.S. states and the District of Columbia.

[2] The most common reliability standard in use in North America is the "1-in-10" standard, usually interpreted to mean that a loss of load event due to insufficient generation capacity will occur on no more than one day in ten years on average [2, 3]. It is also sometimes interpreted to mean no more than 24 hours of loss of load due to supply shortages will occur in ten years on average [4]; various reliability regions have other interpretations [5].

[3] Current resource adequacy planning procedures for several control areas in the United States may be found in the following sources: [14–19].

generating unit to a single number. Because all temporal information is discarded, it implicitly assumes failures are independent among generating units [20]. However, failures could be correlated for a number of reasons, including common weather events, fuel supply disruptions, or a common vintage of defective mechanical components, leading to biased estimates of the level of capacity needed to achieve a reliability standard [21]. For lack of a tractable alternative, the assumption of independent failures is also made by scholars working outside RAM: [22] assume independence when simulating the marginal cost curve for electricity supply in California to test for the exercise of market power. Finally, we note that the availability statistic approach to RAM also implicitly assumes that generator availability is not seasonal.

Here we seek to test the validity of these two assumptions. To do this we devise a novel method for reconstructing the availability history of a generating unit from event records. We demonstrate our method using the Generating Availability Data System (GADS), a proprietary database maintained by the North American Electric Reliability Corporation (NERC) [23]. GADS contains more than two million event records affecting approximately 8,000 generating units between 2012 and 2015. These units represent approximately 85% of generation capacity in the conterminous U.S. and Canadian provinces. We use the GADS data to create time series of unavailable capacity from unscheduled, maintenance, and scheduled events for each unit.

Our primary objective is to use the hourly time series to test both for failure correlation among generating units and for seasonal availability patterns. While we are not the first to recognize the potential challenges posed by correlated failures for RAM, previous research has been hampered by a lack of access to the necessary data [5, 20].

In addition, we use the time series to generate inputs for Markov modeling of power systems. This includes Weibull and lognormal distributions fit to each region's series of unavailable capacity, Weibull distributions fit to unscheduled derating magnitudes by unit type, lognormal distributions fit to unscheduled event durations by event type, and lognormal distributions fit to hourly unscheduled event arrival probabilities. In each case we report the parameters of our fits. We also calculate the mean time between failure (MTBF) and mean time to recovery (MTTR) for every unit in the GADS data and test whether large and small units have different MTBF values by unit type. Finally, we

present hourly time series of unavailable capacity from unscheduled, maintenance, and scheduled events and publish the data. These may be used in conjunction with load data to study resource adequacy risks without assuming either independent failures or constant generator availability, which we believe represents a significant advancement versus current RAM practice.

The paper is organized as follows. Section 2.2 introduces the GADS data. Section 2.3 describes the steps we take to clean the data and generate time series of unavailable capacity. Section 2.4 presents our results. Section 2.5 concludes.

The novel results discussed in Section 2.4 are summarized here. We present the first evidence that correlated failures are present in most NERC regions, even when removing Hurricane Sandy and January 2014 from the data (a map of the NERC regions is Figure A.1 in the supplementary materials). We find that correlated failures can occur in any season. We show distributions of unscheduled unavailable capacity in each NERC region and find that they are reasonably well modeled by Weibull and lognormal distributions. The distributions of normalized derating magnitudes vary by unit type; combined cycle and simple cycle gas units are not well approximated by common parametric fits. Three out of five unit types that we studied show statistically significant differences in mean time between failure for small and large units. The mean time between failure for fossil steam units tends to be shorter for large units, while the mean time between failure for simple cycle and hydroelectric units tends to be shorter for small units.

2.2 Data

A working group of the Institute of Electrical and Electronics Engineers (IEEE) Application of Probability Methods subcommittee began developing generator reliability definitions to support the use of probability methods in bulk power system planning in 1968. This led to the creation of IEEE Standard 762, which provides the basis for generator availability data collection today [24]. NERC, formed in 1968 to develop voluntary standards to support bulk power system reliability following the Northeast blackout of 1965, assumed responsibility for collecting generator availability data from the Edison Electric Institute in 1979, renaming the database GADS [25, 26].

In response to rapid changes in the North American resource mix and NERC's designation as the electric reliability organization in the U.S. in 2006, NERC phased in mandatory reporting to GADS [27]. Beginning in January 2012, all units with nameplate capacities greater than 50 megawatts (MW), other than wind and solar generators, were required to report. This threshold was reduced to 20 MW in January 2013. There are approximately 8,000 units with events logged in GADS, representing approximately 85% of installed capacity in the conterminous U.S. and the Canadian provinces. The present analysis spans January 1, 2012 through December 31, 2015, the full period of mandatory reporting for which complete data were available at the time we began our work.

The GADS database comprises several tables. We use primarily the Units table, which records attributes of each generating unit reporting to GADS, and the Events table, which records each event affecting any generating unit reporting to GADS. Secondarily we use the Performance table, which records monthly summaries of the hours each generating unit spent in different operational states, to validate the Events table data.

Units are required to report nearly every event that affects their ability to generate electricity, even if dispatch requirements can still be met.[4] Approximately 500,000 events are logged each year under mandatory reporting. There are 20 event types in total, including startup failures (where the affected unit is fully unavailable due to a failure that occurred during its startup procedure), outages (where the affected unit is fully unavailable), deratings (where the affected unit is partially unavailable), reserve shutdowns (where the affected unit is offline for economic reasons but is not experiencing any reduction in its ability to generate power), unit retirements, and several others [28]. Each event logged in GADS reports the affected unit, the type of event, the start and end time of the event, and several additional details.

Outages and deratings are further classified as unscheduled, maintenance, or scheduled events based on how much advance notice the unit operator had before the event went into effect (ranging from none to several weeks). We focus on the seven unscheduled (forced) event types: startup

[4]Reporting failures that represent less than 2% of a unit's capacity and last less than 30 minutes is voluntary. Hydroelectric and pumped storage units without automatic data recording equipment are not required to report reserve shutdown events, but as noted above these events do not affect a unit's ability to generate power [28].

failures (the GADS term for these is SF), the three unscheduled outages (U1, U2, and U3), and the three unscheduled deratings (D1, D2, and D3).[5] These are the primary event types considered in RAM. We next describe our methods for processing the raw GADS events data into time series of unavailable capacity.

2.3 Methods

2.3.1 Preprocessing

Both the Units and Events tables required basic cleaning and preprocessing. Preprocessing of the Units table included removing any records missing a nameplate capacity value or having a NERC region code other than the eight corresponding to the conterminous U.S. and Canada.

Derating event do not have their magnitude directly reported. Instead, each derating records the net available capacity (NAC) remaining for the affected generating unit at the start of that event. To ensure that all derating magnitudes will be calculated correctly (Section 2.3.2), we check that each unit's nameplate capacity is greater than its largest reported NAC. An example is shown in Figure 2.1. Approximately 300 units' nameplate capacity values were updated by this procedure. This accounts for unit up-ratings, as GADS nameplate capacity values are not generally kept up to date by operators.

[5] Among the seven unscheduled event types, there are still temporal gradations: SF, U1, and D1 events take effect immediately, U2 and D2 events take effect within six hours, and U3 and D3 events can be postponed beyond six hours but not beyond the end of the upcoming weekend.

Scenario requiring update	***Scenario not requiring update***
Original nameplate capacity: 100 MW	Original nameplate capacity: 100 MW
Derating event 1: NAC 70	Derating event 1: NAC 70
Derating event 2: NAC 30	Derating event 2: NAC 30
Derating event 3: NAC 102	Derating event 3: NAC 100
Final nameplate capacity: 103 MW	Final nameplate capacity: 100 MW

Figure 2.1: Illustration of scenarios for which updating the unit's nameplate capacity is and is not necessary. In the example on the left, the unit experiences a derating event with a net available capacity (NAC) greater than its nameplate capacity so the nameplate is increased in order for all derating event magnitudes to be positive (see Section 2.3.2). In the example on the right, the unit's current nameplate capacity is sufficient to yield a positive magnitude for each derating event.

We also validate the reported time zone for each unit using ABB Velocity Suite [29]. For units whose time zone was updated by this process, we adjust the start and end times of its events accordingly.[6]

In the Events data, we remove any records missing a start or end date, as well as duplicate derating records. These are derating events that match on start time, end time, event type, and NAC. When derating events match on start time, event type, and NAC but have different end times, we keep only the event with the latest end time. These steps are necessary for correctly calculating the magnitudes of overlapping deratings, as described next.

2.3.2 Calculating derating magnitudes

Deratings account for 19-35% of all unscheduled unavailable MWh during our study period, depending on the NERC region. Thus it is important to treat reported deratings rigorously. If deratings never overlapped, each derating magnitude could be calculated as:

$$Magnitude\_of\_event = Nameplate\_capacity\_of\_unit - NAC\_of\_event \tag{2.1}$$

[6]Subsequent conversations with members of the GADS Working Group identified that at least one large utility sets all of its units to adhere to the time zone of headquarters, even when that conflicts with the time observed in the state. We do not account for this as it would be extremely difficult to confirm this behavior for the hundreds of reporting entities, but believe the bias introduced should be small.

However, deratings can overlap and usually the magnitude of the succeeding derating must be calculated as a function of the derating(s) already underway [28]. For example, if just one derating was already in progress, the magnitude of the succeeding derating must be calculated against it rather than against the nameplate capacity of the unit:

$$Magnitude\_of\_event = NAC\_of\_previous\_event - NAC\_of\_event \tag{2.2}$$

Any number of deratings can overlap, which makes determining the correct baseline event difficult. We develop specialized functions to handle all possible configurations of overlapping deratings.

2.3.3 Calculating time series of unscheduled unavailable capacity

With the derating magnitudes calculated, we next build hourly time series of unavailable capacity for each generating unit.[7] For outages and startup failures, unavailable capacity is the unit's nameplate capacity in every hour where an outage event is in effect. For deratings, unavailable capacity is the sum of the magnitudes of events in effect in each hour. We sum outages and deratings for each unit, cap the series at each unit's nameplate capacity, and aggregate the unit-level series to the eight NERC regions.[8] An example time series is shown in Figure 2.2.

[7] Despite event starts and ends reported to the minute, the large plurality of start and end times fall on the hour. Histograms of start and end minute of each unscheduled event are shown in Figure A.9 and Figure A.10.

[8] Because the original purpose of the GADS database was to facilitate unit benchmarking, a derating in progress when an outage occurs is not modified to prevent unavailable capacity from being overstated. Other potential causes of overestimation include events appearing to overlap at the hourly resolution.

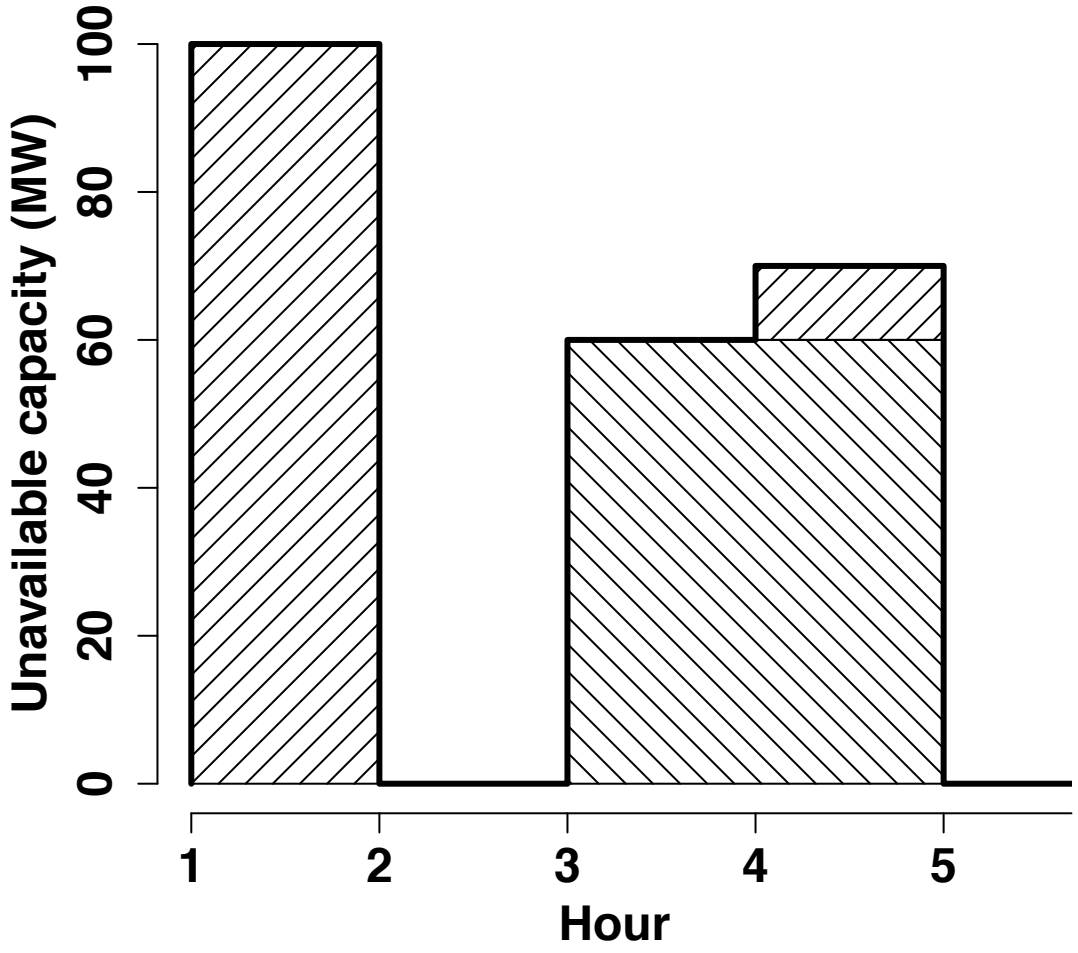

Figure 2.2: Illustrative unavailable capacity time series for one generating unit. In hour 1 the unit experiences an outage and is fully unavailable. In hour 2 the outage has been repaired and the unit is fully available. In hour 3 the unit experiences a 60-MW derating. In hour 4 the unit experiences a second derating (10 MW) in addition to the previous derating, so 70 MW is unavailable. In hour 5 both deratings have been repaired and the unit is fully available.

2.4 Results

We begin by presenting a brief descriptive summary of the GADS data by NERC region (Section 2.4.1). We then test for correlated failures (Section 2.4.2) and for seasonal patterns in unscheduled

unavailable capacity (Section 2.4.3). Finally, we present a set of analyses with direct application for reliability analysis (Section 2.4.4). These include parametric fits to each NERC region's distribution of unscheduled unavailable capacity; parametric fits to each unit type's distribution of normalized derating magnitudes; parametric fits to each region's distribution of event durations by event type; parametric fits to the hourly probability of an unscheduled event arrival by region; parametric fits to the distributions of mean time between failure and mean time to recovery by unit type and region; statistical tests of whether small and large units have different mean times between failure; and time series of unavailable capacity from unscheduled, maintenance, and scheduled events by unit type and region.

2.4.1 Descriptive analysis of unscheduled unavailable capacity

Installed capacity and unit counts for the eight NERC regions spanning the conterminous U.S. and the Canadian provinces are listed in Table 2.1. We show time series of unscheduled unavailable capacity for each NERC region as a percentage of installed capacity in Figure 2.3, using data from ABB Velocity Suite to construct an hourly series of each region's installed capacity. Hourly time series of the percent of units (unweighted by capacity) affected by an unscheduled event are shown in Figure A.2 in the supplementary material.

We report the mean (which may be thought of as the base rate of unscheduled unavailable capacity), median, maximum, and quartile coefficient of dispersion (QCD) for each region's time series in Table A.1. Over the four years we analyzed, the regions' mean unscheduled unavailable capacity ranged from 2.8% of installed capacity in FRCC to 6.3% in SPP. We use the QCD as a measure of the spread of the unscheduled unavailable capacity rather than the standard deviation because the data are asymmetric. The QCD ranges from 0.13 for SERC to 0.30 for FRCC. The ratio of the maximum to the mean ranges from 1.8 in WECC to 4.0 in RFC. There is more than a three-fold difference in the regional maxima, ranging from 7.2% of installed capacity for WECC to 22.6% for RFC.

Table 2.1: Description of the eight NERC regions in the conterminous U.S. and Canada.

Acronym	Definition	Installed capacity, start (MW)[1]	Installed capacity, end (MW)[2]	Unit count[3]
FRCC	Florida Reliability Coordinating Council	60,111	60,300	328
MRO	Midwest Reliability Organization	56,100	56,300	523
NPCC	Northeast Power Coordinating Council	149,700	146,900	1,142
RFC	ReliabilityFirst Corporation	227,800	215,000	1,441
SERC	Southeast Reliability Corporation	264,400	264,300	1,688
SPP	Southwest Power Pool	58,100	59,000	423
TRE	Texas Reliability Entity	80,400	81,900	428
WECC	Western Electricity Coordinating Council	206,900	209,300	1,903

[1] Starting installed capacity is the sum of nameplate capacity of active conventional units with nameplate capacities greater than 20 MW on January 1, 2012; wind and solar units are excluded. Data source: [29].
[2] Ending installed capacity is the sum of nameplate capacity of active conventional units with nameplate capacities greater than 20 MW on June 30, 2015; wind and solar units are excluded. Data source: [29].
[3] The number of units experiencing an unscheduled event during the study period.

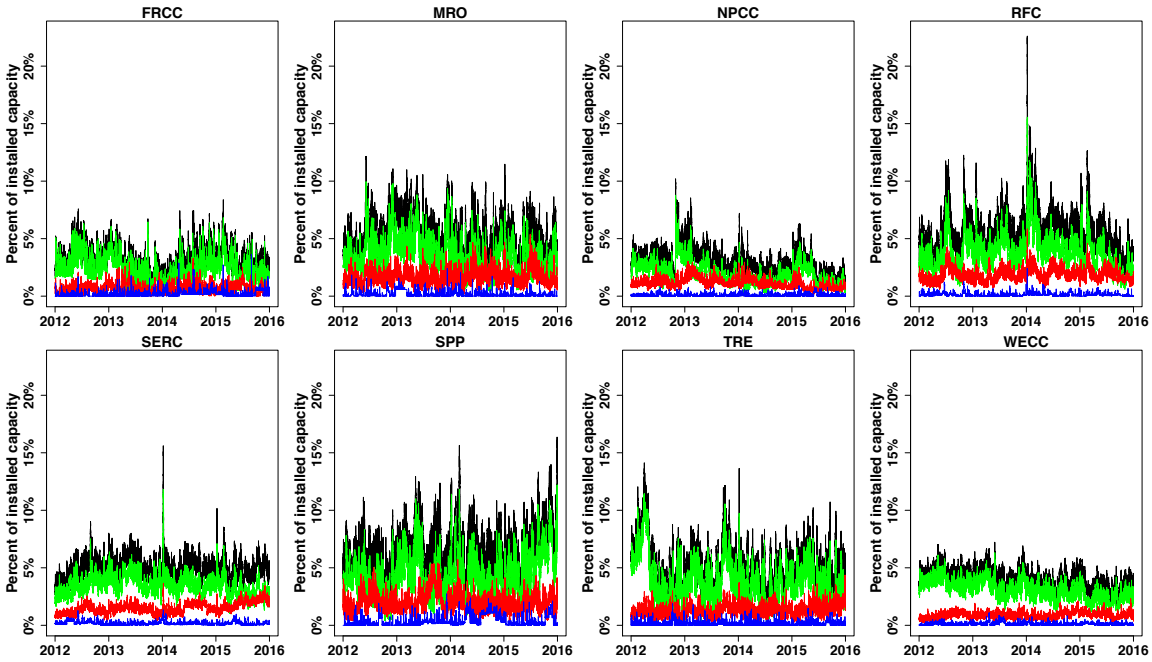

Figure 2.3: Unscheduled unavailable capacity as a percent of installed capacity, by unscheduled event type. Green: unscheduled outages only; red: unscheduled deratings only; blue: start-up failures only; black: all unscheduled events (the sum of green, red, and blue curves).

We compute the breakdown of unscheduled unavailable megawatt-hours (MWh), the sum of the product of each event's magnitude (MW) and duration (hours), by event type. On average, startup failures account for 3% (ranging from 2% to 5% for the eight NERC regions), deratings account for 27% (ranging from 19% to 35%), and outages account for 70% (ranging from 63% to 77%) of the total. These breakdowns are shown in Figure A.3. We present breakdowns of the count of unscheduled events by event type in Figure A.4.

2.4.2 Tests for independence among generators

We wish to test the assumption that generator failures are independent. From Figure 2.3 it is clear that several regions show instances of much greater unavailable capacity than their base rate. These could be due to correlated failures or to random chance.

We test whether the peaks violate the independent failures assumption using two main methods. We describe each briefly here, then give details and results in the next sections. First we apply block subsampling to "shuffle" each unit's observed time series independent of every other unit. Summing over units to regions yields a simulated time series for each region that is representative of each unit's observed performance, but which breaks any correlation among generator failures that may have been present. Repeating this process many times allows us to compare the prevalence and magnitudes of large unavailable capacity instances in the observed time series to what is possible under the null hypothesis of independent failures. We do this by creating confidence bands from each region's subsampled runs and plotting them along with the empirical series as exceedance curves.

As a second test, we model each unit's hourly availability as a binomial random variable using its observed time series to determine the probability of an event arrival in each hour. With these arrival probabilities we then simulate representative time series independently for each unit. As with block subsampling, we then aggregate the unit series to regions. Repeating this process many times allows us to compare the prevalence of large unavailable capacity instances in the observed series to what is possible under the null hypothesis of independent failures.

With each method, we look for violations of the independent failures assumption both with and without Hurricane Sandy and the cold weather events of January 2014 in order to test the possibility that these two well-known events were responsible for all the observed violations during our study period.[9]

[9] We remove all hours from October 29, 2012 through November 30, 2012 for Hurricane Sandy and from January 1, 2014 through January 31, 2014 for the Polar Vortex and the subsequent winter storms of January 2014 to allow time for some unit repairs to be completed. We do this in all regions for consistency.

Test of independent failures method 1: Block subsampling

We first test whether the observed generator failures are independent using block subsampling with replacement on each unit's time series. The time series of unavailable capacity for a generator has significant and important dependence over time. This dependence is preserved by sampling blocks of hours instead of individual hours. Sampling blocks independently ensures independence between distinct generators. In essence, block subsampling allows us to "shuffle" each unit's series independent of every other series, breaking any dependence across units while preserving dependence within units. This allows us to generate new (simulated) regional distributions under the null hypothesis that generator failures are independent. By repeating this process many times, we can trace out the space of distributions that is consistent with independent failures (the null hypothesis) for each region.[10] We reject the hypothesis that the generator failures are independent if a region's empirical distribution exceeds the upper bound of its 99% confidence band at any point above the 50th percentile.

We begin by generating 1000 subsampled series for each region using the full study period. This is consistent with current industry practice: it assumes not only independent failures among units, but also no seasonality in generator performance.[11] We use these series to generate 95% and 99% confidence bands of the distribution of unscheduled unavailable capacity under the null hypothesis, which we plot together with the region's empirical distribution as exceedance curves (termed survival curves in medical and some reliability literature) in Figure A.11. We summarize the percentiles at which each region's empirical distribution exceeds the upper bound of the 99% confidence band, along with the maximum magnitude of exceedance, in the left-hand side of Table A.4.

Six regions (MRO, NPCC, RFC, SERC, SPP, and TRE) show evidence of correlated failures at the 99% confidence level. FRCC and WECC are the only two regions whose empirical distributions do not exceed the upper bounds of their 99% confidence bands at any point in their respective domains. As a measure of whether the exceedances we observe in these regions represent a resource adequacy

[10] The block length is a function of the autocovariance sequence, the spectral density function, and the length of the time series [30]. We compute each unit's block length using the np library in R [31, 32]. Subsampling is carried out using the boot library in R [33, 34].

[11] Because there is no requirement that, for example, a winter observation be selected when populating winter hours in the subsampled series.

risk, we determine the amount of capacity that must be procured in order to achieve the 1-in-10 loss of load expectation (LOLE) standard under the assumption of independent failures. The "one day in ten years" interpretation of this rule translates to 2.4 hours of loss of load expectation per year, denoted 2.4 LOLH [2] as used in the SPP region; other NERC regions use slightly different interpretations. 2.4 LOLH is indicated via the dashed horizontal line in Figure A.11; the corresponding amount of capacity required at 95% and 99% confidence is indicated by the dashed vertical lines, drawn where the dashed horizontal line intersects the upper bound of each region's confidence bands. We define a region as having "managerially significant" correlated failures if its empirical distribution exceeds the amount of capacity required to meet the 2.4 LOLH criterion at the 99% confidence level, at an incidence greater than that corresponding to 2.4 LOLH. Using this definition, we conclude that managerially significant correlated failures are present in NPCC, RFC, SERC and TRE during the full study period.

Hurricane Sandy in 2012 and the two cold events in January 2014 were responsible for the largest violations of the independence assumption in our study period. To see if other correlated failures exist, we remove October 29-November 30, 2012 and January 2014 and repeat our analysis. As before, we plot exceedance curves (Figure A.12) and summarize the instances where each region's empirical distribution exceeds the upper bound of its 99% confidence band, along with the maximum magnitude of exceedance (right-hand side of Table A.4).

Even without Hurricane Sandy and January 2014, five regions (NPCC, RFC, SERC, SPP, and TRE) show evidence of correlated failures at the 99% confidence level. When considering the 2.4 LOLH resource adequacy requirement, we conclude that managerially significant correlated failures were present at the 99% confidence level in only NPCC, RFC, and TRE.

Test of independent failures method 2: Modeling hourly availability as a binomial random variable

We next test whether the observed peaks in unavailable capacity are due to correlation or to random chance by modeling each unit's hourly availability as a binomial random variable. We

estimate the probability of an unscheduled event arrival at each unit in a given hour as:

$$P(arrival_i) = \frac{C(events_{D1:D3,SF,U1:U3_i})}{C(hours_{1:T}) - C(hours_{SF,U1:U3_i})} \tag{2.3}$$

where C indicates the count of the elements taken in its argument, i indexes generating units, T indicates the final hour of the study period, and $D1:D3$, SF, and $U1:U3$ refer to the seven unscheduled event types.[12] When calculating this probability, we subtract the number of hours in which the unit is fully unavailable from the total period hours because no additional event arrivals can occur during these times. We retain only the units that are at least partially available for at least 1,000 hours (approximately 6 weeks) during the study period; this removes nine units from the analysis. Histograms of the estimated probabilities are reported in Figure A.14. Parameters from lognormal fits to the estimated probabilities are reported in Table A.5.

With the event arrival probabilities calculated for each unit, we then draw from each unit's parameterized binomial distribution as many times as there are hours in the study period to create a simulated series of event arrivals for each unit. We populate each event's magnitude and duration by sampling uniformly with replacement from the unscheduled events experienced by that unit. After completing this process, we cap each unit's series of unavailable capacity at its nameplate capacity and aggregate the unit-level time series to the regions. We show exceedance curves in Figure 2.4.

We again adopt the convention of rejecting the hypothesis of independent failures if a region's empirical distribution exceeds the upper bound of its 99% confidence band at any point above the 50th percentile. We conclude that all regions except FRCC violate the independence assumption. While this finding for WECC differs from the corresponding block subsampling result, we note that our definition of statistical significance ignores the magnitude of exceedance and that the results are qualitatively quite similar. When considering the 2.4 LOLH resource adequacy requirement, only NPCC, RFC, SERC, and TRE exhibit resource adequacy risk for the full study period, in agreement with block subsampling.

[12] Assuming constant failure probabilities, as we do here, is again consistent with typical RAM practice in the U.S. which implicitly assumes no seasonality in generator availability.

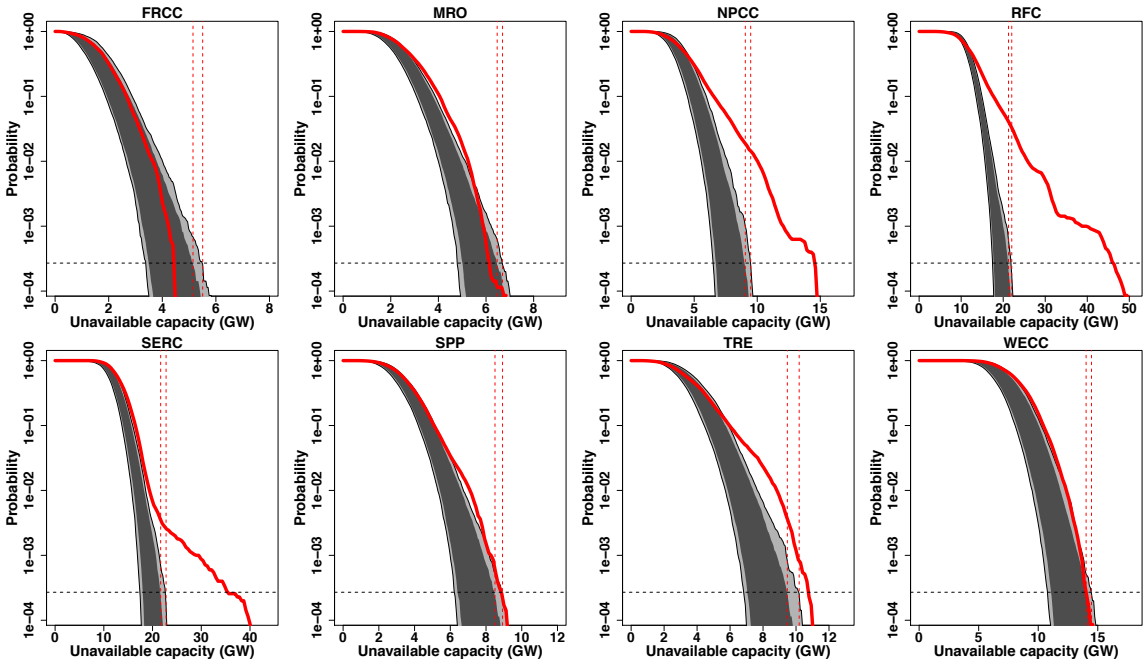

Figure 2.4: 95% and 99% confidence bands from 1000 binomial simulation runs shown in dark and light gray, respectively; empirical distributions from full study period shown in red. Dashed horizontal line indicates 2.4 LOLH threshold; dashed vertical lines indicate intersection of 2.4 LOLH threshold with the upper bound of each confidence band.

We again remove the hours corresponding to our definition of Hurricane Sandy and January 2014 and repeat our analysis. Now when creating the simulated series, we exclude events that start inside either deleted period; events that start prior to and continue into or beyond either period are not removed or altered. Exceedance curves of the results are shown in Figure 2.5.

Without these two months of data we again conclude that all regions except FRCC exhibit violations of the independent failures assumption. When considering the 2.4 LOLH resource adequacy requirement, NPCC, RFC, SERC, and TRE were the only regions to exhibit resource adequacy risk. While this finding for SERC differs from the corresponding block subsampling result, we again note that our definition of statistical significance ignores the magnitude of exceedance and that the results are qualitatively quite similar.

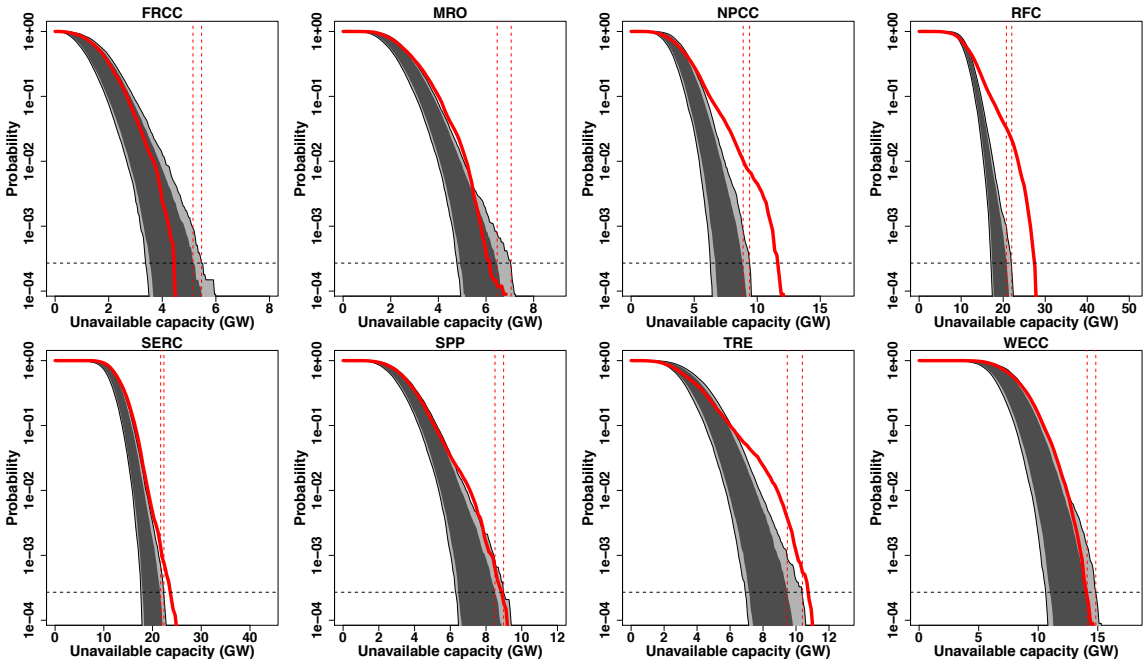

Figure 2.5: 95% and 99% confidence bands from 1000 binomial simulation runs shown in dark and light gray, respectively; empirical distributions from removing Hurricane Sandy and January 2014 shown in red. Dashed horizontal line indicates 2.4 LOLH threshold; dashed vertical lines indicate intersection of 2.4 LOLH threshold with the upper bound of each confidence band.

There is reasonable agreement between the block subsampling and the binomial results. In the full study period we conclude that six and seven regions, respectively, exhibit violations of the independent failures assumption under our basic definition of correlated failures. When removing Hurricane Sandy and January 2014 from the study period, five and seven regions, respectively, exhibit violations under this definition. By either method, NPCC, RFC, SERC, SPP, and TRE show clear evidence of violating the independent failures assumption, even when Hurricane Sandy and January 2014 are removed. We also see reasonable agreement between block subsampling and binomial results for both study periods when applying our managerially significant correlated failures definition. We summarize these results in Table 2.2.

While FRCC, MRO, and WECC show little to no evidence of violating the independent failures assumption over the period examined, we caution that four years of data is not sufficient to conclude that no such violations are possible in these regions. For example, on September 8, 2011 WECC experienced system disturbances that resulted in a loss of 7 GW of capacity, representing a 4-sigma

Table 2.2: Summary of correlated failure test results. "–" indicates no correlated failures; "*" indicates correlated failures according to our basic definition (the region's empirical trace exceeds the upper bound of its 99% confidence band above the 50^{th} percentile); "**" indicates correlated failures according to our definition of managerial significance (the region's empirical trace exceeds the level of capacity corresponding to the intersection of the upper bound of the 99% confidence band with 2.4 LOLH, with greater incidence than allowed under the 2.4 LOLH resource adequacy requirement). The definitions are nested such that a region cannot satisfy the second definition without also satisfying the first.

Region	Full period		Excluding Hurricane Sandy and January 2014	
	Block subsampling	Binomial	Block subsampling	Binomial
FRCC	–	–	–	–
MRO	*	*	–	*
NPCC	**	**	**	**
RFC	**	**	**	**
SERC	**	**	*	**
SPP	*	*	*	*
TRE	**	**	**	**
WECC	–	*	–	*

event for our study period, larger than any event we observed in the region during the four years we studied [35].

2.4.3 Seasonality

We next wish to test whether there are intra-annual patterns in unscheduled unavailable capacity. Our goals are to understand whether violations of the independent failures assumption occur in only particular seasons and whether particular seasons experience more unscheduled unavailable capacity on average, more variability in unscheduled unavailable capacity, or a greater number of large unavailable capacity events than others. Systematic patterns in any of these attributes would support improved forecasting and could provide insight into whether reserve margins should be computed seasonally.

When do correlated failures occur?

For this analysis, we adopt NERC's definition of the seasons: winter is December through February, spring is March through May, summer is June through September, and fall is October

through November [36]. When considering winter and fall with and without January 2014 and Hurricane Sandy, respectively, we test six seasons in total.

We use block subsampling to generate 1000 simulated runs of unscheduled unavailable capacity for each region in each season.[13] From this, we compute 95% and 99% confidence bands and plot them as exceedance curves along with the corresponding empirical distributions (Figure A.15 through Figure A.20). We find violations of the independent failures assumption in all seasons using our basic definition of correlated failures: four regions (NPCC, RFC, SERC, and SPP) in each winter definition, four regions (MRO, RFC, SERC, and TRE) in spring, three regions (NPCC, RFC, and TRE) in summer, five regions (MRO, NPCC, RFC, SPP, and TRE) in the full fall definition, and three regions (RFC, SPP, and TRE) in the shortened fall definition. We conclude that violations of the independent failures assumption can likely occur in any season in any region.

Seasonality in average unavailable capacity

Since violations of the independent failures assumption are observed in all seasons, we next examine whether there are recurrent patterns in average unavailable capacity by month. We compute the average unscheduled unavailable capacity in each month for each region and plot autocorrelation functions for each region (Figure A.21).

Significant seasonality would manifest as a lag-12 peak (corresponding to a one-year lag) that exceeds the 95% confidence bands. Except for FRCC, we see that each region's lag-12 peak is not significant. However, every region shows a significant 1-month lag, suggesting that unscheduled unavailable capacity can be thought of as an autoregressive process of order 1 (AR(1) process).[14] This is intuitive: failures can occur anytime during the year and require time to repair, so our best prediction of average unavailable capacity next month is the average unavailable capacity this month. As a robustness check, we repeat this analysis by NERC season; the results are consistent with the monthly result (Figure A.22). From these results we conclude that generally we cannot support the hypothesis of seasonality in average unavailable capacity from unscheduled events.

[13] We use only winter observations for winter, only spring observations for spring, and so on.

[14] A weakly stationary AR(1) model can be written $x_t = \mu + \rho x_{t-1} + \epsilon_t$ where ϵ_t is an independent and identically distributed zero-mean process with variance σ^2 and $|\rho| < 1$. The temporal dependence in x_t is completely summarized by conditioning on only its previous value.

As a complementary approach, we make exceedance curves for the empirical distribution of unscheduled unavailable capacity in each of the 17 seasons fully or partially covered by our study period (Figure A.23).[15] We observe significant overlap of the seasonal exceedance curves in most regions, indicating that periods of low and high unscheduled unavailable capacity can occur in any season.

Heteroskedasticity

We next study whether certain times of the year have more variability in unscheduled unavailable capacity. If so, these periods could represent elevated resource adequacy risks. We test for the presence of heteroskedasticity at the monthly level by fitting AR(1) terms to each region's monthly series of average unavailable capacity and plotting the residuals (Figure A.24). The residuals resemble white noise and appear to be homoskedastic. Autocorrelation functions of the residuals show no significant remaining structure (Figure A.25). Values and t-statistics for the AR(1) parameters are reported in Table A.8. From these results we conclude that we cannot generally support the hypothesis that certain times of the year systematically have more variability in unscheduled unavailable capacity than do others.

Would seasonal availability statistics improve RAM?

Current RAM practice in North America calculates a single availability statistic for each generating unit using five years of historical data. This implicitly assumes that generator availability is constant throughout the year. If instead generator availability was seasonal, calculating availability statistics separately for each season could improve the accuracy of each season's probability distribution of available capacity. To assess these potential benefits we combine the seasonal block subsampling results and plot the results as exceedance curves for both the full study period (Figure A.26) and when excluding Hurricane Sandy and the Polar Vortex (Figure A.27). Consistent with the results from our previous tests of seasonality, we find minimal benefits from calculating availability statistics separately for each season.

[15] The first winter includes only January and February 2012 (i.e. no December 2011); the fifth winter includes only December 2015 (i.e. no January and February 2016).

In summary, even with just four years of data we see violations of the independent failures assumption in all seasons. We do not see recurrent seasonal patterns in unscheduled unavailable capacity on average or in terms of variance. Finally, we do not find evidence to suggest that seasonal availability statistics would significantly improve the accuracy of RAM. With a longer study period it is possible that more intra-annual structure would emerge (for example from hurricanes), thus we recommend that system planners repeat this analysis to assess implications of seasonality for RAM in their control areas.

2.4.4 Reliability applications

We next present a set of results that can be used to populate Markov models of the generating units in each NERC region. We report: (1) Weibull and lognormal distributions fit to each region's series of unscheduled unavailable capacity; (2) Weibull distributions fit to each unit type's normalized derating magnitudes; (3) lognormal distributions fit to each region's event durations by event type; (4) lognormal fits to the hourly probability of an unscheduled event arrival by region; (5) mean time between failure and mean time to recovery values for each region and unit type, with fitted Weibull and gamma distributions; and (6) time series plots of unavailable capacity from unscheduled, maintenance, and scheduled events.

Markov models of generator availability have long been employed in reliability analyses. In a standard two-state model, a unit is assumed to be either fully available or fully unavailable, with failure rate $\lambda = 1/MTBF$ and recovery rate $\mu = 1/MTTR$ [37].[16] These values can be used to define the steady-state availability and unavailability of a generating unit. Unit availability models are also implicitly employed in current RAM practice in the definition of the availability statistic computed for each unit [12]. Many extensions have been made to improve the applicability of these models. For example, additional Markov states have been added to model partial unit availability, maintenance and planned outages, and whether periods of unit unavailability coincide with periods of system need. Models have also been extended to sets of units, allowing for both independent and common-mode failure states [38, 39].

[16] For completeness we note that sometimes the term mean time to failure (MTTF) is used to indicate the same concept as we are terming MTBF [37].

The primary challenge for populating Markov models is data availability. NERC GADS and Strategic Power Systems' Operational Reliability Analysis Program (ORAP) are the main sources for reliability data in the U.S., but neither makes sufficiently disaggregated data publicly available. Representative examples of reliability metrics published in the literature include MTBF values for 10 power stations [40], MTTR and MTTF values for a single coal-fired generating unit modeled with 10 availability states [41], MTBF values for seven gas turbine units in India [42], MTBF and MTTR values for 11 gas turbine units in Nigeria [43], and MTTF values for a single combined-cycle unit modeled with 8 availability states in Israel [44].

We are not aware of any published source of MTBF and MTTR data for all of the generating units in a large power system. We provide below the first such data for the vast majority of generation capacity in the U.S. and Canada. In conjunction with the fit parameters for time series of unavailable capacity from unscheduled events, the normalized magnitudes of unscheduled deratings, unscheduled event durations, and hourly event arrival probabilities, these data can be used to significantly improve the numeric accuracy of reliability modeling.

Parametric fits to distributions of unscheduled unavailable capacity

We fit Weibull and lognormal distributions to each region's distribution of unscheduled unavailable capacity, both for the full study period (Figure A.28) and with January 2014 and Hurricane Sandy removed (Figure A.29). We report the parameters of each fit in Table A.9.

Parametric fits to distributions of normalized derating magnitudes

We fit Weibull distributions to each unit type's distribution of normalized derating magnitudes (Figure A.30). We report the parameters of each fit in Table A.10.

Parametric fits to unscheduled event durations by event type

We present histograms of event durations by event type and overlaid with lognormal fits in Figure A.6 through A.8. We report the fit parameters in Table A.2.

Parametric fits to hourly probabilities of unscheduled event arrivals

We fit lognormal distributions to each region's distribution of hourly probabilities of an unscheduled event arrival, calculated according to Equation 2.3. We present histograms of the results in Figure A.14. We report the parameters of each fit in Table A.5.

Mean time between failure and mean time to recovery

We determine the mean time between failure and mean time to recovery for each generating unit and fit Weibull and gamma distributions to the results. We define the mean time between failure (MTBF) as the average number of service hours that elapse between unscheduled reductions of availability of any magnitude.[17] To do this, we first process the 1.6 million reserve shutdown (RS) events reported during our study period into hourly time series for each unit. Any hour when an RS event is in effect is removed from the unit's corresponding time series of unscheduled unavailable capacity. We then calculate an MTBF for the unit by averaging the durations of all instances where it is fully available.

We generate capacity-weighted histograms of these values by associating each unit's nameplate capacity (reported in MW) with its MTBF (Figure 2.6 through Figure 2.10). In each of these plots we construct histograms with 50 bins. The heading of each plot reports the number of units for which an MTBF value can be calculated (numerator) and the number of units reporting at least a single unscheduled event during our study period (denominator), which serves as a proxy for the sample size.[18] We exclude units with significant discrepancies in RS reporting between the Events and Performance tables, taking that to indicate that RS hours may be incompletely reported on those units and thus that our estimate of the MTBF restricted to service hours would be unreliable. Table A.16 summarizes the proportion of capacity that appears to incompletely report RS events. We report selected percentiles of MTBF values for each unit type in Table 2.3 to facilitate comparison. Larger MTBF values indicate greater reliability.

[17] Restricting our attention to service hours is important since peaking units are likely to be offline for economic reasons for large portions of the year.

[18] We use this as an estimate of the number of units that were active during the study period as GADS does not always correctly record commercialization and retirement dates for units that are sold.

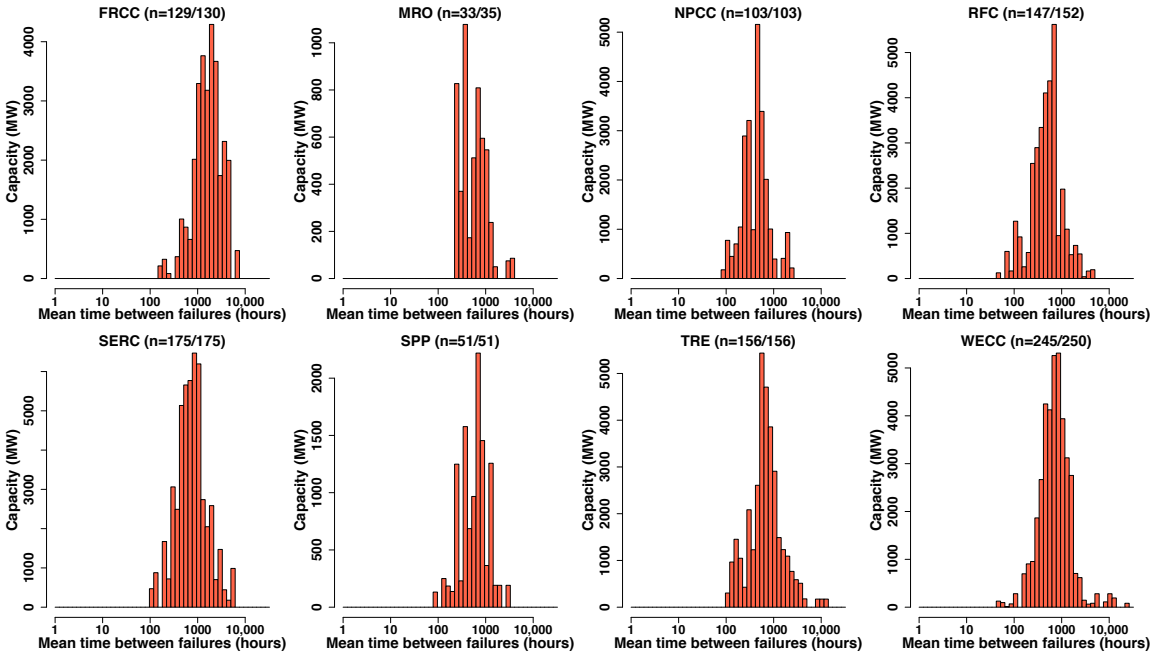

Figure 2.6: Capacity weighted mean time between failure (MTBF) values for combined cycle gas units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

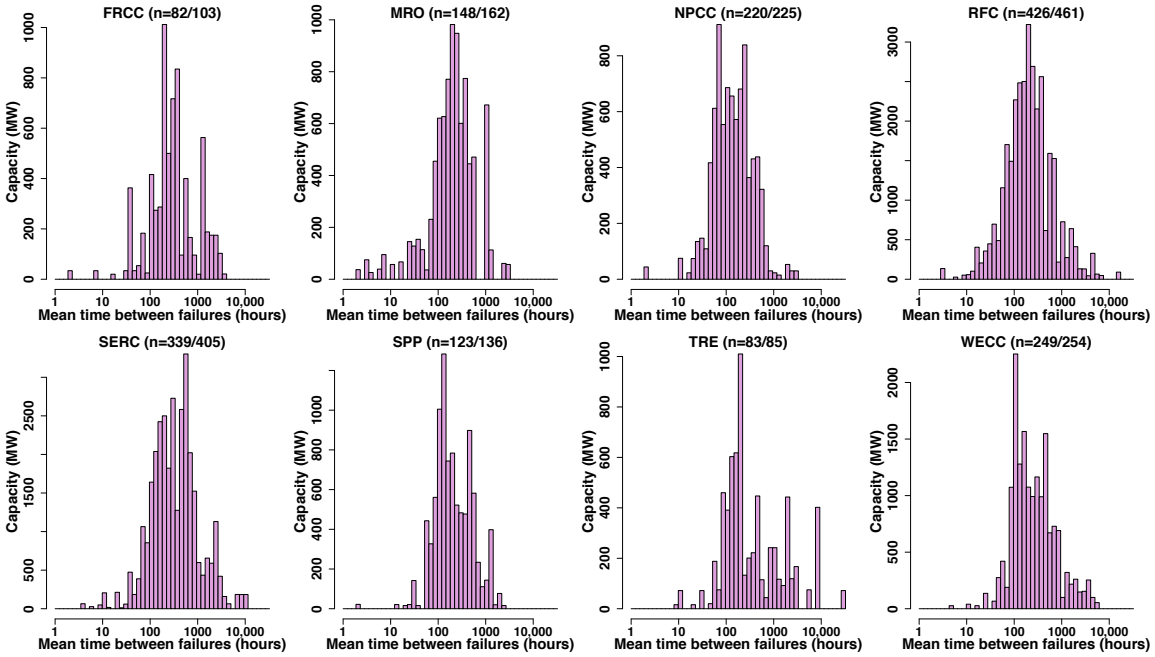

Figure 2.7: Capacity weighted mean time between failure (MTBF) values for simple cycle gas units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

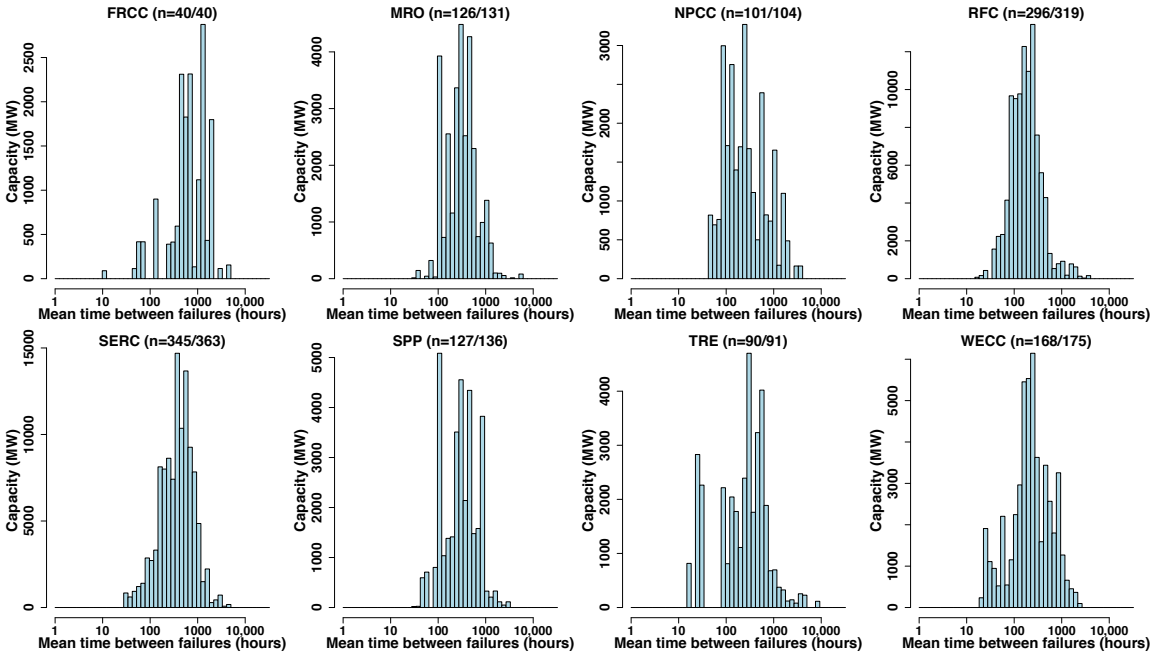

Figure 2.8: Capacity weighted mean time between failure (MTBF) values for fossil steam and fluidized bed units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

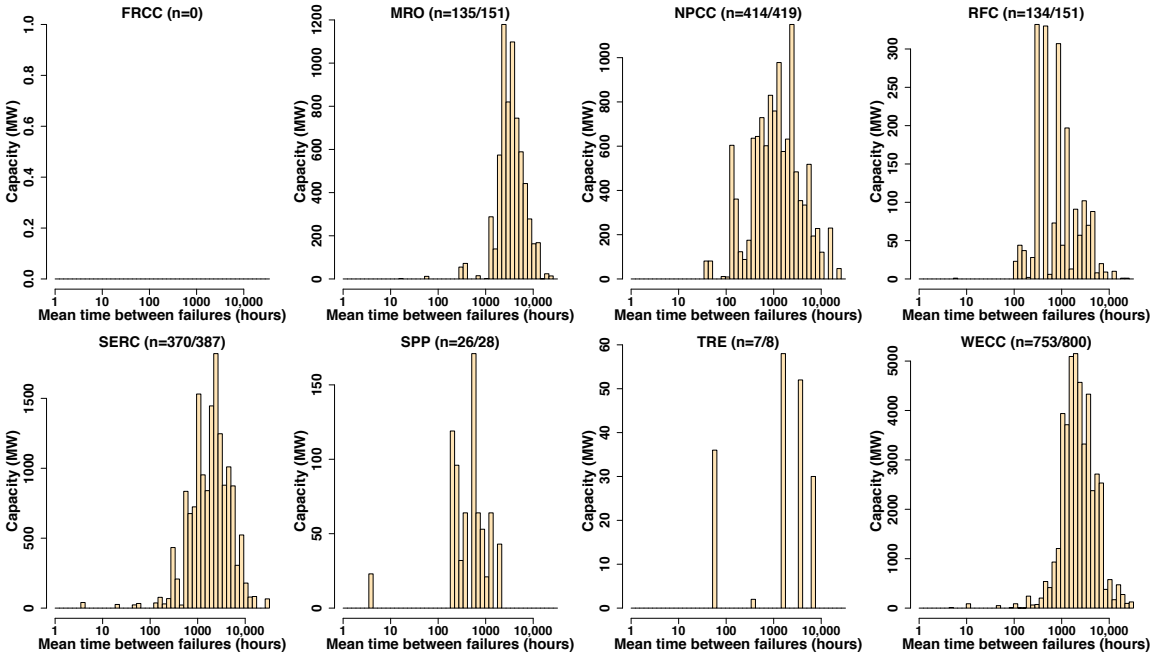

Figure 2.9: Capacity weighted mean time between failure (MTBF) values for hydroelectric units. There are no such units in FRCC. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

Table 2.3: Selected percentiles of capacity-weighted mean time between failure values (hours) by unit type. Values have been calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16). Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

Type	10^{th}	20^{th}	30^{th}	40^{th}	50^{th}	60^{th}	70^{th}	80^{th}	90^{th}
CC	250	362	464	572	669	808	1,002	1,336	1,984
CT	66	102	135	176	220	297	410	589	1,076
FSFB	84	119	167	213	263	335	445	576	853
HY	573	1,013	1,328	1,698	2,113	2,495	3,298	4,286	6,107
NU	702	1,097	1,503	1,839	2,032	2,412	2,850	3,629	4,889

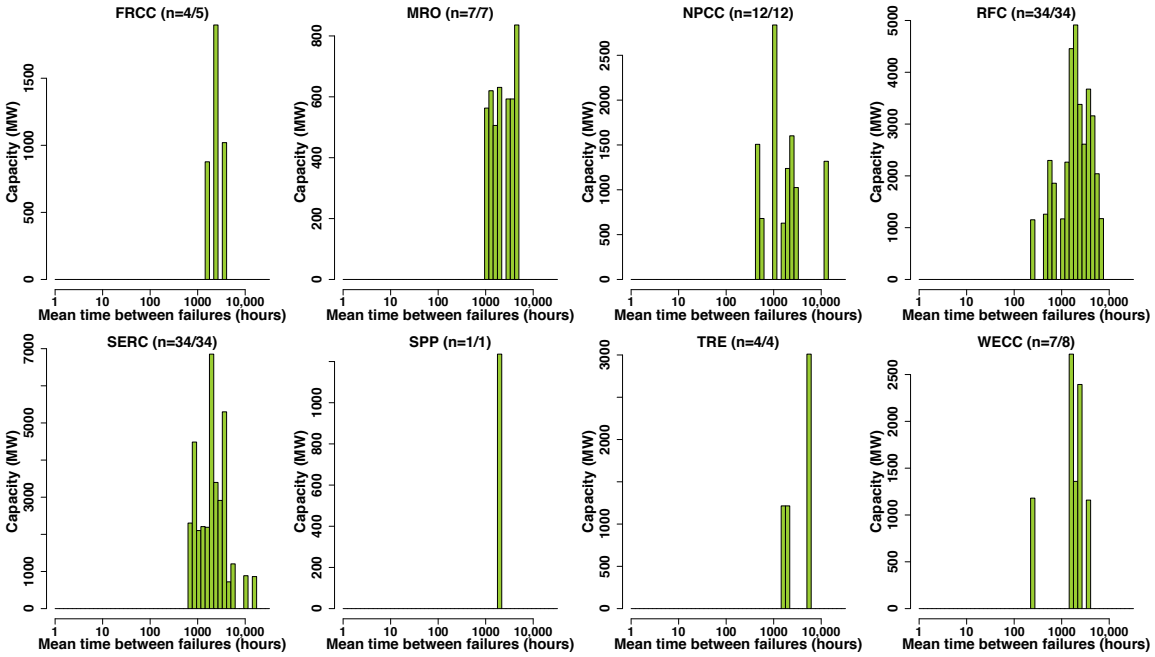

Figure 2.10: Capacity weighted mean time between failure (MTBF) values for nuclear units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). No nuclear units had significant reserve shutdown reporting discrepancies (see Table A.16).

We conclude that nuclear, hydro, and combined cycle units tend to run longer before failing than simple cycle and fossil steam units.[19] We fit Weibull and gamma distributions to the MTBF values

[19]We note that our results do not control for the type of failures, the age of the units, operations and maintenance expenditures, or other variables that may affect the MTBF.

by unit type and report the parameters in Table A.11 and Table A.12.

We note that these results depend to some degree on our definition of a failure. While we have defined a failure as any reduction from full availability, any desired threshold could be used. We include a sensitivity analysis over a range of failure definitions (Figure A.36). The MTBF results are quite insensitive to alternative failure definitions. We present histograms of the number of between-failure periods used to calculate each unit's MTBF in Figure A.31 through Figure A.35. We note that some units' MTBFs are calculated based on only a single between-failure period. With a longer time series, the proportion of units with MTBFs based on very few between-failure periods would decrease, increasing confidence in the robustness of these results. We also note that metrics such as the equivalent forced outage rate (EFOR) can complement the MTBF by summarizing the average availability of a unit over a desired study period, rather than just the frequency of reductions in availability.

We define the mean time to recovery (MTTR) as the average number of hours that elapse while a unit experiences some reduction in availability—i.e. the average duration of failure periods. In contrast to MTBF, we do not need to remove RS hours prior to calculating MTTR, so no units are excluded on the basis of their RS reporting fidelity. We present capacity-weighted histograms of the MTTR results (Figure A.37 through Figure A.41). The heading of each plot reports the number of units for which an MTTR value can be calculated (numerator) and the number of units reporting at least a single unscheduled event during our study period (denominator), which again serves as a proxy for the sample size. We summarize selected percentiles of MTTR values for each unit type in Table 2.4. Smaller MTTR values indicate shorter average repair durations.

We conclude that combined cycle units typically have among the lowest MTTR values while nuclear units have among the highest.[20] We fit Weibull and gamma distributions to the MTTR values by unit type and report the parameters in Table A.13 and Table A.14. We present histograms of the number of failure periods used to calculate each unit's MTTR in Figure A.42 through Figure A.46. We note that some units' MTTRs are calculated based on only a single failure period. With a longer time series, the proportion of units with MTTRs based on very few failures periods would

[20] We note that our results do not control for the type of failures, the age of the units, operations and maintenance expenditures, or other variables that may affect the MTTR.

Table 2.4: Selected percentiles of capacity-weighted mean time to recovery values (hours) by unit type. Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

Type	10^{th}	20^{th}	30^{th}	40^{th}	50^{th}	60^{th}	70^{th}	80^{th}	90^{th}
CC	6	10	13	17	21	28	38	60	115
CT	5	9	14	20	31	50	79	161	409
FSFB	15	21	27	34	44	56	74	107	240
HY	4	6	9	14	22	32	57	134	370
NU	49	69	81	97	126	176	332	680	1,058

decrease, increasing confidence in the robustness of these results.

Testing whether MTBF values differ for small and large units

We next test whether the MTBF values of large and small units differ. For each of the five unit types we consider, we define "large" units as those with nameplate capacities greater than or equal to the median value for that unit type, and "small" otherwise. We employ two tests of stochastic dominance: the Mann-Whitney U test (two-sample unpaired Wilcoxon test) and the two-sample Kolmogorov-Smirnov test.

For each of the 40 region-by-unit-type cases we examine, the Mann-Whitney U test sorts the MTBF values and uses the resulting arrangement of the labels (i.e. whether each data point represents a small or large unit) to compute a test statistic representing how "separated" the large and small units are. If the separation is great enough, we reject the null hypothesis that there is no statistically significant difference. We conduct both directions of a one-sided test for each of the 40 cases (i.e. testing both that small units have lower MTBF values than large units, and that large units have lower MTBF values than small units)—only one of which can be significant—and report the results in Table 2.5. We also conduct the test when aggregating over regions. Table 2.6 provides selected percentiles of nameplate capacity by unit type for reference. We present histograms with the MTBF values of small and large units overlaid in Figure A.47 through Figure A.51.

For combined cycle gas units, large units have statistically significantly lower MTBF values than small units in four regions (MRO, RFC, SERC, and TRE), FRCC exhibits the reverse relationship,

Table 2.5: Mann-Whitney U test results for stochastic dominance comparing MTBF values of small (S) and large (L) units by unit type and region. Large units are those with nameplate capacities greater than or equal to the median value for that unit type. $L < S$ indicates that large units have a statistically significantly lower MTBF than small units at the indicated significance level. Reserve shutdown hours have been removed. Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16). Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

Type	**CC**	**CT**	**FSFB**	**HY**	**NU**
FRCC	$S < L$ ****	–	$L < S$ *	N/A[1]	N/A[2]
MRO	$L < S$ *	$S < L$ ****	$L < S$ ***	–	N/A[2]
NPCC	–	$S < L$ ****	$L < S$ ***	$S < L$ ****	$S < L$ *
RFC	$L < S$ **	$S < L$ ****	$L < S$ ****	$L < S$ ***	$L < S$ *
SERC	$L < S$ ***	$S < L$ ****	$L < S$ ***	$S < L$ ***	–
SPP	–	–	$L < S$ ****	–	N/A[2]
TRE	$L < S$ ****	$S < L$ **	$L < S$ **	–	N/A[2]
WECC	–	$S < L$ ***	$L < S$ ****	$S < L$ ****	N/A[2]
Combined	–	$S < L$ ****	$L < S$ ****	$S < L$ ****	–

Significance levels: – ≥ 0.1; * < 0.1; ** < 0.05; *** < 0.01; **** < 0.001

[1] FRCC has no hydroelectric units, so no test could be conducted.

[2] Five regions do not have any nuclear units in one size category, so no test could be conducted.

Table 2.6: Selected percentiles of nameplate capacity values by unit type. Here units with significant reserve shutdown reporting discrepancies are included, so median nameplate capacity values do not always match the threshold used to define "small" and "large" units for the MTBF comparisons. Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

Type	**10th**	**20th**	**30th**	**40th**	**50th**	**60th**	**70th**	**80th**	**90th**
CC	60	91	155	173	182	191	217	269	370
CT	19	24	40	49	57	66	85	100	165
FSFB	50	80	111	150	199	265	394	527	661
HY	2	3	8	14	24	34	47	70	113
NU	663	867	887	946	1,022	1,155	1,175	1,207	1,274

and the remaining three regions show no significant difference. For simple cycle gas units, small units have statistically significantly lower MTBF values than large units in six regions, while FRCC and SPP showed no statistically significant difference. For fossil steam units, all eight regions show statistically significantly lower MTBF for large units than small units. For hydroelectric units, three of the regions (NPCC, SERC, and WECC) show smaller units having lower MTBF values than large units, RFC exhibits the reverse relationship, three regions (MRO, SPP, and TRE) showed no significant difference, and FRCC had no hydroelectric units. Finally, for nuclear units, five of the regions did not have representation from both "small" and "large" categories so no test could be conducted, SERC showed no significant difference, NPCC showed smaller units having statistically significantly lower MTBF values than large units, and RFC exhibited the reverse relationship. In the aggregate only simple cycle, fossil team, and hydroelectric units showed statistically significant MTBF values between small and large units.

As a robustness check, we also conduct two-sample Kolmogorov-Smirnov tests for each region-by-unit-type case. This test similarly seeks to determine whether two data samples come from the same population. The test statistic is the greatest discrepancy between the empirical distribution functions of the small and large units for the current region-by-unit-type case. The test statistic is compared to a critical value defined by the sample sizes and desired significance level. Results were consistent with the Mann-Whitney U test and are reported in Table A.15.

Time series of unavailable capacity from unscheduled, maintenance, and scheduled events

In addition to the seven unscheduled event types, GADS also includes maintenance and scheduled outages and deratings. Maintenance events have flexible start dates and are less urgent than unscheduled events. Scheduled (planned) events are set well in advance and are of predetermined durations.[21] We present time series of these events, overlaid with our time series of unscheduled events for context, by unit type (Figure 2.11 through Figure 2.15). We also publish these data so

[21] NERC defines maintenance events as those that can be deferred beyond the end of the upcoming weekend (i.e. beyond Sunday 2400 hours) if the event occurs prior to Friday 2400 hours, or beyond the end of the subsequent weekend if the event occurs after Friday 2400 hours [28]. While NERC does not give a precise definition for scheduled events, system operators typically require requests for scheduled events to be submitted with at least 30 days' notice [45].

that others may work with them. These time series can be used to model the statistical relationship of unavailable capacity and relevant features such as weather, attributes of the generating unit, and system load. Such a model could then be used to determine capacity requirements for a power system without making any assumptions about independence or seasonality.

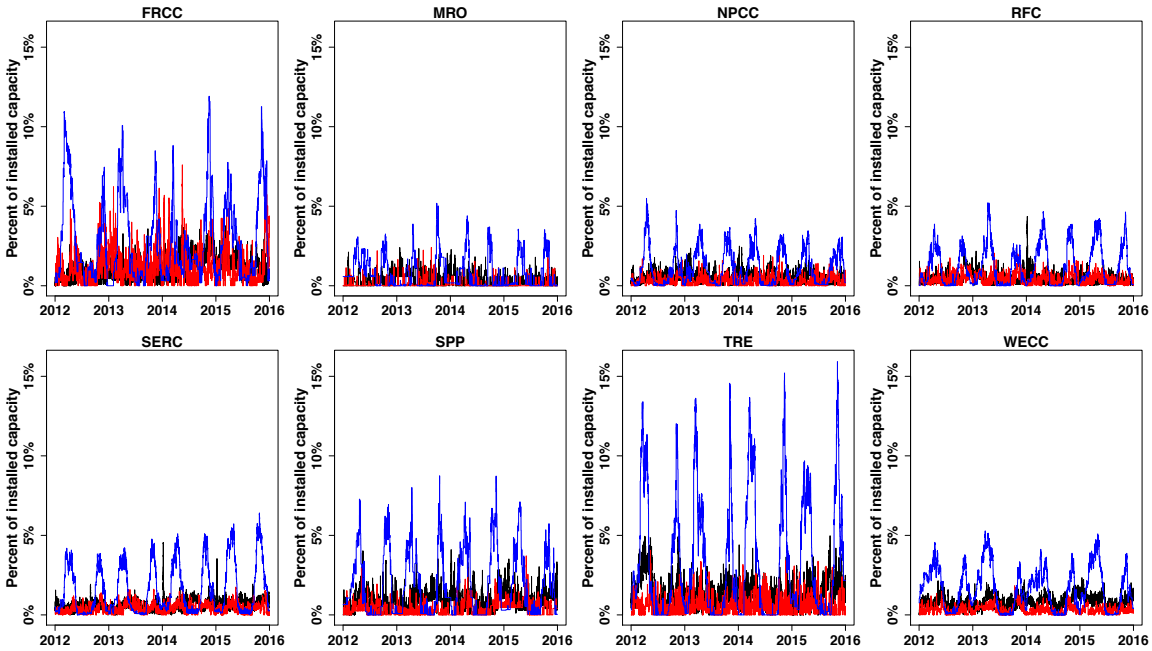

Figure 2.11: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for combined cycle units.

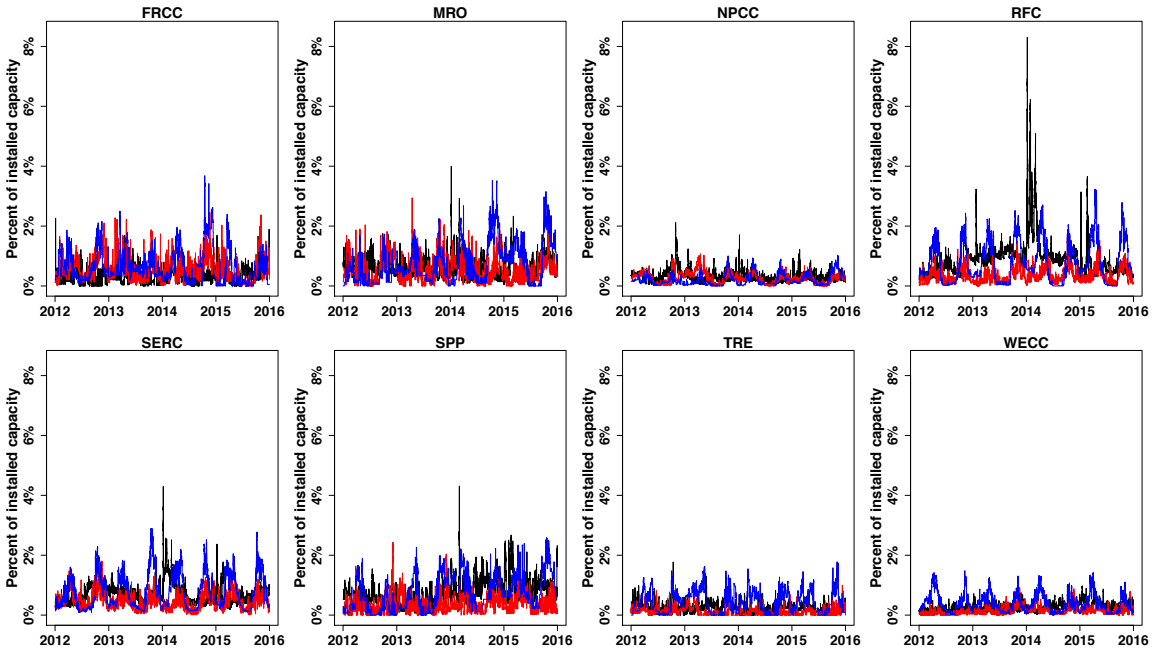

Figure 2.12: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for simple cycle units.

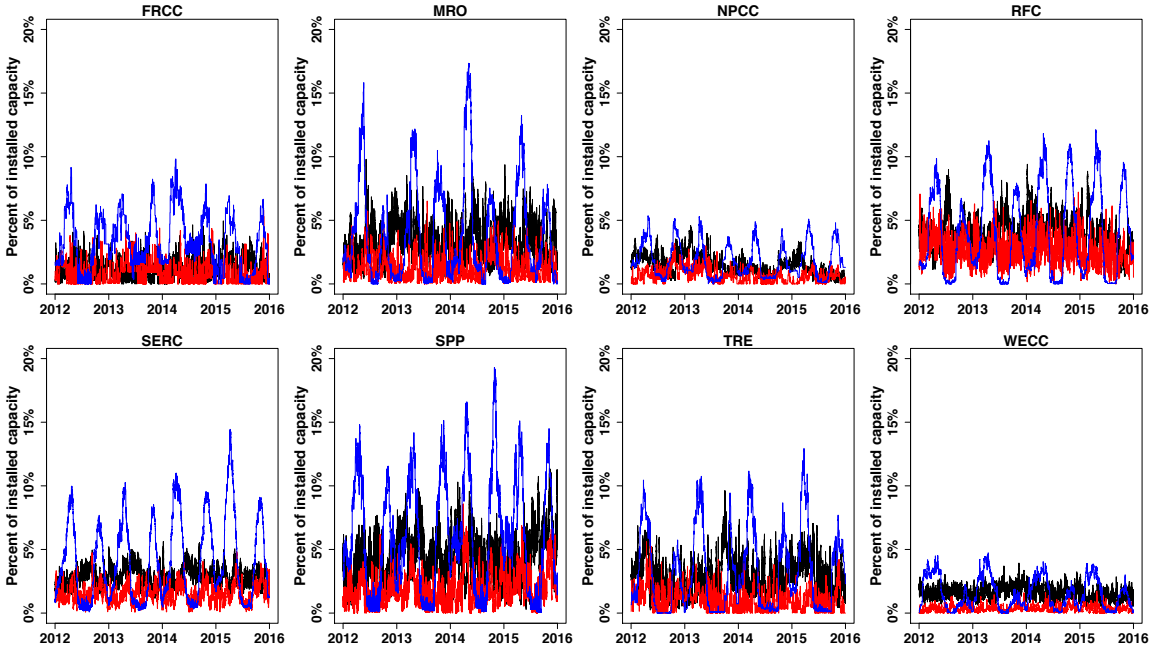

Figure 2.13: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for fossil steam and fluidized bed units.

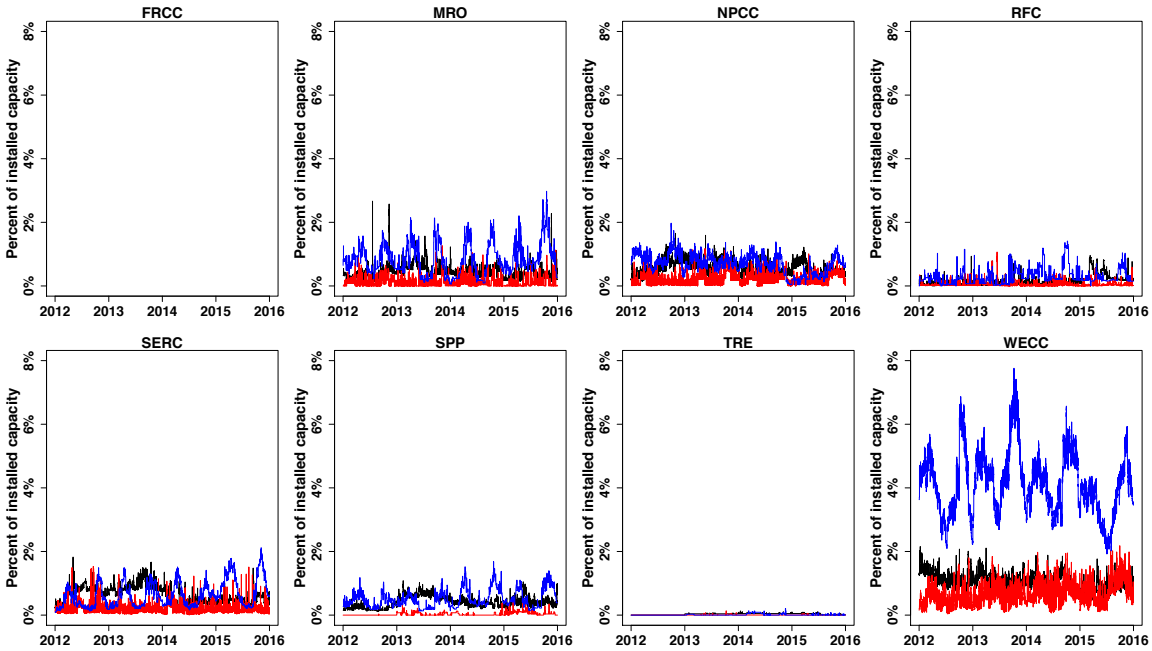

Figure 2.14: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for hydroelectric units.

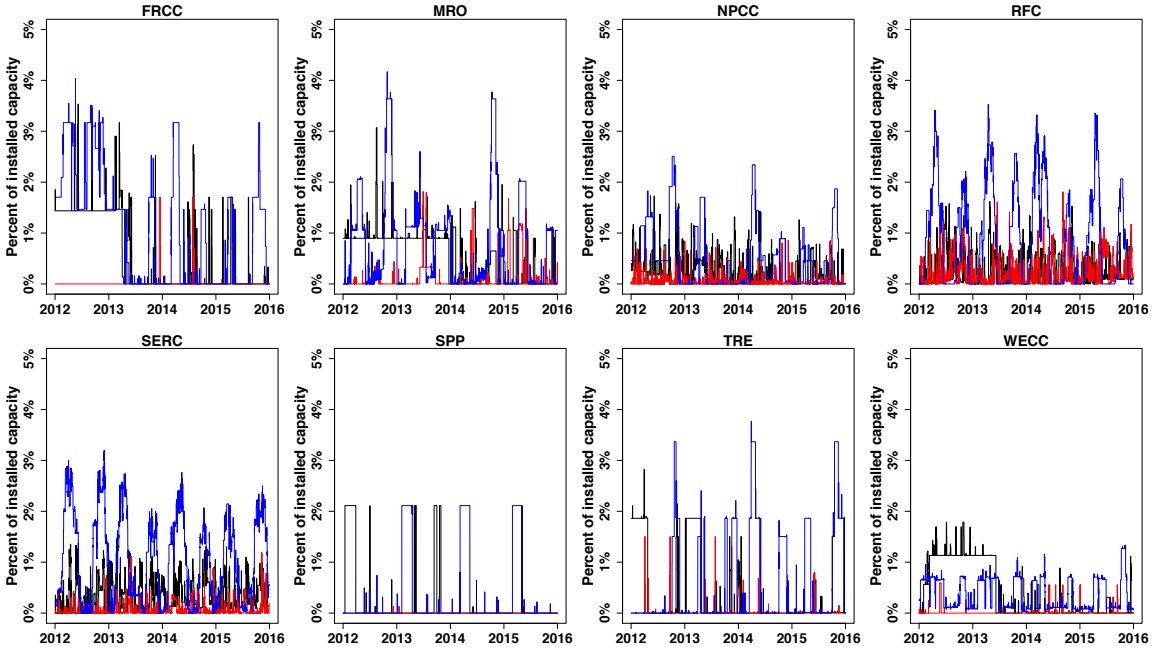

Figure 2.15: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for nuclear units.

2.5 Conclusions

Using four years of generator availability data for approximately 85% of installed capacity in the conterminous U.S. and the Canadian provinces, we have shown that correlated failures represent a significant resource adequacy risk. While FRCC and WECC were exceptions in this analysis[22], we note that neither region is likely immune to correlated failures; they simply did not experience correlated failures during our four-year study period.

We found little evidence for seasonal patterns in unscheduled unavailable capacity in the eight NERC regions. Instead we found that large unavailable capacity events can occur in any season and that unavailable capacity should be thought of as an AR(1) process, where the best prediction of average unavailable capacity this month is average unavailable capacity last month. These findings suggest that a seasonal resource adequacy construct, whereby an availability statistic is calculated for each unit in each season, may not meaningfully reduce resource adequacy risk. However, these conclusions may change with a longer study period.

Our findings highlight an important limitation of current resource adequacy modeling (RAM) practice: distilling the availability history of a generating unit to a single value (e.g. EFORd, the equivalent forced outage rate during times of high demand) discards important information about when units in a power system fail in relation to one another. Only by incorporating the full availability history of each unit into RAM can we account for correlations among generator failures when determining the capacity needs of a power system. We strongly recommend that system planners incorporate correlated failure analysis into their RAM practice.

Noting that the largest correlated failure instances were caused by extreme weather (Hurricane Sandy and the cold weather events of January 2014), we further recommend that unscheduled unavailable capacity be modeled as a function of relevant features (e.g. temperature and other weather variables, unit age, maintenance histories, system load). In conjunction with temperature and load forecasts for a desired planning year, system planners could likely compute improved estimates of capacity requirements. We are currently pursuing this research.

[22]FRCC showed no evidence of correlated failures during our study period. WECC showed evidence of correlated failures under our binomial modeling approach but not under our block subsampling approach.

In addition to our study of correlated failures and seasonality, we reported the results of a set of analyses that can be used to populate Markov models of generator availability for a power system. These included parametric fits to distributions of unscheduled unavailable capacity, derating magnitudes, event durations, hourly failure probabilities, and mean time between failures (MTBF) and mean time to recovery (MTTR) values. In each case we report the fit parameters for use by reliability practitioners. The final component needed to allow for correlated failures under a Markov model is a correlation matrix for the generating units. We do not report this due to confidentiality requirements, but note that it can be readily calculated using time series of unscheduled unavailable capacity; it cannot be computed when using the availability statistic approach to RAM. In addition we tested for statistically significant differences in the MTBF between small and large units. We found many significant differences when looking at regions individually; when aggregating over regions, simple cycle gas, fossil steam, and hydroelectric units showed statistically significant differences while combined cycle gas and nuclear units did not.

The requirement for mandatory reporting to GADS has allowed the development of the analyses presented. As additional years of data accumulate, the techniques used here will allow more robust results that may differ from the conclusions reached based on only four years of data.

References

[1] PJM Interconnection, "Analysis of operational events and market impacts during the January 2014 cold weather events." http://www.pjm.com/~/media/library/reports-notices/weather-related/20140509-analysis-of-operational-events-and-market-impacts-during-the-jan-2014-cold-weather.ashx, 2014.

[2] J. Pfeifenberger, K. Spees, K. Carden, and N. Wintermantel, "Resource adequacy requirements: Reliability and economic implications." https://www.ferc.gov/legal/staff-reports/2014/02-07-14-consultant-report.pdf, 2013.

[3] North American Electric Reliability Corporation, "2014 probabilistic assessment." http://www.nerc.com/pa/RAPA/ra/ReliabilityAssessmentsDL/2014ProbAAprilReportFinal_Final.pdf, 2015.

[4] K. Spees, S. Newell, and J. Pfeifenberger, "Capacity markets—Lessons learned from the first decade," *Economics of Energy & Environmental Policy*, pp. 1–26, 2013.

[5] R. Lueken, J. Apt, and F. Sowell, "Robust resource adequacy planning in the face of coal retirements," *Energy Policy*, vol. 88, pp. 371–388, 2016.

[6] S. Smith, "Service reliability measured by probabilities of outage," *Electrical World*, vol. 103, pp. 371–374, 1934.

[7] P. Benner, "The use of theory of probability to determine space capacity," *General Electric Review*, vol. 37, no. 7, pp. 345–348, 1934.

[8] G. Calabrese, "Generating reserve capability determined by the probability method," *AIEE Transactions on Power Apparatus and Systems*, vol. 66, pp. 1439–1450, 1947.

[9] W. Lyman, "Calculating probability of generating capacity outages," *AIEE Transactions on Power Apparatus and Systems*, vol. 66, pp. 1471–1477, 1947.

[10] E. Loane and C. Watchorn, "Probability methods applied to generating capacity problems of a combined hydro and steam system," *Transactions of the American Institute of Electrical Engineers*, vol. 66, pp. 1645–1657, 1947.

[11] H. Seylee, "Outage expectancy as a basis for generator reserve," *Transactions of the American Institute of Electrical Engineers*, vol. 66, pp. 1483–1488, 1947.

[12] M. Bhavaraju, J. Hynde, and G. Nunan, "A method for estimating equivalent forced outage rates of multistate peaking units," *IEEE Transactions on Power Apparatus and Systems*, vol. 97, no. 8, pp. 1224–1232, 1978.

[13] R. Billinton and W. Li, *Reliability assessment of electric power systems using Monte Carlo methods*. Springer Science and Business Media, first ed., 1994.

[14] ISO New England, "ISO New England installed capacity requirement, local sourcing requirements and capacity requirement values for the system-wide capacity demand curve for the 2019/20 capacity commitment period." `https://www.iso-ne.com/static-assets/documents/2016/01/icr_values_2019_2020_report_final.pdf`, 2016.

[15] MISO, "Business practices manual 11: Resource adequacy." `https://www.misoenergy.org/Library/BusinessPracticesManuals/Pages/BusinessPracticesManuals.aspx`, 2016.

[16] NYISO, "Installed capacity manual." `http://www.nyiso.com/public/webdocs/markets_operations/documents/Manuals_and_Guides/Manuals/Operations/icap_mnl.pdf`, 2017.

[17] New York State Reliability Council, "New York control area installed capacity requirement." `http://www.nysrc.org/pdf/Reports/2017%20IRM%20Study%20Report%20Final%2012-2-16%20(002).pdf`, 2016.

[18] PJM Interconnection, "PJM manual 20: PJM resource adequacy analysis." `http://www.pjm.com/~/media/documents/manuals/m20.ashx`, 2016.

[19] PJM Interconnection, "PJM manual 18: PJM capacity market." https://www.pjm.com/~/media/documents/manuals/m18.ashx, 2017.

[20] F. Felder, "Incorporating resource dynamics to determine generation adequacy levels in restructured bulk power systems," *KIEE International Transactions on Power Engineering*, vol. 4, pp. 100–105, 2004.

[21] Fegan, GR, "EL-3669: Reliability calculations for interdependent plant outages," 1984.

[22] S. Borenstein, J. Bushnell, and F. Wolak, "Measuring market inefficiencies in California's restructured wholesale electricity market," *American Economic Review*, vol. 92, no. 5, pp. 1376–1405, 2002.

[23] North American Electric Reliability Corporation, "The Generating Availability Data System." http://www.nerc.com/pa/RAPA/gads/Pages/default.aspx.

[24] IEEE Reliability Risk and Probability Methods Subcommittee, "History of the Application of the Probability Methods (APM) and Reliability, Risk and Probability Methods (RRPA) subcommittees." http://sites.ieee.org/pes-rrpasc/files/2015/09/IEEE_PES_RRPA_History_Final_08_24_2015.pdf, 2015.

[25] North American Electric Reliability Council, "The Generating Availability Data System (GADS): Applications and benefits." http://www.nerc.com/pa/RAPA/gads/Publications/GADS-Applications-and-Benefits.pdf, 1995.

[26] North American Electric Reliability Corporation, "NERC operating manual." http://www.nerc.com/comm/OC/Pages/Operating-Manual.aspx, 2016.

[27] North American Electric Reliability Corporation, "Generating Availability Data System: Mandatory reporting of conventional generation performance data." http://www.nerc.com/pa/RAPA/gads/MandatoryGADS/Revised_Final_Draft_GADSTF_Recommendation_Report.pdf, 2011.

[28] North American Electric Reliability Corporation, "Generating Availability Data System data reporting instructions." http://www.nerc.com/pa/RAPA/gads/Pages/DataReportingInstructions.aspx, 2017.

[29] ABB, "Velocity Suite." http://new.abb.com/enterprise-software/energy-portfolio-management/market-intelligence-services/velocity-suite.

[30] D. Politis and H. White, "Automatic block-length selection for the dependent bootstrap," *Econometric Reviews*, vol. 23, pp. 53–70, 2004.

[31] T. Hayfield and J. Racine, "Nonparametric econometrics: The np package," *Journal of Statistical Software, Articles*, vol. 27, pp. 1–32, 2008.

[32] A. Patton, D. Politis, and H. White, "Correction to 'Automatic block-length selection for the dependent bootstrap' by D. Politis and H. White," *Econometric Reviews*, vol. 28, pp. 372–375, 2009.

[33] Canty, A and Ripley, B, "boot: Bootstrap R (S-Plus) Functions. R package version 1.3-16." , 2015.

[34] A. Davison and D. Hinkley, *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press, 1997.

[35] Federal Energy Regulatory Commission and North American Electric Reliability Corporation, "Arizona-Southern California outages on September 8, 2011: Causes and recommendations." `http://www.nerc.com/pa/rrm/ea/Pages/September-2011-Southwest-Blackout-Event.aspx`, 2012.

[36] North American Electric Reliability Corporation, "2016 summer reliability assessment." `http://www.nerc.com/pa/RAPA/ra/ReliabilityAssessmentsDL/2016SRAReport_Final.pdf`, 2016.

[37] R. Billinton and R. Allan, *Reliability evaluation of engineering systems*. Springer Science and Business Media, second ed., 1992.

[38] W. Li, *Risk assessment of power systems: Models, methods, and applications*. IEEE Press, first ed., 2005.

[39] M. Papic, S. Agarwal, R. Allan, R. Billinton, C. Dent, S. Ekisheva, D. Gent, K. Kiang, W. Li, J. Mitra, A. Pitto, A. Schneider, C. Singh, V. Vadlamudi, and M. Varghese, "Research on common-mode and dependent (CMD) outage events in power systems: A review," *IEEE Transactions on Power Systems*, vol. 32, pp. 1528–1536, 2017.

[40] L. Sugianto and W. Mielczarski, "A fuzzy logic approach to optimise inventory," in *Proceedings of the 8th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 419–424, 1995.

[41] D. Elmakias, ed., *New computational methods in power system reliability*, vol. 111 of *Studies in computational intelligence*. Springer Science and Business Media, 2008.

[42] A. Sarkar, D. Kumar, S. Kumar, and M. Singha, "Reliability assessment of Rukhia gas turbine power plant in Tripura," *International Journal of Current Engineering and Technology*, vol. 2, pp. 184–195, 2012.

[43] S. Oyedepo, R. Fagbenle, and S. Adefila, "Assessment of performance indices of selected gas turbine power plants in Nigeria," *Energy Science & Engineering*, vol. 3, pp. 239–256, 2015.

[44] A. Lisnianski, D. Laredo, and H. Benhaim, "Multi-state Markov model for reliability analysis of a combined cycle gas turbine power plant," in *Proceedings of the 2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management*, pp. 131–135, 2016.

[45] PJM Interconnection, "PJM manual 10: Pre-scheduling operations." `https://www.pjm.com/-/media/documents/manuals/m10.ashx`, 2016.

[46] S. Paul and N. Ho, "Estimation in the bivariate poisson distribution and hypothesis testing concerning independence," *Communications in Statistics - Theory and Methods*, vol. 18, pp. 1123–1133, 1989.

2.6 Appendix A: Supplementary figures and tables

2.6.1 Characteristics of the generator availability data

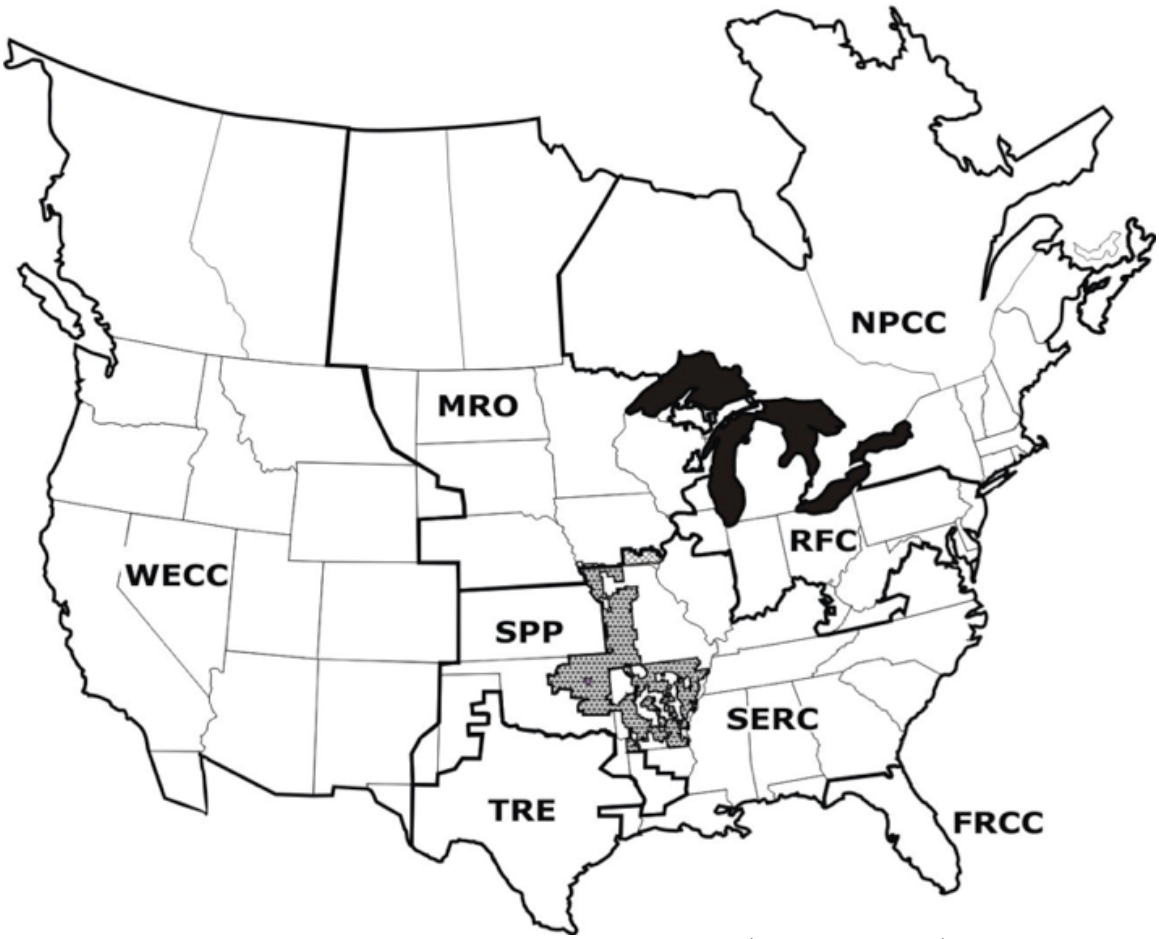

Figure A.1: The NERC regions. (Source: NERC)

Table A.1: Summary statistics for the regional time series of unscheduled unavailable capacity.

Region	**Mean (%)**	**Median (%)**	**Maximum (%)**	**QCD**[1]
FRCC	2.8	2.8	8.4	0.30
MRO	4.9	4.7	12.2	0.24
NPCC	2.9	2.8	10.2[2]	0.25
RFC	5.7	5.5	22.6[3]	0.19
SERC	5.0	5.0	15.6[3]	0.13
SPP	6.3	6.1	16.4	0.21
TRE	4.8	4.5	14.1	0.28
WECC	4.0	3.9	7.2	0.15

[1] The quartile coefficient of dispersion (QCD) is calculated as $(Q3 - Q1)/(Q3 + Q1)$, where $Q1$ and $Q3$ refer to the first and third quartiles, respectively.

[2] Hurricane Sandy, October 2012.

[3] Polar Vortex, January 2014.

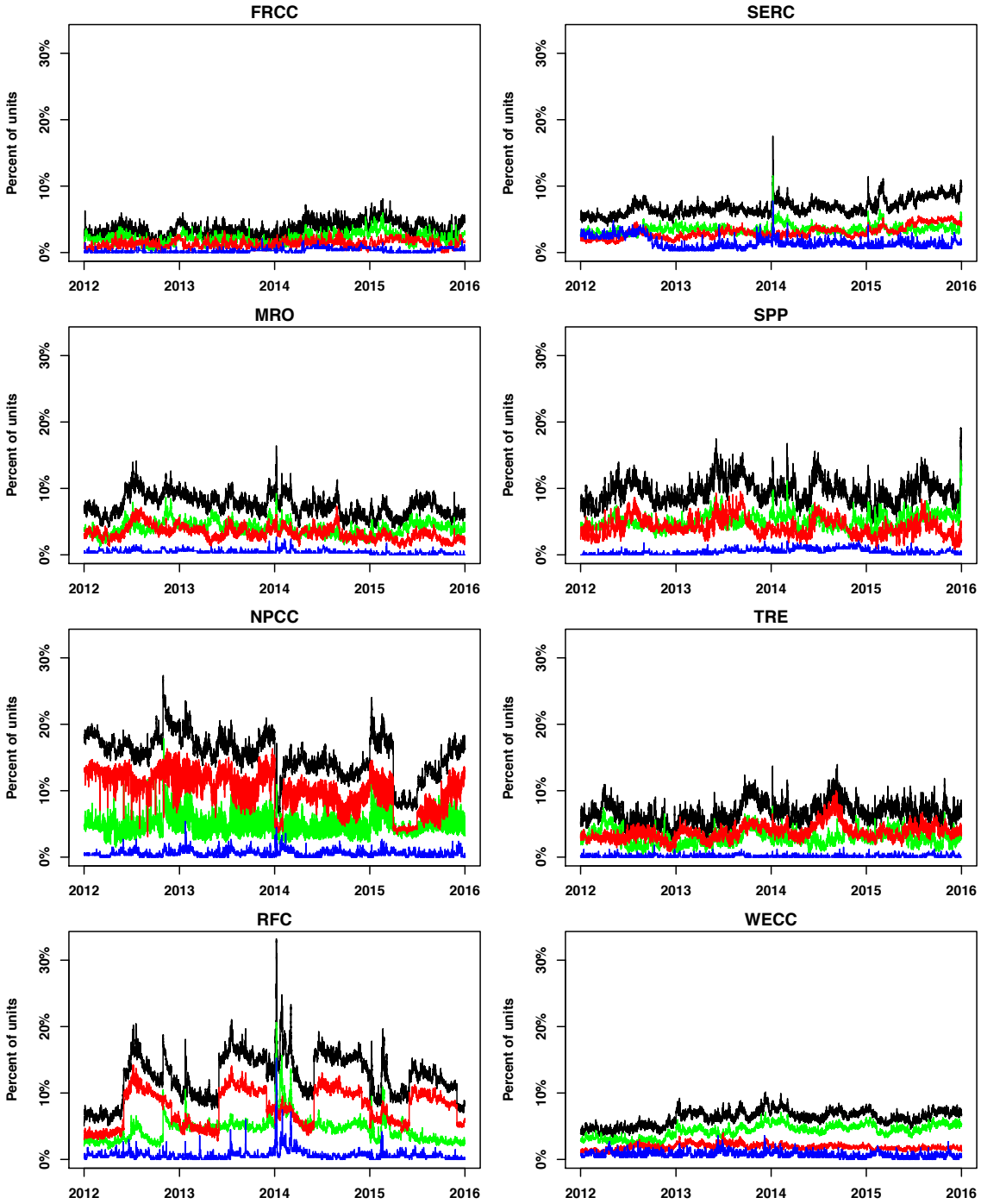

Figure A.2: Percent of units in each region reporting an unscheduled event in each hour of the time series. Green is unscheduled outages, red is unscheduled deratings, blue is start-up failures, and black is the sum of all unscheduled events.

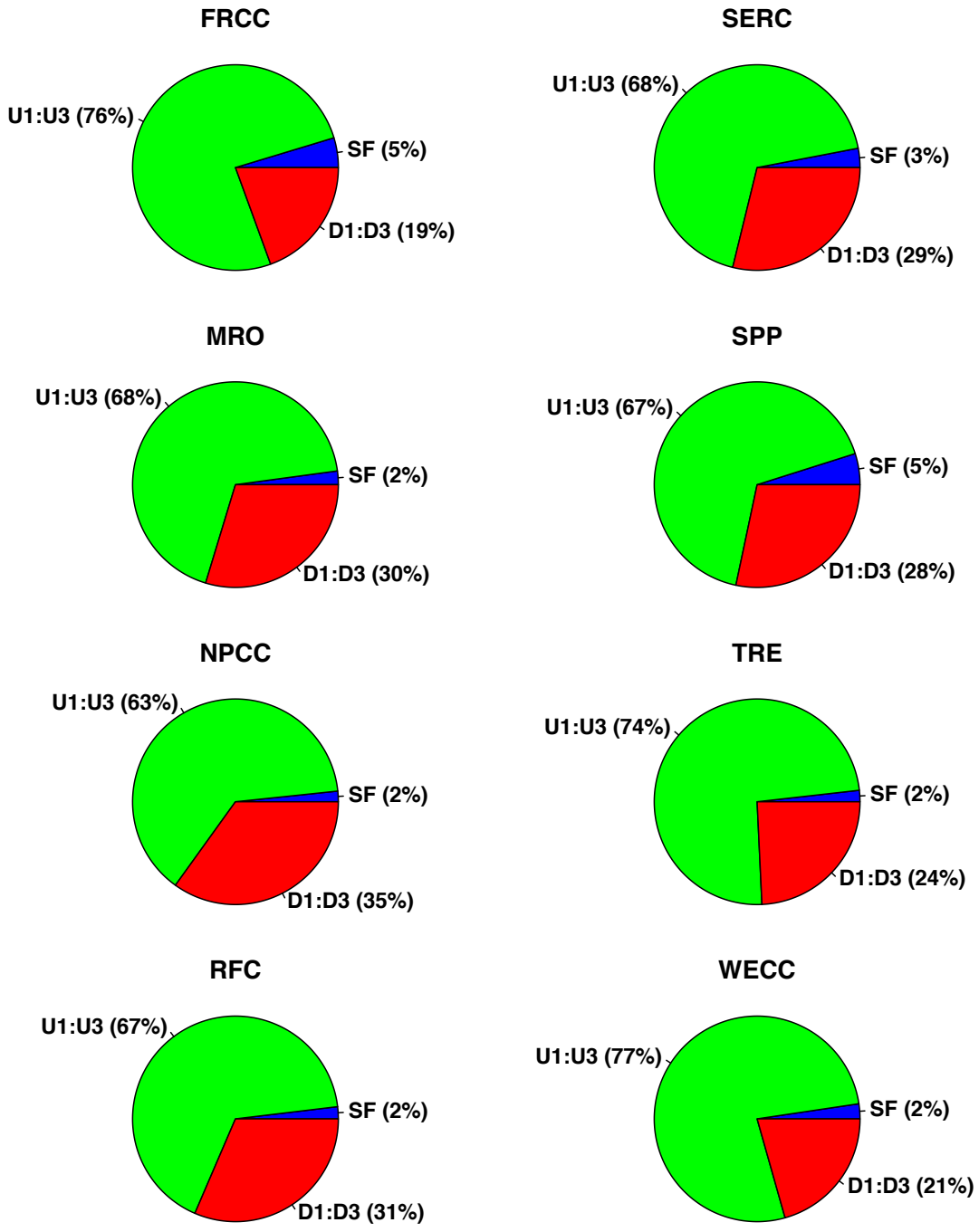

Figure A.3: Proportion of unscheduled unavailable capacity (MWh) by event type category.

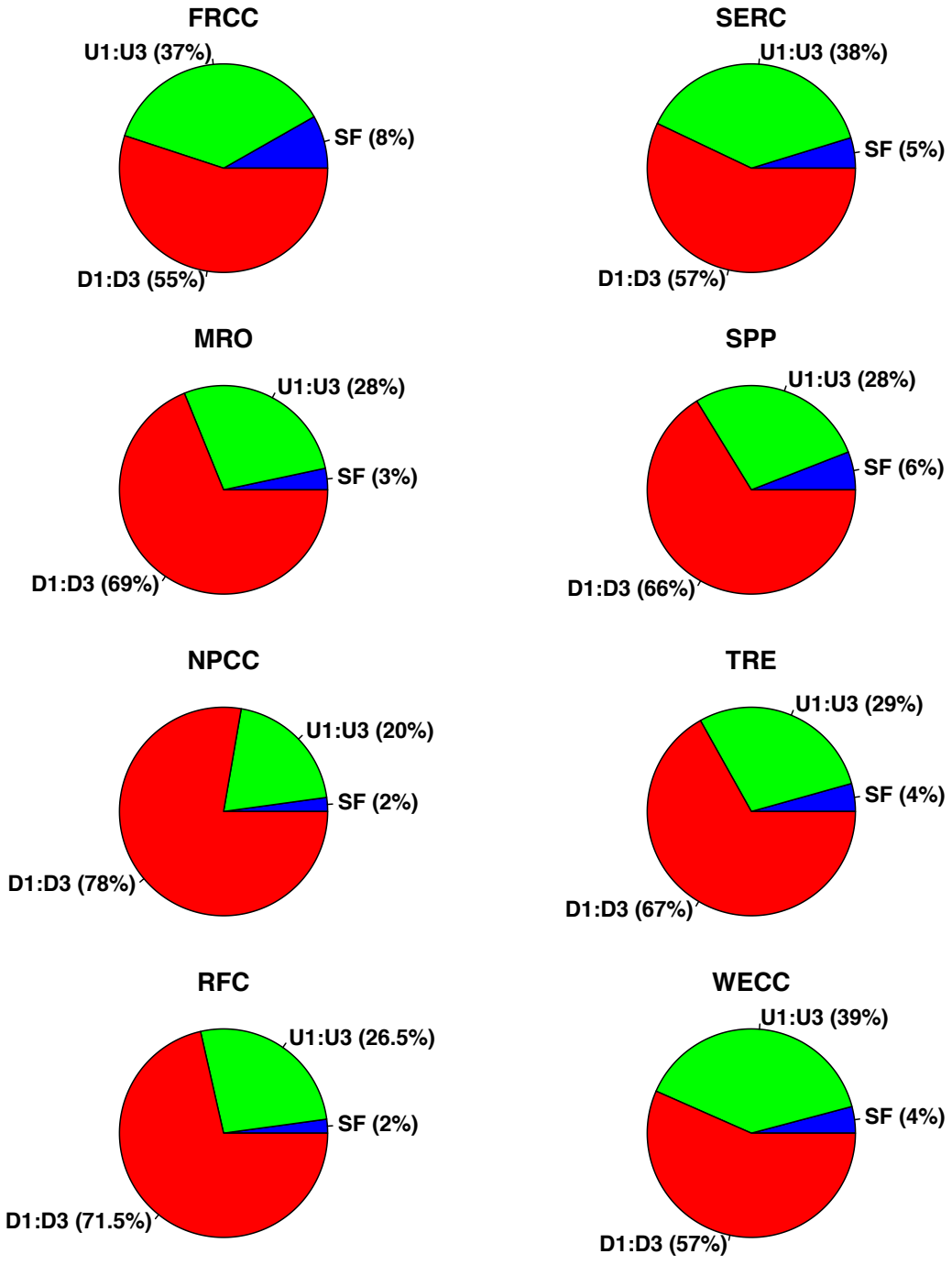

Figure A.4: Proportion of unscheduled event counts by event type category.

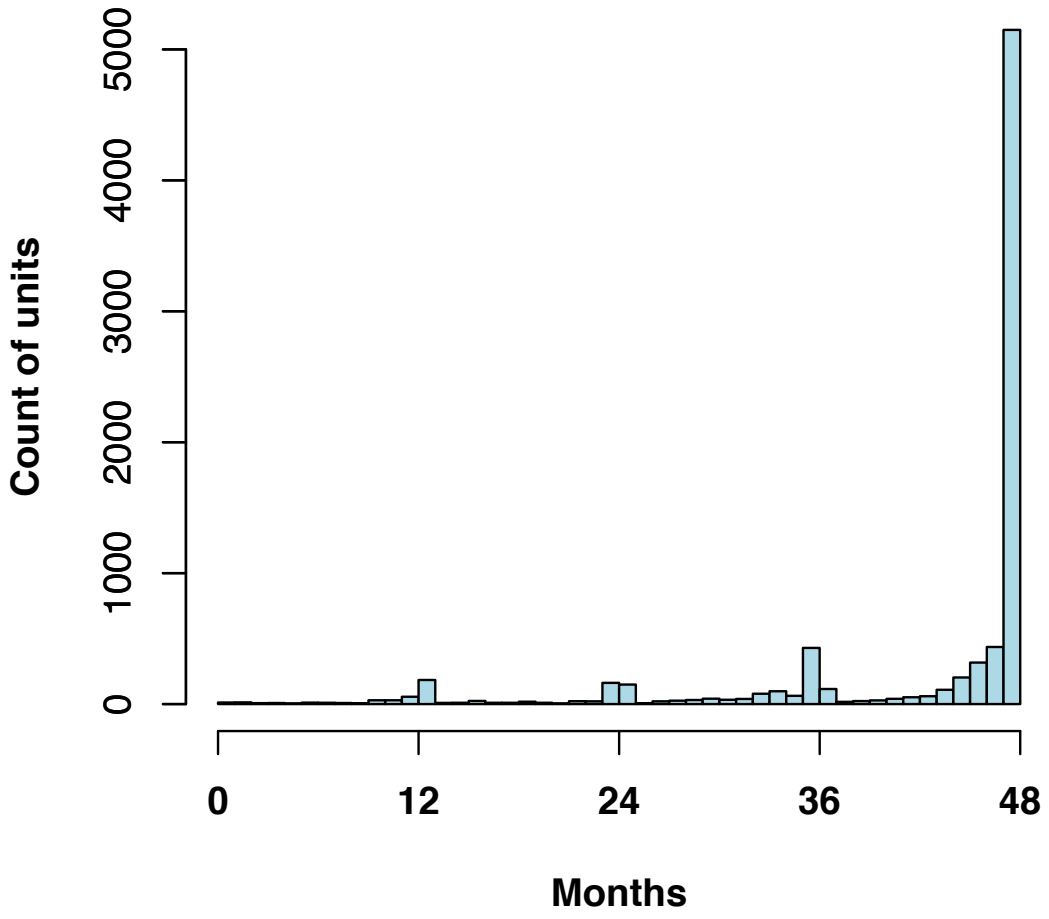

Figure A.5: Histogram of the number of months each unit was in operation (2012-2015).

In order to facilitate some forms of statistical analysis of the durations of unscheduled events, we show histograms of the duration of unscheduled events by event type in Figures A.6, A.7, and A.8 for the full four-year study period. In all regions the distributions are long-tailed: while the 90th percentile duration is less than 100 hours in each region, the maximum is always a full calendar year (i.e., either 8760 or 8784 hours). Events that span multiple calendar years are broken into calendar year components automatically in GADS, though this does not affect any of our primary results. We note that in NPCC the most common event duration is 24 hours—representing 60% of

all unscheduled events reported in the region. More than 99% of these 24-hour events are deratings. When derating magnitudes vary over time, NERC allows generating unit operators to either report one derating corresponding to the average availability reduction or to report separate deratings each time the unit's availability changes [28]. This suggests that generating unit operators in NPCC are more likely than in other regions to report separate deratings each time the unit's availability changes, and to do so on a daily basis. Again, this does not affect any of our primary results but it does illustrate a human element of GADS reporting.

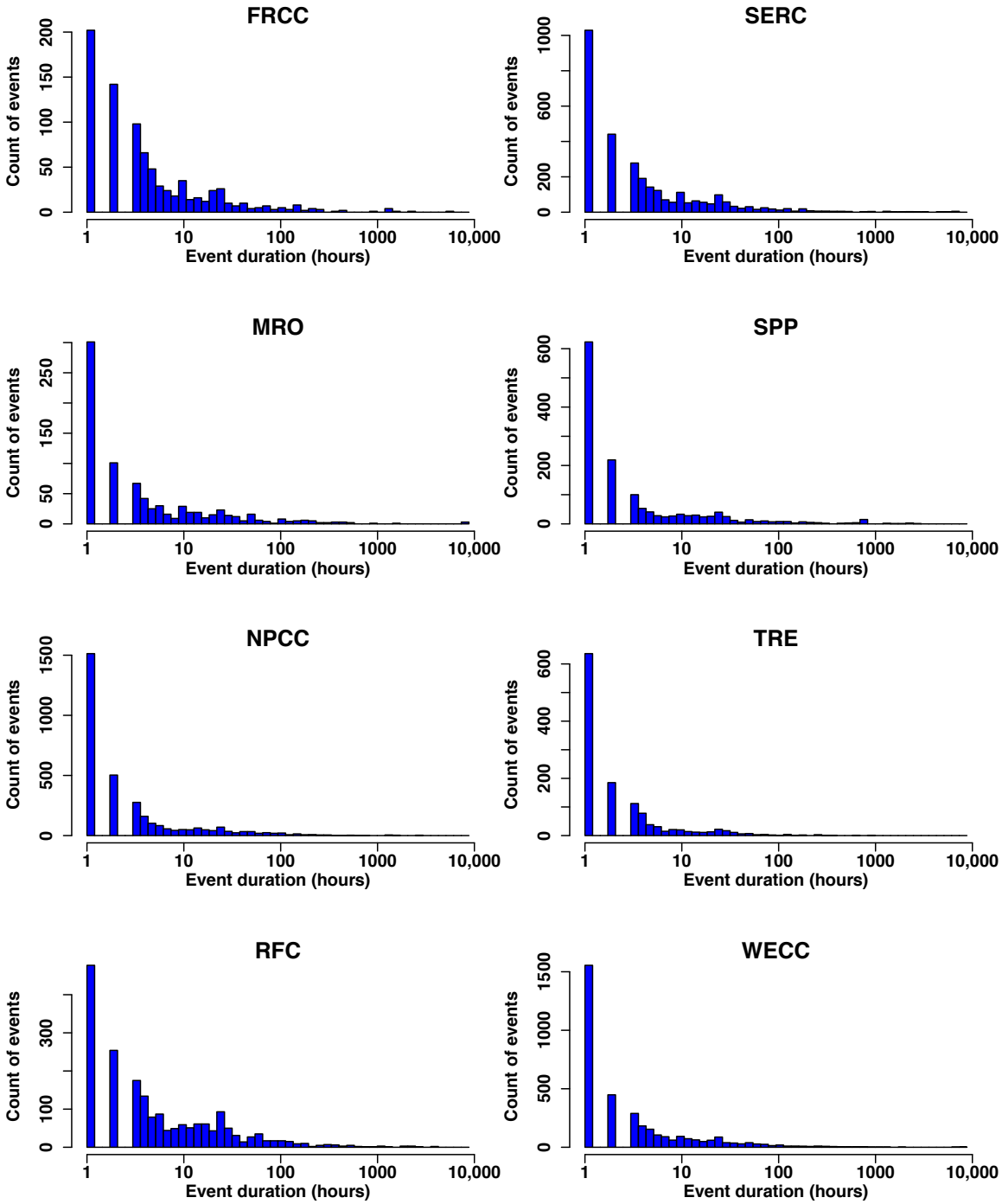

Figure A.6: Histogram of event durations for start-up failures. Full study period. Note the log scale.

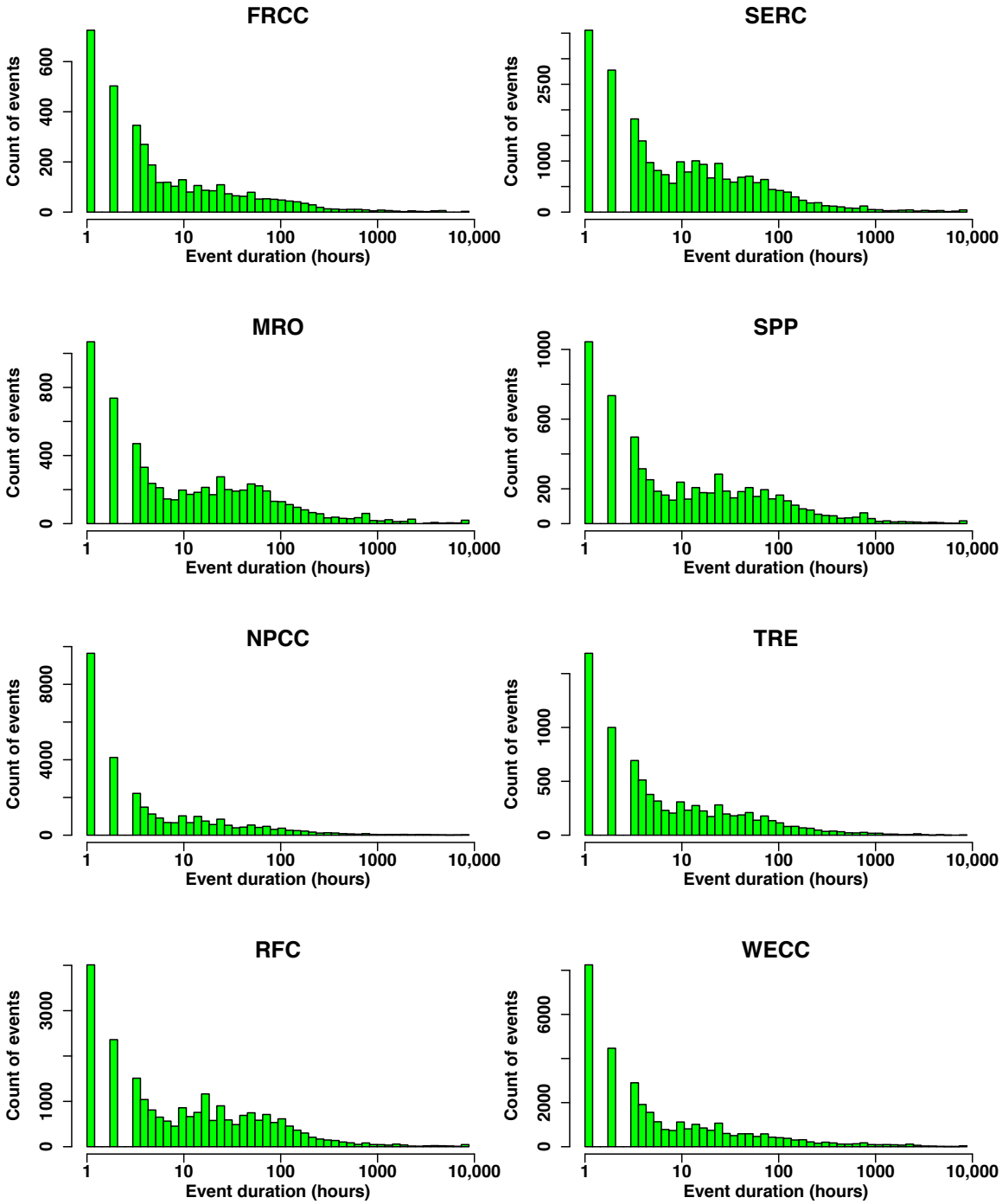

Figure A.7: Histogram of event durations for unscheduled outages. Full study period. Note the log scale.

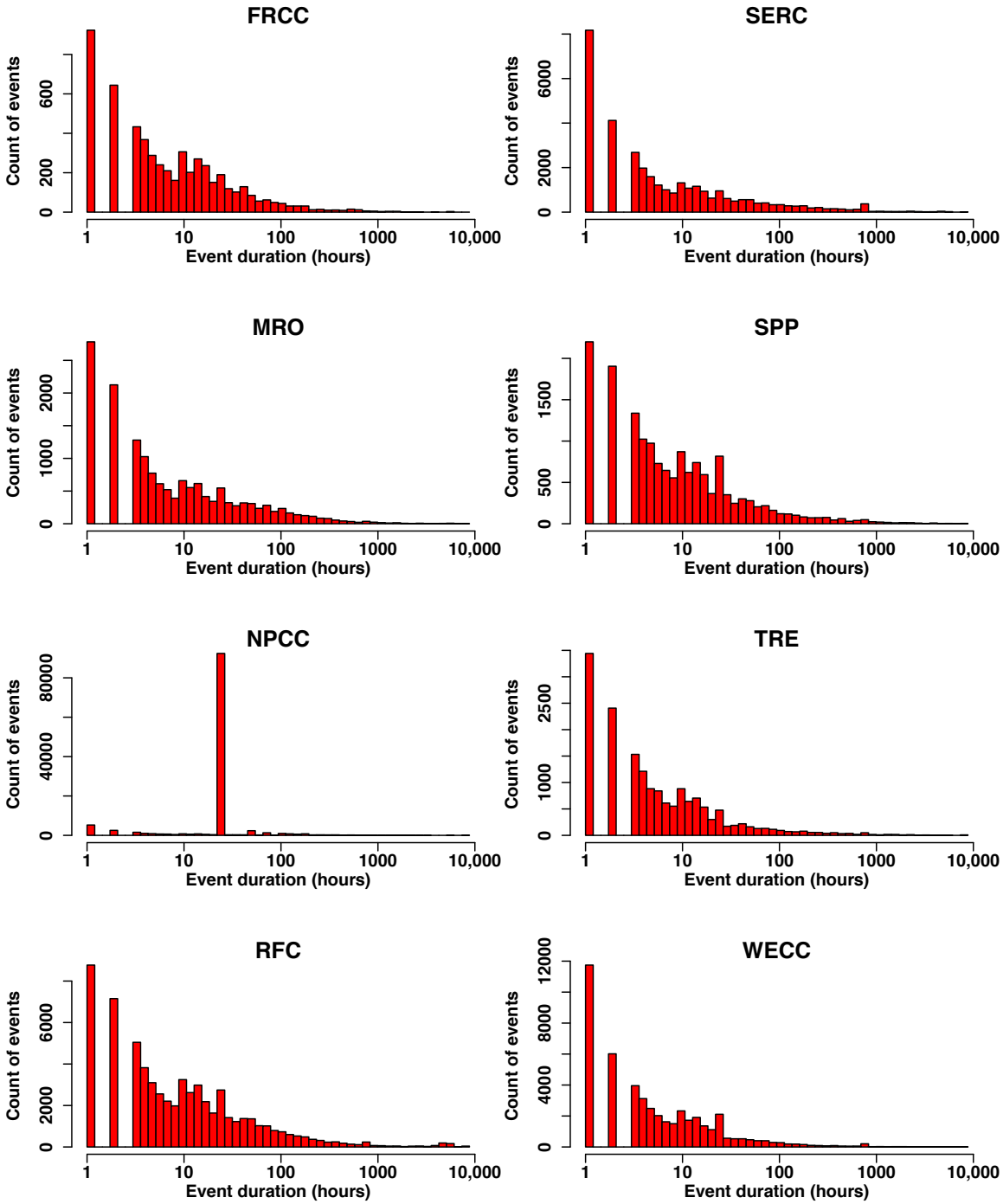

Figure A.8: Histogram of event durations for unscheduled deratings. Full study period. Note the log scale.

Table A.2: Parameters of lognormal distributions fit to each region's unscheduled event durations by event type. Standard errors in parentheses.

Region	Startup failures		Unscheduled outages		Unscheduled deratings	
	meanlog	**sdlog**	**meanlog**	**sdlog**	**meanlog**	**sdlog**
FRCC	1.47 (0.049)	1.42 (0.035)	1.93 (0.028)	1.70 (0.020)	1.89 (0.019)	1.43 (0.014)
MRO	1.38 (0.056)	1.58 (0.039)	2.45 (0.023)	1.89 (0.016)	1.96 (0.013)	1.60 (0.0090)
NPCC	1.04 (0.023)	1.35 (0.016)	1.65 (0.0097)	1.71 (0.0069)	3.01 (0.0028)	0.97 (0.0012)
RFC	1.69 (0.035)	1.53 (0.025)	2.39 (0.012)	1.83 (0.0084)	2.14 (0.0065)	1.63 (0.0046)
SERC	1.36 (0.027)	1.48 (0.019)	2.32 (0.011)	1.76 (0.0079)	1.86 (0.0094)	1.73 (0.0066)
SPP	1.22 (0.041)	1.57 (0.029)	2.45 (0.023)	1.90 (0.016)	2.03 (0.012)	1.50 (0.0084)
TRE	0.84 (0.032)	1.14 (0.023)	2.01 (0.018)	1.70 (0.013)	1.69 (0.011)	1.43 (0.0079)
WECC	1.14 (0.023)	1.38 (0.016)	1.93 (0.0099)	1.84 (0.0070)	1.61 (0.0065)	1.43 (0.0046)
All	1.24 (0.011)	1.44 (0.008)	2.07 (0.0048)	1.81 (0.0034)	2.31 (0.0026)	1.49 (0.0019)

Another human element of the GADS data is the time resolution of reported events. While GADS allows the precise start and end time of an event to be recorded, the observed data shows that nearly 40% of event starts and event ends are recorded on the hour. Event starts and ends are next most commonly logged on the half-hour, with significant representation in all other multiples of five minutes.

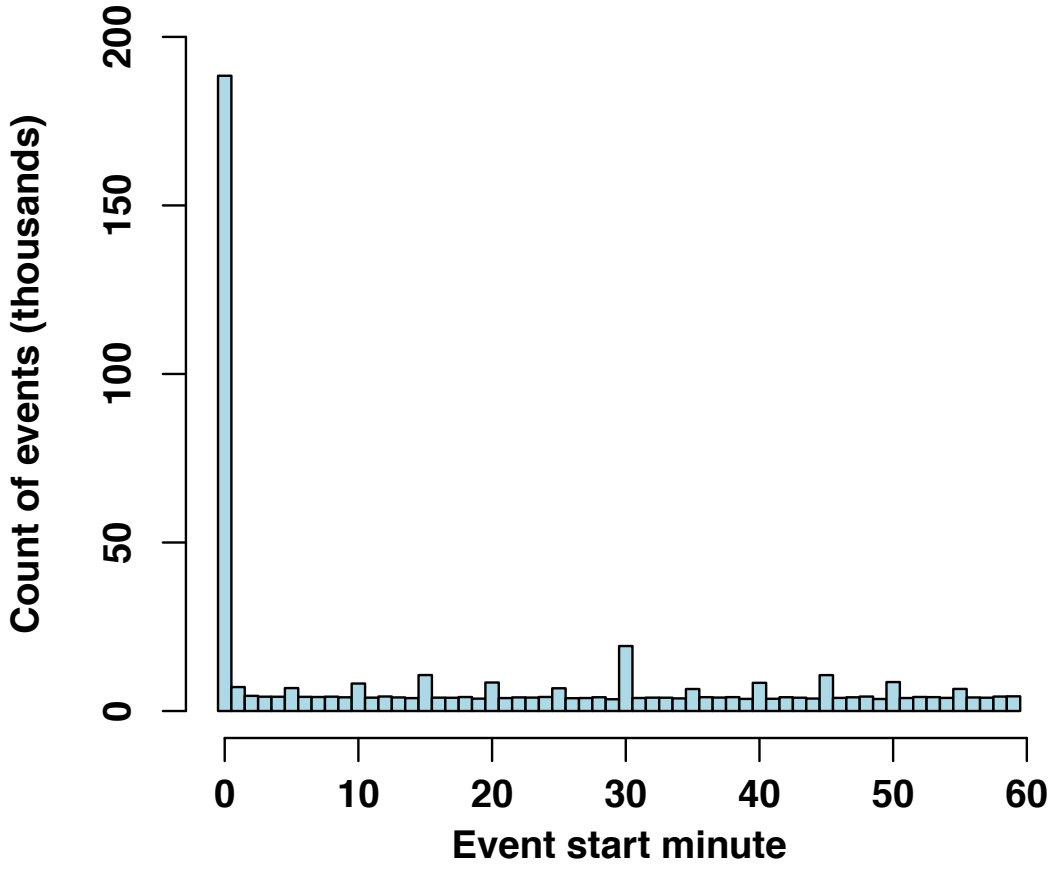

Figure A.9: Histogram of recorded start minute of unscheduled events.

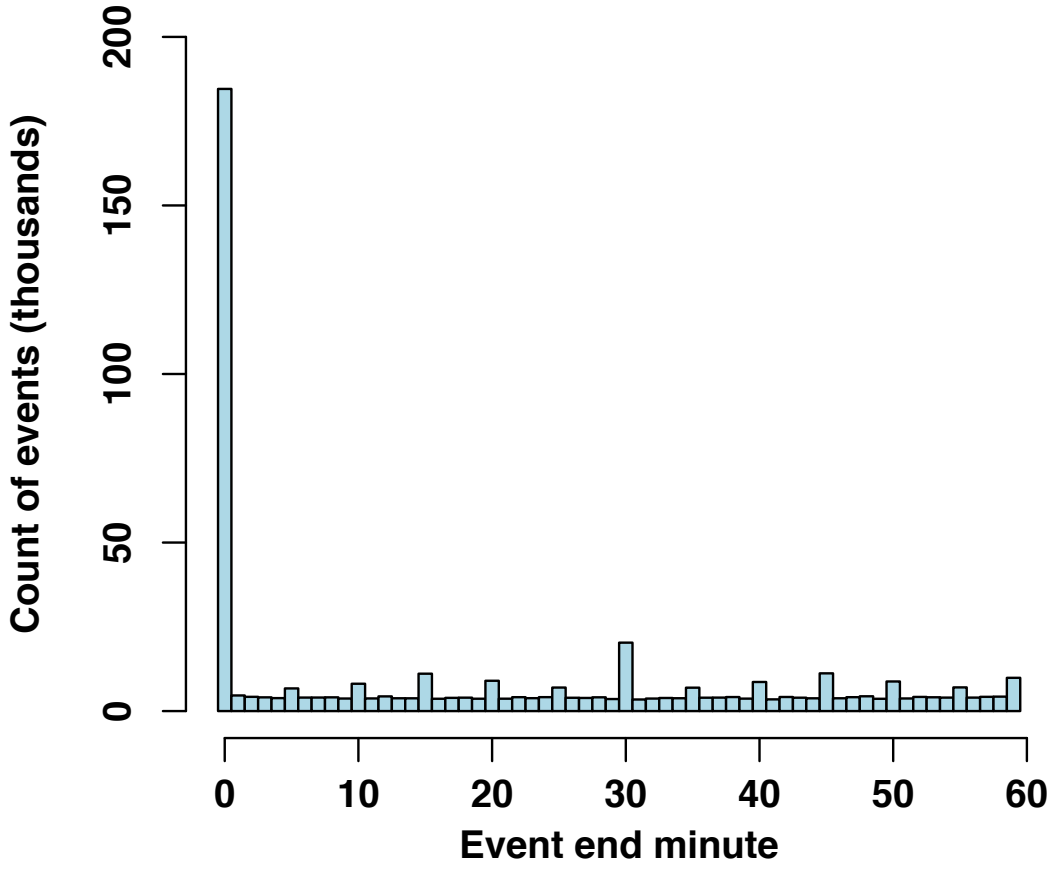

Figure A.10: Histogram of recorded end minute of unscheduled events.

Table A.3: Selected percentiles of the number of unscheduled events experienced during our 4-year study period in each NERC region.

Region	10^{th}	20^{th}	30^{th}	40^{th}	50^{th}	60^{th}	70^{th}	80^{th}	90^{th}
FRCC	3	5	7	9	13	19	25	32	43
MRO	2	4	6	9	11	16	22	38	99
NPCC	8	15	21	27	36	47	69	131	491
RFC	3	6	9	13	17	26	41	80	179
SERC	3	5	7	10	13	18	27	42	84
SPP	3	7	10	15	22	29	42	62	130
TRE	5	10	15	20	27	33	44	56	83
WECC	4	6	9	14	18	23	30	31	65

2.6.2 Tests for independence among generators

Block subsampling supplementary data

Figure A.11 shows the distribution of each region's empirical series of unscheduled unavailable capacity along with confidence intervals generated from 1000 block subsampling runs as exceedance curves for the full study period. Figure A.12 shows the same with the months containing Hurricane Sandy and January 2014 removed. We summarize the percentiles at which each region's empirical distribution exceeds the upper bound of the 99% confidence band, along with the maximum magnitude of exceedance, in the left-hand side of Table A.4. The right two columns show the same results when removing the periods that encompass Hurricane Sandy and the extreme cold month of January 2014. As can be seen from the figures and table, all regions except FRCC and WECC exceed the upper bound of the 99% confidence band at some point in each region's domain during the study period. When removing Hurricane Sandy and January 2014, MRO no longer exceeds the upper bound of its 99% confidence band. As a measure of whether the exceedances we observe in these six (five) regions represent a resource adequacy risk, we determine the amount of capacity needed to be procured in order to achieve the 1-in-10 loss of load expectation (LOLE) standard under the assumption of independent failures. Using the "one day in ten years" interpretation of this rule translates to 2.4 hours loss of load expectation per year, denoted 2.4 LOLH [2]. 2.4 LOLH is indicated via the dashed horizontal line in Figure A.11 and Figure A.12; the corresponding amount of capacity required at

95% and 99% confidence is indicated by the dashed vertical lines, drawn where the dashed horizontal line intersects the upper bound of each regions confidence bands. Using this approach, we conclude that managerially significant correlated failures are present in NPCC, RFC, SERC and TRE during the full study period, and only in NPCC, RFC, and TRE when removing Hurricane Sandy and January 2014.

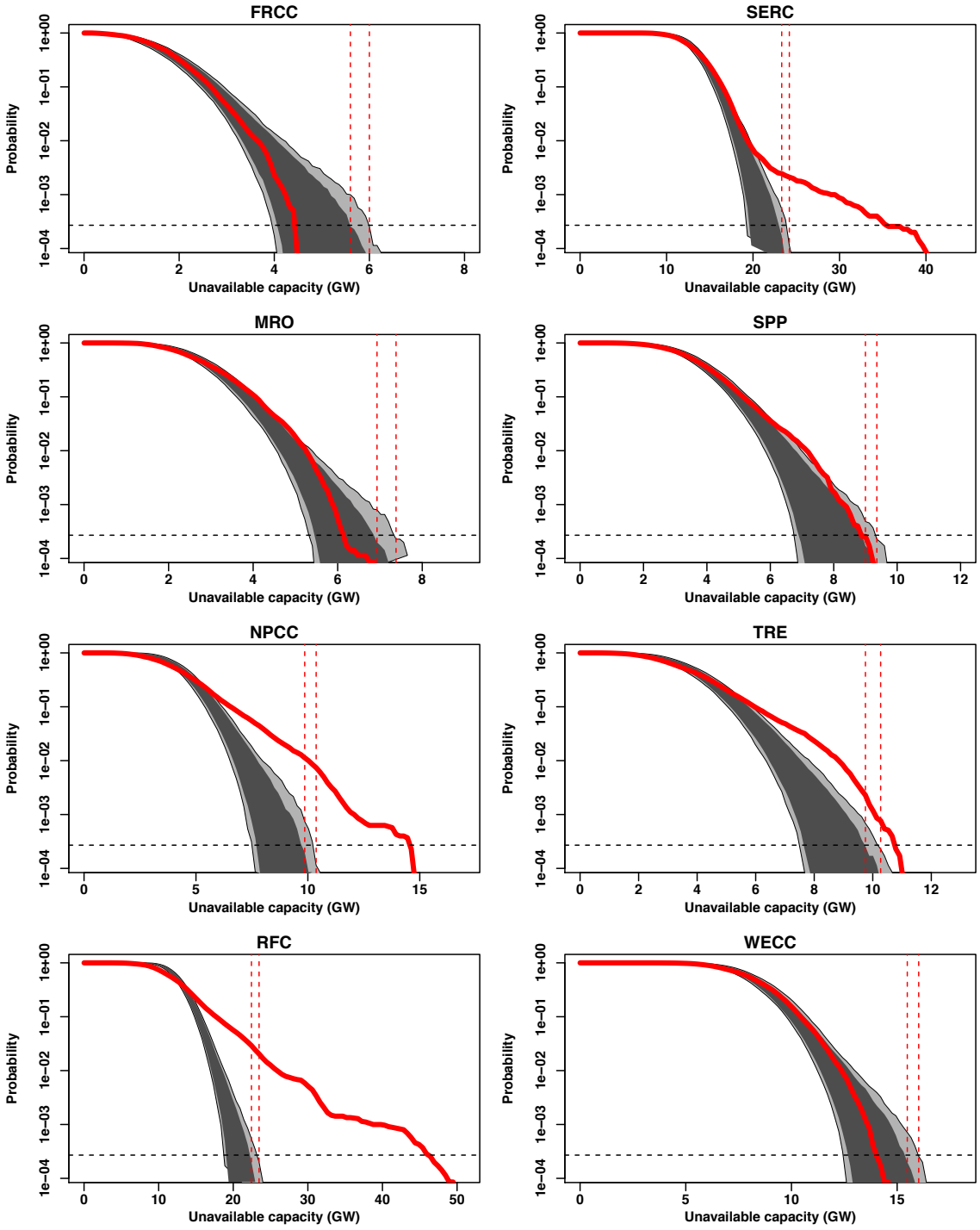

Figure A.11: 95% and 99% confidence bands from 1000 block subsampling runs (non-seasonal subsampling, full 2012-2015 period) in dark and light gray, respective; full empirical distributions in red. Dashed horizontal line indicates 2.4 loss-of-load-hours (LOLH) resource adequacy threshold; dashed vertical lines indicate intersection of 2.4 LOLH threshold with the upper bound of each confidence band.

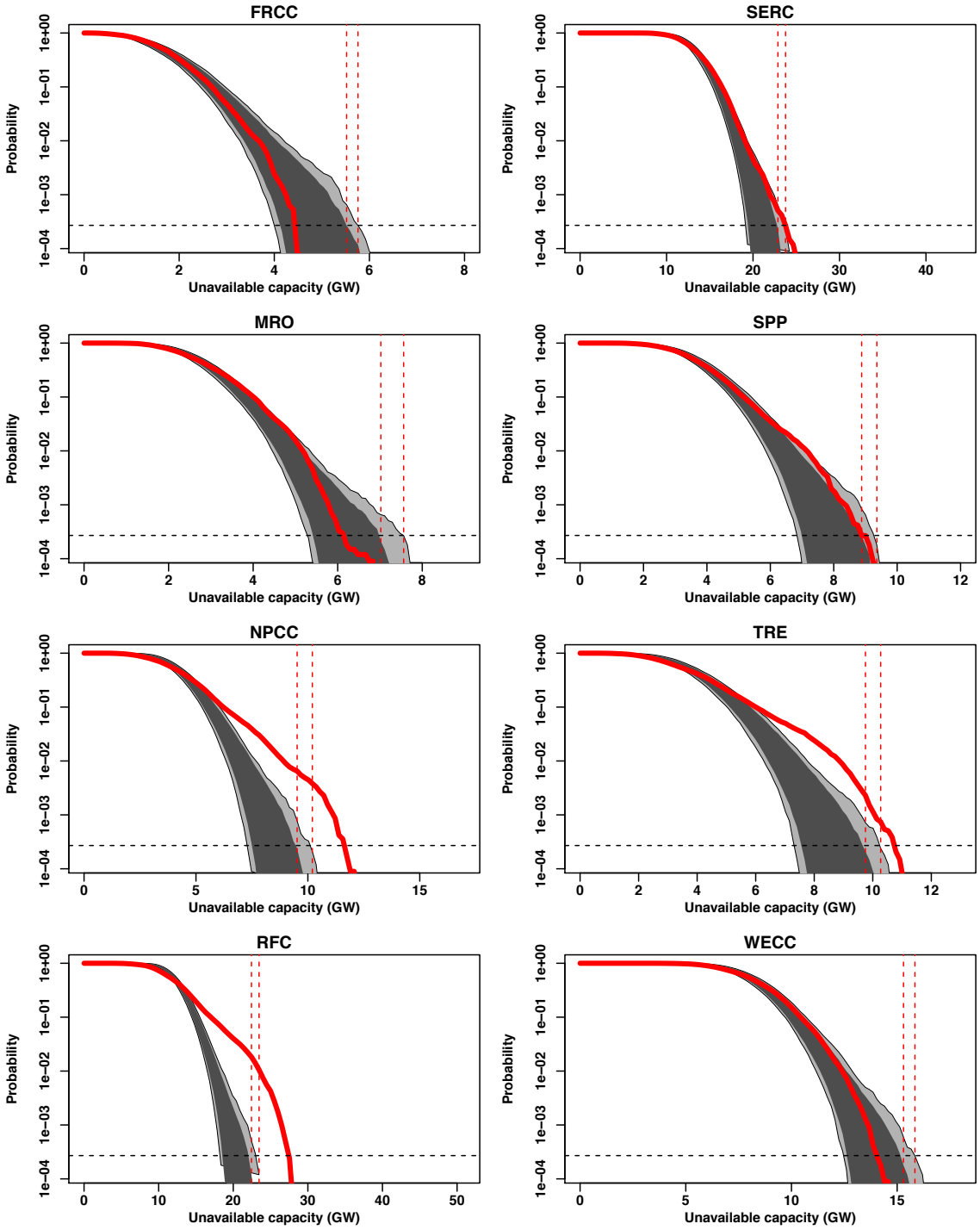

Figure A.12: 95% and 99% confidence bands from 1000 block subsampling runs (non-seasonal subsampling, Hurricane Sandy and January 2014 removed) in dark and light gray, respective; empirical distributions (excluding Hurricane Sandy and January 2014) in red. Dashed horizontal line indicates 2.4 loss-of-load-hours (LOLH) resource adequacy threshold; dashed vertical lines indicate intersection of 2.4 LOLH threshold with the upper bound of each confidence band.

Table A.4: Summarizing percentiles for which empirical distribution exceeds block subsampling confidence band and magnitude of exceedance via block subsampling, for both the full study period (left) and when removing Hurricane Sandy and January 2014 (right).

Region	**Full period**		**Excluding Sandy and Jan 2014**	
	Percentiles	**Max divergence**	**Percentiles**	**Max divergence**
FRCC	–	–	–	–
MRO	97	2%	–	–
NPCC	79-100	47%	79-100	28%
RFC	73-100	115%	70-100	29%
SERC	100	73%	100	14%
SPP	98-99	5%	99	4%
TRE	87-100	17%	89-100	16%
WECC	–	–	–	–

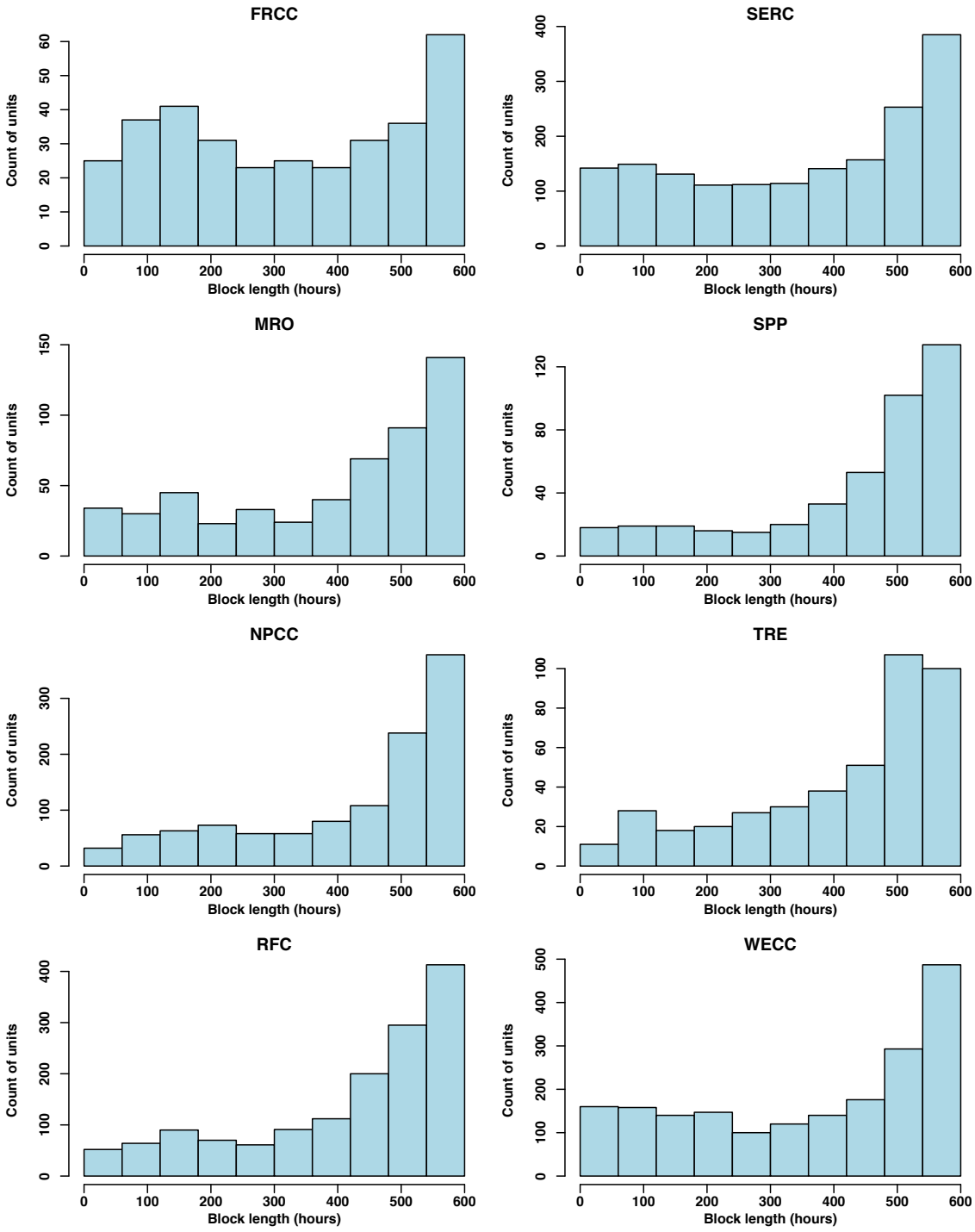

Figure A.13: Histograms of optimal block lengths used in the block subsampling analysis.

Binomial event arrivals supplementary data

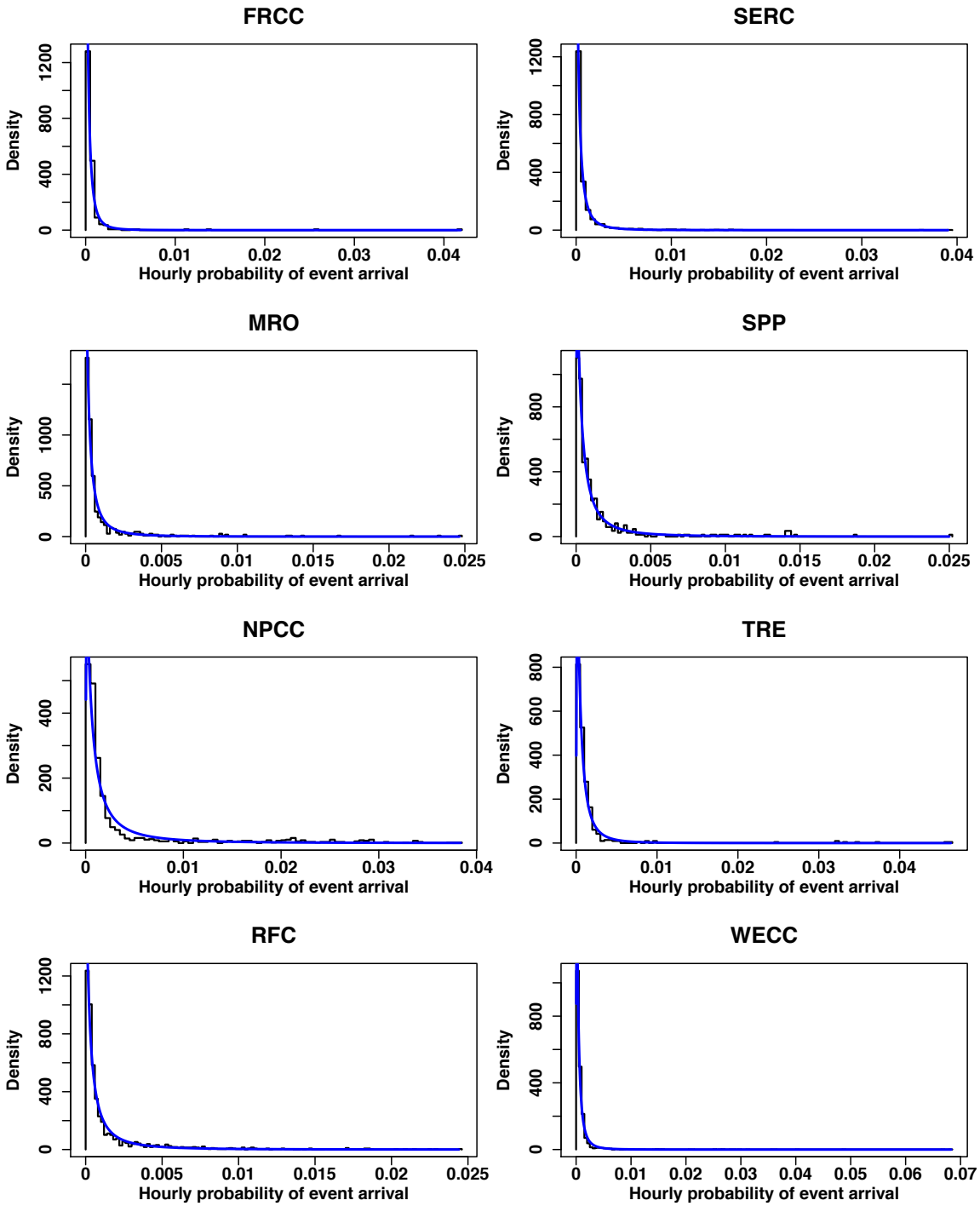

Figure A.14: Histograms of event arrival probabilities by region with fitted lognormal distribution overlaid. Probabilities estimated via Equation 2-3 in the main text. Full study period. Excludes the nine units that are not at least partially available for 1000 or more hours during the study period.

Table A.5: Parameters of lognormal fits to distribution of unscheduled event arrival probabilities, by region. Full period. Standard errors in parentheses.

Region	meanlog	sdlog
FRCC	-8.10 (0.065)	1.19 (0.046)
MRO	-7.94 (0.061)	1.40 (0.043)
NPCC	-6.77 (0.044)	1.49 (0.031)
RFC	-7.53 (0.040)	1.52 (0.028)
SERC	-7.93 (0.032)	1.33 (0.23)
SPP	-7.50 (0.069)	1.42 (0.049)
TRE	-7.37 (0.057)	1.19 (0.041)
WECC	-7.73 (0.027)	1.20 (0.019)
Combined	-7.60 (0.016)	1.41 (0.011)

A supplemental statistical test for failure correlation

We test for positive correlation for every possible pair of units in each region using a t-test as described in [46]. For each pairwise comparison, we compare the test statistic to the critical value for a one-sided 5% test (1.645) and then record the percent of test statistics that are statistically significant in each region. For this test, each unit's series has been normalized by its nameplate capacity. Because each unit's series of arrivals is used for many pairwise comparisons, we establish the rejection rate under the null hypothesis for each region through simulation. We use the same simulated series of normalized unscheduled unavailable capacity generated for the binomial simulation analysis presented in the main text. Here, for each iteration of the simulation, we compute a t-test for all possible pairs of generating units in each region and record the percent of test statistics that are statistically significant. We repeat this process 1000 times and determine the 95th and 99th percentile of rejection rates for each region. These are the thresholds of significance we use to say whether each region demonstrates correlated failures under this method. We present results for the full study period in Table A.6 and when removing Hurricane Sandy and January 2014 in Table A.7. Using this method we find evidence of correlated failures in all eight regions for the full study period and when removing the two months of data corresponding to Hurricane Sandy and January 2014, at both the 95% and 99% confidence levels. However we emphasize that as this test does not weight generating units by their size, statistical significance need not imply meaningful resource adequacy

risk. Rather, we refer the reader to our conclusions in Section 2.4 in the main text, where we found meaningful violations of the independence assumption only in four regions: NPCC, RFC, SERC, and TRE.

Table A.6: Pairwise binomial test results for empirical and simulated series by region for the full period.

	FRCC	**MRO**	**NPCC**	**RFC**	**SERC**	**SPP**	**TRE**	**WECC**
Empirical	13.3%	20.5%	29.0%	30.7%	18.8%	25.0%	20.3%	17.7%
99% simulation	12.1%	17.4%	23.9%	21.8%	15.5%	22.5%	19.1%	16.1%
95% simulation	11.8%	17.2%	23.7%	21.7%	15.4%	22.2%	18.8%	15.9%

Table A.7: Pairwise binomial test results for empirical and simulated series by region when excluding Hurricane Sandy and January 2014.

	FRCC	**MRO**	**NPCC**	**RFC**	**SERC**	**SPP**	**TRE**	**WECC**
Empirical	13.2%	19.7%	28.3%	27.0%	17.4%	24.7%	20.0%	17.6%
99% simulation	12.0%	17.0%	23.5%	21.0%	15.1%	22.5%	18.9%	15.9%
95% simulation	11.8%	16.8%	23.3%	20.8%	15.0%	22.2%	18.7%	15.8%

2.6.3 Seasonality

When do correlated failures occur?

To determine whether correlated failures are restricted to particular seasons, we generate exceedance curves for each NERC season and overlay 95% and 99% confidence bands generated via block subsampling (Figures A.15 through A.20). When removing Hurricane Sandy from fall and January 2014 from winter, there are six seasons in total. We find evidence of correlated failures in all seasons: four regions (NPCC, RFC, SERC, and SPP) in each winter definition, four regions (MRO, RFC, SERC, and TRE) in spring, three regions (NPCC, RFC, and TRE) in summer, five regions (MRO, NPCC, RFC, SPP, and TRE) in the full fall definition, and three regions (RFC, SPP, and TRE) in the shortened fall definition. We conclude that violations of the independent failures assumption can likely occur in any season in any region.

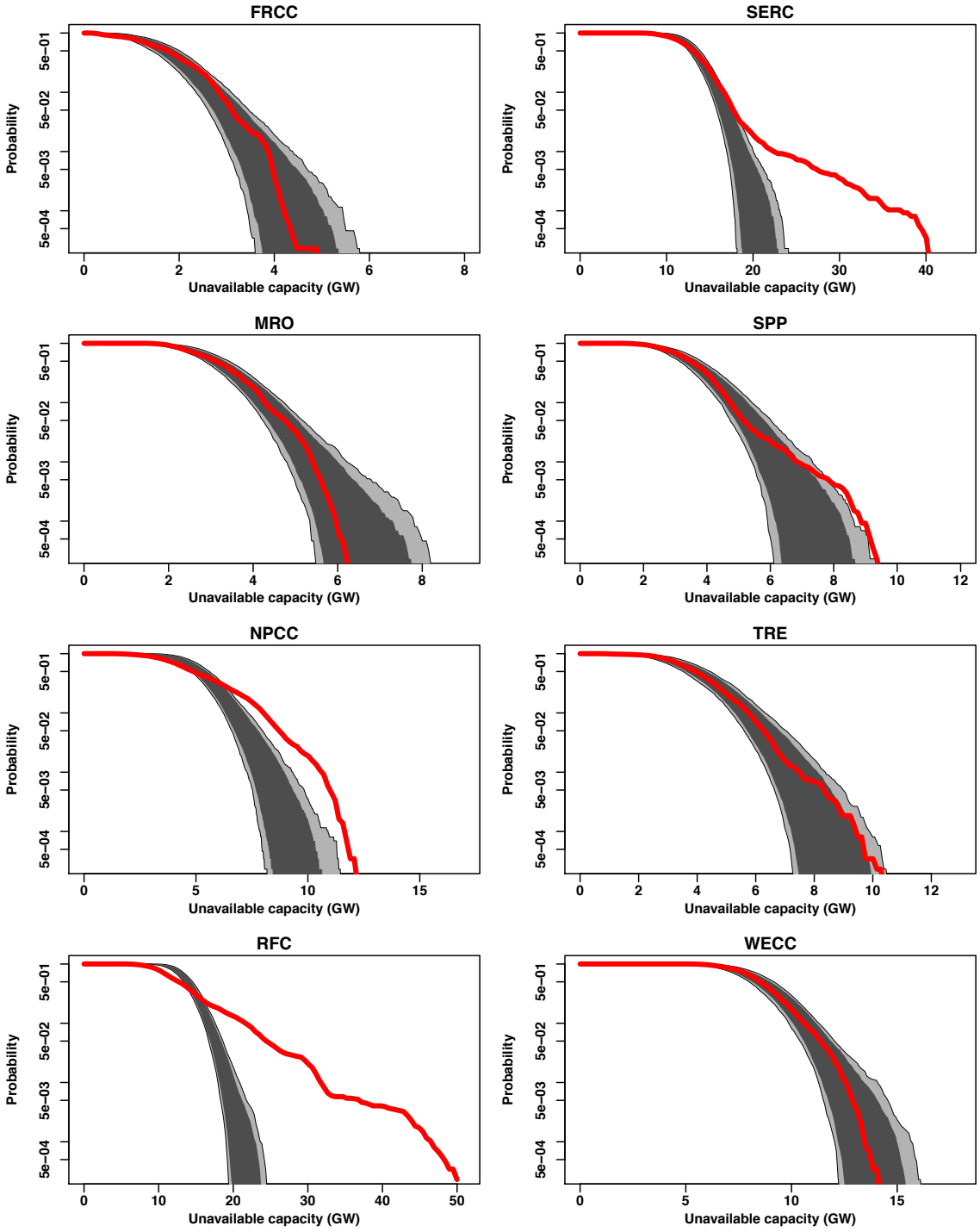

Figure A.15: 95% and 99% confidence bands from 1000 block subsampling runs for the full winter period (December, January, and February 2012-2015) in black, empirical distributions for only winter months in red.

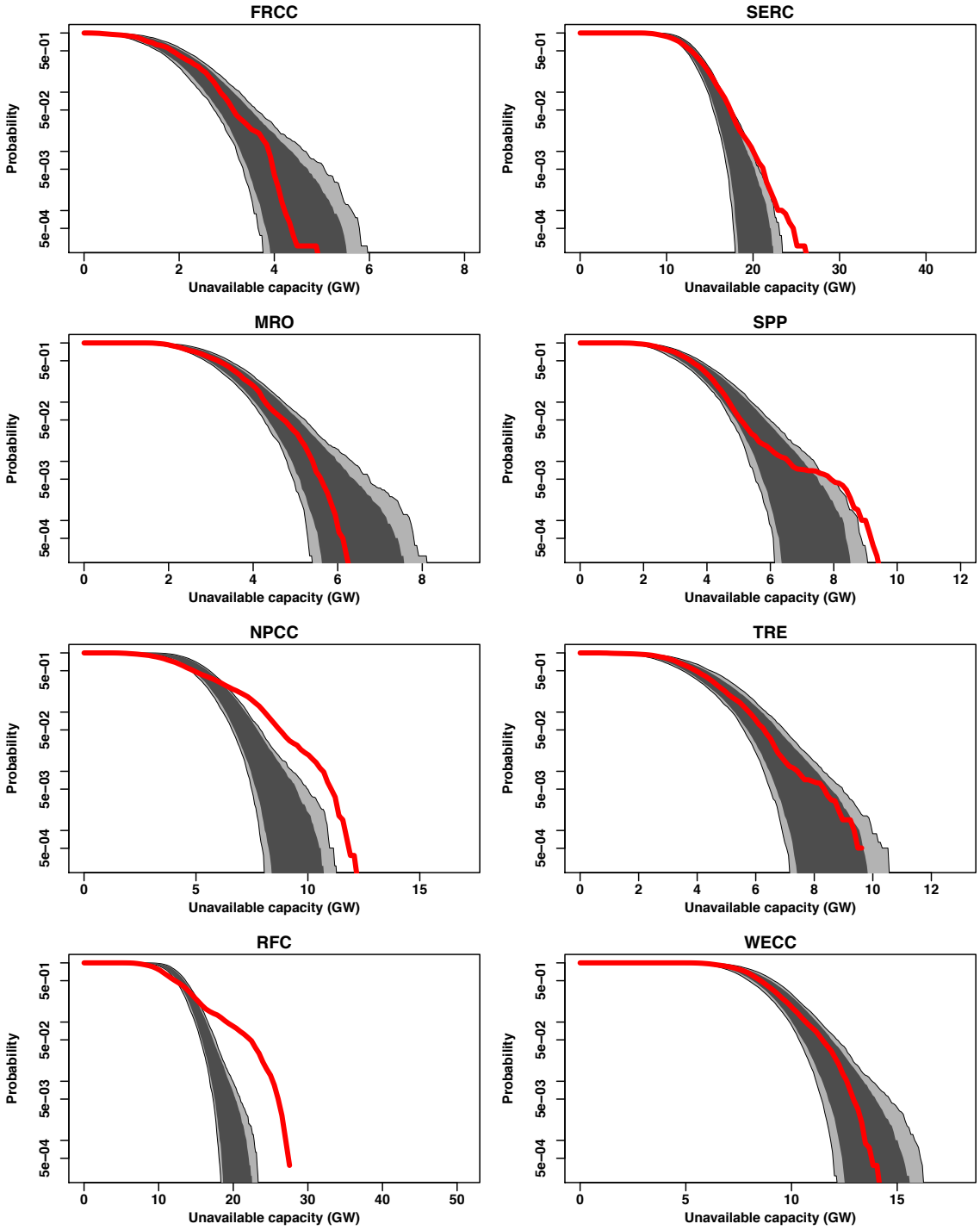

Figure A.16: 95% and 99% confidence bands from 1000 block subsampling runs for the winter period except January 2014 in black, empirical distributions for corresponding winter months in red.

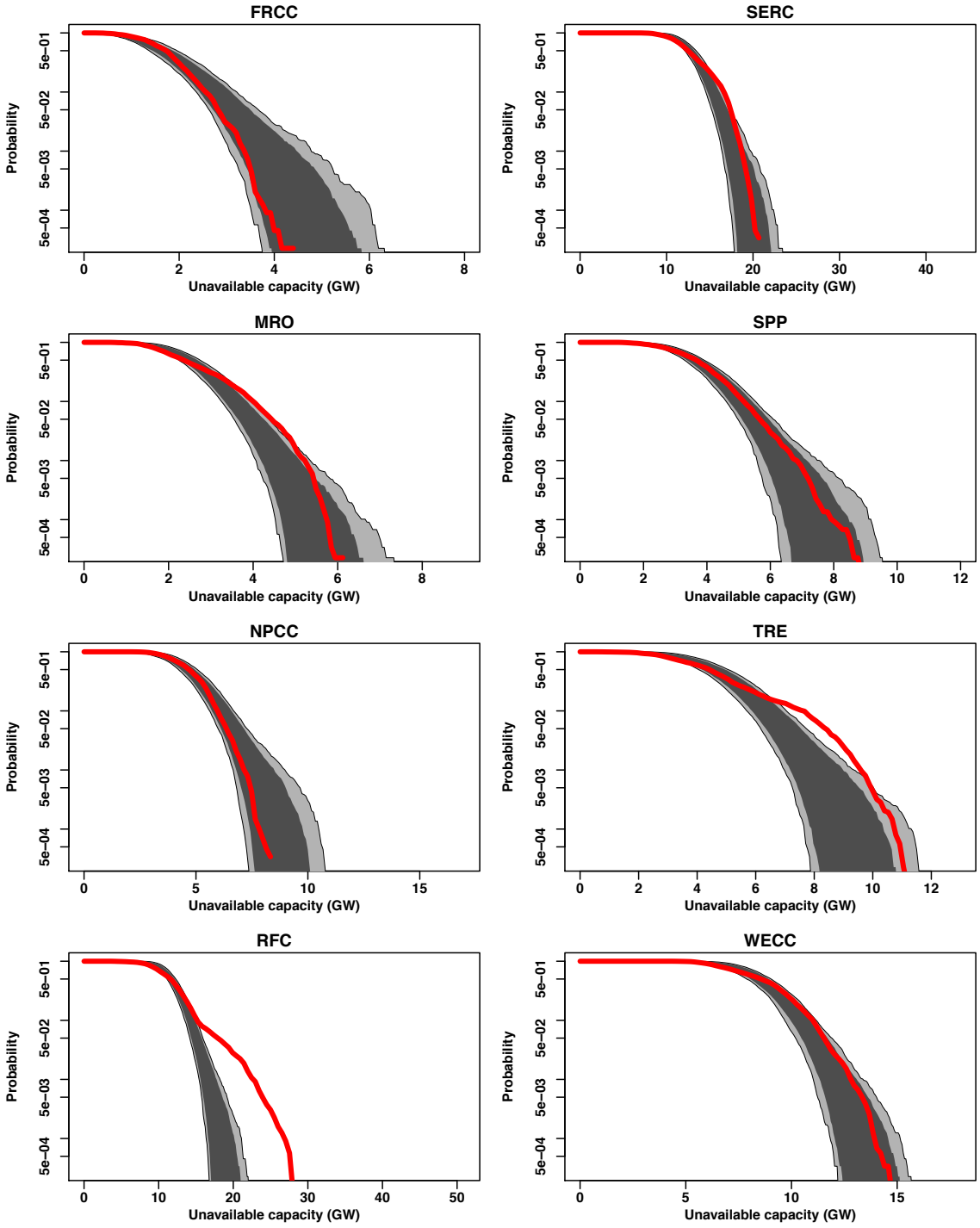

Figure A.17: 95% and 99% confidence bands from 1000 block subsampling runs for the spring period (March, April, and May 2012-2015) in black, empirical distributions for corresponding months in red.

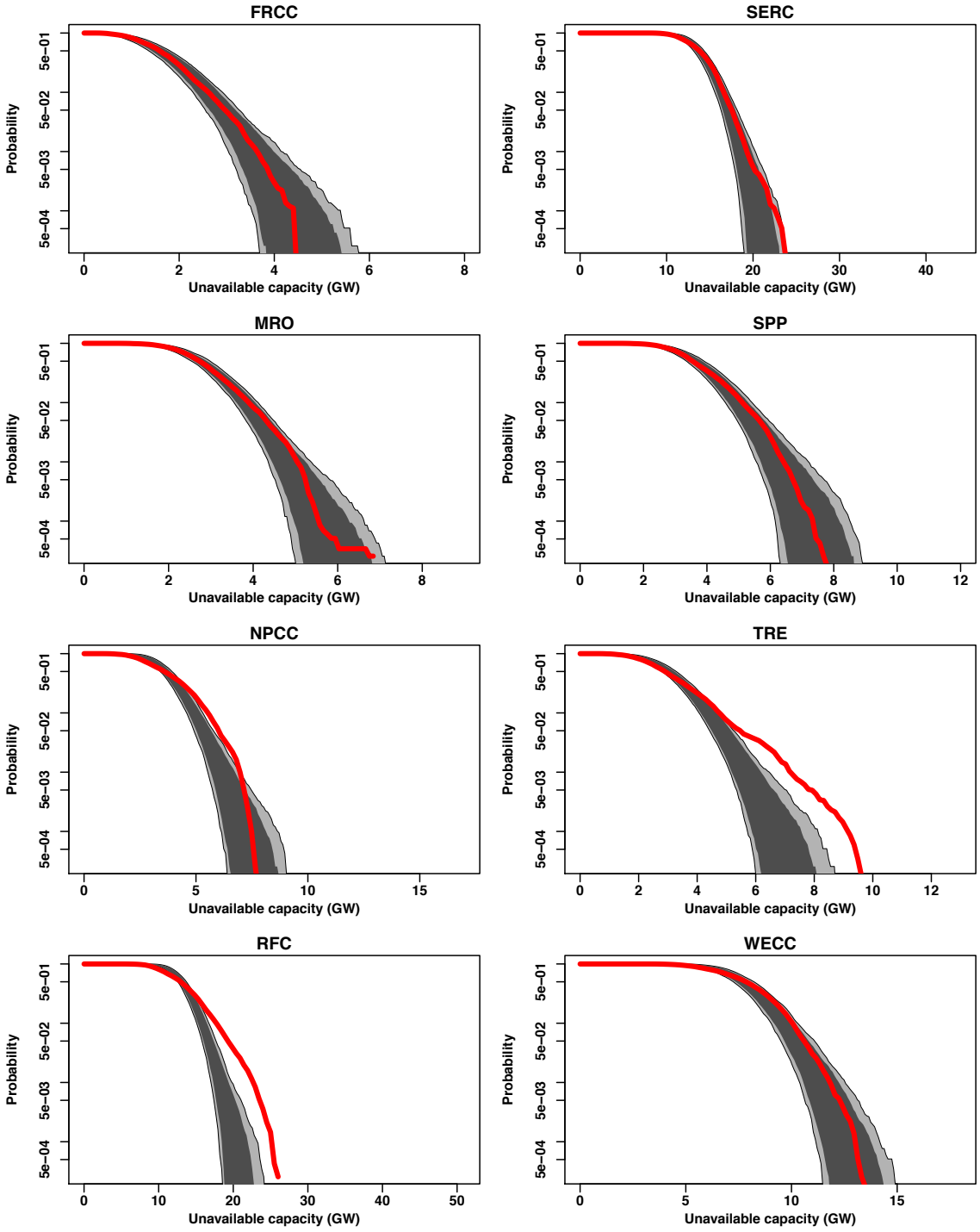

Figure A.18: 95% and 99% confidence bands from 1000 block subsampling runs for the summer period (June, July, August, and September 2012-2015) in black, empirical distributions for corresponding months in red.

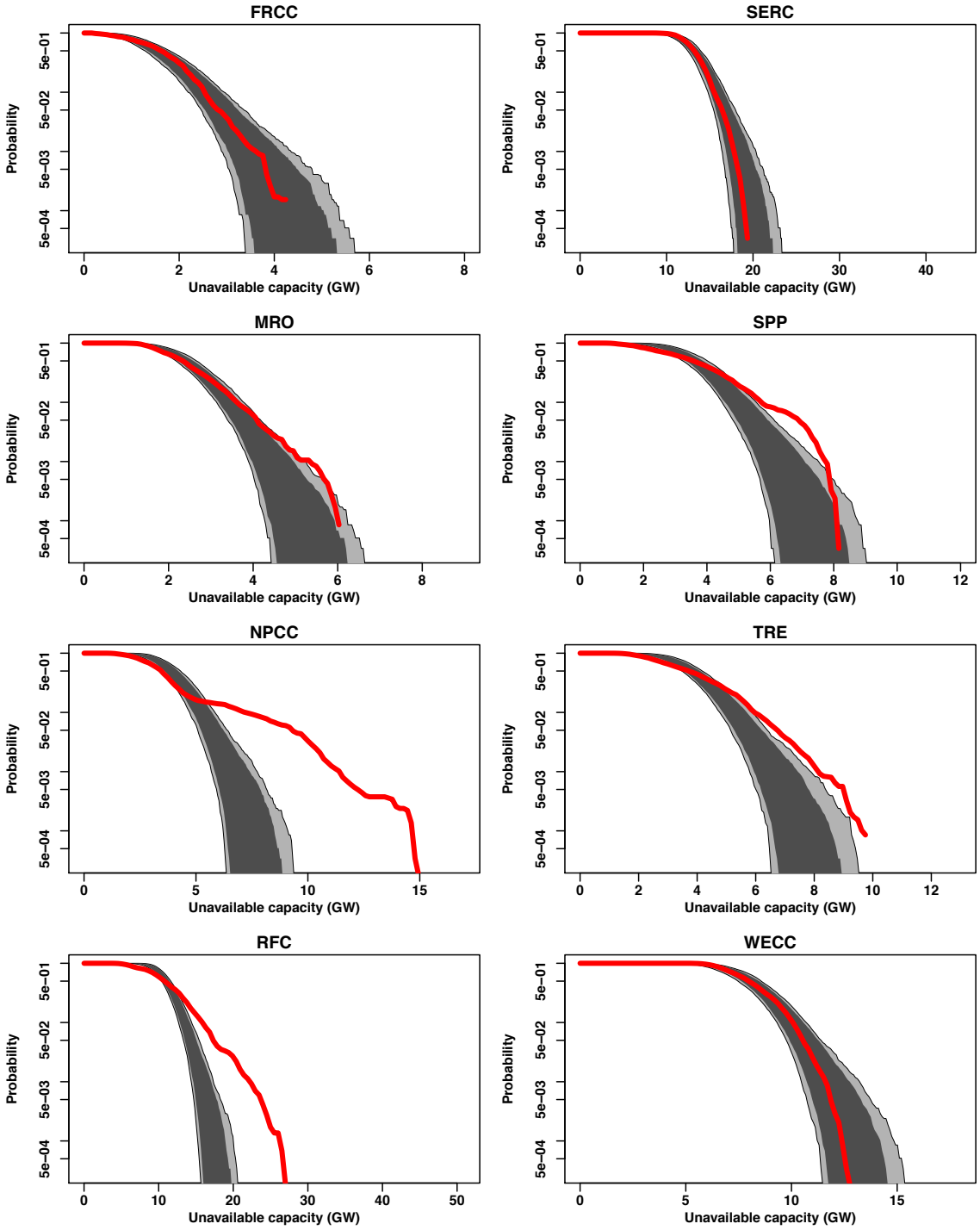

Figure A.19: 95% and 99% confidence bands from 1000 block subsampling runs for the fall period (October and November 2012-2015) in black, empirical distributions for corresponding months in red.

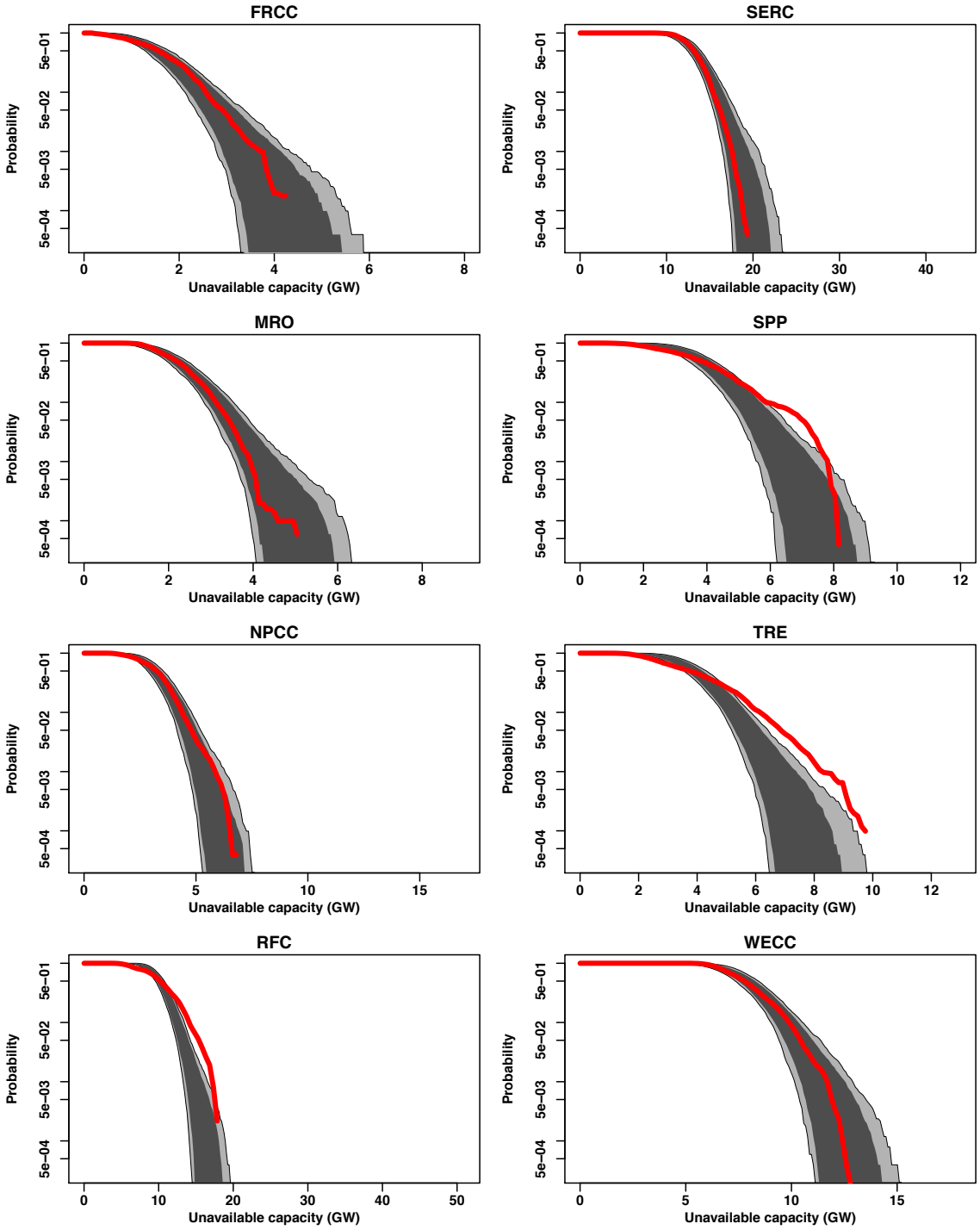

Figure A.20: 95% and 99% confidence bands from 1000 block subsampling runs for the fall period except Hurricane Sandy (October and November 2012-2015 except for October 29-November 30, 2012) in black, empirical distributions for corresponding months in red.

Supplementary data for seasonality in average unavailable capacity

To test for recurrent patterns in average unscheduled unavailable capacity, we compute monthly autocorrelation functions of unscheduled unavailable capacity by region (Figure A.21). We generate these by computing average unscheduled unavailable capacity in each month for each region. Significant seasonality would manifest as a lag-12 peak (corresponding to a one-year lag) that exceeds the 95% confidence bands. Except for FRCC, we see that each region's lag-12 peak is not significant. As a sensitivity analysis, we repeat this analysis by season (Figure A.22). Here significant seasonality would manifest as a lag-4 peak (again corresponding to a one-year lag) that exceeds the 95% confidence bands. Except for FRCC, we see that each region's lag-4 peak is not significant.

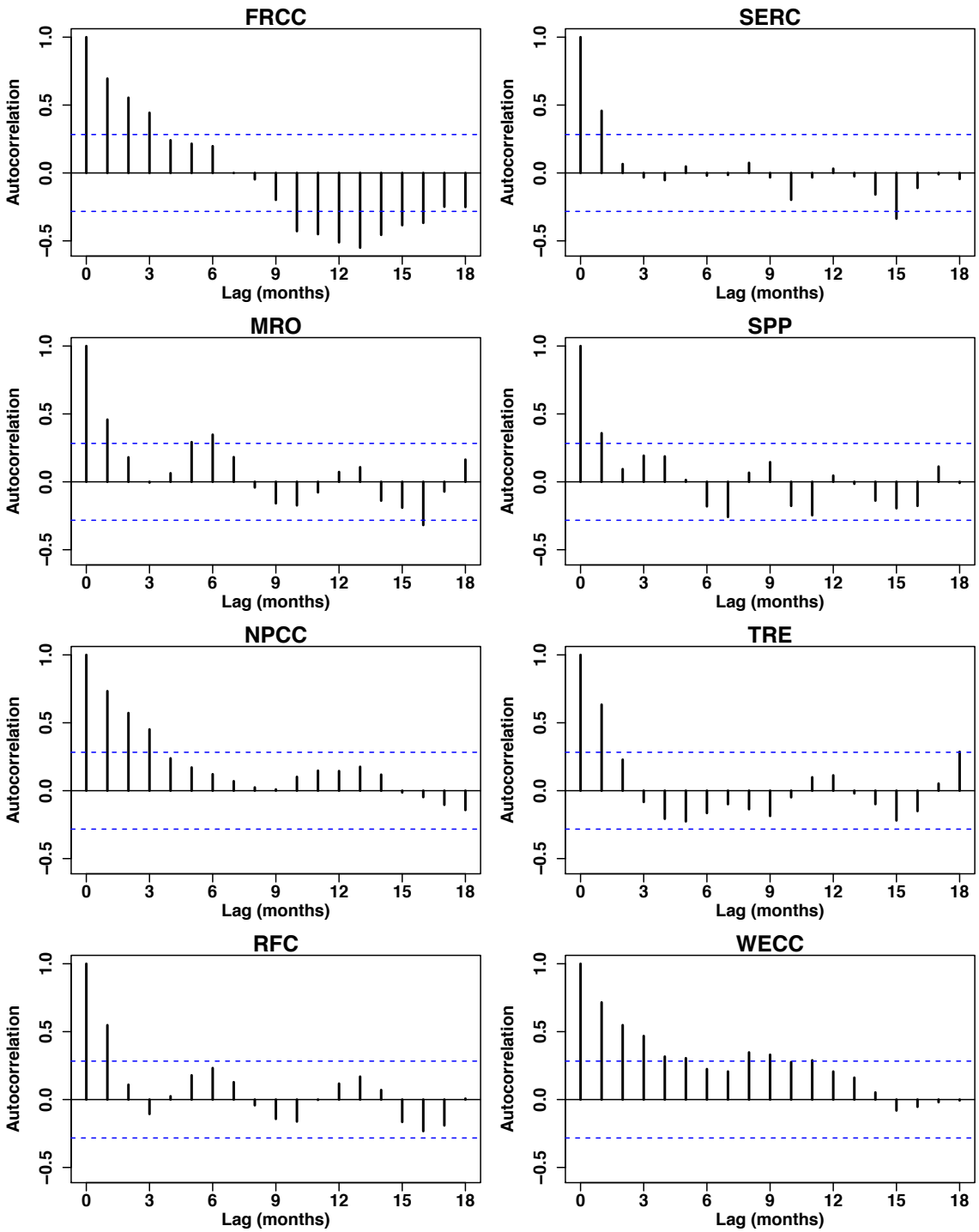

Figure A.21: Autocorrelation functions of monthly average unscheduled unavailable capacity by region with 95% confidence interval for assessing statistical significance of lags (dashed blue lines).

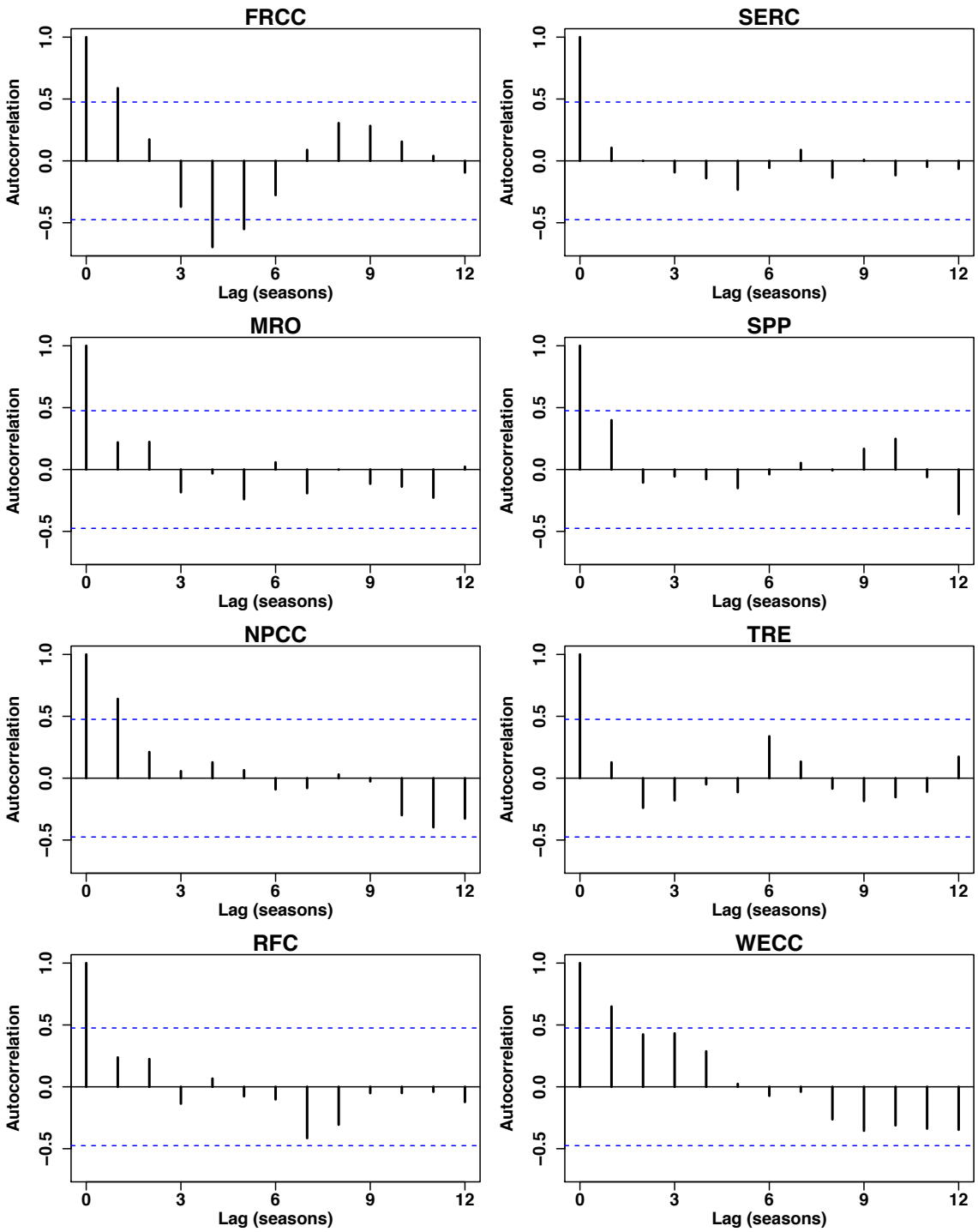

Figure A.22: Autocorrelation functions of seasonal average unscheduled unavailable capacity by region with 95% confidence interval for assessing statistical significance of lags (dashed blue lines).

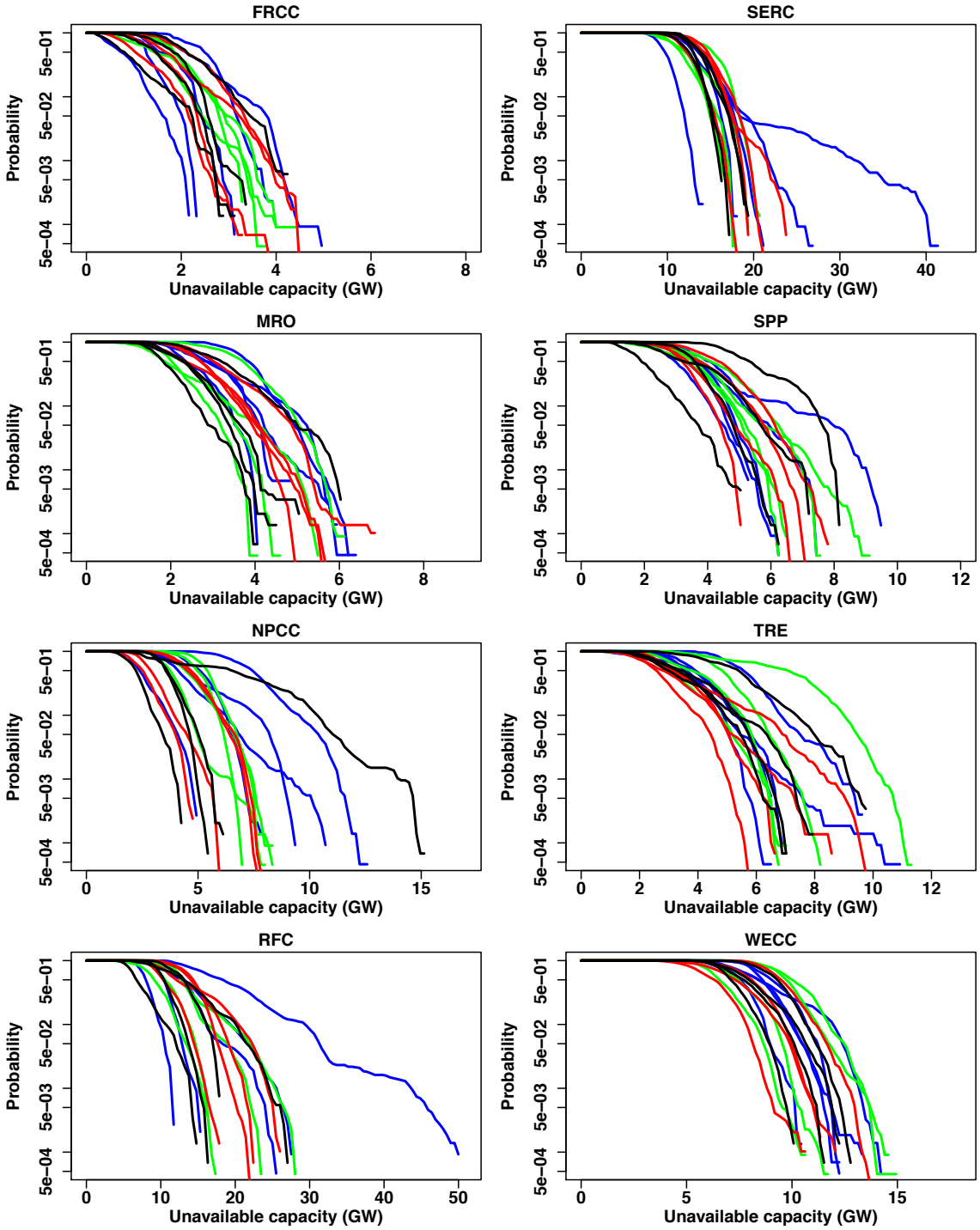

Figure A.23: Seasonal distributions of unscheduled unavailable capacity. Winter is shown in blue, spring in green, summer in red, and fall in black for each of the four years analyzed.

Supplementary data for heteroskedasticity

Given that we previously identified the one-month lag as being statistically significant, to test for heteroskedasticity at the monthly level we first fit AR(1) terms to each region's monthly series of average unavailable capacity and then examine the residuals (Figure A.24). The residuals resemble white noise and appear to be homoskedastic. Autocorrelation functions of the residuals show no significant remaining structure (Figure A.25). We report AR(1) coefficients, standard errors, and t-statistics in Table A.8. From these results we conclude that we cannot generally support the hypothesis that certain times of the year systematically have more variability in unscheduled unavailable capacity than do others.

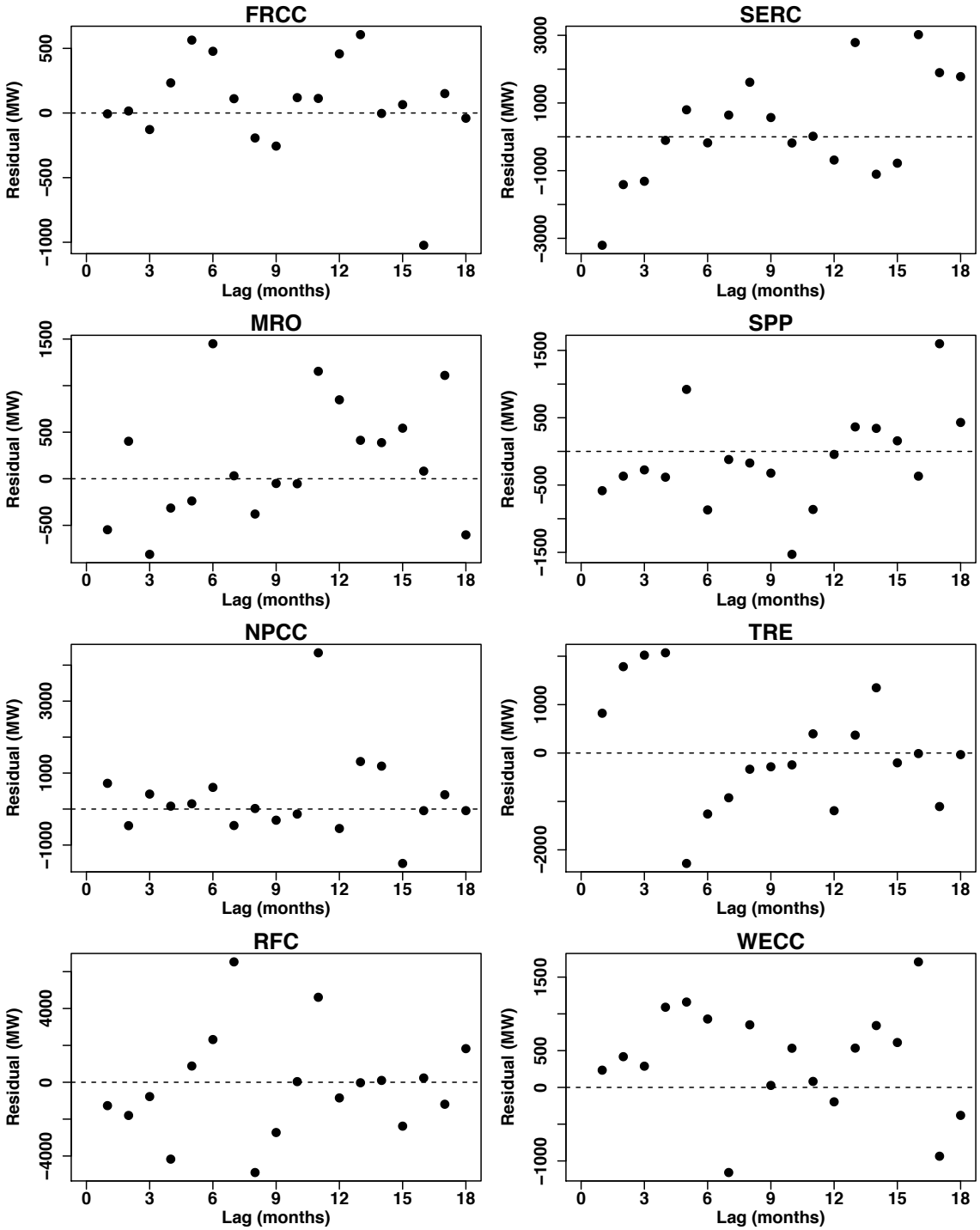

Figure A.24: AR(1) residuals for each NERC region.

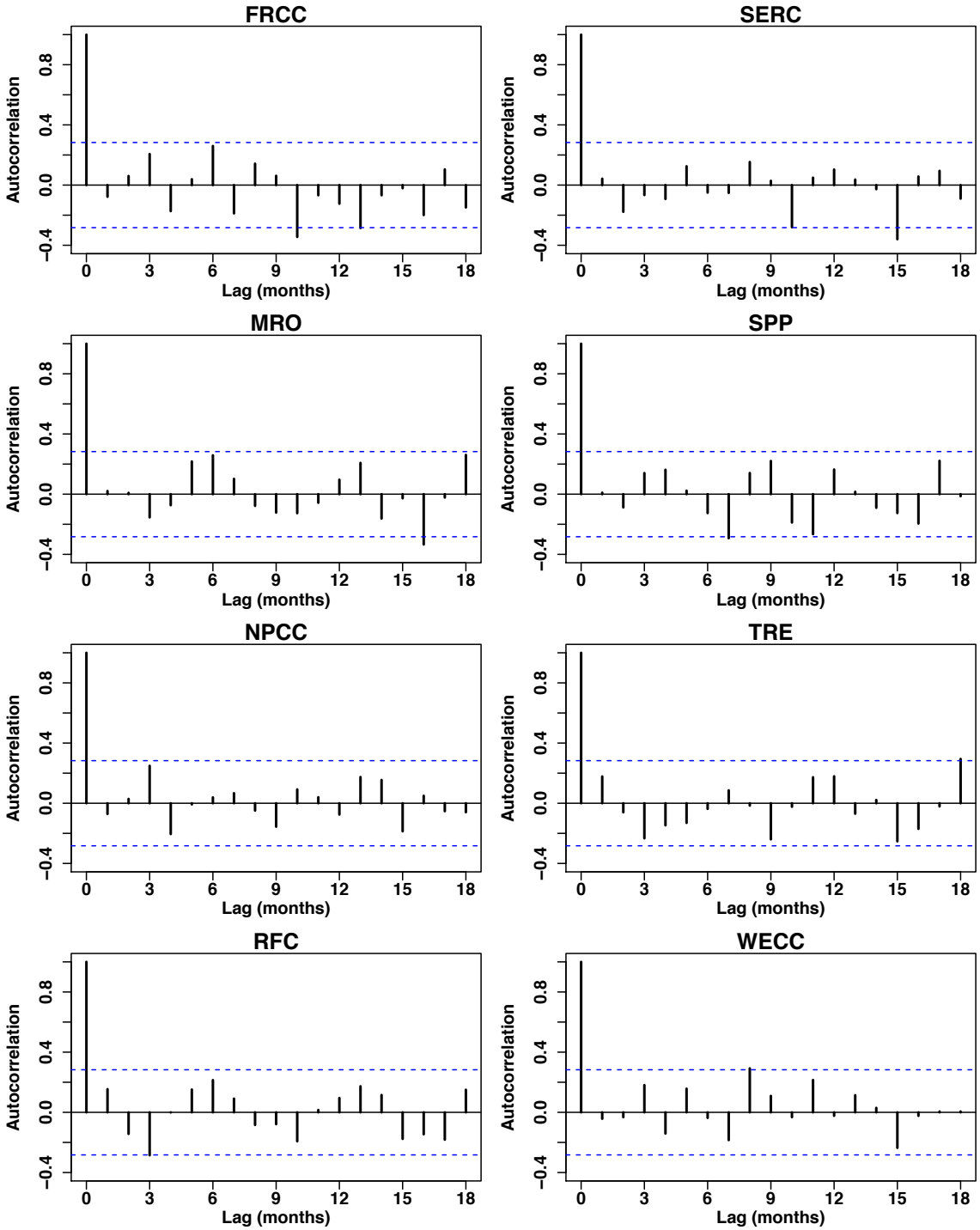

Figure A.25: Autocorrelation functions of the AR(1) residuals for each NERC region with 95% confidence interval for assessing statistical significance of lags (dashed blue lines).

Table A.8: AR(1) coefficients, standard errors, and t-statistics by region.

	FRCC	**MRO**	**NPCC**	**RFC**	**SERC**	**SPP**	**TRE**	**WECC**
Coefficient	0.689	0.458	0.751	0.563	0.502	0.359	0.632	0.707
Standard error	0.101	0.127	0.094	0.120	0.131	0.134	0.109	0.098
t-statistic	6.82	3.61	7.99	4.69	3.83	2.68	5.80	7.21

Supplemental data on the potential benefits of seasonal availability statistics

Current resource adequacy modeling practice in North America calculates an availability statistic using five full years of data. This implicitly assumes that generator availability is constant throughout the year. If instead generator availability was consistently seasonal, calculating availability statistics separately for each season could improve the accuracy of the probability distribution of different contingencies by season. To assess these potential benefits we combine the seasonal block subsampling results from Section 2.4 in the main text and plot the results as exceedance curves for both the full study period (Figure A.26) and when excluding Hurricane Sandy and the Polar Vortex (Figure A.27). We find minimal benefits from using seasonal availability statistics for resource adequacy modeling.

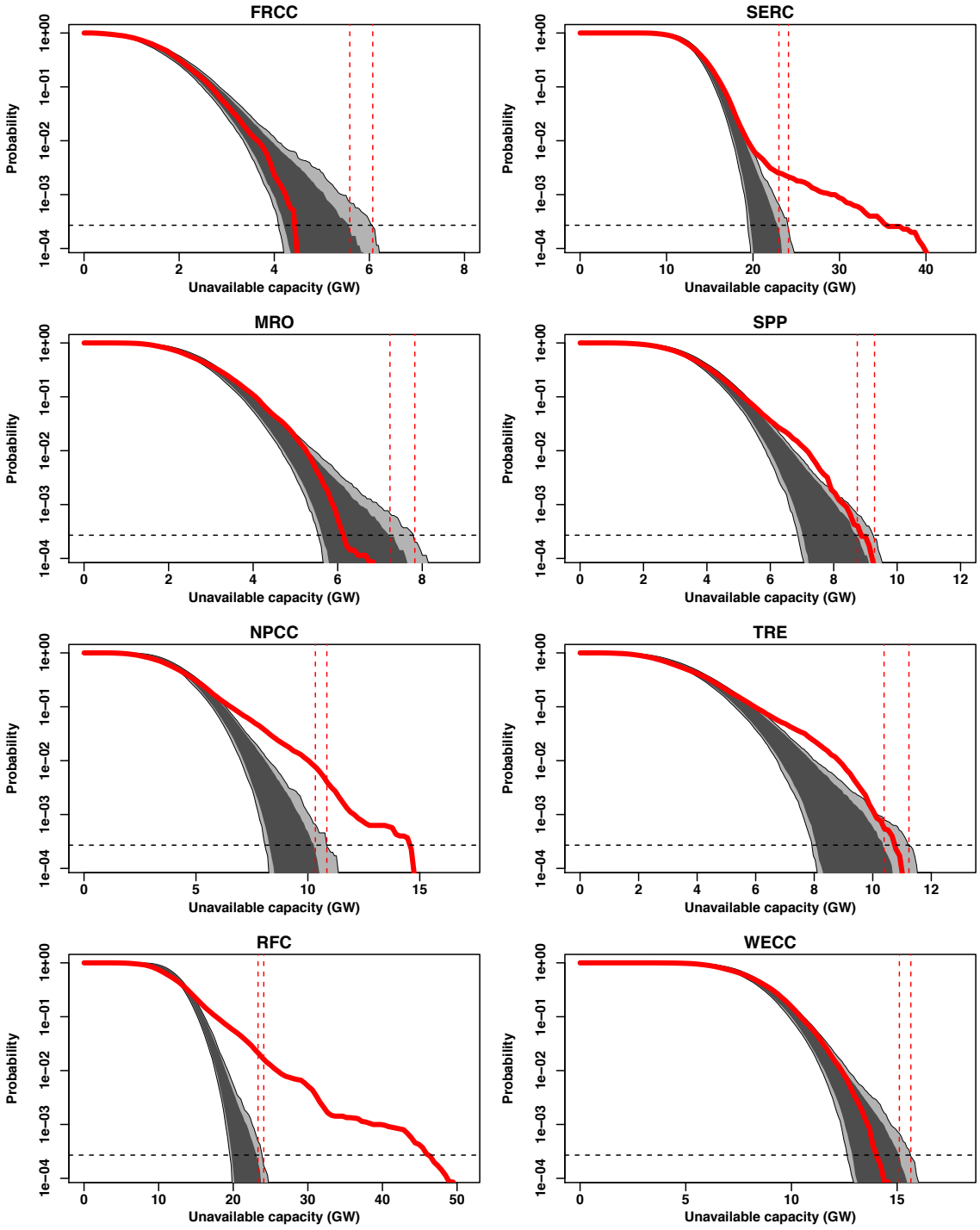

Figure A.26: 1000 seasonal block subsampling runs for the full period 2012-2015.

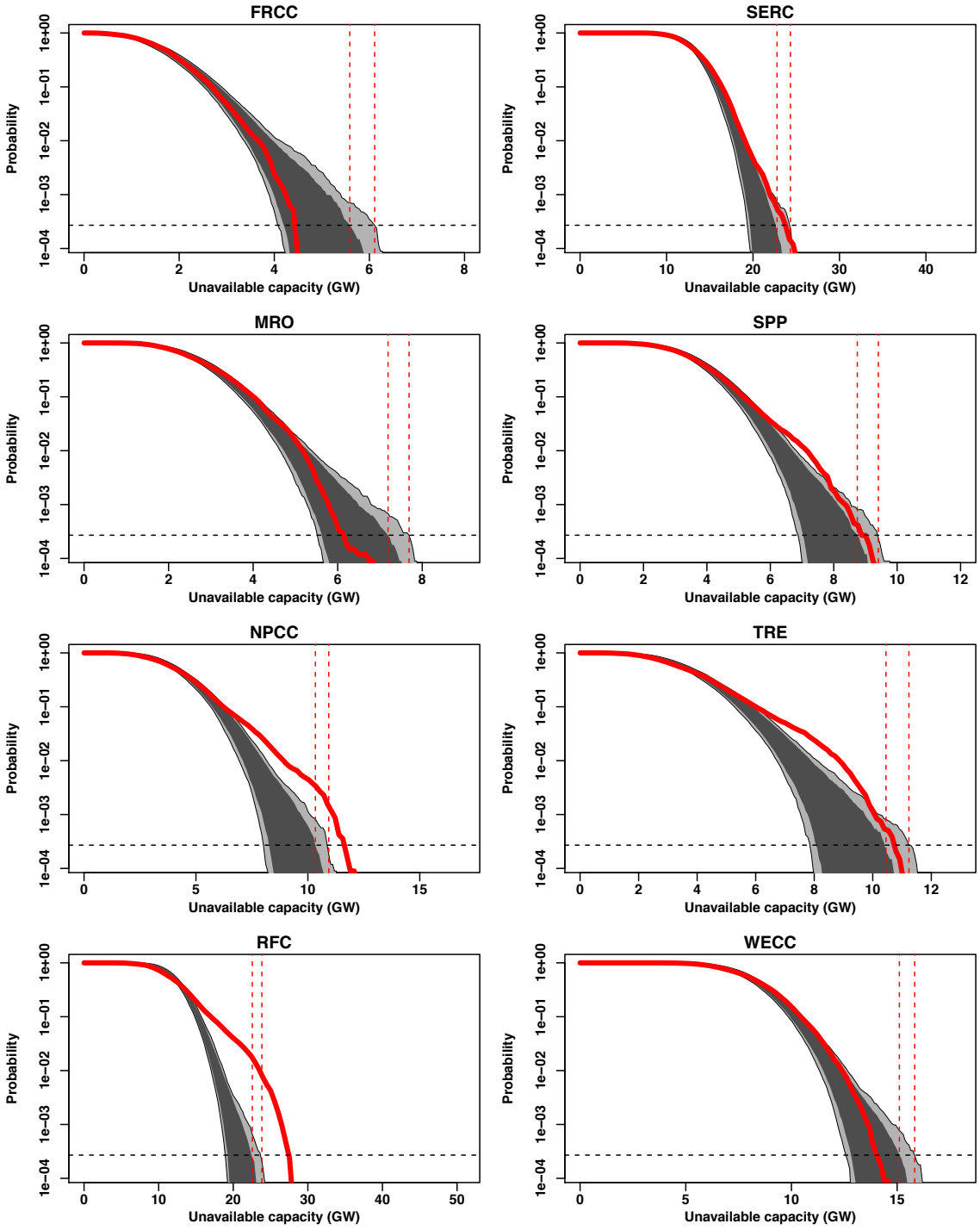

Figure A.27: 1000 seasonal block subsampling runs for the period 2012-2015 excluding Hurricane Sandy and January 2014.

2.6.4 Reliability applications

Parametric fits to distributions of unscheduled unavailable capacity

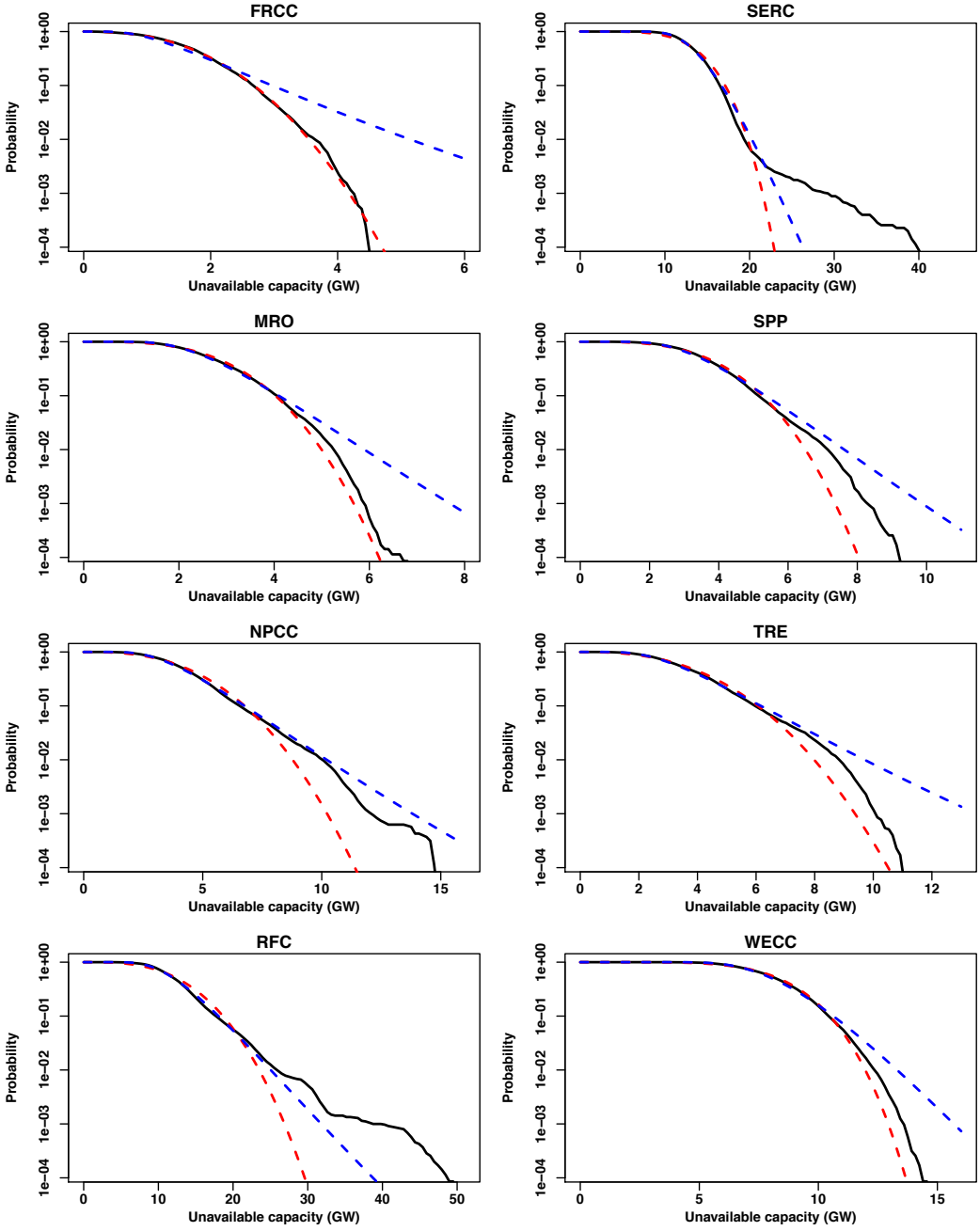

Figure A.28: Overlay of Weibull (red) and lognormal (blue) fits to observed exceedance curves (black) for the 2012-2015 period. Weibull and lognormal parameters are given in Table A.9.

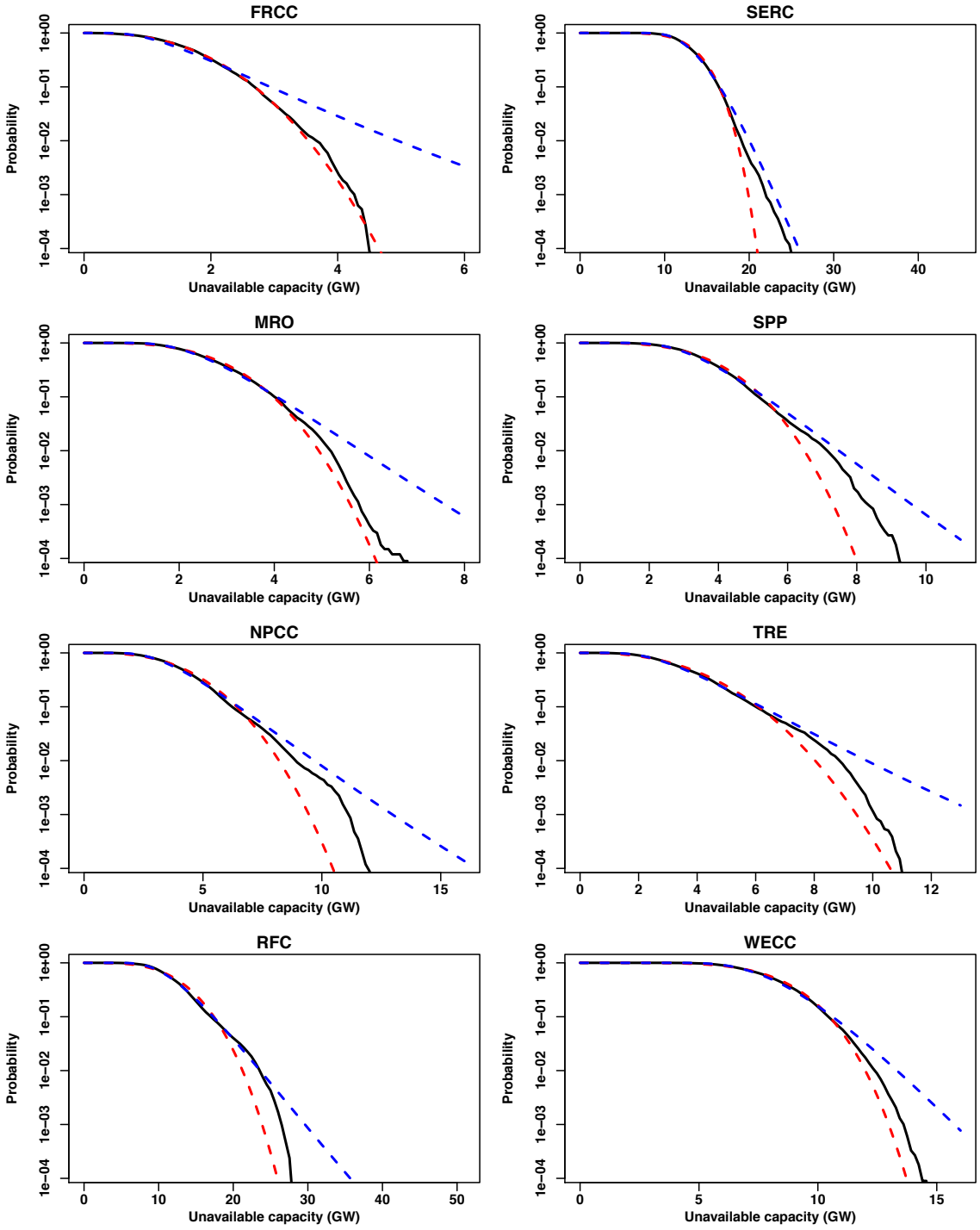

Figure A.29: Overlay of Weibull (red) and lognormal (blue) fits to observed exceedance curves (black), removing Hurricane Sandy and January 2014. Weibull and lognormal parameters are given in Table A.9.

Table A.9: Parameters for Weibull and lognormal distributions of unscheduled unavailable capacity by region for the fully study period (left) and when excluding Hurricane Sandy and January 2014 (right). Standard errors in parentheses.

Region	Full period				Excluding Sandy and Jan 2014			
	Weibull		Lognormal		Weibull		Lognormal	
	shape	scale	meanlog	sdlog	shape	scale	meanlog	sdlog
FRCC	2.47 (0.010)	1,910 (4.35)	7.32 (0.0028)	0.526 (0.0020)	2.54 (0.011)	1,940 (4.39)	7.34 (0.0027)	0.500 (0.0019)
MRO	3.20 (0.013)	3,100 (5.47)	7.87 (0.0019)	0.348 (0.0013)	3.22 (0.013)	3,070 (5.51)	7.86 (0.0019)	0.347 (0.0013)
NPCC	2.65 (0.010)	4,940 (10.53)	8.31 (0.0021)	0.396 (0.0015)	2.84 (0.011)	4,790 (9.74)	8.29 (0.0021)	0.382 (0.0015)
RFC	3.00 (0.011)	14,130 (26.77)	9.40 (0.0017)	0.313 (0.0012)	3.48 (0.014)	13,690 (22.77)	9.38 (0.0016)	0.297 (0.0011)
SERC	4.78 (0.015)	14,370 (16.95)	9.48 (0.0010)	0.187 (0.00071)	5.81 (0.023)	14,270 (14.20)	9.48 (0.0010)	0.184 (0.00071)
SPP	3.27 (0.013)	4,070 (7.044)	8.15 (0.0018)	0.339 (0.0013)	3.34 (0.013)	4,100 (7.11)	8.16 (0.0018)	0.326 (0.0013)
TRE	2.52 (0.010)	4,360 (9.77)	8.17 (0.0023)	0.435 (0.0016)	2.50 (0.010)	4,360 (10.09)	8.17 (0.0024)	0.439 (0.0017)
WECC	5.22 (0.021)	8,940 (9.65)	8.99 (0.0012)	0.216 (0.00081)	5.15 (0.021)	8,920 (10.01)	8.99 (0.0012)	0.218 (0.00084)
Combined	1.41 (0.0020)	6,989 (9.91)	8.46 (0.0015)	0.798 (0.0011)	1.24 (0.0018)	5,880 (9.74)	8.25 (0.0016)	0.850 (0.0012)

Parametric fits to distributions of normalized derating magnitudes

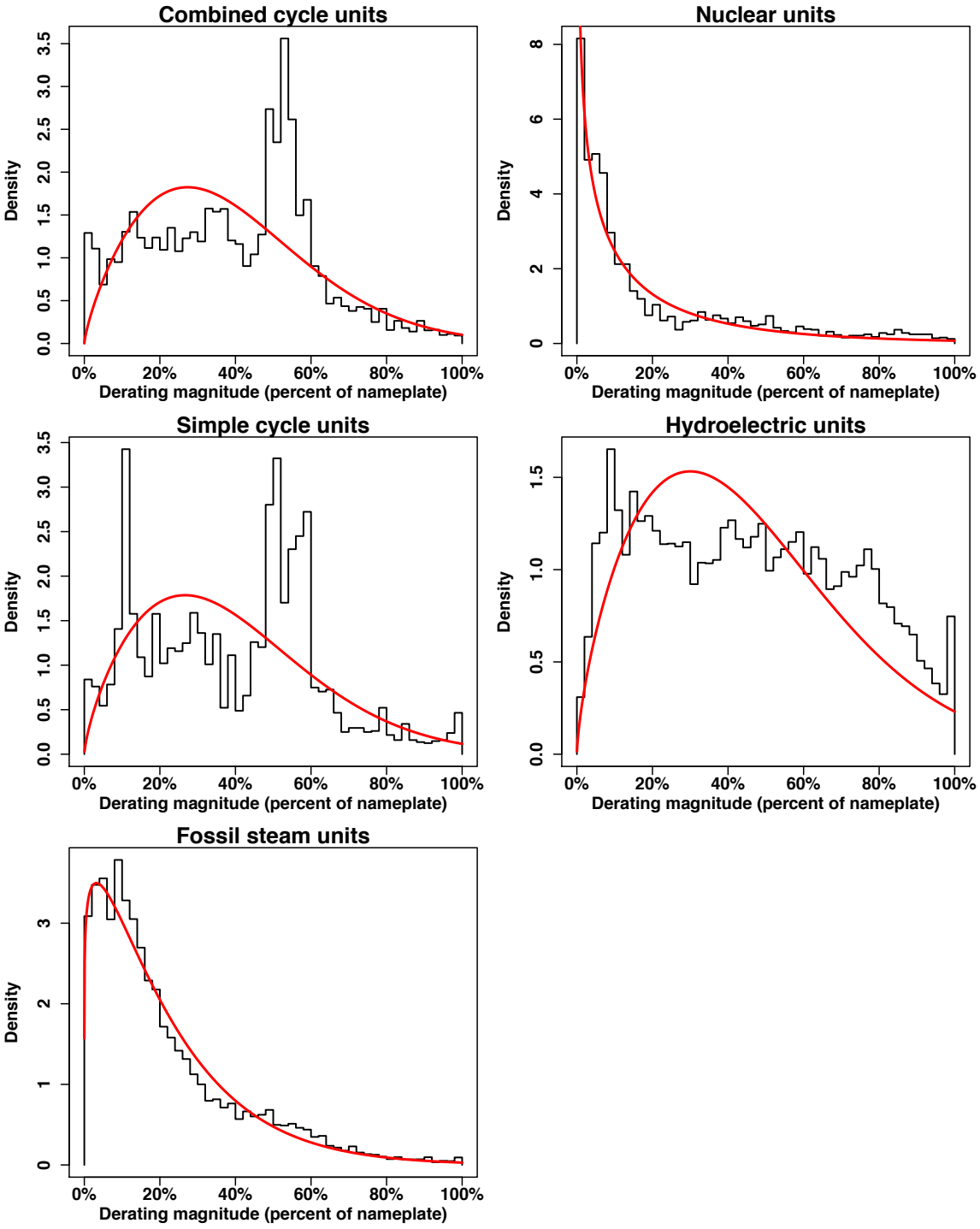

Figure A.30: Distribution of derating magnitudes as a percentage of nameplate capacity, with fitted Weibull distributions.

Table A.10: Parameters for Weibull fits to derating magnitudes by unit type. Standard errors in parentheses.

Combined cycle		Simple cycle		Fossil steam		Hydroelectric		Nuclear	
Shape	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**
1.77 (0.010)	0.44 (0.0018)	1.72 (0.022)	0.44 (0.0040)	1.12 (0.0022)	0.23 (0.00053)	1.69 (0.0042)	0.51 (0.00095)	0.72 (0.011)	0.17 (0.0047)

Mean time between failure and mean time to recovery

Table A.11: Parameters for Weibull fits to capacity-weighted MTBF values excluding reserve shutdown hours, by unit type. Standard errors in parentheses.

Combined cycle		Simple cycle		Fossil steam		Hydroelectric		Nuclear	
Shape	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**
1.12 (0.0016)	1,060 (2.12)	0.78 (0.0015)	440 (1.71)	1.08 (0.112)	420 (0.65)	1.13 (0.0029)	3,120 (10.35)	1.34 (0.0029)	2,880 (7.03)

Table A.12: Parameters for gamma fits to capacity-weighted MTBF values excluding reserve shutdown hours, by unit type. Standard errors in parentheses.

Combined cycle		Simple cycle		Fossil steam		Hydroelectric		Nuclear	
Shape	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**
1.48 (0.004)	680 (2.19)	0.75 (0.0026)	710 (3.41)	1.30 (0.0026)	310 (0.76)	1.34 (0.006)	2,220 (12.04)	1.94 (0.0079)	1,350 (6.20)

Histograms of the number of between-failure periods used to calculate each unit's MTBF are shown in Figure A.31 through Figure A.35. We note that some units' MTBFs are calculated based upon only a single between-failure period. With a longer time series, the proportion of units with MTBFs based on very few between-failure periods would decrease, increasing confidence in the robustness of these results.

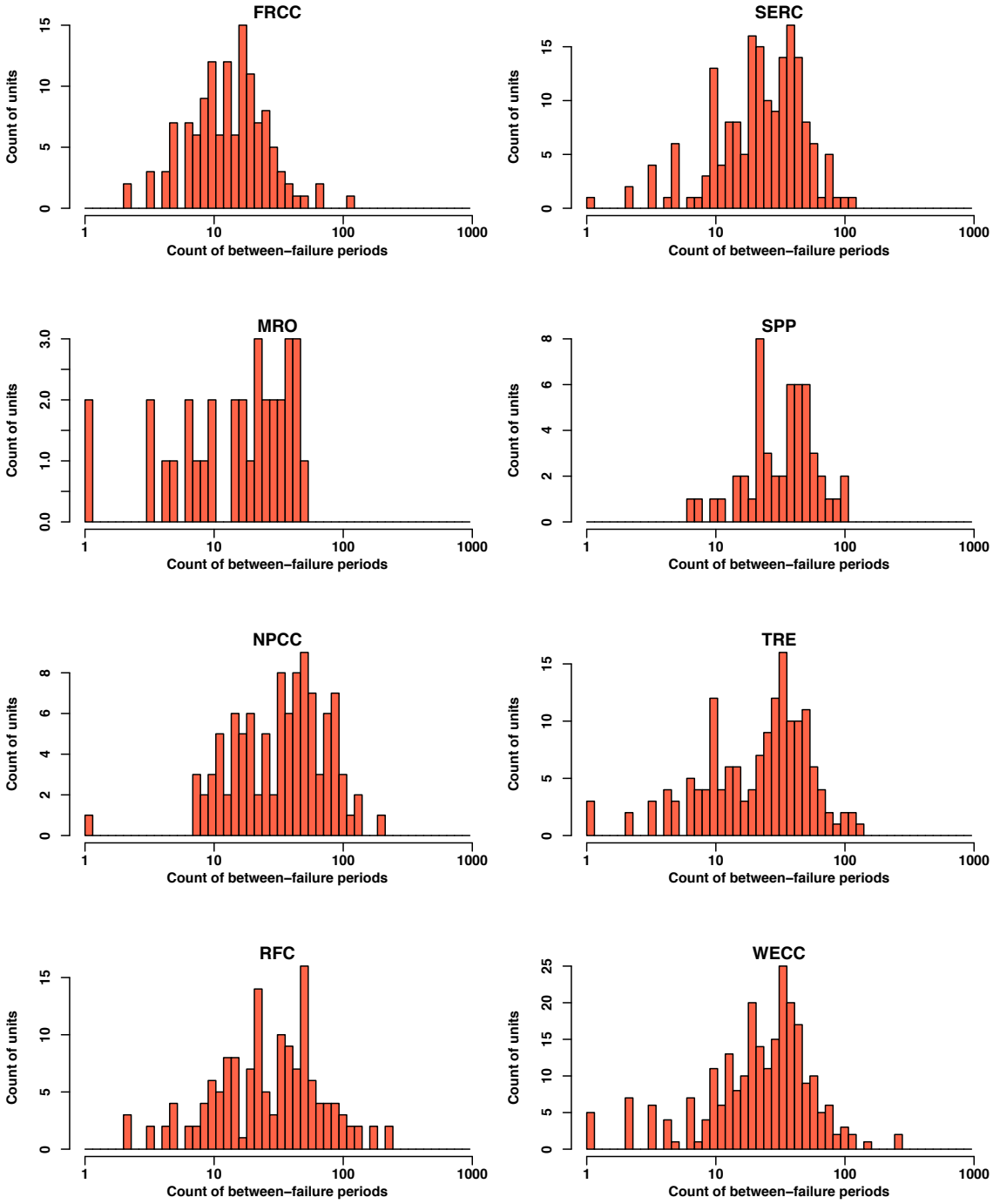

Figure A.31: Count of between-failure periods used to calculate mean time between failure for combined cycle units. Units with significant reserve shutdown reporting discrepancies are excluded.

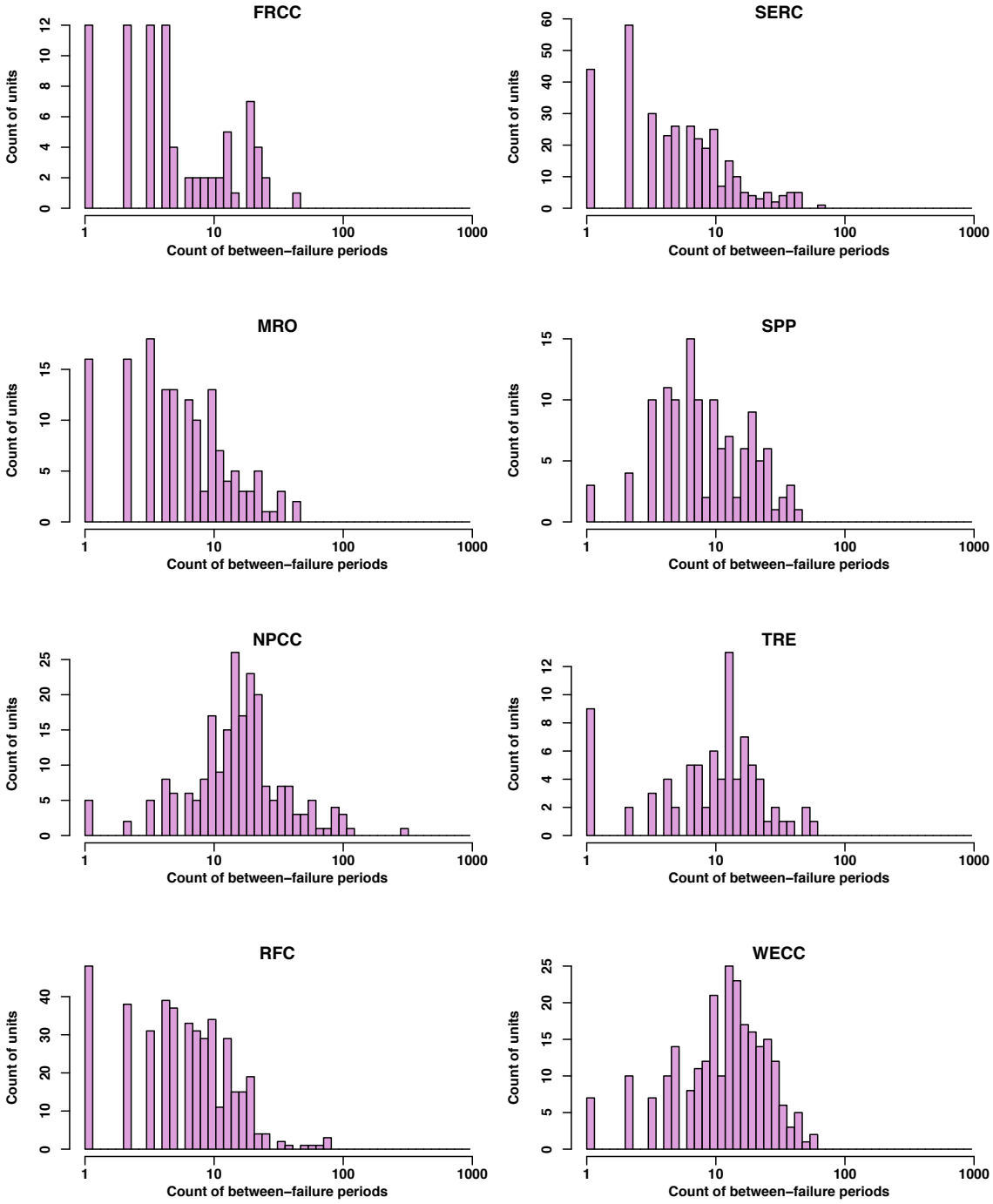

Figure A.32: Count of between-failure periods used to calculate mean time between failure for simple cycle units. Units with significant reserve shutdown reporting discrepancies are excluded.

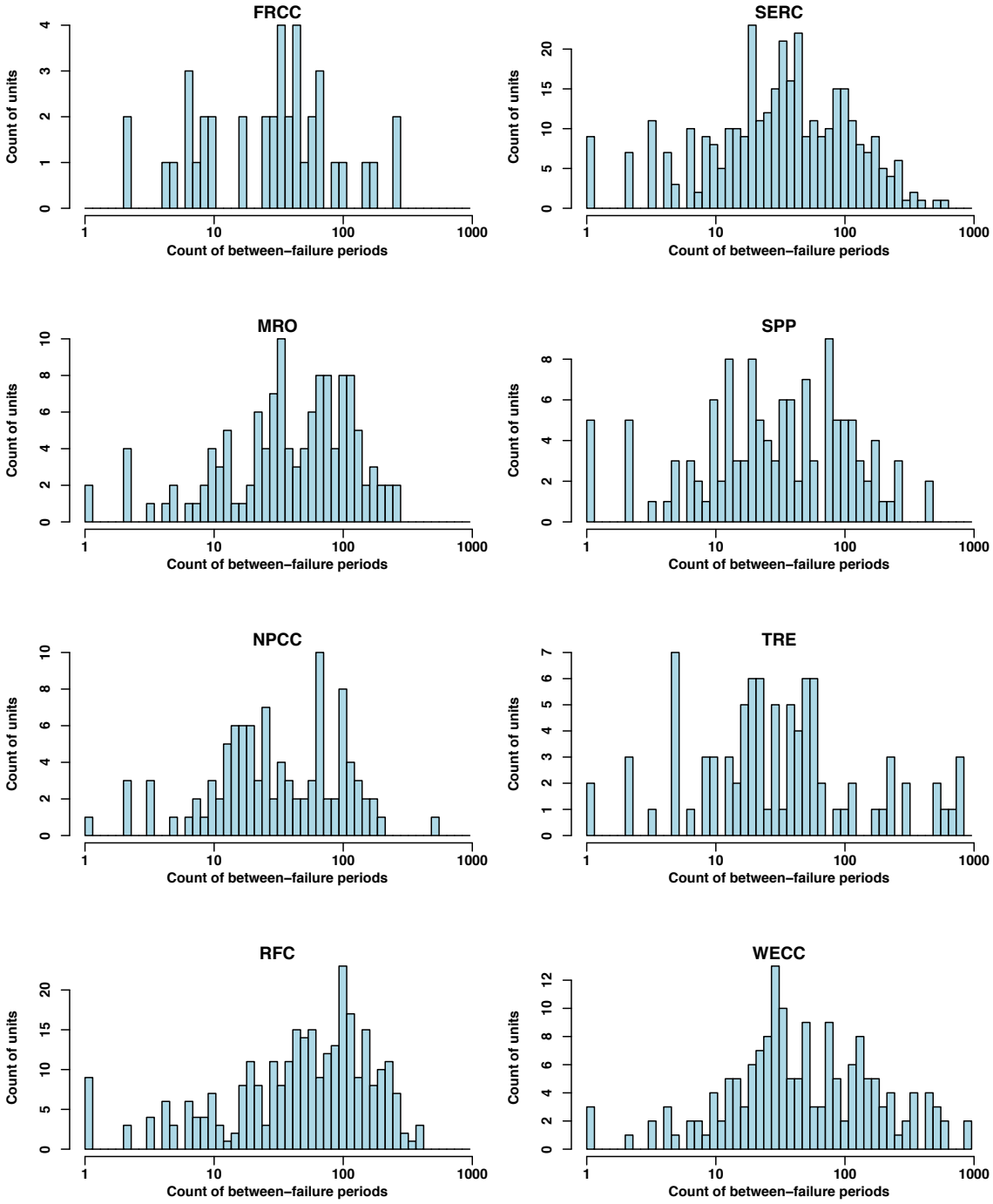

Figure A.33: Count of between-failure periods used to calculate mean time between failure for fossil steam and fluidized bed units. Units with significant reserve shutdown reporting discrepancies are excluded.

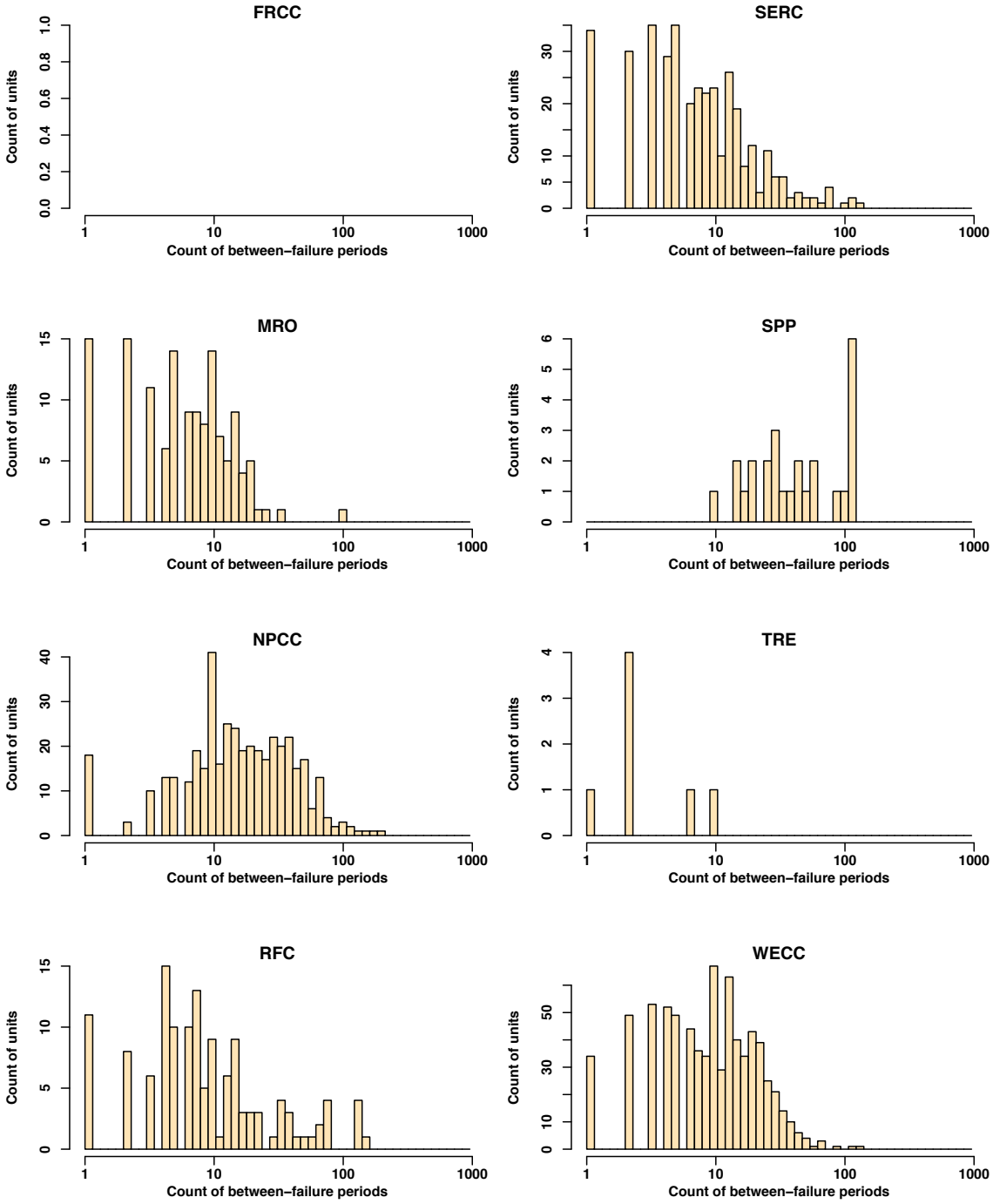

Figure A.34: Count of between-failure periods used to calculate mean time between failure for hydroelectric units. FRCC has no such units. Units with significant reserve shutdown reporting discrepancies are excluded.

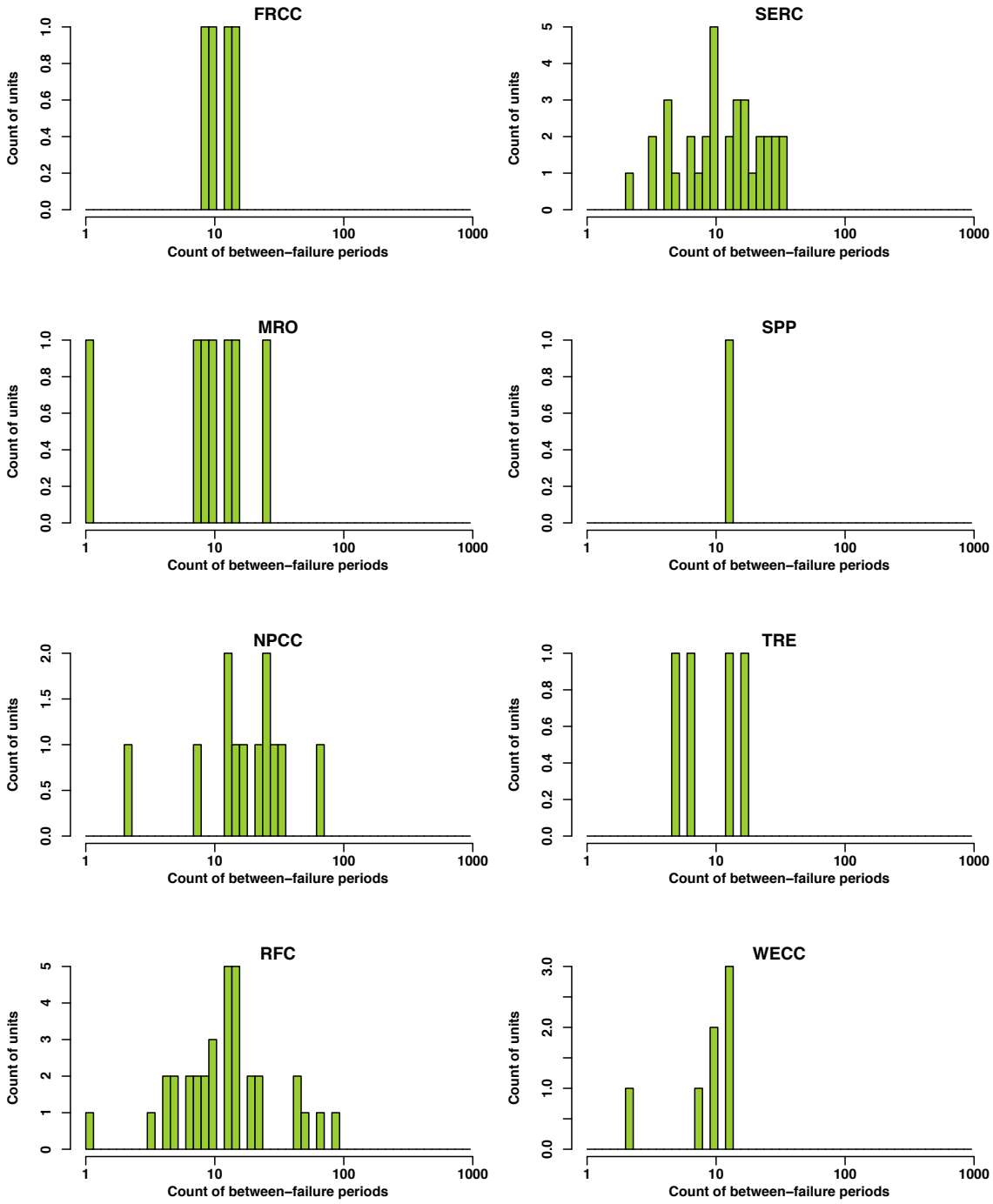

Figure A.35: Count of between-failure periods used to calculate mean time between failure for nuclear units. Units with significant reserve shutdown reporting discrepancies are excluded.

Not every unit has a calculated MTBF value because calculation of the MTBF requires at least

two non-overlapping failures (i.e. at least one fully contained non-failure period). We disregard the hours occurring before the first failure and after the last failure, as applicable, because we cannot say how long these non-failure periods would extend. An alternative approach would be to average the durations of all non-failure periods. This would not exclude any units, but would be guaranteed to underestimate the MTBF for those units excluded by our method. With either approach, one could also apply a Bayesian prior on the MTBF to perhaps make the results more robust for units that have few unscheduled events. However, selection of an informative prior would be difficult. Some units' MTBF values are based on very few non-failure periods. These units may be very reliable or may have just performed well during our study period. We recommend that system planners repeat this analysis with longer time series to increase confidence in the result. One could also look at the MTBF and MTTR for different primary cause codes. We note that our MTBF results depend to some degree on our definition of a failure. While we have defined a failure as any reduction from full availability, any desired threshold could be used. We include a sensitivity analysis over a range of failure definitions (Figure A.36). We conclude that the calculation of MTBF is not very sensitive to the selection of the minimum percent of a unit's nameplate capacity that must be unavailable to constitute a "failure".

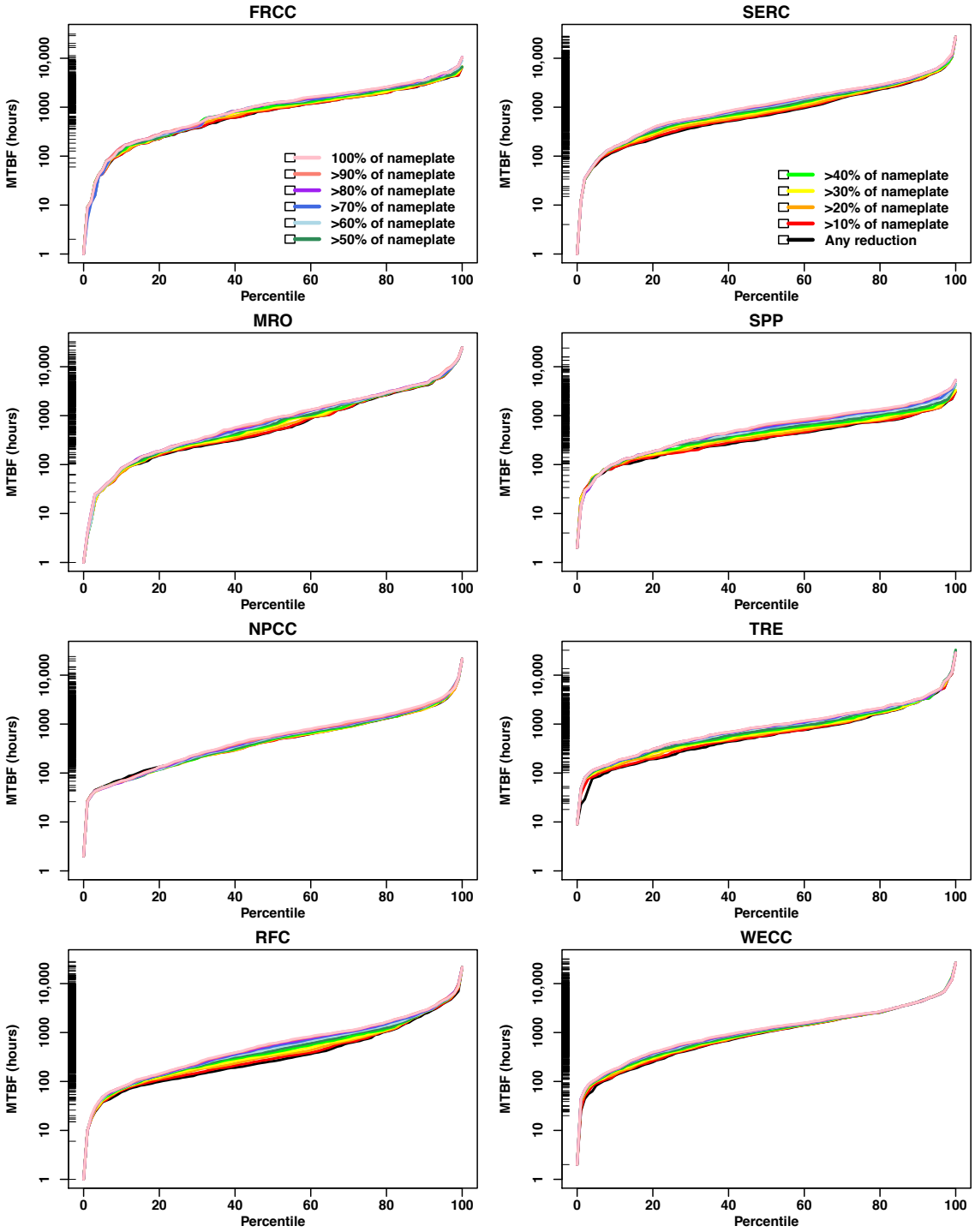

Figure A.36: Sensitivity of MTBF to the minimum percent of a unit's nameplate capacity that must be unavailable to constitute a failure. The legend has been split between the top two panels for readability. The rug along the y-axis indicates each region's MTBF values when a failure is defined as any reduction in availability. Values have been calculated with all reserve shutdown hours removed. Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

We define the mean time to recovery (MTTR) as the average number of hours that elapse while a unit experiences some reduction in availability—i.e., the average duration of failure periods. In contrast to MTBF, we do not need to remove RS hours prior to calculating MTTR. We present capacity-weighted histograms of the MTTR results (Figures A.37 through A.41). In each of these plots we construct histograms with 50 bins. The heading of each plot reports the number of units for which an MTTR value can be calculated (numerator) and the number of units reporting at least a single unscheduled event during our study period (denominator), which again serves as a proxy for the sample size. Smaller MTTR values indicate shorter average repair durations. Histograms of the number of failure periods used to calculate each unit's MTTR are shown in Figure A.42 through Figure A.46. We note that some units' MTTRs are calculated based upon only a single failure period. With a longer time series, the proportion of units with MTTRs based on very few failure periods would decrease, increasing confidence in the robustness of these results. We present the parameters for Weibull and gamma distributions fit to each unit type's distribution of MTTR values in Tables A.13 and A.14. We do not report parameters at the region-by-unit-type level due to small sample sizes in several instances.

Table A.13: Parameters for Weibull fits to capacity-weighted MTTR values by unit type. Standard errors in parentheses.

Combined cycle		**Simple cycle**		**Fossil steam**		**Hydroelectric**		**Nuclear**	
Shape	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**
0.69 (0.00085)	45 (0.14)	0.53 (0.00090)	95 (0.48)	0.60 (0.00056)	106 (0.28)	0.47 (0.00092)	77 (0.50)	0.66 (0.0013)	386 (1.90)

Table A.14: Parameters for gamma fits to capacity-weighted MTTR values by unit type. Standard errors in parentheses.

Combined cycle		**Simple cycle**		**Fossil steam**		**Hydroelectric**		**Nuclear**	
Shape	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**	**Shape**	**Scale**
0.58 (0.0014)	125 (0.45)	0.37 (0.0010)	644 (3.21)	0.44 (0.00076)	528 (1.51)	0.30 (0.00098)	914 (5.62)	0.55 (0.0020)	1,071 (5.84)

Similar to the considerations mentioned for the MTBF calculations above, we require that units have at least one completed failure period in order for us to calculate an MTTR. While this is a less restrictive requirement than that needed to calculate an MTBF, it still excludes 10 units from the analysis.

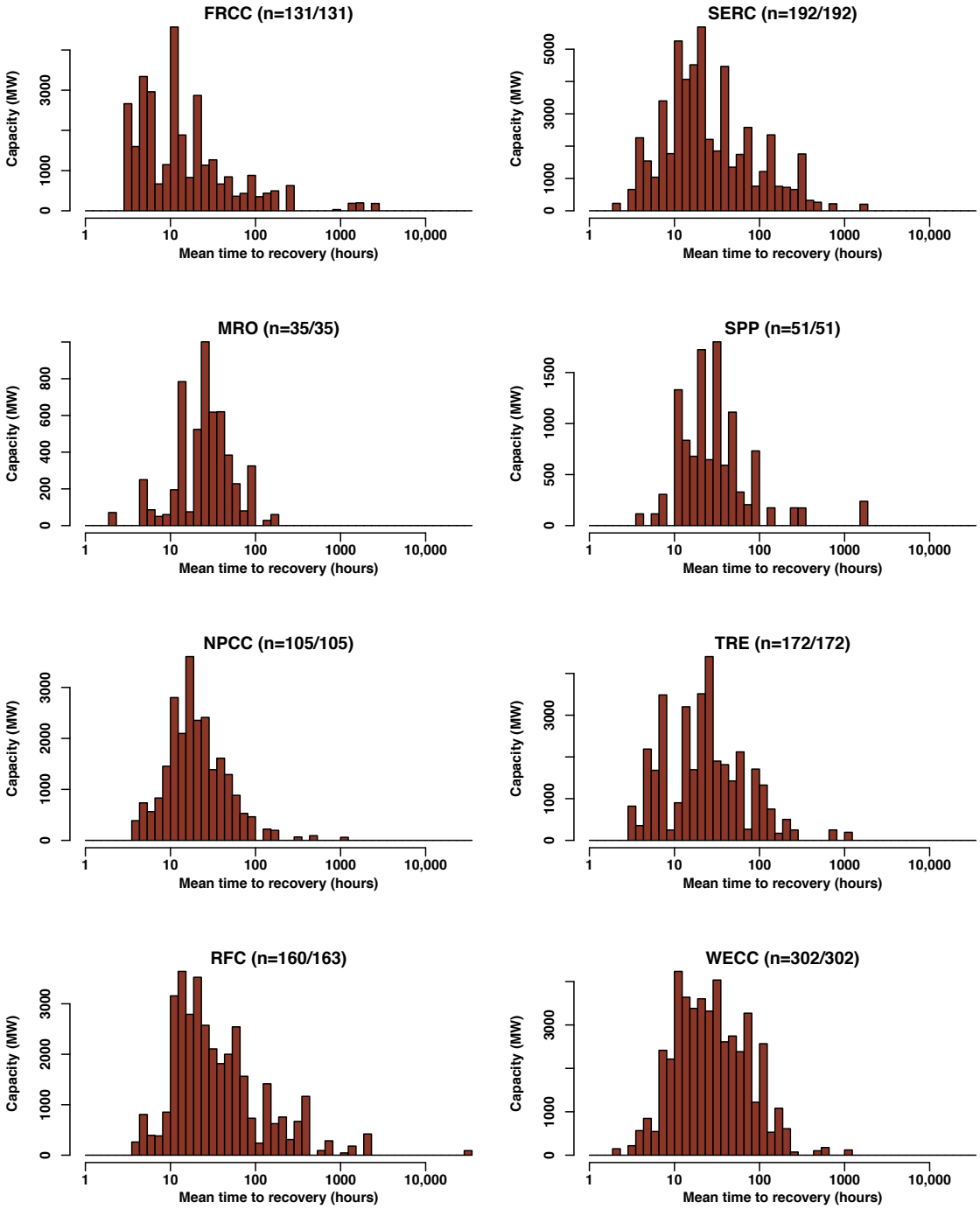

Figure A.37: Capacity-weighted mean time to recovery (MTTR) values for combined cycle gas units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units experiencing at least one unscheduled event during the study period (as a proxy for total count of active units during the study period).

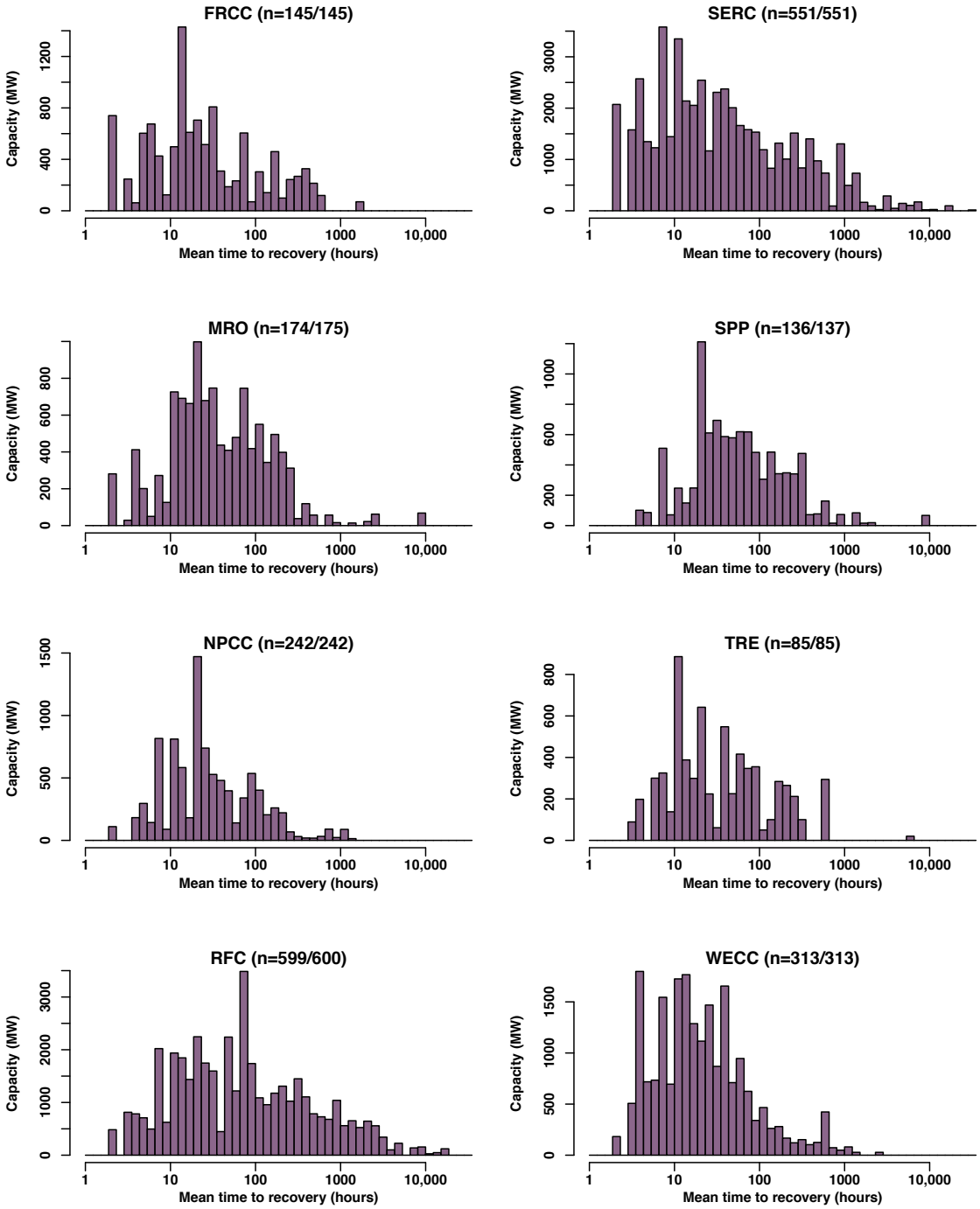

Figure A.38: Capacity-weighted mean time to recovery (MTTR) values for simple cycle gas units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units experiencing at least one unscheduled event during the study period (as a proxy for total count of active units during the study period).

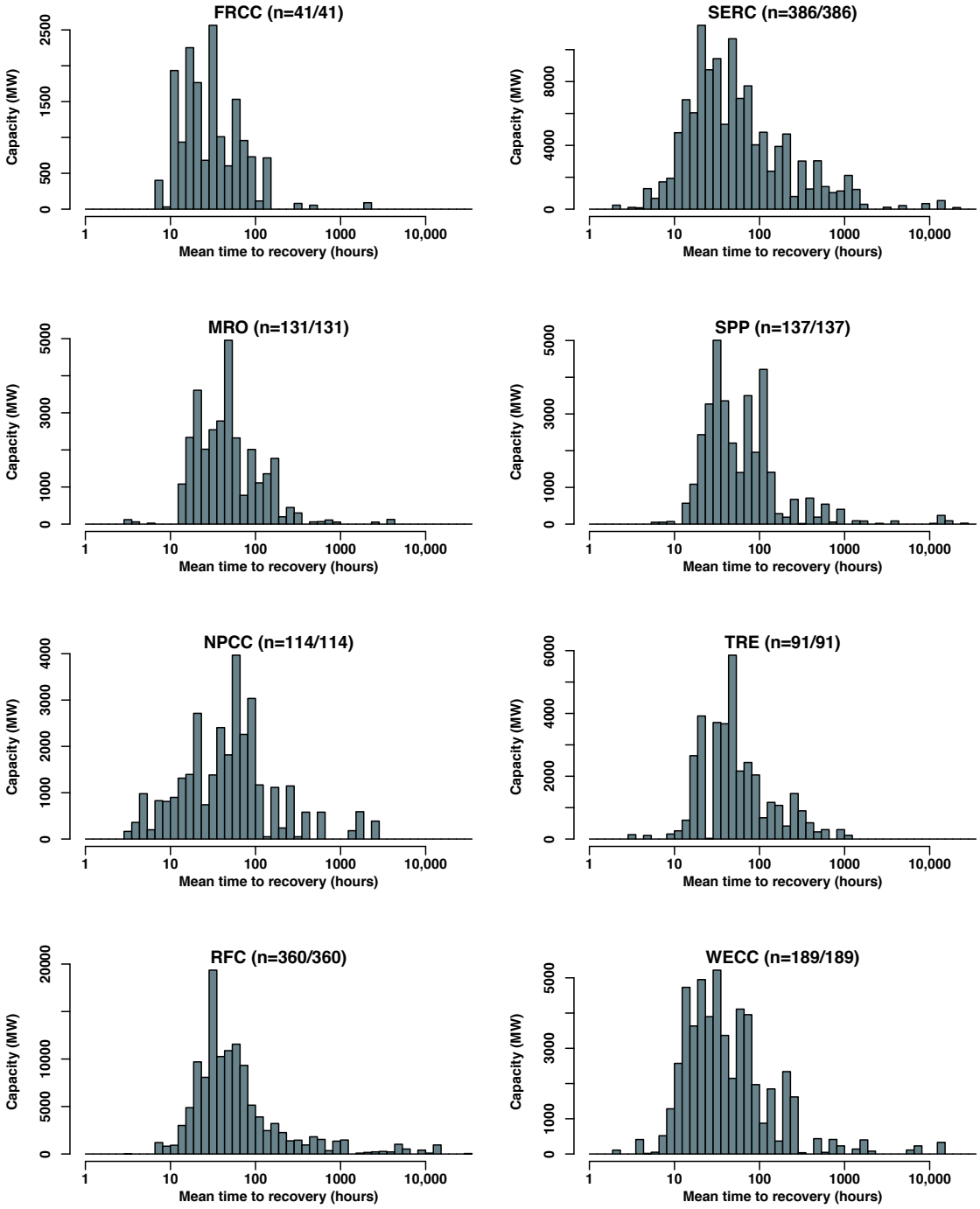

Figure A.39: Capacity-weighted mean time to recovery (MTTR) values for fossil steam and fluidized bed units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units experiencing at least one unscheduled event during the study period (as a proxy for total count of active units during the study period).

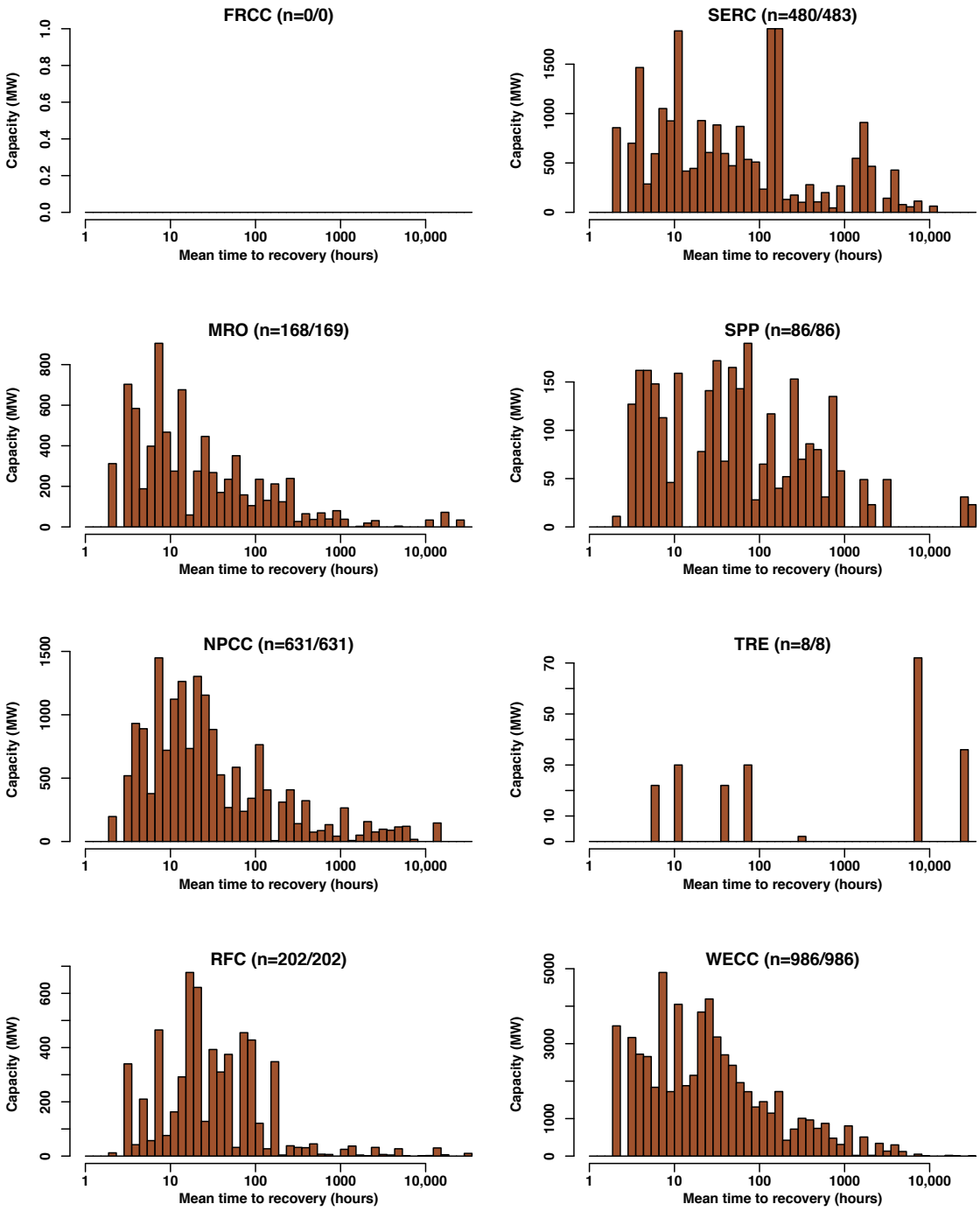

Figure A.40: Capacity weighted mean time to recovery (MTTR) values for hydroelectric units. FRCC has no such units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units experiencing at least one unscheduled event during the study period (as a proxy for total count of active units during the study period).

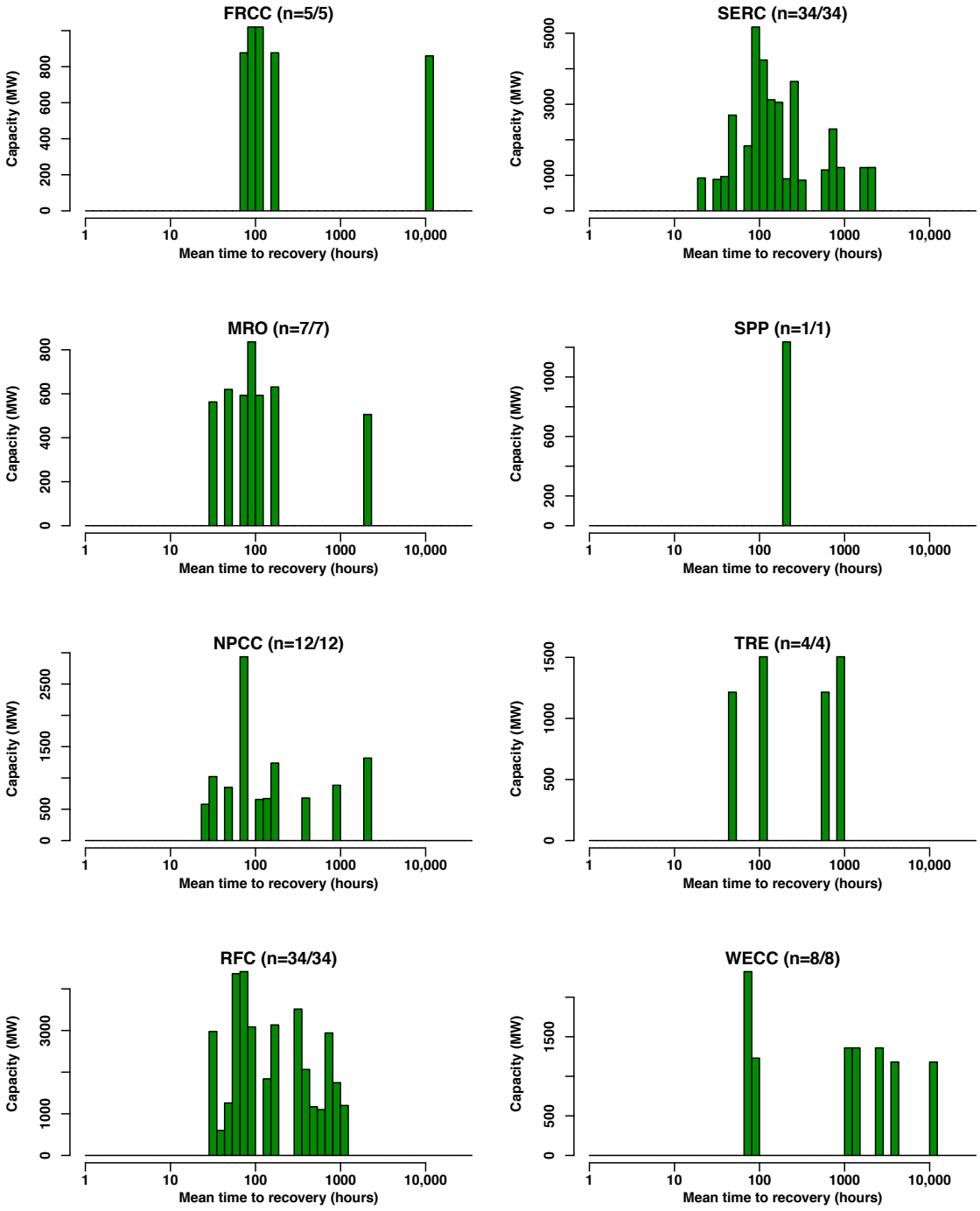

Figure A.41: Capacity-weighted mean time to recovery (MTTR) values for nuclear units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units experiencing at least one unscheduled event during the study period (as a proxy for total count of active units during the study period).

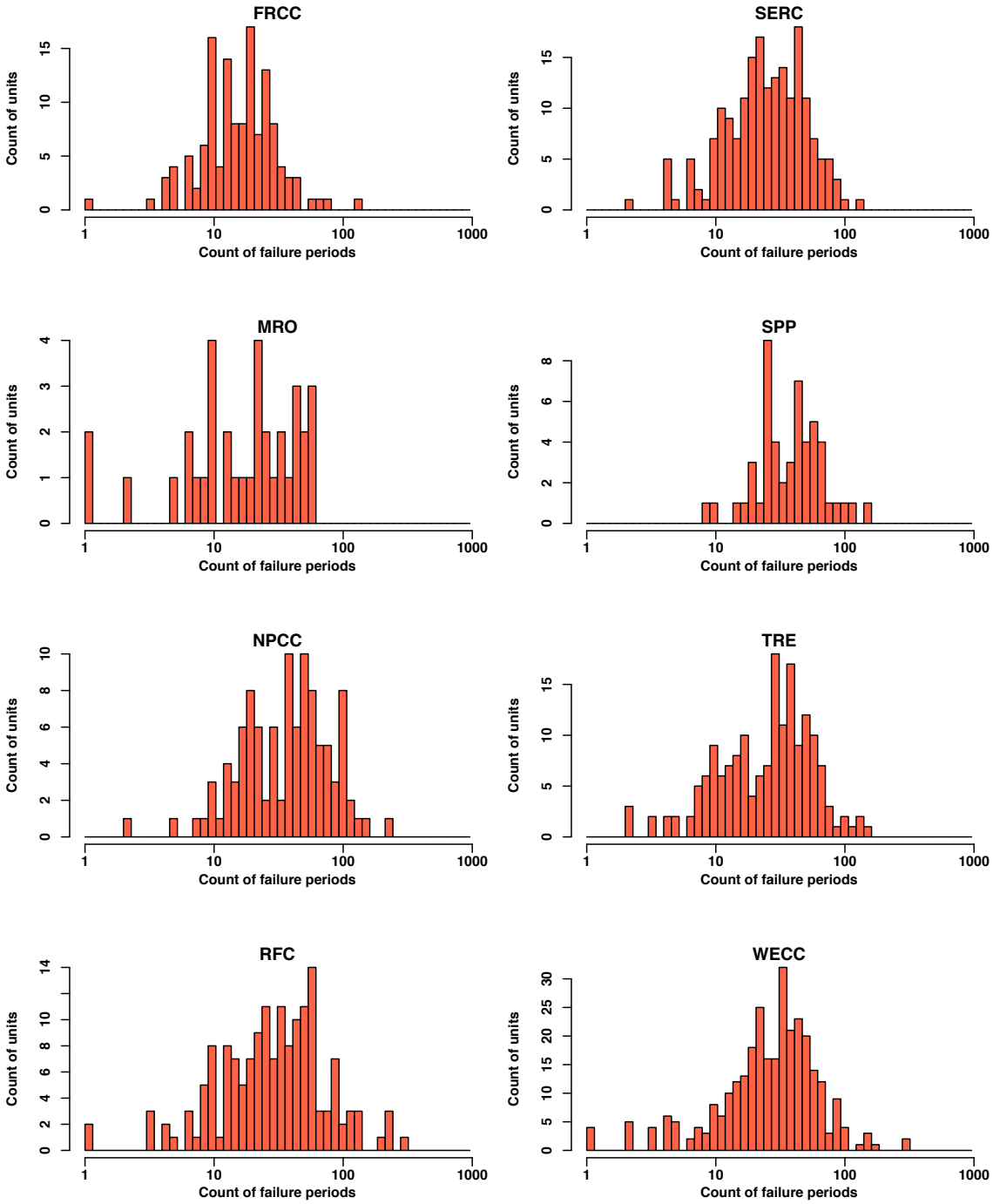

Figure A.42: Count of failure periods used to calculate mean time to recovery for combined cycle units.

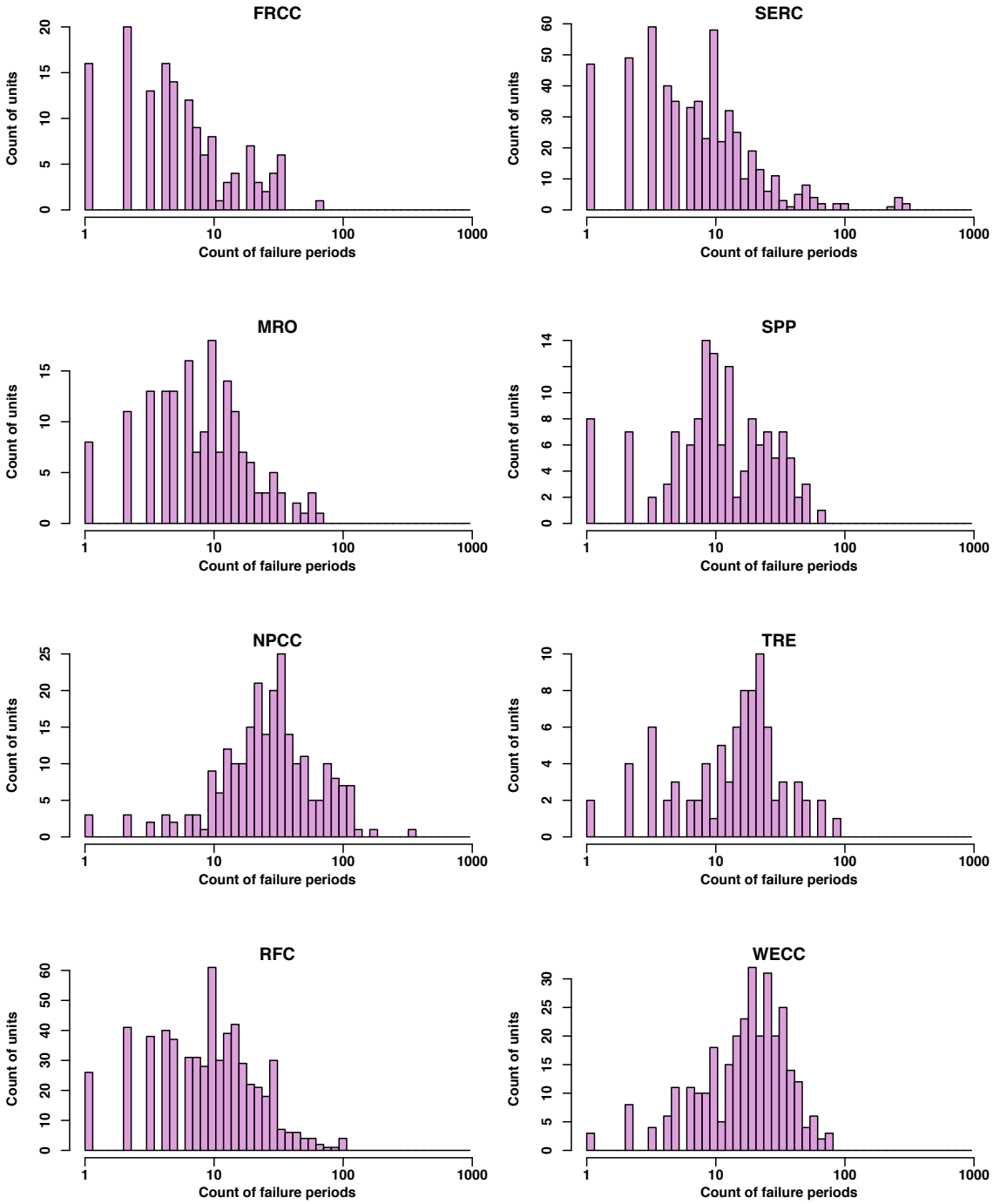

Figure A.43: Count of failure periods used to calculate mean time to recovery for simple cycle units.

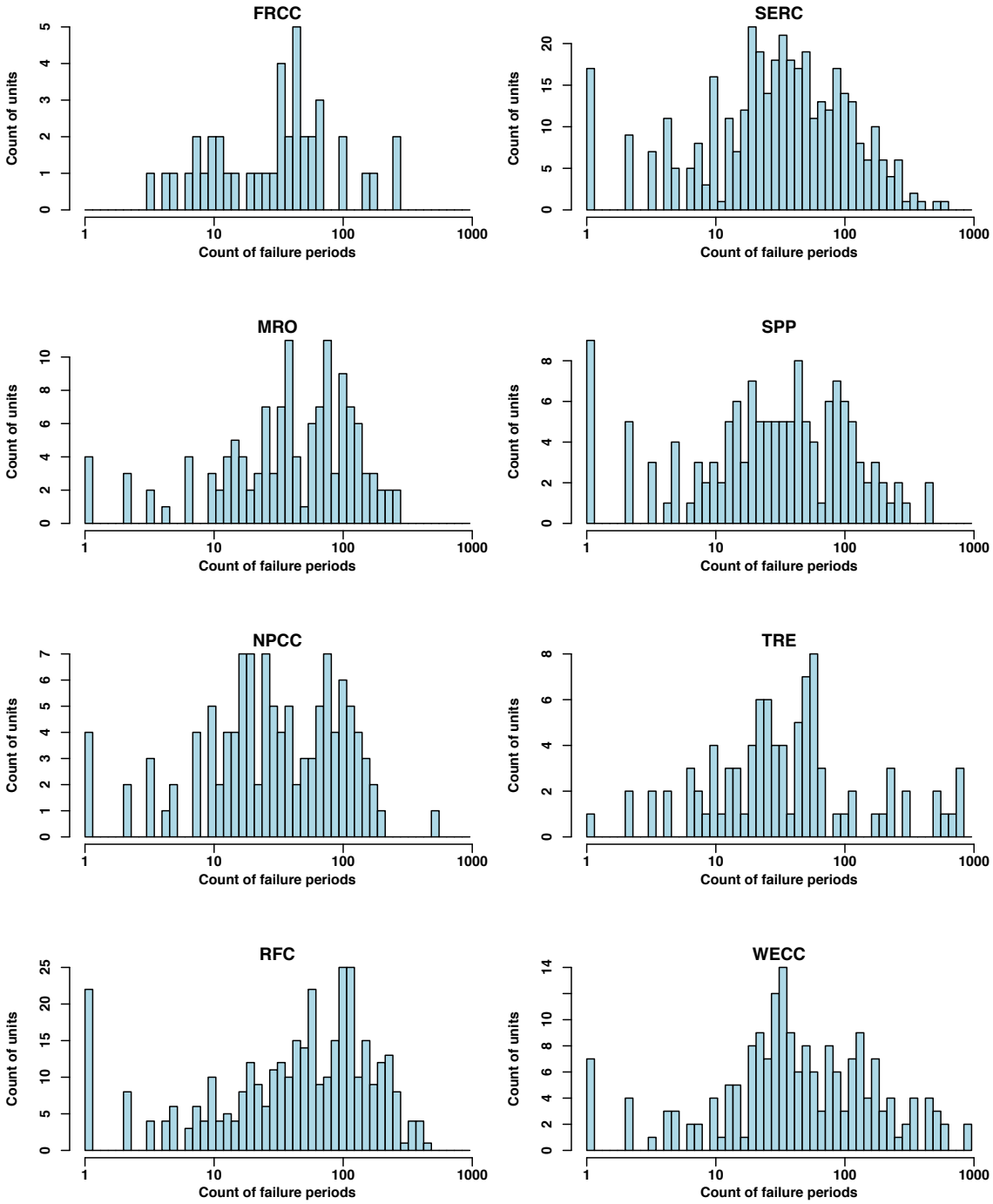

Figure A.44: Count of failure periods used to calculate mean time to recovery for fossil steam and fluidized bed units.

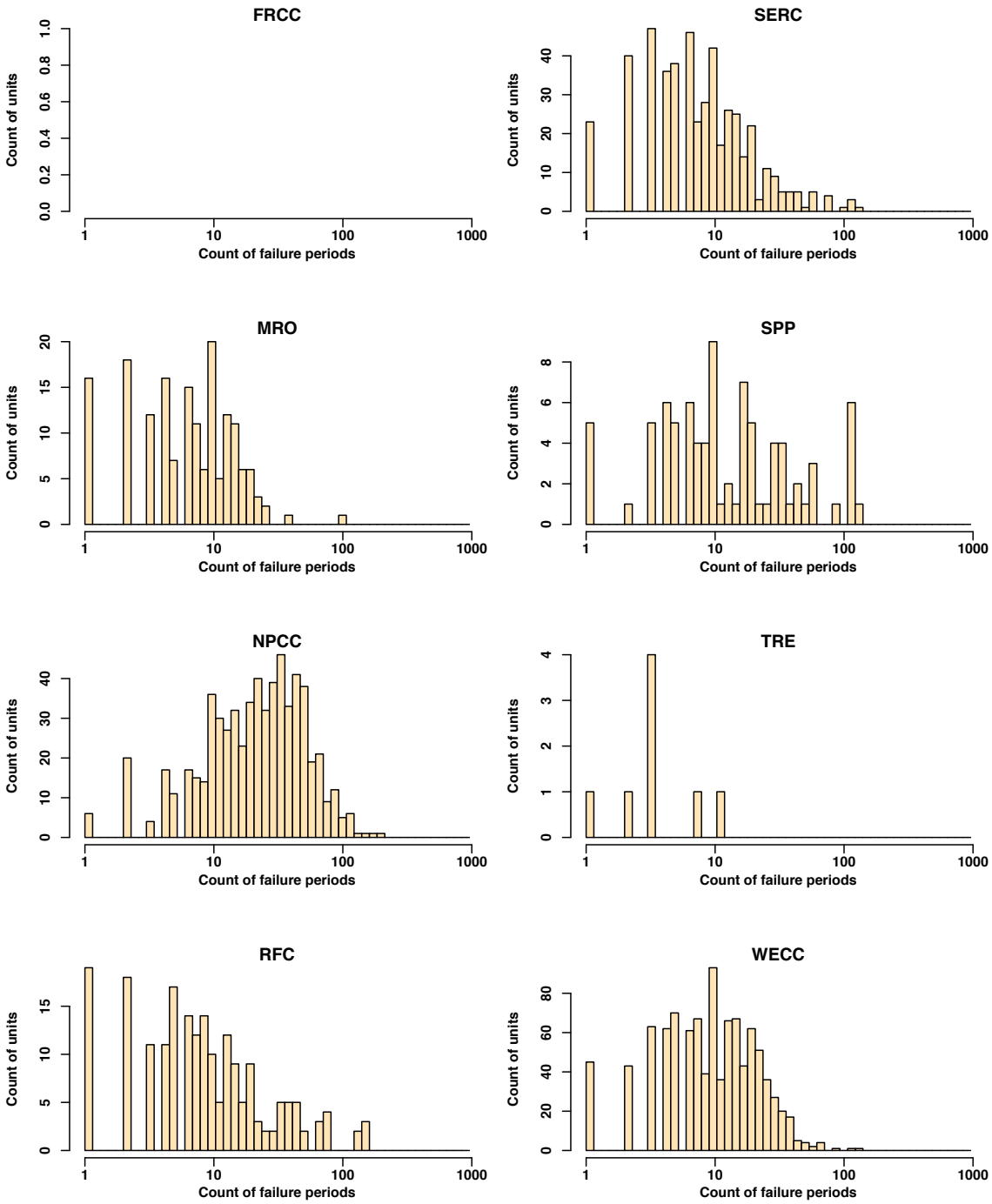

Figure A.45: Count of failure periods used to calculate mean time to recovery for hydroelectric units. FRCC has no such units.

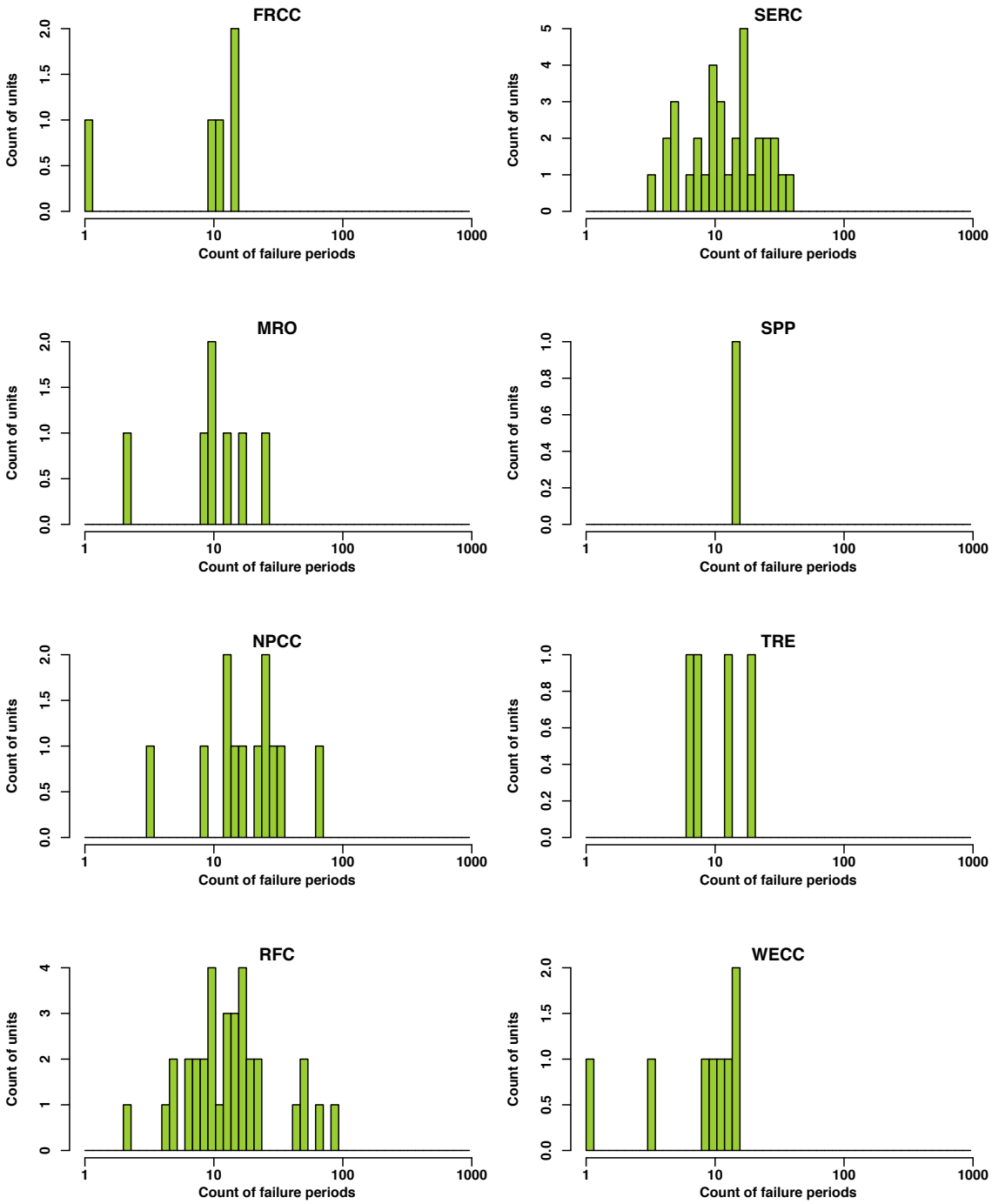

Figure A.46: Count of failure periods used to calculate mean time to recovery for nuclear units.

In section 2.4 of the main text, we sought to determine whether the MTBF of large and small units differed using the Mann-Whitney U test and the two-sample Kolmogorov-Smirnov test. For

each of the five unit types considered, we defined "large" units as those with nameplate capacities greater than or equal to the median value for that unit type, and "small" otherwise. We exclude units with significant discrepancies in RS reporting between the GADS Events and GADS Performance tables, which can lead to median nameplate values that differ slightly from those shown in Table 2.6 in the main text. Here we present the two-sample Kolmogorov-Smirnov test results (Table A.15). Results are highly consistent with the Mann-Whitney U test results (Table 2.5 in the main text). Note the test does not report directionality, unlike the Mann-Whitney U test; we presume directionality is consistent with those results.

Table A.15: Two-sample Kolmogorov-Smirnov test for statistically significant differences in MTBF between small and large units by region and unit type. Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

Type	**CC**	**CT**	**FSFB**	**HY**	**NU**
FRCC	****	–	***	N/A[1]	N/A[2]
MRO	–	****	***	–	N/A[2]
NPCC	–	****	–	****	–
RFC	*	****	****	***	**
SERC	***	****	***	***	–
SPP	–	***	***	****	N/A[2]
TRE	****	***	****	N/A[3]	N/A[2]
WECC	–	–	***	****	N/A[2]
Combined	–	****	****	****	–

Significance levels: – ≥ 0.1; * < 0.1; ** < 0.05; *** < 0.01; **** < 0.001

[1] FRCC has no hydroelectric units, so no test could be conducted.

[2] Five regions do not have any nuclear units in one size category, so no test could be conducted.

[3] There are not enough hydroelectric units in TRE to conduct this test.

As a complement to those results, we present histograms of the MTBF for small versus large units as Figure A.47 through Figure A.51. We again exclude units with significant discrepancies in RS reporting between the GADS Events and GADS Performance tables.

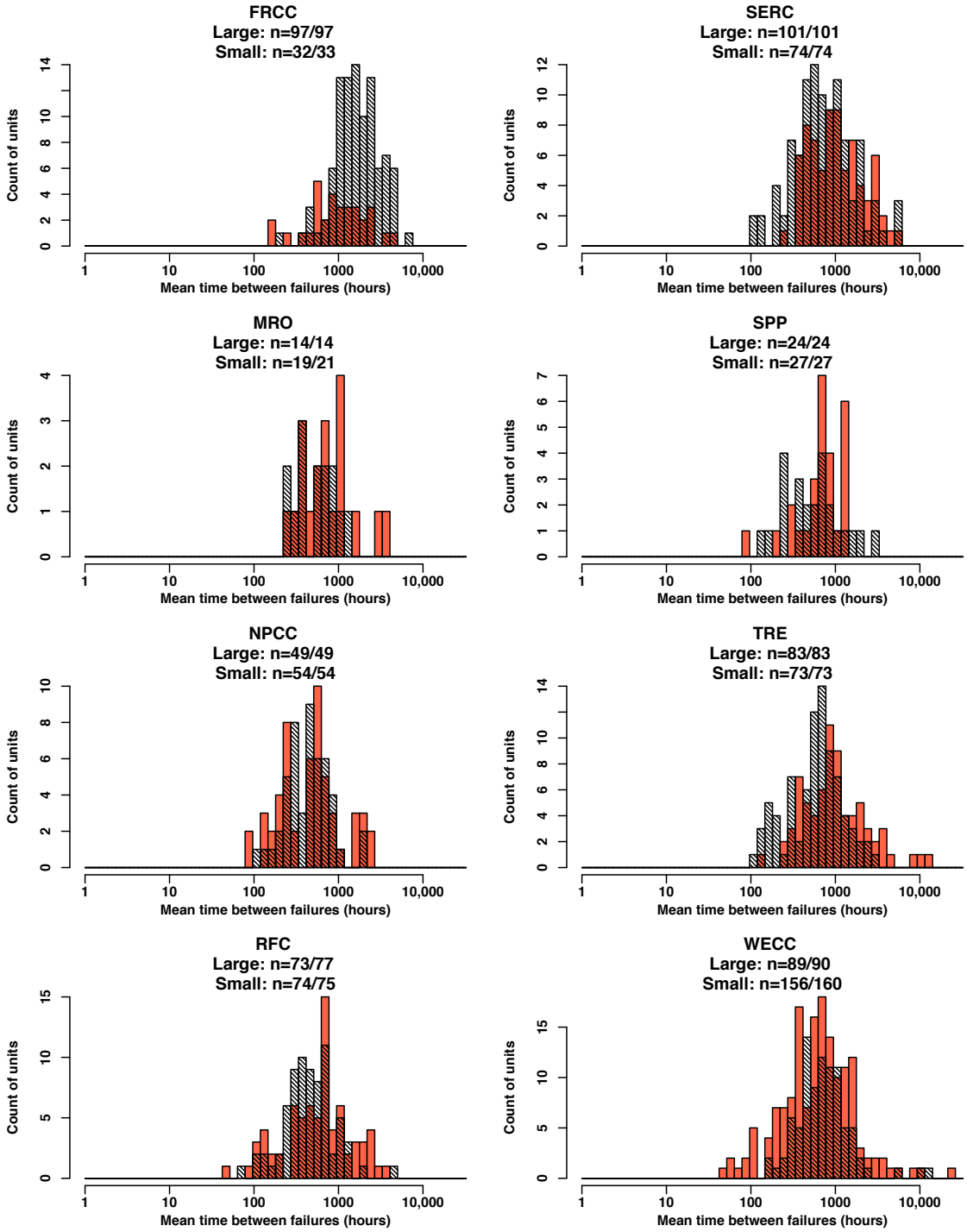

Figure A.47: Mean time between failure values for small (orange) versus large (black) combined cycle gas units; threshold is 185 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

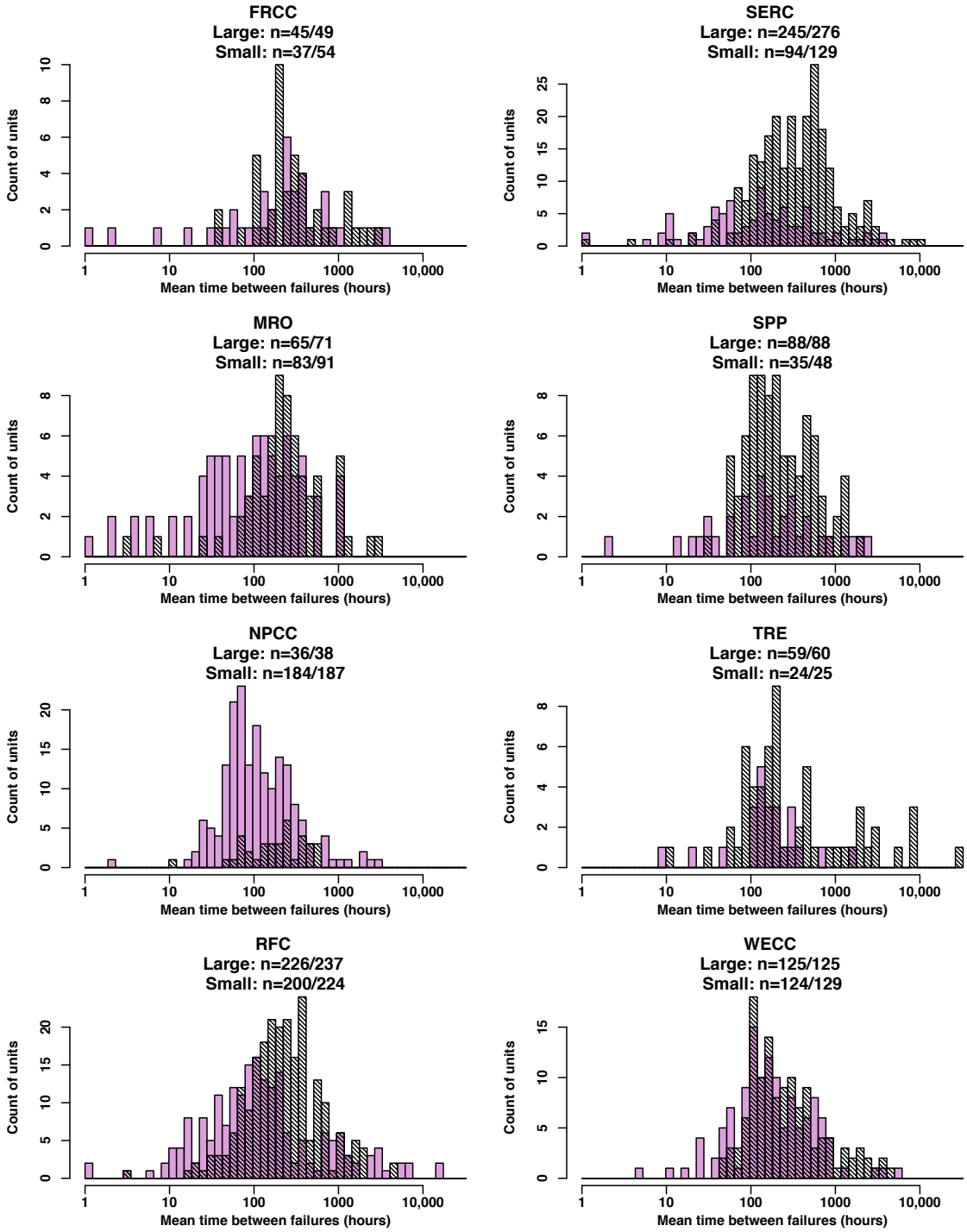

Figure A.48: Mean time between failure values for small (purple) versus large (black) simple cycle gas units; threshold is 57 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

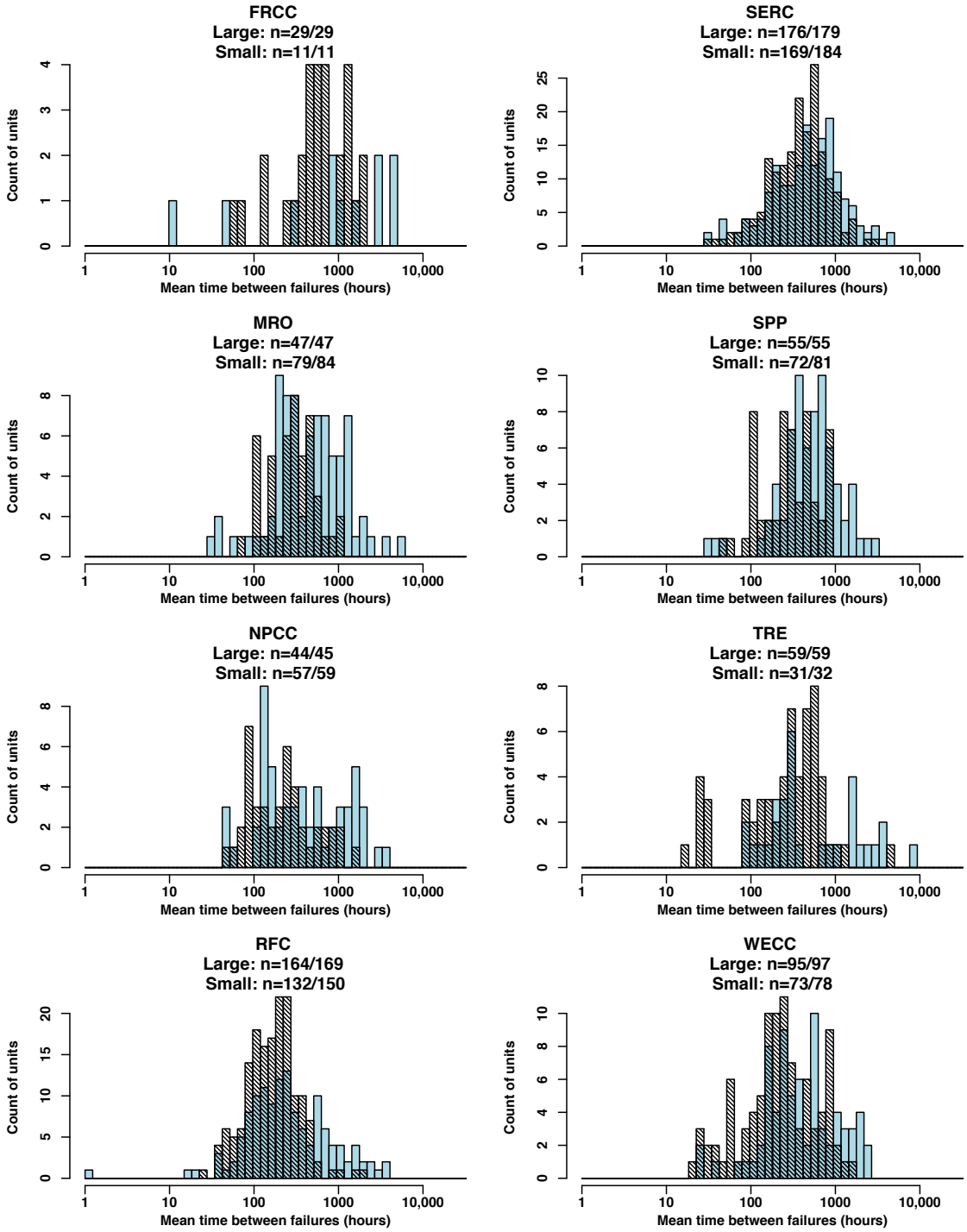

Figure A.49: Mean time between failure values for small (blue) versus large (black) fossil steam and fluidized bed units; threshold is 212 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

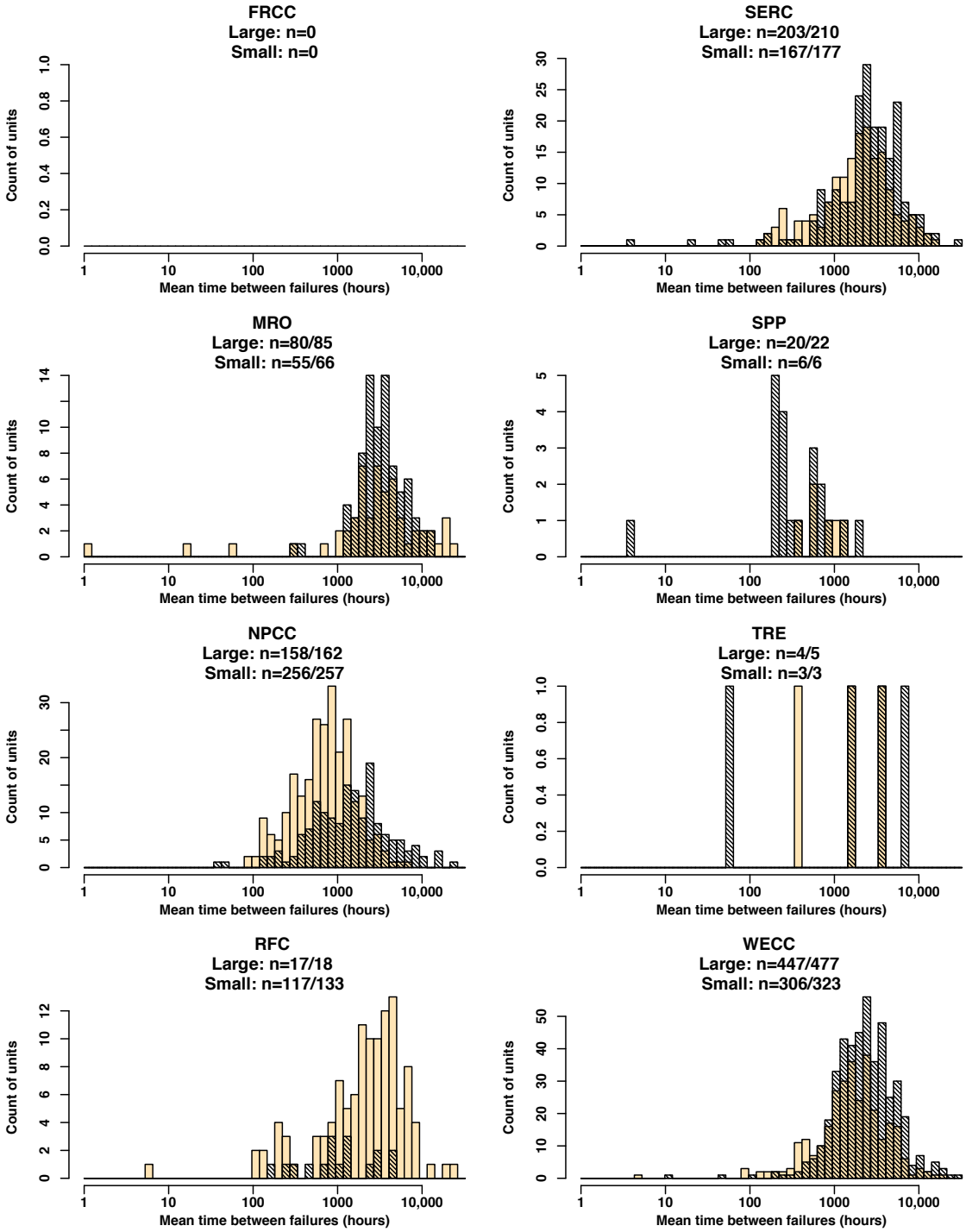

Figure A.50: Mean time between failure values for small (tan) versus large (black) hydroelectric units; threshold is 23 MW. FRCC has no such units. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). Units with significant reserve shutdown reporting discrepancies are excluded (see Table A.16).

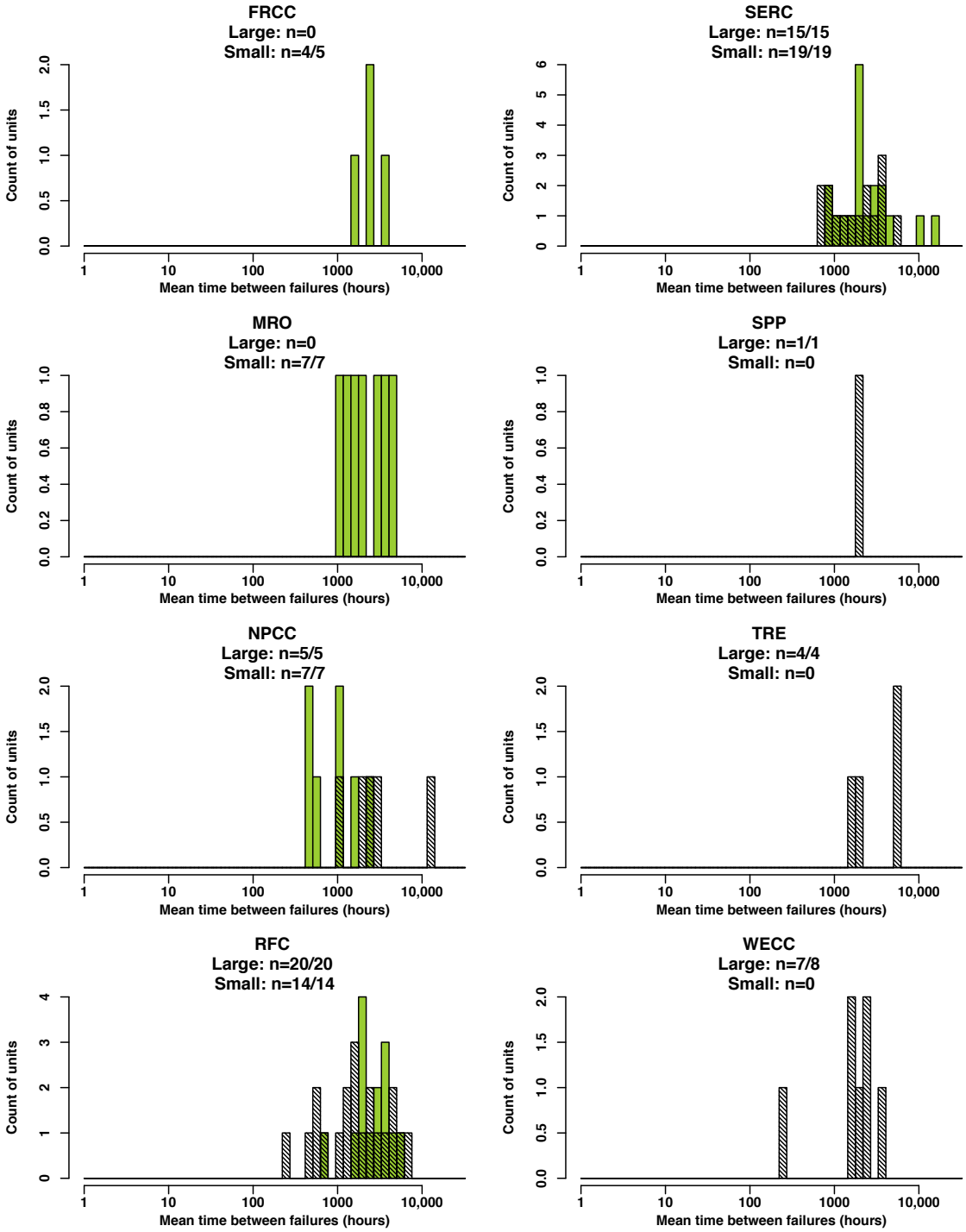

Figure A.51: Mean time between failure values for small (green) versus large (black) nuclear units; threshold is 1022 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event during the study period (proxy for total count of active units during the study period). No nuclear units had significant reserve shutdown reporting discrepancies (see Table A.16).

Some generator types are most typically employed as shoulder or peaking units, while others are most typically employed as baseload units. Non-baseload units will not always be required by a power system even when physically able to generate power, and these service requirements may vary by region due to market structure. Thus, when comparing the MTBF, both across and within unit types, it is important to consider only service hours. To do this, we remove RS hours from each unit's time series prior to calculating the MTBF. The validity of our MTBF analysis therefore hinges upon robust reporting of RS events. While RS reporting is mandatory for all units except those hydroelectric units without automatic data recording equipment, we sought to verify reporting prior to attempting the analysis [28]. We tabulate the total RS hours reported under the Events and Performance tables for each unit that reported an unscheduled event during our study period. Based on these results, we divide units into three categories: those that do not report RS events at all, those for whom total RS hours differ by 100 hours or less between the two tables, and those for whom total RS hours differ by more than 100 hours between the two tables. Unit counts are converted to installed capacity values and summarized by unit type in Table A.16.

Table A.16: Summary of capacity falling into each of three reserve shutdown reporting categories, by unit type. Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

Category	**CC**	**CT**	**FSFB**	**HY**	**NU**	**Total**
Capacity not reporting RS events	11,660	4,840	64,670	32,740	103,710	217,620
Capacity with discrepancy $>$ 100 hours	16,060	29,970	30,080	36,790	0	112,900
Capacity with discrepancy $\leq$ 100 hours	215,690	129,480	348,530	51,440	3,630	748,770
Percent of capacity included in MTBF analysis	93.4%	81.8%	93.2%	69.6%	100%	89.5%

The majority of capacity (70%) reports RS events with high fidelity, 10% of capacity has lower fidelity reporting, and 20% does not report RS events at all. In some cases the complete lack of RS reporting is not surprising. Large fossil steam units, nuclear units, geothermal units, and hydroelectric units could all have very low operating costs and never be "out of the money". (And, as mentioned previously, hydroelectric units without automatic data recording equipment are not required to report RS events to GADS.) In other cases—particularly for simple cycle gas turbines, diesel units, and most combined cycle units—this seems much more likely to be the result of incomplete data

reporting. However for the purposes of this analysis we only exclude those units that fall into the lower fidelity RS reporting category. Sensitivity analysis on the threshold used to delineate higher and lower fidelity RS reporting could also be done.

Chapter 3 A model of correlated generator failures and recoveries*

Abstract

Most current approaches to resource adequacy modeling assume that each generator in a power system fails and recovers independently of other generators. This assumption is inconsistent with recent empirical work. Here we present a statistical model that allows for correlated failures and recoveries. The probability of a transition is modeled as a time-dependent coin flip. Transition probabilities are a function of exogenous variables; as an example we use temperature and system load. We model the dependence of transition probabilities on temperature and load using logistic regression. The resulting time-varying probabilities imply a nonhomogeneous Markov model of generator failures and recoveries. Model parameters are estimated using 23 years of generator-level data for 1,845 generators in the USA's largest electricity market. Temperature and load covariates are found to be statistically significant for both generator failures and recoveries. Temperature dependencies are observed in all generator types, but are most pronounced for diesel and natural gas generators at low temperatures and nuclear generators at high temperatures. Load dependencies are observed in nuclear and coal generators at high temperatures. Our approach yields significant improvements in predictive performance compared to current practice, suggesting that explicit models of generator transitions using jointly experienced stressors should help resource planners more precisely manage their systems.

* Under peer review as Murphy, S, Sowell, F., & Apt, J. A time-dependent model of generator failures and recoveries captures correlated events and quantifies temperature dependence. Working paper CEIC-18-02 available at `https://ceic.tepper.cmu.edu/publications/working-papers`

3.1 Introduction

It is well known that severe environmental conditions can lead to elevated failure probabilities for exposed power system components [1–3]. Yet most current approaches to resource adequacy modeling (RAM) are unable to account for these risks because they assume that each generator fails and recovers with fixed probabilities, independently of other generators [1, 4–6].[1] This assumption is inconsistent with results from recent empirical work using four years of Generating Availability Data System (GADS) data from the North American Electric Reliability Corporation (NERC) [7]. The statistical evidence of dependence between generators leaves open the question of how to model correlated failures and recoveries.

Here we propose an answer with a nonhomogeneous (time-varying) Markov model. The nonhomogeneous Markov model's probabilities of transitioning, e.g. from fully available to derated, depend on exogenous variables. Many factors could affect transition probabilities. However, if transition probabilities depend on variables that are jointly experienced by many generators, such an approach could potentially capture the observed correlated failures. Understanding the causes of correlated failures and recoveries can help in the procurement of reserves, payments for which amount to billions of dollars per year in the USA [8].

Markov models are widely used in power system reliability analyses [9]. The traditional two-state model assumes generators are either fully available or fully unavailable with constant transition probabilities [10]. Common generalizations allow different two-state models over a discrete set of environments: e.g. "normal weather" versus "adverse weather", or generator "in demand" versus "not in demand" [11–14], where transition probabilities are constant over each environment. To model correlated failures, a new state must be created for each combination of generators failing simultaneously; the state space therefore grows geometrically as the number of generators increases [15, 16]. The intractability of applying this to a large power system has led researchers to define states in terms of system capabilities or to merge states [17, 18].

[1] Standard RAM practice in the USA is as follows. First, the most recent five years of historical availability data are used to calculate an availability statistic for each generator. Second, the availability statistics are combined to calculate a distribution of available capacity for a future planning year for the power system. RAM assumes that the availability statistic corresponds to the generator's probability of being unavailable due to an unscheduled failure in every hour of the planning year.

This paper takes a different approach to generalizing the two-state model: we employ only two states, but allow transition probabilities to depend on exogenous variables. To examine the suitability of such an approach, we create hourly time series of transitions for 1,845 generators in the eastern USA using 23 years of GADS data from the PJM Interconnection (PJM). For each generator, the two-state Markov model's time-varying probabilities are modeled as functions of exogenous variables using logistic regression. We model transition probabilities as a function of temperature and load, though the model can be extended to include additional covariates. Both temperature and load vary with time and are jointly experienced by many generators, thus transition probabilities in generators' Markov chains can be correlated.

This paper proceeds as follows. Section 3.2 describes our modeling approach. Section 3.3 describes the datasets and the steps we take to prepare them for analysis. Section 3.4 discusses our findings. Section 3.5 concludes and makes suggestions for future work.

3.2 Method

We use logistic regression to model each generator's transition probabilities as a function of covariates. We fit these models using the GLM library in R, with default initial values. While there are many binary classification algorithms, logistic regression is relatively insensitive to unbalanced data [19]. This is an important attribute for this analysis, as most generators fail infrequently. Unbalanced data makes accurately estimating transition probabilities more difficult [20].

We employ a two-state Markov model wherein each generator is treated as either fully available (subsequently referred to as available and abbreviated A) or at least partially unavailable (subsequently referred to as derated and abbreviated D). For each generator we separately model two pairs of transition probabilities: the probability of an available generator remaining available in the next hour versus becoming derated (failing), and the probability of a derated generator remaining derated in the next hour versus becoming available (recovering).

As in [21], we allow transition probabilities to be a function of covariates. We consider temperature and load because they have time series dependence and affect multiple generators simultaneously.

As a result, if they are found to have statistically significant associations with changes in transition probabilities, our model may be able to explain the correlated failures identified in [7]. If no covariates are statistically significant, this model reduces to the familiar homogeneous (time-invariant) Markov model of [9] (Figure 3.1).

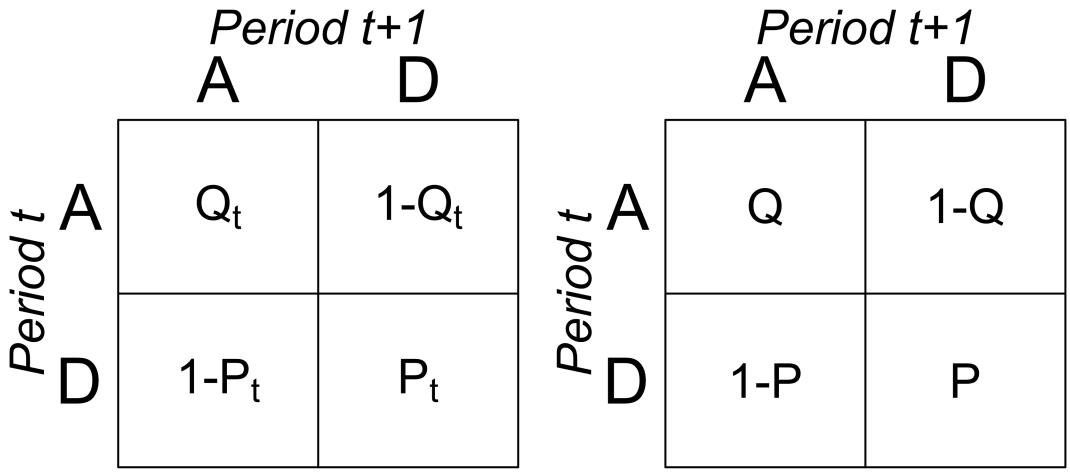

Figure 3.1: Nonhomogeneous (left) and homogeneous (right) two-state Markov models. A indicates the available state, D indicates the derated state. Q and Q_t are the constant and time-varying probabilities of an AA transition, respectively. P and P_t are the constant and time-varying probabilities of a DD transition, respectively.

We fit our models using maximum likelihood estimation (iteratively reweighted least squares). The estimation procedure is conducted on each generator, using its hourly series of Markov state transitions and covariate data, described below. If the transition probabilities were constant, this would be equivalent to determining the probability of a coin coming up heads. The likelihood functions are:

$$\mathcal{L}(\beta^A) = \prod_{i=1}^{count(A)} Q_i(\beta^A)^{AA_i} * (1 - Q_i(\beta^A))^{1-AA_i} \tag{3.1}$$

$$\mathcal{L}(\beta^D) = \prod_{i=1}^{count(D)} P_i(\beta^D)^{DD_i} * (1 - P_i(\beta^D))^{1-DD_i} \tag{3.2}$$

where β^D and β^D are vectors of parameters for the failure and recovery models, respectively, Q_i is

the probability of the generator remaining available in the next hour when it is currently available, P_i is the probability of the generator remaining derated in the next hour when it is currently derated, $count(A)$ is the sum of the number of available-to-available (AA) and available-to-derated (AD) transitions experienced by the generator, $count(D)$ is the sum of the number of derated-to-derated (DD) and derated-to-available (DA) transitions experienced by the generator, $AA_i = 1$ if the ith AA or AD transition is AA and 0 otherwise, and $DD_i = 1$ if the ith DD or DA transition is DD and 0 otherwise. The sum of $count(A)$ and $count(D)$ equals the total number of Markov state transitions in the period of reporting for the current generator. The failure and recovery models are fit separately for each generator (Figure 3.2).

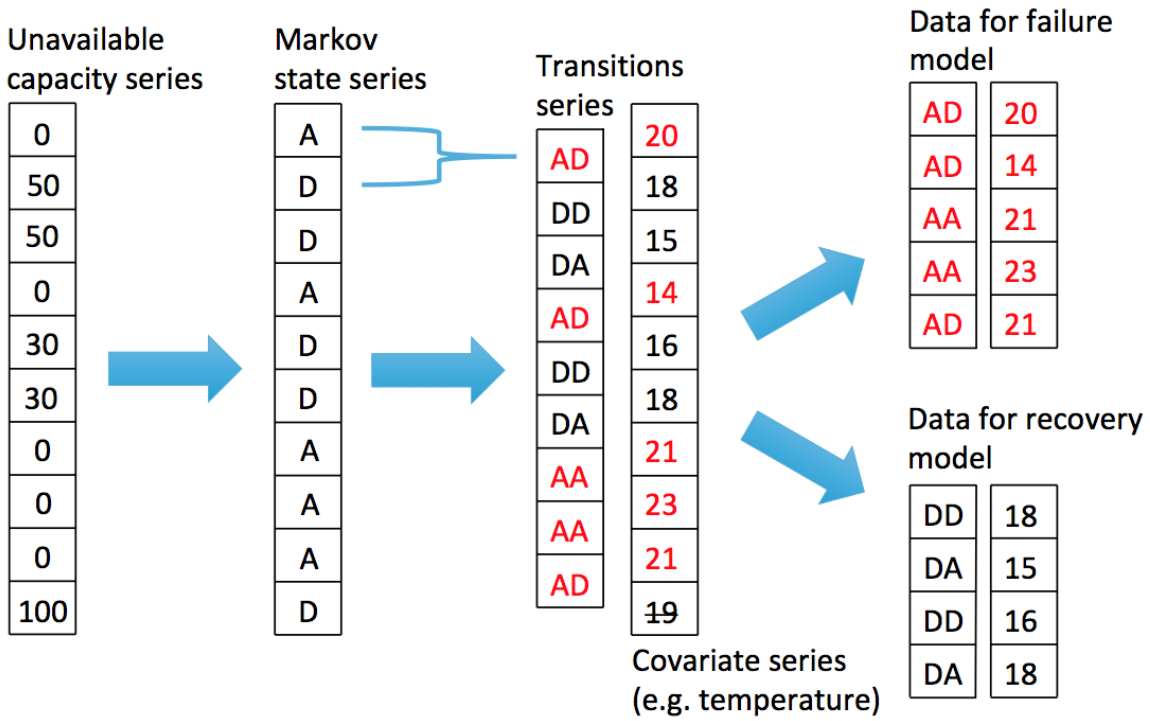

Figure 3.2: Defining a generator's time series of transitions and allocating them to the failure and recovery models. The generator's hourly time series of unavailable capacity is first used to determine which Markov state the generator is in in each hour. The series of hour-over-hour state transitions is then determined. These observations, along with our covariates (illustrated as a single vector of hourly temperatures for clarity of presentation) are then allocated to the failure and recovery models. Any observation in which a generator begins in the A state is assigned to the failure model, whereas any observation in which a generator begins in the D state is assigned to the recovery model. Note that there are one fewer transitions than original observations, so the final covariate observation is not used.

We allow Q_i and P_i to be functions of covariates while still ensuring all transition probabilities are bounded by $[0, 1]$ by employing the logistic function:

$$Q_i(\beta^A) = 1/(1 + \exp(-\beta^A X_i)) \tag{3.3}$$

$$P_i(\beta^D) = 1/(1 + \exp(-\beta^D X_i)) \tag{3.4}$$

where X_i is a vector of covariate observations in hour i, with as many elements as the number of constants and covariates in the model.

We consider the following model specification for both failure and recovery models for each generator:

$$\begin{aligned} Index_i = {} & \beta_1 * constantHot_i + \beta_2 * constantCool_i \\ & + \beta_3 * degreesHot_i + \beta_4 * degreesHot_i^2 + \beta_5 * degreesCool_i \\ & + \beta_6 * degreesCool_i^2 + \beta_7 * systemLoad_i \end{aligned} \tag{3.5}$$

where $\beta X_i = Index_i$ (linking Equations 3.3 through 3.5), $degreesHot_i = max(temperature_i - 18.3, 0)$, $degreesCool_i = max(18.3 - temperature_i, 0)$, $systemLoad_i$ is the load residual in hour i, $constantHot_i = 1$ if $temperatureCool_i = 0$ (and 0 otherwise), $constantCool_i = 1$ if $temperatureCool_i > 0$ (and 0 otherwise), and $temperature_i$ is the temperature in hour i, reported in degrees Celsius.[2] This specification allows for an asymmetric response to hot and cold temperature.

So that our model can better generalize to temperatures and loads not observed in the data, we employ stepwise regression (backward elimination) as described in Procedure 3.1, selecting a significance level of 0.05. To reduce bias, we then eliminate any generator having fewer than 10 DA or AD transitions per statistically significant model covariate [23].[3]

[2] 18.3 degrees Celsius is approximately 65 degrees Fahrenheit. This corresponds to the flattest region of the temperature-load relationship in the PJM area found by [22].

[3] DA or AD is always the least-experienced transition.

Procedure 3.1: Adaptive logistic regression model fitting:
For each generator, do:

- For each model (i.e., failure and recovery), do:
  - Fit full model specification (Equation 3.5)
  - While model has one or more linearly dependent or statistically insignificant covariates, do:
    - If model has one or more linearly dependent covariates, remove one linearly dependent covariate and re-estimate model
    - Else, remove covariate with smallest absolute value of t-statistic and re-estimate model
  - Save final model

Remove generators that do not have at least 10 AD and 10 DA transitions per final model parameter

3.3 Data

3.3.1 GADS data description

The GADS database records availability and design information for all generators serving the PJM control area, with the exception of wind, solar, and behind-the-meter generation. Reporting to GADS is mandatory, regardless of generator size [24]. We work primarily with the Events, Units, and Performance tables. The Events table reports any event affecting the ability of a generator to produce electricity, as well as other event types defined by the Institute of Electrical and Electronics Engineers (IEEE) Standard 762 [25]. The Units table reports design details of each generator, such as generator type and nameplate capacity.[4] The Performance table reports monthly summary statistics of each generator's operating and non-operating time. We analyze data from January 1, 1995 (database inception) through March 31, 2018. Over this period 1,845 generators representing 267 gigawatts (GW) of capacity have reported to GADS.

[4]The generator types include combined cycle gas (abbreviated as CC in figures and tables), simple cycle gas (CT), diesel (DS), hydroelectric and pumped storage (HD), nuclear (NU), and steam turbine (ST). In 2017, the vast majority (95%) of ST generation in PJM was from coal, thus we use the two terms interchangeably [26].

3.3.2 GADS data processing

Obtaining time series of availability state transitions

PJM's GADS database is virtually identical to that of NERC (albeit covering many more years), thus we prepare it for analysis as described in [7]. We calculate the magnitude of each derating event and then process events into time series of unavailable capacity. We restrict each generator's time series to complete calendar years. We then use each generator's time series of unavailable capacity to define a corresponding time series of hour-over-hour Markov state transitions (Figure 3.2). For example, an AA transition occurs when the generator is available in two adjacent hours.

Determining when a generator is available to transition

Our model assumes each generator is able to transition out of its current state in each hour (i.e., the generator can experience a failure if it is currently available and recover if currently derated). We attempt to exclude hours in which this assumption is violated in order to minimize bias. When fitting the failure model, we remove mothball, inactive reserve, and all scheduled outage events because the generator cannot be operating when these events are underway [27]. The generator can still operate when a scheduled derating is in effect, so these hours are not removed.

When fitting the recovery model, we remove only mothball and inactive reserve events. This is because no repair work is allowed to occur when these events are in progress [27]. Repair work on unscheduled failures can occur during scheduled outage and scheduled derating events, so these hours are not removed. In addition, some failures are catastrophic and take many months to repair. Including these events would bias recovery probabilities downward. To correct for this, we remove hours when a generator remains in the derated state without interruption for more than six months.

A note on reserve shutdown events

Reserve shutdown events are used to indicate when a generator is offline for economic reasons but is capable of coming online within its normal startup time if needed. With the exception of hydroelectric and pumped storage generators without automatic reporting equipment, all conventional

generators participating in the PJM market became obligated to report reserve shutdown events to GADS in January 2004, nine years after the beginning of our data.

When a reserve shutdown event is underway, a generator should neither be in service or have repair work conducted. If one assumes that the incidence of a failure while a generator is not operating and not being repaired is much lower than when operating or when being repaired, reserve shutdown hours should also be excluded from both failure and recovery model fits. However, given that most generators fail infrequently and that we require a minimum of 10 AD and DA transitions per statistically significant covariate to keep a generator in our analysis, eliminating the first nine years of data results in significantly fewer generators retained, particularly for CTs.

As a result, we fit our models twice: first using the full data period (1995-2018) ignoring reserve shutdown events, and second restricting to 2004-2018 and removing reserve shutdown hours from both failure and recovery model fits. Results based upon the former are presented in the main text and in Appendix B (figures and tables numbered B.1, B.2, etc.), while results based upon the latter are included in Appendix C (figures and tables numbered C.1, C.2, etc.). In general, we find reasonable agreement between the two sets of results.

Calculating the average derating magnitude for each generator

Because derating magnitudes can take any value up to a generator's nameplate capacity, but our model allows only one derated state, we calculate the average failure magnitude for each generator (Figure B.1). We calculate this as a duration-weighted average of all unscheduled events experienced by the generator, excluding any hour removed when fitting either the failure or recovery model. The average and median failure magnitudes are 78% and 96% of nameplate capacity, respectively.

3.3.3 Geographic, weather, and load data processing

Geocoding generators

To determine the location of each generator, we match the GADS data to the Emissions and Generation Resource Integrated Database (eGRID), maintained by the USA Environmental Protection Agency [28–31]. This task was completed using a combination of automated and manual

matching using generator names and other descriptive fields. We manually confirm each automated match and then associate the eGRID latitude/longitude data with the generator.

Weather station data

We obtain temperature data from the Global Surface Hourly database, maintained by the USA National Oceanic and Atmospheric Administration [32]. We include all weather stations active for the full study period in any state containing or adjacent to any generator. We process these data into hourly time series for each weather station by first rounding observations to the nearest hour and then removing observations with duplicate time stamps. We discard any weather station missing more than 100 sequential observations or more than 5,000 total observations over the 23 years, with three exceptions to increase coverage in Pennsylvania.[5] We then fill missing observations by propagating forward the most recent non-missing observation.[6] Finally, we link each generator to its nearest weather station meeting our data criteria. We map the retained generators and matched weather stations (Figure 3.3 and Figure C.1).

Load data

Finally we obtain hourly metered load data by PJM transmission zone for the full study period. We sum over all zones that have been part of the control area since January 1995 to develop an hourly load series for the system.[7] To account for non-stationarities in that series, we fit a time trend (Equation 3.6) and recover the residuals, which we use as the load signal experienced by each generator.

$$PredictedLoad_i = \beta_0^L + \beta_1^L * hour_i + \beta_2^L * hour_i^2 \tag{3.6}$$

[5]These three stations had 268, 65, and 103 sequential missing observations and 2937, 8962, and 1370 total missing observations.

[6]We initially filled missing observations by propagating forward the most recent non-missing observation at the same hour of the day, but discovered that several weather stations were systematically missing observations at particular times of the day over long durations.

[7]We include: Allegheny Power, Atlantic City Electric Company, Baltimore Gas and Electric Company, Delmarva Power and Light Company, Jersey Central Power and Light Company, Metropolitan Edison Company, PPL Electric Utilities Corporation, Pennsylvania Electric Company, Philadelphia Electric Company, Potomac Electric Power Company, and UGI.

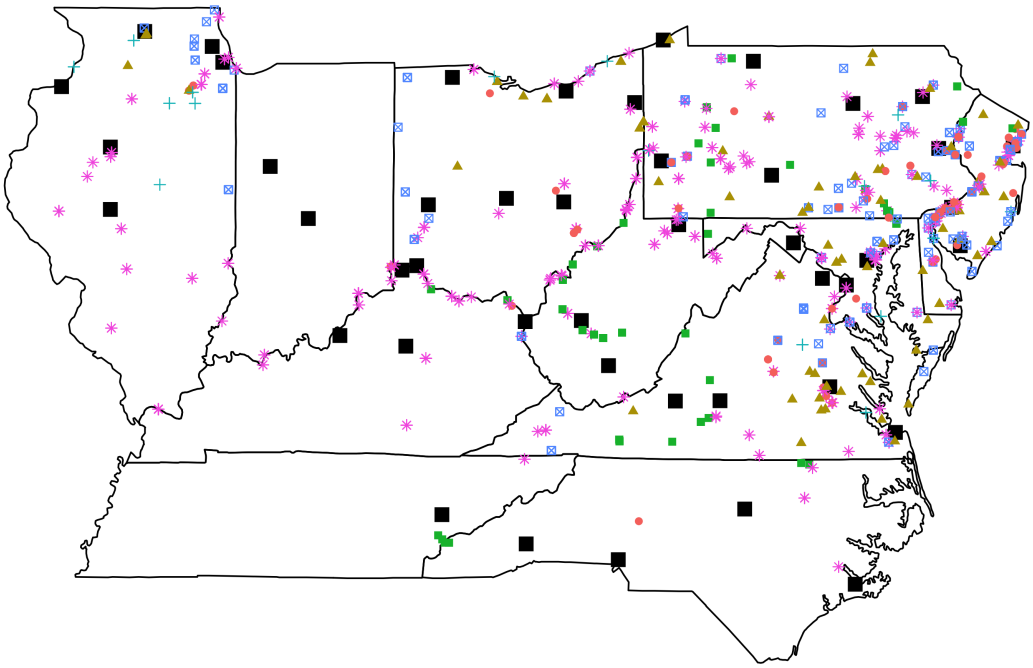

Figure 3.3: Locations of 1,111 retained generators and linked weather stations, overlaid on corresponding U.S. states (1995-2018 model fits). Only generators with at least 10 failure and recovery transitions per statistically significant model parameter are retained. All generators in multi-generator power plants have identical locations. Orange circles indicate combined cycle, blue squares with inset 'x' indicate simple cycle, yellow triangles indicate diesel, small green squares indicate hydroelectric and pumped storage, teal 'plus' signs indicate nuclear, purple asterisks indicate steam turbine, and large black squares indicate weather stations. A small number of retained generators are not shown for presentation considerations: Alabama (3), Louisiana (5), Michigan (23), Mississippi (3), South Carolina (1), Texas (8).

where β^L are the parameters for the load model. We plot the load time series with regression trend and residuals (Figure B.2 and Figure C.2).

3.3.4 Model significance summaries

When fitting models on the full dataset, we retain 1,111 of 1,845 generators, representing 78% of the capacity that has ever reported to GADS (Figure B.3); when restricting to 2004-2018, we retain 748 generators representing 67% of capacity (Figure C.3). While failures and recoveries for the remaining generators may indeed be influenced by temperature and/or load, they have so few

transitions that we would not have confidence in the fitted models. We summarize the count and capacity of these generators (Table 3.1 and Table C.1).

Generator type	Total count	Retained count (%)	Total capacity	Retained capacity (%)
CC	224	148 (66)	53.4	33.5 (63)
CT	663	274 (41)	44.9	16.5 (37)
DS	236	132 (56)	0.8	0.5 (58)
HD	244	125 (51)	11.0	8.3 (75)
NU	35	35 (100)	37.2	37.2 (100)
ST	443	397 (90)	119.5	113.1 (95)
All	1,845	1,111 (60)	266.8	209.0 (78)

Table 3.1: Summary of total and retained generator counts and capacity, by generator type (1995-2018 model fits). CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

We summarize marginal statistical significance of the covariates by plotting parameter t-values by generator type (Figures B.4-B.5 and Figures C.4-C.5) and reporting the number of times each model term is statistically significant at the 95% level by generator type (Tables 3.2-3.3 and Tables C.2-C.3). We include corresponding summaries of model coefficients (Figures B.6-B.7 and Figures C.6-C.7).

Generator type	Count	Mean hot	Mean cool	Temp hot	Temp hot^2	Temp cool	Temp cool^2	Load
CC	148	148	148	25	26	75	115	41
CT	274	274	274	59	53	110	203	228
DS	132	131	132	32	30	59	29	104
HD	125	125	125	16	14	22	35	43
NU	35	35	35	10	10	9	7	15
ST	397	397	397	70	61	103	134	285
All	1,111	1,110	1,111	212	194	378	523	716

Table 3.2: Number of times each model term is statistically significant at the 95% level for the failure model (1995-2018 model fits). Results reported for the 1,111 generating generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

When fitting on the full dataset, linear and quadratic hot-temperature variables are statistically significant for 19% and 17% of generators' failure models; linear and quadratic cold-temperature

Generator type	Count	Mean hot	Mean cool	Temp hot	Temp hot^2	Temp cool	Temp cool^2	Load
CC	148	147	148	38	35	61	52	100
CT	274	270	272	65	54	104	124	242
DS	132	131	130	40	31	56	67	113
HD	125	124	124	24	16	39	41	93
NU	35	35	35	13	11	12	6	10
ST	397	397	397	73	79	125	101	192
All	1,111	1,104	1,106	253	226	397	391	750

Table 3.3: Number of times each model term is statistically significant at the 95% level for the recovery model (1995-2018 model fits). Results reported for the 1,111 generating generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

variables are statistically significant for 34% and 47% of generators' failure models; and load is statistically significant for 64% of generators' failure models. For the recovery model, linear and quadratic hot-temperature variables are statistically significant for 23% and 20% of generators; linear and quadratic cold-temperature variables are statistically significant for 36% and 35% of generators; and load is statistically significant for 68% of generators.

We summarize the joint statistical significance of model covariates by creating scatterplots of parameter t-values between all non-orthogonal covariate pairs, excluding constants (Figures B.8-B.9 and Figures C.8-C.9). We observe systematic joint statistical significance between linear and quadratic temperature parameters in both sets of models, suggesting true temperature dependence rather than individual temperature parameters being significant by random chance. We include corresponding bivariate summaries of model coefficients (Figures B.10-B.11 and Figures C.10-C.11).

We report the number of statistically significant parameters for each generator (Table 3.4 and Table C.4). We report similar information when restricting attention to linear and quadratic temperature parameters (Table 3.5 and Table C.5). When fitting on the full dataset, 69% of generators have at least one statistically significant temperature covariate for the failure model; 67% do for the recovery model. These results demonstrate that temperature and load can have independent effects on transition probabilities. We compactly summarize variation in model predictions over the experienced covariate observations for each generator (Figure B.12 and Figure C.12).

Model	0	1	2	3	4	5	6	7
Failure	0	1	125	338	338	236	64	9
Recovery	0	2	130	366	294	235	72	12

Table 3.4: Number of statistically significant parameters (including constants) for the 1,111 generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models (1995-2018 model fits).

Model	0	1	2	3	4
Failure	346	341	323	84	17
Recovery	365	351	288	88	19

Table 3.5: Number of statistically significant temperature parameters (excluding constants) for the 1,111 generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models (1995-2018 model fits).

3.4 Results and discussion

3.4.1 Modeling correlated failures

Our descriptive results have demonstrated that temperature and load can predict state transitions at the generator level. We next use Monte Carlo simulation to demonstrate that the models can also predict correlated failures (Procedure 3.2). This procedure simulates time series of unavailable capacity for each generator according to the hourly failure and recovery probability distributions defined by the historical series of covariate values. Any hour that was ignored when fitting a generator's failure or recovery model is set to zero in both the empirical and simulated series. In order to have a true out-of-sample test of model performance, we refit the models using only 1995-2015 data (rather than 1995-2018) and retain just the 1,047 generators that have sufficient transitions over the shortened time series. This leaves 2016-2018 as test data. We carry out this procedure 5,000 times and generate pointwise median and 95% confidence intervals from the result, which we plot along with the empirical time series (Figure 3.4). We repeat the process fitting on

2004-2015 data, again leaving 2016-2018 as test data (Figure C.13). For reference, we report annual installed capacity values for these generators (Table 3.6 and Table C.6).

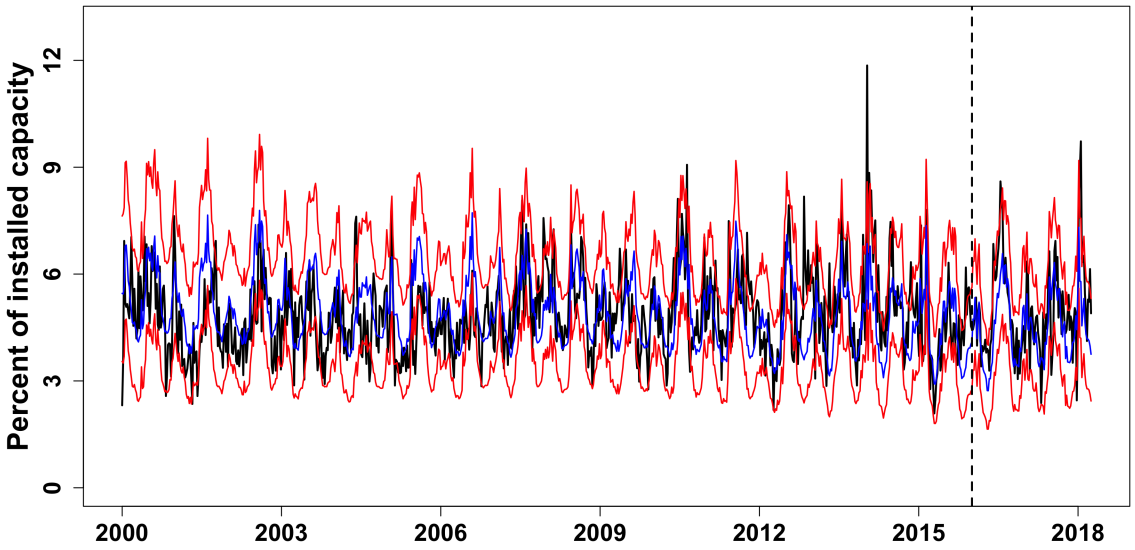

Figure 3.4: Simulated time series from logistic regression model (1995-2015 model fits). Results presented for 1,047 generators with at least 10 failure and recovery transitions per statistically significant model parameter when fitting on 1995-2015; 2016-2018 used as test of model performance. The split between training and testing periods is denoted with a dashed vertical line. Presented for 2000-2018 for consistency with Figure 3.5. Weekly averages rather than hourly series. 5000 simulations conducted. Refer to Table 3.6 for installed capacity by calendar year. Black trace is the empirical time series; blue trace is the concatenation of pointwise median simulation values; red traces are the concatenation of pointwise 2.5% and 97.5% simulation values.

Year	CC	CT	DS	HD	NU	ST	All
2000	4.2	7.1	0.1	3.3	19.5	69.2	103.5
2001	5.2	8.0	0.1	6.4	32.7	79.4	131.8
2002	6.3	9.8	0.1	6.4	32.7	80.4	135.7
2003	9.1	10.2	0.1	6.5	32.7	83.0	141.5
2004	15.6	11.1	0.1	7.2	34.9	83.8	152.9
2005	20.3	12.9	0.1	6.3	36.3	96.5	172.4
2006	21.2	12.9	0.2	5.9	36.3	96.6	173.1
2007	23.1	13.2	0.2	5.9	36.3	103.8	182.5
2008	24.9	13.4	0.3	6.0	36.3	103.3	184.1
2009	26.0	13.5	0.3	6.0	36.3	102.9	184.9
2010	25.8	13.6	0.3	6.0	36.3	103.1	185.1
2011	25.8	13.6	0.3	6.1	36.3	103.0	185.1
2012	28.1	13.3	0.4	6.1	36.3	95.7	179.8
2013	28.3	13.6	0.4	6.1	36.3	95.7	180.3
2014	29.1	13.0	0.4	6.1	36.3	94.0	178.9
2015	29.1	11.4	0.4	6.1	36.3	86.2	169.5
2016	29.1	10.9	0.4	6.1	36.3	85.2	167.9
2017	28.5	10.8	0.4	6.1	36.3	83.1	165.0
2018	28.5	10.8	0.4	6.1	36.3	82.9	164.9

Table 3.6: Installed capacity (GW) of 1,047 retained generators by year and generator type (1995-2015 model fits). For use with Figure 3.4 and Figure 3.5. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

Procedure 3.2: Simulating unavailable capacity from nonhomogeneous Markov models:
For each simulation, do:

- For each generator, do:
  - Initialize the state of the generator to match its reported state during its first hour of data reporting
  - For each subsequent hour of the generator's reporting period, do:
    - Use the current state of the generator and the current values of all model covariates to define the current transition probability distribution (i.e., over AA/AD if the generator is currently available; over DD/DA if the generator is currently unavailable)
    - Draw 0 or 1 using the probability distribution defined above, where 0 indicates the generator is available and 1 indicates the generator is unavailable
  - Replace all 1s with the generator's average unscheduled capacity reduction to yield a time series of unscheduled unavailable capacity
  - Zero out any unavailable capacity occurring during hours removed during model fitting
- Sum over generators' time series to obtain one simulated system-level time series

Compute desired quantiles from simulation results (e.g., 2.5%, 50%, 97.5%) and save

Even with our simple model specification, we find that the median simulation generally tracks the empirical time series quite well. The correlation between weekly average median simulation values and weekly average empirical values is 0.47 and 0.67 over the training and testing periods, respectively, for the 1995-2015 model fits and 0.47 and 0.69 during training and testing periods for the 2004-2015 fits. Furthermore, it is rare for an empirical event to exceed the upper confidence band. The largest instances of under-prediction by our model occurred during two known events where significant generator outages were due to causes not included as covariates: the 2014 Polar Vortex (due to fuel unavailability events, which increase non-linearly in cold weather) and Hurricane Sandy (an extreme weather event but not with regard to temperature). While many other factors may contribute to generator failures and recoveries [33–35], these results demonstrate that temperature and load are strongly correlated with system-level unavailable capacity dynamics.

We next compare the performance of our model to that of current RAM practice. This entails computing an availability statistic for each generator in each planning year (Equation 3.7), and then using those statistics in Monte Carlo simulations (Procedure 3.3).

We compute the equivalent forced outage factor ($EFOF$[8]) as follows:

$$EFOF = (FOH + EFDH)/PH \tag{3.7}$$

where FOH (forced outage hours) is the sum of hours where the generator experiences a forced outage, $EFDH$ (equivalent forced derating hours) is the sum of hours where the generator experiences a forced derating, reported in full-outage-equivalent hours, and PH (period hours) is the total number of hours in the period of interest. In accord with current RAM practice, we define the period supporting each planning year as the preceding five calendar years. For consistency with the logistic regression results, we carry out the procedure for the 1,047 generators retained when fitting models on 1995-2015 data and we ignore contributions to FOH and $EFDH$ that occur during any hour

[8] More commonly, the equivalent forced outage rate ($EFOR$) is used. $EFOR = (FOH + EFDH)/(FOH + SH + Synch + Pump + EFDHRS)$, where SH (service hours) is the total number of hours the generator produces electricity, $Synch$ is the number of hours the generator operates in synchronous condensing mode, $Pump$ is the number of hours a pumped-storage hydroelectric generator operates in pumping mode, and $EFDHRS$ (equivalent forced derating hours during reserve shutdown) is the number of hours the generator experiences a forced derating during a reserve shutdown event, reported in full-outage-equivalent hours [27]. However, using $EFOF$ allows us to not worry about incomplete reporting of reserve shutdown events prior to 2004.

removed during model fitting.

We plot the pointwise median and 95% confidence intervals from 5,000 simulations (Figure 3.5). As anticipated, the current practice approach does not capture system-level dynamics (e.g. correlated failures) because the distribution of unavailable capacity is the same in every hour of a given planning year. The correlation between weekly average median simulation values and weekly average empirical values is 0.15 over 18 years and 0.11 during the testing period.[9] In addition, the pointwise 95% confidence intervals are wider, averaging 5% of installed capacity over 18 years compared to 3.1% of installed capacity for the logistic regression model. We summarize the divergence between the empirical series and the upper bound of the 95% pointwise confidence interval in each hour (i.e., the magnitude the hourly overpredictions and underpredictions) for the logistic regression model and current practice in Figure B.13.

Procedure 3.3: Simulating unavailable capacity from homogeneous Markov models:
Define duration of data period supporting each planning year (e.g., five years)
For each simulation, do:

- For each planning year (e.g., 2000-2018), do:
  - For each generator, do:
    - If the generator was active during period supporting current planning year and does not retire prior to planning year, do:
      - Compute $EFOF$ (Equation 3.7) using all of generator's data supporting current planning year, except for hours removed during model fitting
      - For each hour in planning year, draw a 1 with probability equal to generator's $EFOF$ and 0 otherwise, where 0 indicates the generator is available and 1 indicates the generator is unavailable
      - Replace all 1s with the generator's nameplate capacity
  - Sum over generators' time series to obtain one simulated system-level series for current planning year

Compute desired quantiles from simulation results (e.g., 2.5%, 50%, 97.5%) and save

[9]Note that the predictions of the current practice model are always out of sample, in contrast with those of the logistic regression model prior to 2016.

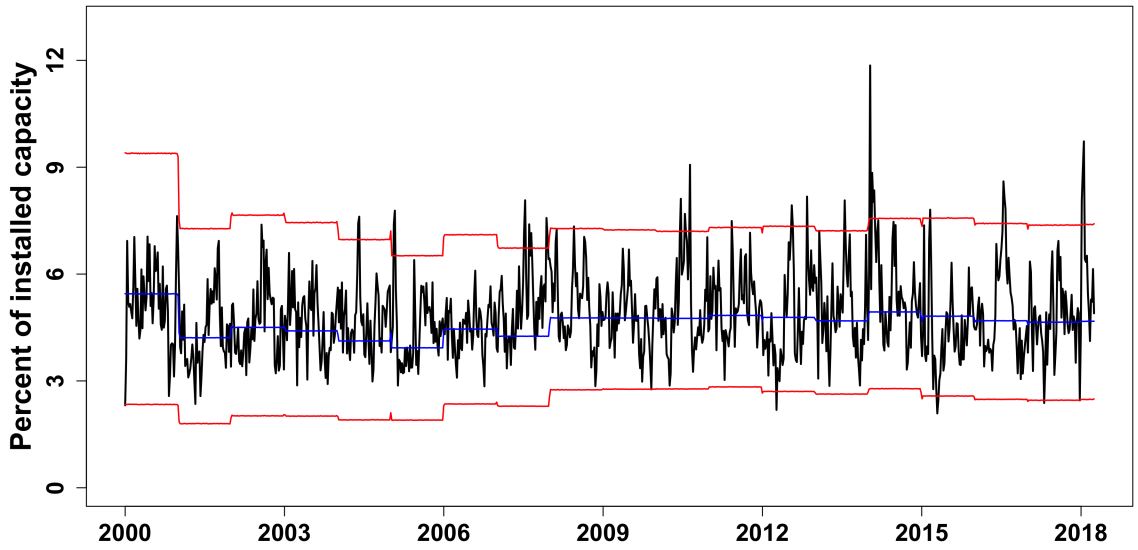

Figure 3.5: Simulated time series from current practice model (1995-2015 model fits). Results presented for the same set of 1,047 generators as in Figure 3.4. Time series restricted to 2000-2018 because five years of data are used to calculate the availability statistic. Traces are flat within each calendar year because current practice model assumes failure probabilities are constant in each hour of a given year. Small discontinuities at year boundaries are due to weekly averaging not respecting calendar year boundaries, in conjunction with capacity additions and retirements occurring at the start of the year. Weekly averages rather than hourly series. 5000 simulations conducted. Refer to Table 3.6 for installed capacity by calendar year. Black trace is the empirical time series; blue trace is the concatenation of pointwise median simulation values; red traces are the concatenation of pointwise 2.5% and 97.5% simulation values.

3.4.2 Characterizing resource adequacy risk as a function of temperature and load

We next examine resource adequacy risks for the modeled fleet. For fixed values of temperature and load, each generator's failure and recovery models imply a stationary distribution over the available and derated states. We make use of this fact to determine the proportion of the time each generator is unavailable in expectation. By calculating this result over a range of temperature values, we determine expected unavailable capacity as a function of temperature for the modeled fleet (Procedure 3.4). We determine the analogous result under current modeling practice by first computing an unconditional transition probability matrix for each generator using all available years of data and then following the remainder of the inner loop of Procedure 3.4. We present results by generator type (Figure 3.6 and Figure C.14). We report the prevalence of temperatures experienced

by the fleet of modeled generators (Figure B.14 and Figure C.15).

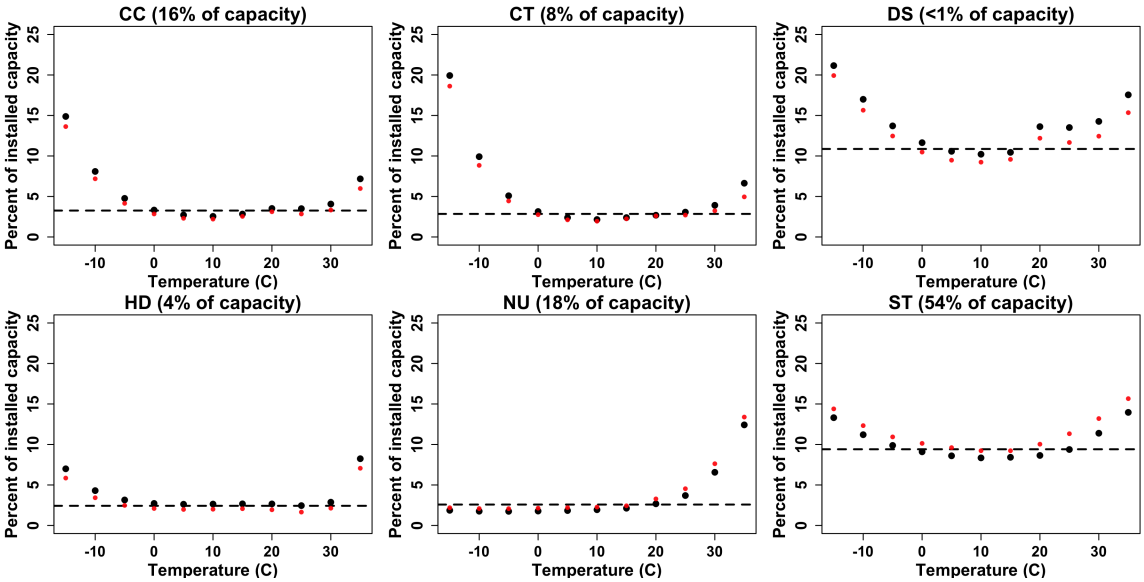

Figure 3.6: Expected levels of unavailable capacity as a function of temperature under logistic regression (dots) and current practice (dashed horizontal line) (1995-2018 model fits). Black dots calculated using median load from temperature neighborhood, red dots calculated using 90^{th} percentile load from temperature neighborhood. Temperature neighborhood is defined as +/- 10 degrees. Not all generators experience full temperature range; see Figure B.14 for prevalence of temperatures. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

Procedure 3.4: Characterizing unavailable capacity as a function of temperature

For each desired quantile of load (e.g., 50^{th}, 90^{th}), do:

- For each desired temperature value (e.g., spanning the range of temperatures experienced by the fleet, in 5-degree intervals), do:
  - Fix the value of temperature
  - Fix the value of load at the current load quantile, calculated on observations within the "neighborhood" of the current temperature value (e.g., within +/- 10 degrees)
  - For each generator, do:
    - Compute predicted transition probabilities using generator's failure and recovery model and current temperature and load values
    - Define transition probability matrix as the transpose of Figure 3.1
    - Normalize the first eigenvector of the eigendecomposition of the transition probability matrix to obtain the proportion of the time the generator is unavailable in expectation
    - Multiply result by generator's nameplate capacity and its average unscheduled capacity reduction to obtain expected unavailable capacity
    - Compute variance of expected unavailable capacity by treating generator as a Bernoulli random variable
  - Sum expected unavailable capacity values over generators and save
  - Sum variance values over generators and use to calculate desired confidence interval (e.g., 95%) and save

With the exception of nuclear, all generator types perform worse in very cold weather than recognized under current modeling practice. This result is consistent with analysis conducted by PJM [2]. Poor cold-weather performance is particularly pronounced for gas and diesel generators. In addition, all generator types perform worse in very hot weather than recognized under current practice. Because loads are high at both temperature extremes, the resource adequacy risk implied by these performance penalties is compounded: less generation capacity is available when demand is greatest.

Finally, we repeat the preceding analysis switching the role of temperature and load in order to visualize resource adequacy risk as a function of load. Because the relationship between load and unavailable capacity could be different at high and low temperatures, we generate two sets of results: one for observations where the temperature is below 18.3 degrees, and one for observations where temperature is above 18.3 degrees. With these modifications, we repeat Procedure 3.4. We again present results by generator type (Figures 3.7-3.8 and Figures C.16-C.17).

In Figure 3.7, at median temperature values, only coal generators at very high loads show

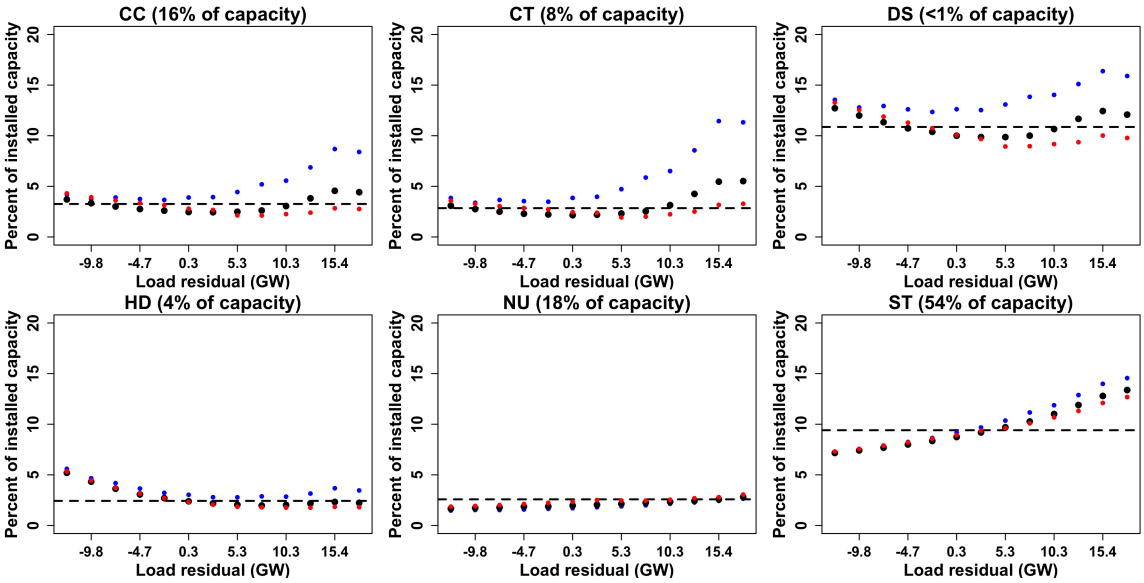

Figure 3.7: Expected levels of unavailable capacity as a function of load under logistic regression (dots) and current practice (dashed horizontal line) when restricting to temperatures below 18.3 degrees Celsius (1995-2018 model fits). Black dots computed at median temperature from load neighborhood; blue and red dots correspond to 10^{th} and 90^{th} percentile temperatures from load neighborhood, respectively. Load neighborhood defined analogously to temperature neighborhood of Figure 3.6. Current practice dashed line matches that of Figure 3.6. Plot domain defined using only observations below 18.3 degrees. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

noticeable divergence from the unconditional level of unavailable capacity. When considering low-percentile temperatures, gas and diesel generators also exhibit divergence from the unconditional result at higher loads. Nuclear generators show no load response for cold-temperature observations, regardless of load level or temperature quantile, consistent with Figure 3.6. In Figure 3.8, coal and nuclear generators diverge from their respective unconditional levels of unavailable capacity at high loads regardless of temperature percentile considered. Diesel generators show some divergence at very low loads.

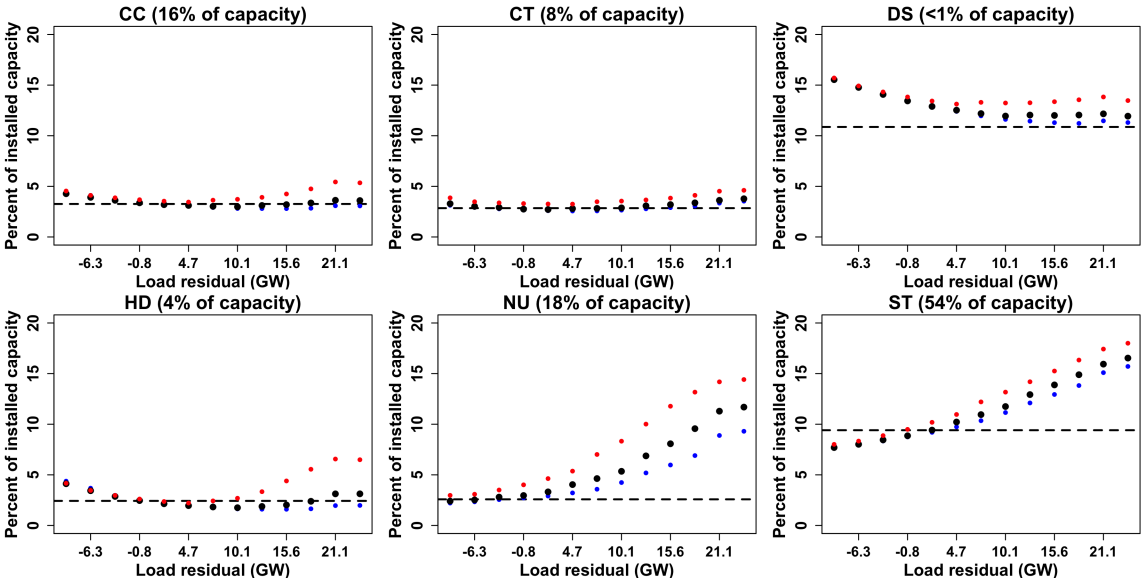

Figure 3.8: Expected levels of unavailable capacity as a function of load under logistic regression (dots) and current practice (dashed horizontal line) when restricting to temperatures above 18.3 degrees Celsius (1995-2018 model fits). Black dots computed at median temperature from load neighborhood; blue and red dots correspond to 10^{th} and 90^{th} percentile temperatures from load neighborhood, respectively. Load neighborhood defined analogously to temperature neighborhood of Figure 3.6. Current practice dashed line matches that of Figure 3.6. Plot domain defined using only observations above 18.3 degrees. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

3.5 Conclusions

In this paper we present a model of how correlated failures previously identified in the North American power system can occur [7]. Our approach is a novel, computationally tractable generalization of the traditional two-state Markov model widely used in power system reliability analyses [9]. We demonstrate a simple specification in which transition probabilities between the available and derated states are modeled as a function of temperature and load, but note that any desired covariates could be employed.

We fit these models using logistic regression with 23 years of availability data for 1,845 generators serving the PJM control area. To reduce bias, we discard any generator with fewer than 10 failure or recovery events per statistically significant covariate. We retain 78% of the generation capacity that has ever reported to PJM GADS. We find that temperature and load can predict generator

transitions: temperature and load are each statistically significant for two-thirds of the retained generators.

We demonstrate that our model specification captures most of the correlated failures observed in PJM since 2000 and that it significantly outperforms the homogeneous Markov model underlying current resource adequacy modeling practice. The correlation of our median simulation with the observed series of unavailable capacity at the weekly level is 0.47 over 18 years, whereas that of the median simulation from current practice is 0.15. Our model also has narrower confidence intervals, averaging 3.1% of installed capacity compared to 5% for current practice.

We demonstrate that all generator types are susceptible to extreme temperatures. With the exception of nuclear generators, which have reduced availability only during hot weather, all generator types have reduced availability at both temperature extremes. The cold-weather penalty for gas and diesel generators is particularly pronounced, as is the hot-weather penalty for nuclear generators. Finally, we demonstrate that nuclear and coal generators experience an availability penalty at high loads; for nuclear generators this penalty is present only in conjunction with high temperatures. These risks are not captured in current approaches to resource adequacy modeling.

Taken together, our results demonstrate that there are systematic relationships between temperature, load, and generator availability. Accounting for these relationships, as we have done here, may enable more efficient procurement of reserve capacity. Future work should examine the specific causes of the temperature dependence of generator availability and what improvements in reserves procurement can be achieved now that correlated failures can be successfully modeled.

References

[1] D. Gaver, F. Montmeat, and A. Patton, "Power system reliability I: Measures of reliability and methods of calculation," *IEEE Transactions on Power Apparatus and Systems*, vol. 83, no. 7, pp. 727–737, 1964.

[2] The PJM Interconnection, "Problem statement on PJM Capacity Performance definition," 2014. Draft report.
`https://www.pjm.com/-/media/library/reports-notices/capacity-performance/20140801-problem-statement-on-pjm-capacity-performance-definition.ashx`.

[3] The PJM Interconnection, "PJM cold snap performance: December 28, 2017 to January 7, 2018," 2018. https://www.pjm.com/-/media/library/reports-notices/weather-related/20180226-january-2018-cold-weather-event-report.ashx.

[4] R. Billinton and K. Bollinger, "Transmission system reliability evaluation using Markov processes," *IEEE Transactions on Power Apparatus and Systems*, no. 2, pp. 538–547, 1968.

[5] R. Billinton and R. Allan, *Reliability evaluation of power systems*. Plenum, second ed., 1996.

[6] Energy Management Associates, Inc., "Reliability calculations for interdependent plant outages," 1984. EPRI report EL-3669.

[7] S. Murphy, J. Apt, J. Moura, and F. Sowell, "Resource adequacy risks to the bulk power system," *Applied Energy*, vol. 212, pp. 1360–1376, 2018.

[8] K. Spees, S. Newell, and J. Pfeifenberger, "Capacity markets - Lessons learned from the first decade," *Economics of Energy and Environmental Policy*, vol. 2, 2013.

[9] R. Billinton and R. Allan, *Reliability evaluation of engineering systems*. Springer Science+Business Media, LLC, second ed., 1992.

[10] IEEE Probability Application for Common Mode Events Working Group of the Reliability Risk and Probability Applications Subcommittee, "Research on common-mode and dependent (CMD) outage events in power systems: A review," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1528–1536, 2017.

[11] M. Bhavaraju, J. Hynds, and G. Nunan, "A method for estimating equivalent forced outage rates of multistate peaking units," *IEEE Transactions on Power Apparatus and Systems*, vol. 6, pp. 2067–2075, 1978.

[12] R. Billinton and A. Chowdhury, "A model for peaking units using the Canadian Electrical Association data base," *IEEE Transactions on Power Apparatus and Systems*, vol. 104, no. 11, pp. 2972–2979, 1985.

[13] R. Billinton, "Basic models and methodologies for common mode and dependent transmission outage events," in *Proceedings of the IEEE Power and Energy Society General Meeting, San Diego, CA, USA*, pp. 1–8, 2012.

[14] IEEE Task Group on Models for Peaking Service of the Application of Probability Methods Subcommittee, "A four-state model for estimation of outage risk for units in peaking service," *IEEE Transactions on Power Apparatus and Systems*, vol. 91, pp. 618–627, 1972.

[15] Probability Applications for Common Mode Events Task Force of the IEEE PES Risk, Reliability and Probability Applications (RRPA) Subcommittee, "Overview of common mode outages in power systems," in *Proceedings of the IEEE Power and Energy Society General Meeting, San Diego, CA, USA*, pp. 1–8, 2012.

[16] W. Li and R. Billinton, "Common cause outage models in power system reliability evaluation," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 966–968, 2003.

[17] F. Felder, "Top-down composite modeling of bulk power systems," *IEEE Transactions on Power Systems*, vol. 20, no. 3, pp. 1655–1656, 2005.

[18] K. Hou, H. Jia, X. Xu, Z. Liu, and Y. Jiang, "A continuous time Markov chain based sequential analytical approach for composite power system reliability assessment," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 738–748, 2016.

[19] S. Crone and S. Finlay, "Instance sampling in credit scoring: An empirical study of sample size and balancing," *International Journal of Forecasting*, vol. 28, no. 1, pp. 224–238, 2012.

[20] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, pp. 137–163, 2001.

[21] S. Ertekin, C. Rudin, and T. McCormick, "Reactive point processes: A new approach to predicting power failures in underground electrical systems," *Annals of Applied Statistics*, vol. 9, no. 1, pp. 122–144, 2015.

[22] R. Lueken, J. Apt, and F. Sowell, "Robust resource adequacy planning in the face of coal retirements," *Energy Policy*, vol. 88, pp. 371–388, 2016.

[23] P. Peduzzi, J. Concato, E. Kemper, T. Holford, and A. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373–1379, 1996.

[24] Integ Enterprise Consulting, *PowerGADS 3.0 User Manual*, 2015. `https://www.pjm.com/~/media/etools/egads/egads-user-guide.ashx`.

[25] IEEE Power Engineering Society, *Standard 762-2006: Standard definitions for use in reporting electric generating unit reliability, availability, and productivity*, 2007.

[26] Monitoring Analytics, "2017 State of the Market Report for PJM," 2018. `https://www.monitoringanalytics.com/reports/PJM_State_of_the_Market/2017.shtml`.

[27] North American Electric Reliability Corporation, *Generating Availability Data System: Data reporting instructions*, 2018. `https://www.nerc.com/pa/RAPA/gads/Pages/Data%20Reporting%20Instructions.aspx`.

[28] United States Environmental Protection Agency, "Emissions and Generation Resource Integrated Database (eGRID)," 1996. `https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid`. Accessed September 3, 2017.

[29] United States Environmental Protection Agency, "Emissions and Generation Resource Integrated Database (eGRID)," 2000. `https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid`. Accessed September 3, 2017.

[30] United States Environmental Protection Agency, "Emissions and Generation Resource Integrated Database (eGRID)," 2014. `https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid`. Accessed September 3, 2017.

[31] United States Environmental Protection Agency, "Emissions and Generation Resource Integrated Database (eGRID)," 2016. https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid. Accessed July 24, 2018.

[32] NOAA National Centers for Environmental Information, "Global Surface Hourly (DS3505)," 2001. https://www7.ncdc.noaa.gov/CDO/cdopoemain.cmd?datasetabbv=DS3505andcountryabbv=andgeoregionabbv=andresolution=40. Accessed May 28, 2018.

[33] North American Electric Reliability Council, "Predicting unit availability: Top-down analyses for predicting electric generating unit availability," 1991. https://www.nerc.com/pa/RAPA/gads/Publications/Predicting-Unit-Availability.pdf.

[34] North American Electric Reliability Council, "Seasonal performance trends: Peak season equivalent forced outage rate trend evaluations for fossil-steam generating units," 1991. https://www.nerc.com/pa/RAPA/gads/Publications/Seasonal-Performance-Trends.pdf.

[35] North American Electric Reliability Council, "Predicting generating unit reliability: A methodology for predicting generating unit reliability based on design characteristics, operational factors, and maintenance and plant betterment activities," 1995. https://www.nerc.com/pa/RAPA/gads/Publications/Predicting-Generating-Unit-Reliability.pdf.

3.6 Appendix B: Supplementary figures for 1995-2018 model fits

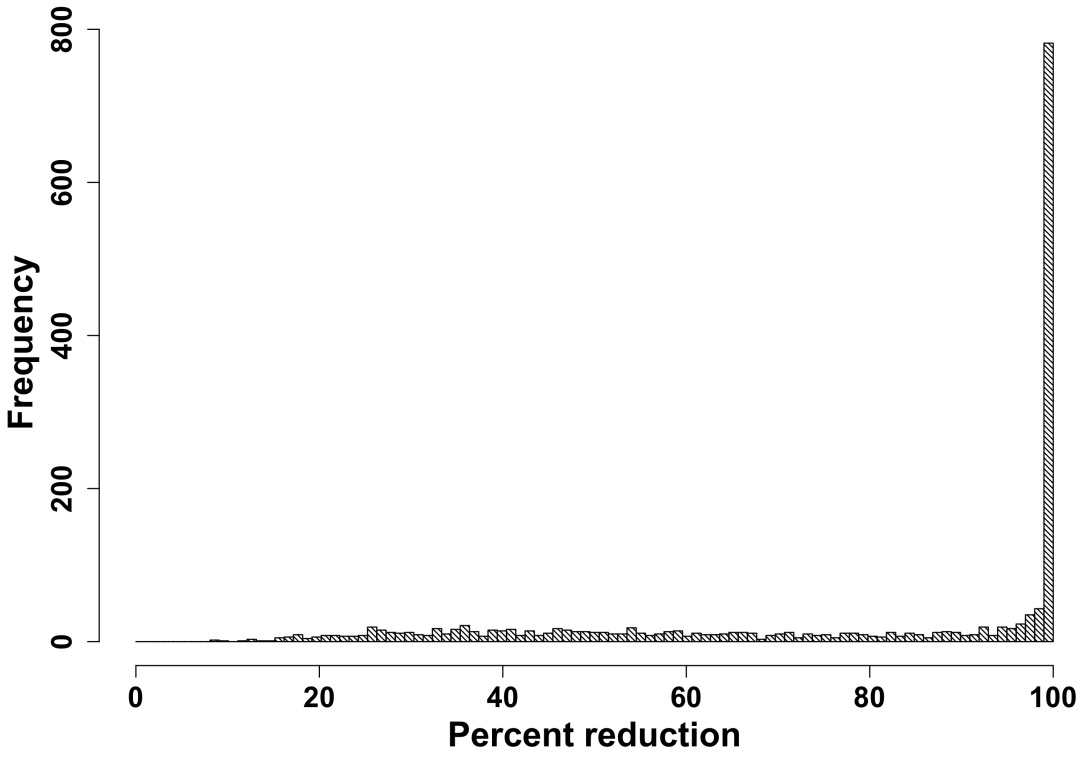

Figure B.1: Histogram of average generator failure magnitudes. Calculated for the 1,748 generators with at least one full calendar year of data reporting and at least one unscheduled transition during 1995-2018. Values calculated as a duration-weighted average of the magnitudes all unscheduled events experienced by the generator, excluding any hour removed when fitting either the failure or recovery model.

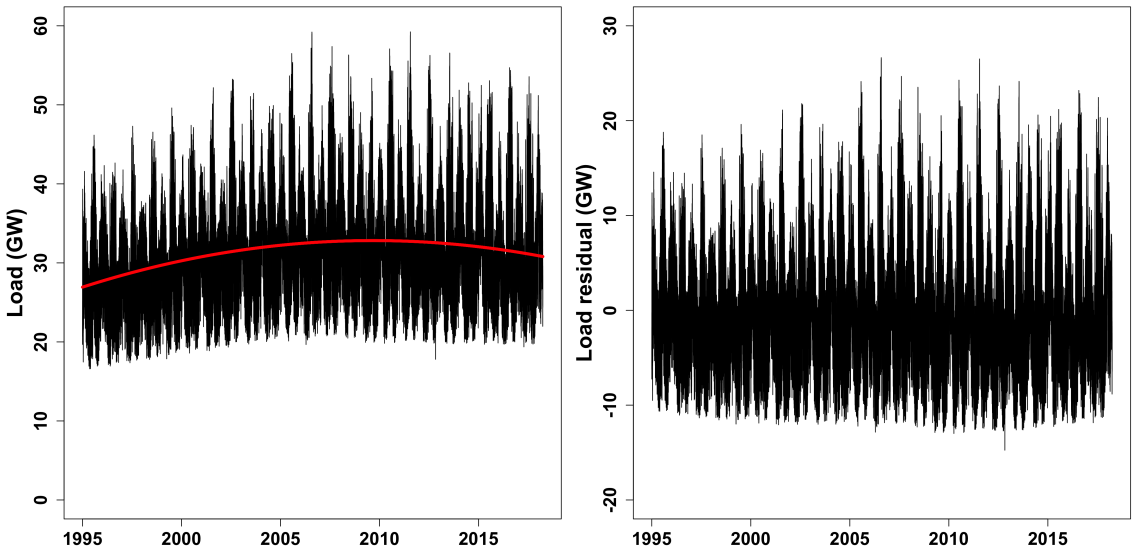

Figure B.2: Metered load time series (1995-2018 model fits). Left: hourly time series of metered system load for PJM transmission zones that have been part of PJM since database inception, with time trend (red curve) as given by Equation 3.6. Right: residuals from fitting time trend given by Equation 3.6 to the original series.

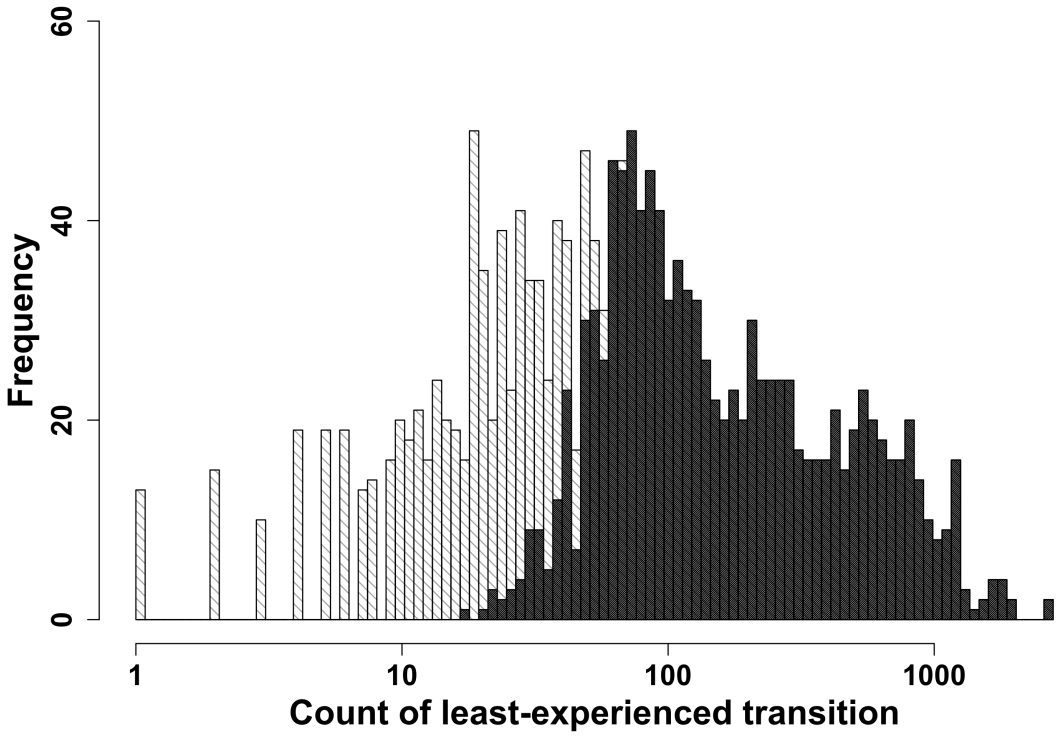

Figure B.3: Histogram of the count of each generator's least frequently experienced transition (1995-2018 model fits). The 1,748 generators with at least one full calendar year of data reporting and at least one unscheduled transition during 1995-2018 are plotted in light gray. Of these, the 1,111 generators with at least 10 failure and recovery transitions per statistically significant parameter are then overplotted in dark gray. Above 70 such transitions the full model specification (Equation 3.5) can be supported, so no generators in corresponding bins are discarded. Note the log scale on the abscissa.

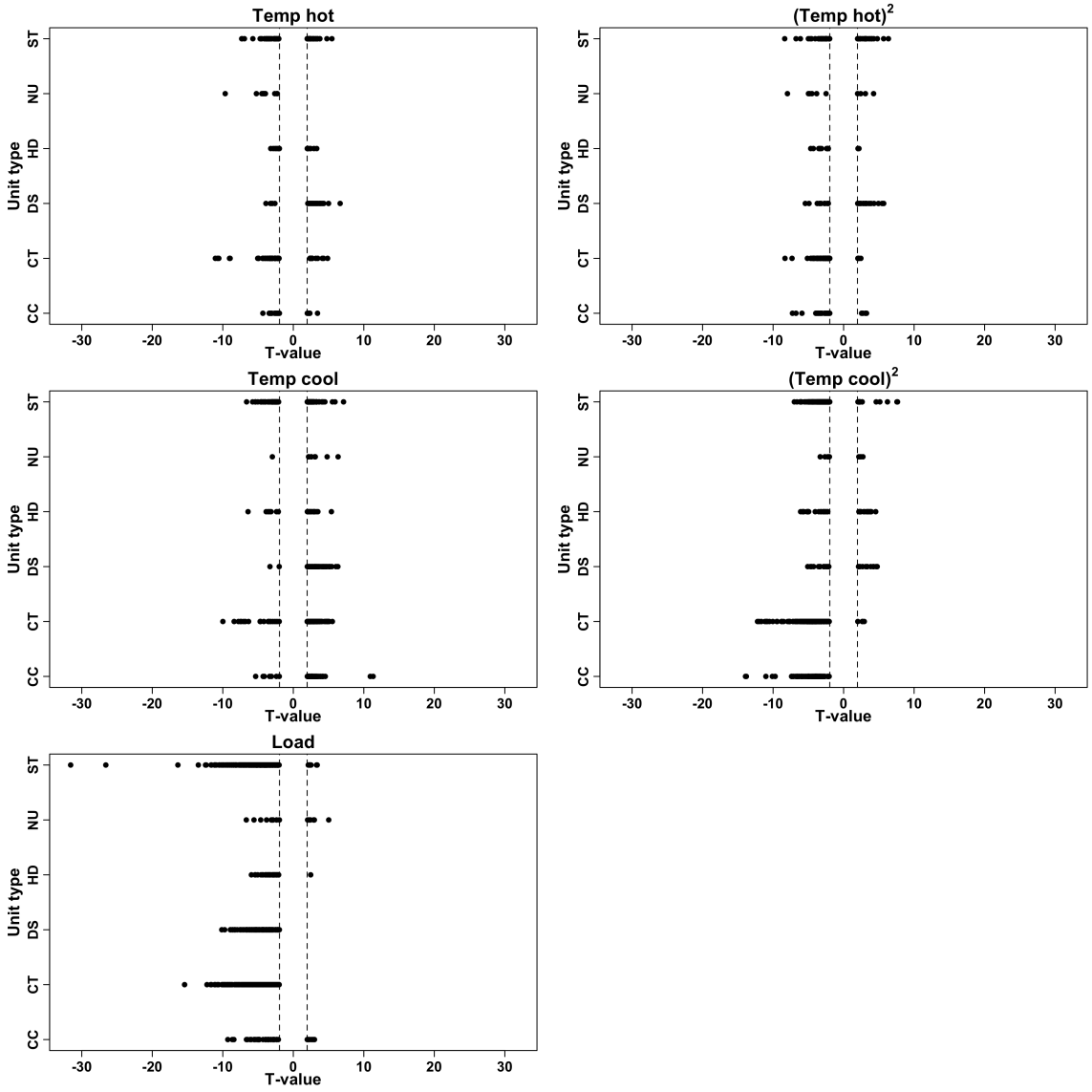

Figure B.4: Summarizing t-values for the failure model by covariate and generator type (1995-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Thresholds for significance (+/-1.96) indicated by dashed vertical lines. Constants are excluded. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

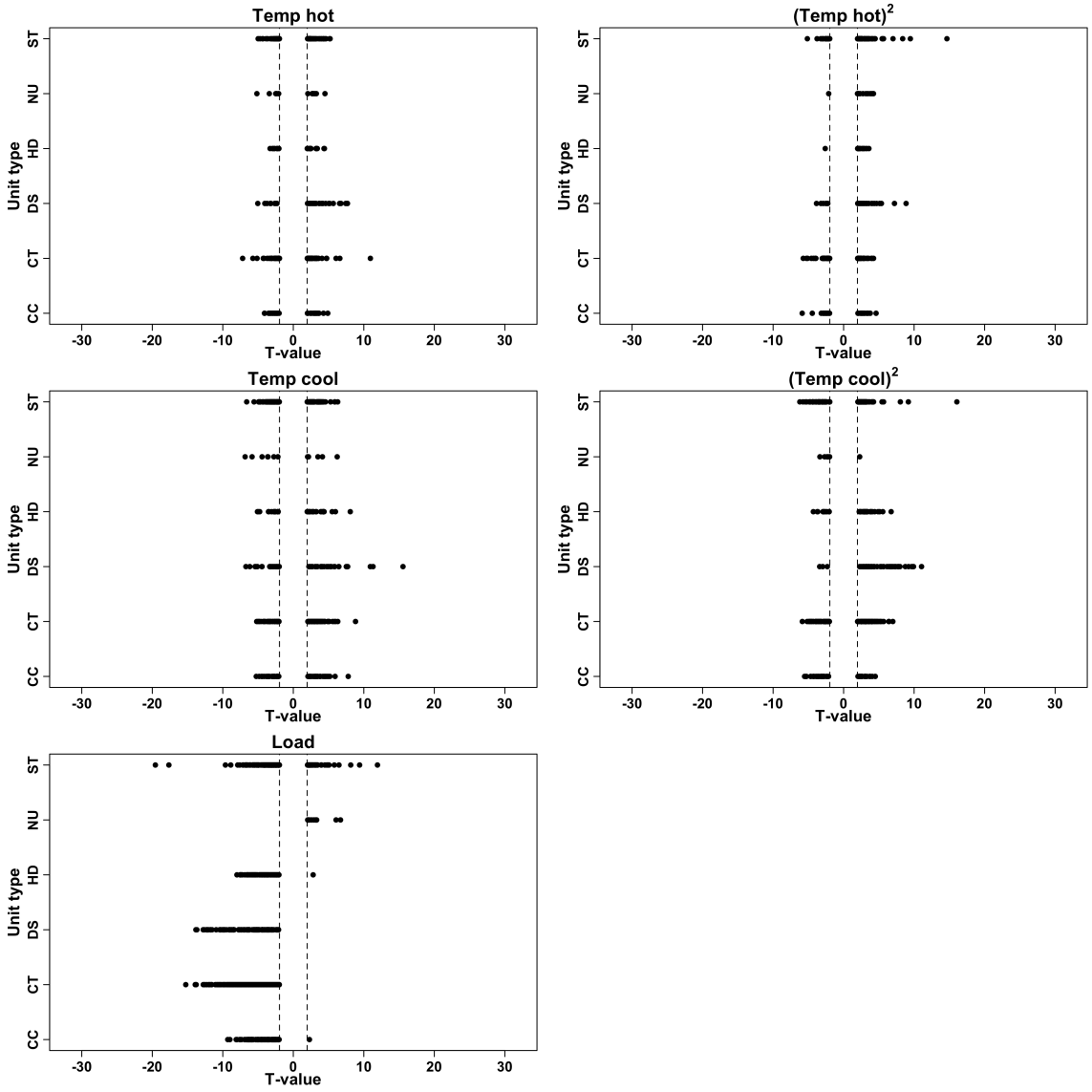

Figure B.5: Summarizing t-values for the recovery model by covariate and generator type (1995-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Thresholds for significance (+/-1.96) indicated by dashed vertical lines. Constants are excluded. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

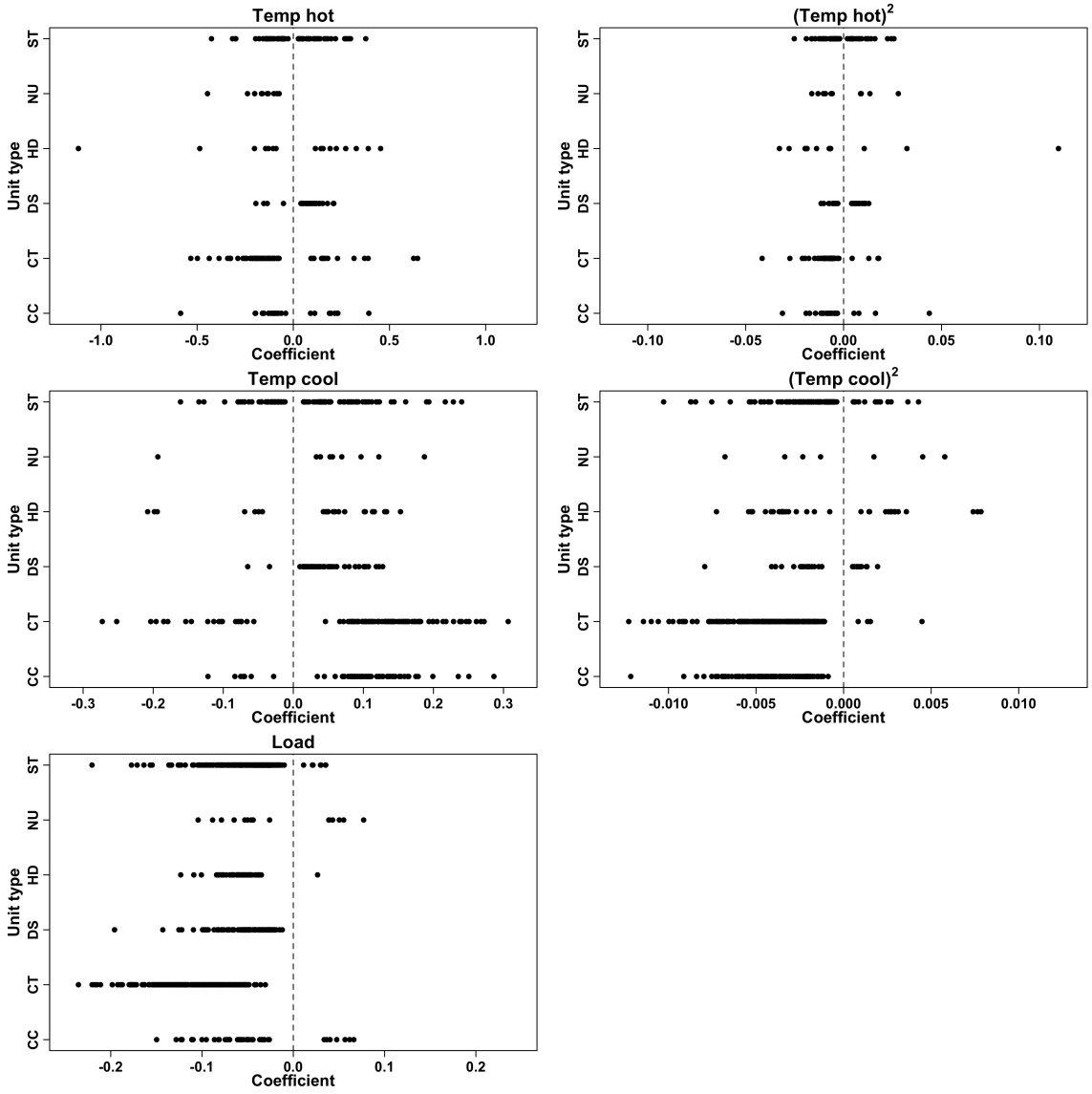

Figure B.6: Summarizing coefficients for the failure model by covariate and generator type (1995-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Dashed vertical lines indicate 0. Constants are excluded. Temperature units are degrees Celsius. Load units are GW. Abscissa scales set independently. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

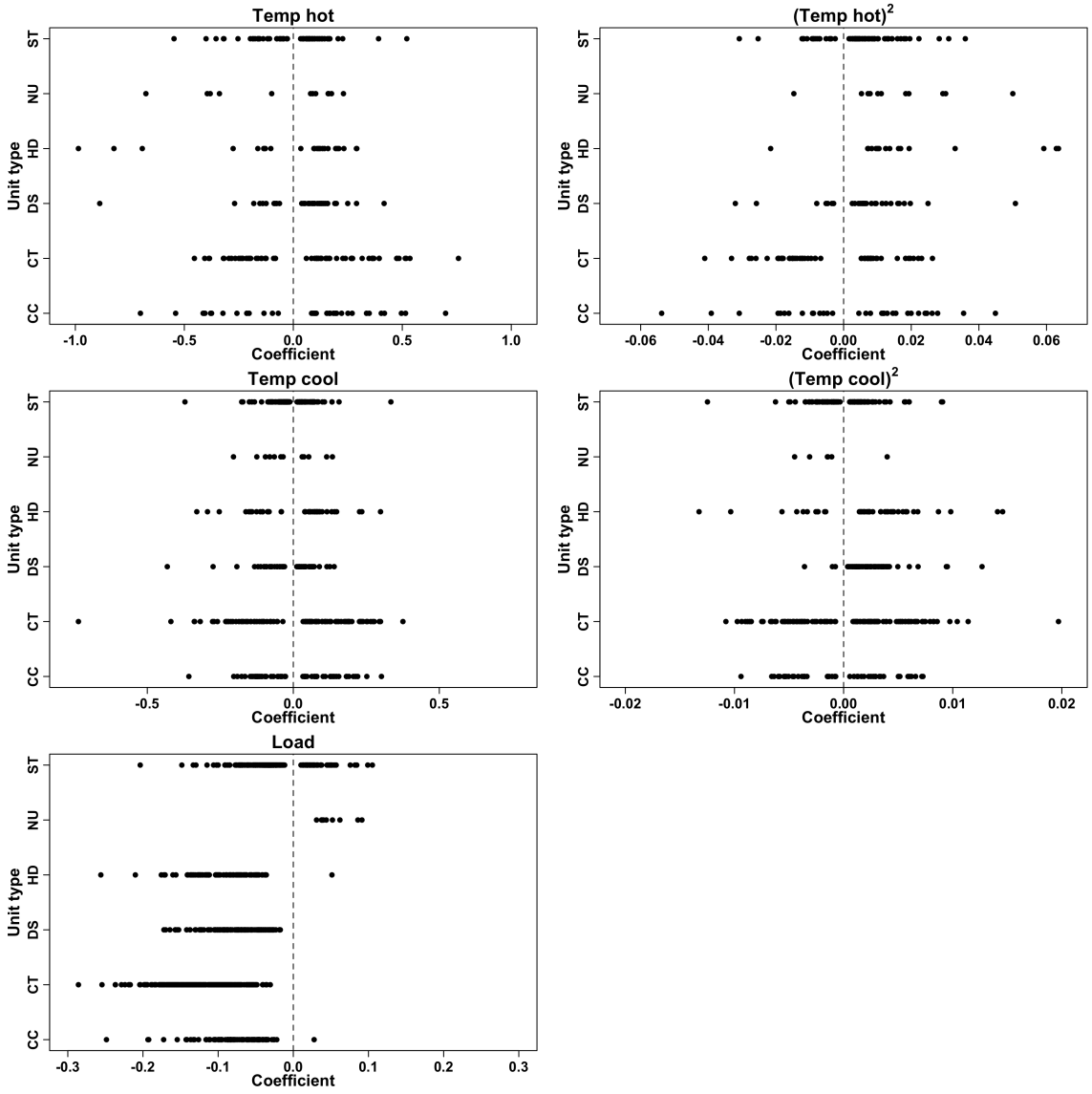

Figure B.7: Summarizing coefficients for the recovery model by covariate and generator type (1995-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Dashed vertical lines indicate 0. Constants are excluded. Temperature units are degrees Celsius. Load units are GW. Abscissa scales set independently. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

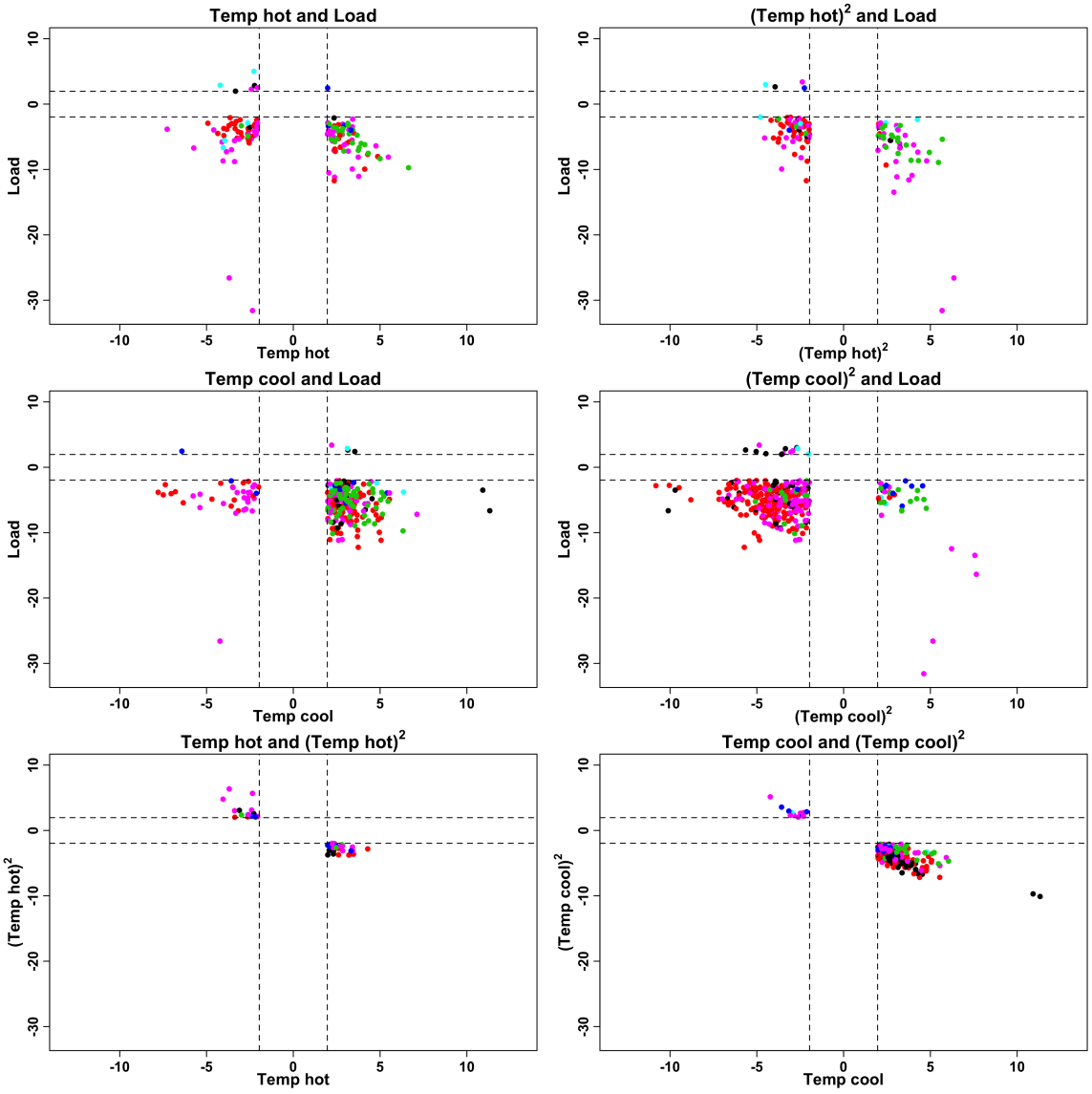

Figure B.8: Summarizing t-value relationships for non-orthogonal covariate pairs for the failure model (1995-2018 model fits). Thresholds for significance at 0.05 level (+/-1.96) indicated by dashed lines. To be included in a plot in this figure, both relevant covariates must be present in a generator's final failure model. Black is combined cycle gas, red is simple cycle gas, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine.

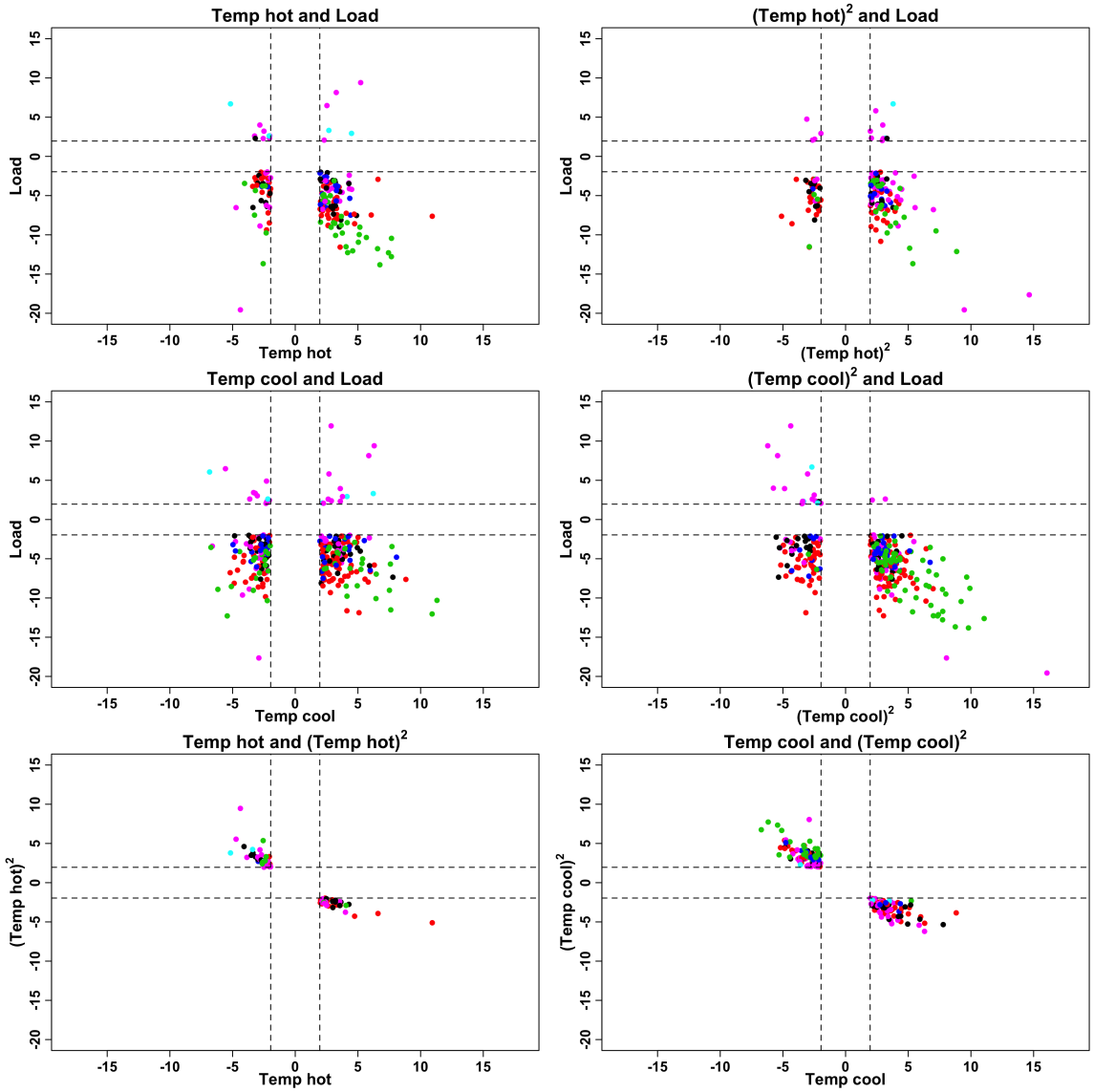

Figure B.9: Summarizing t-value relationships for non-orthogonal covariate pairs for the recovery model (1995-2018 model fits). Thresholds for significance at 0.05 level (+/- 1.96) indicated by dashed lines. To be included in a plot in this figure, both relevant covariates must be present in a generator's final recovery model. Black is combined cycle, red is simple cycle, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine.

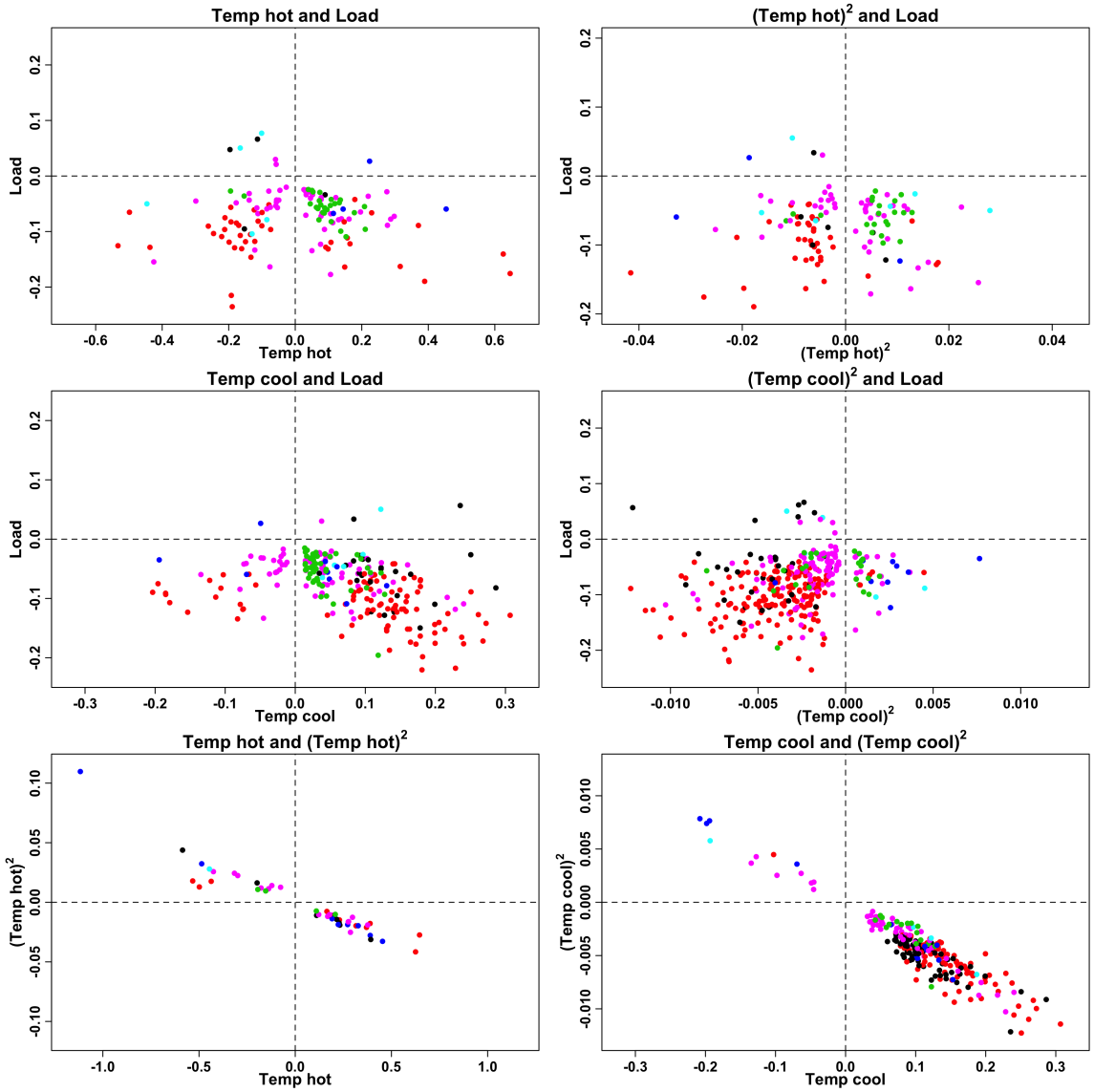

Figure B.10: Summarizing coefficient relationships for non-orthogonal covariate pairs for the failure model (1995-2018 model fits). To be included in a plot in this figure, both relevant covariates must be present in a generator's final failure model. Black is combined cycle gas, red is simple cycle gas, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine. Dashed lines indicate 0. Temperature units are degrees C. Load units are GW. Axis scales set independently in each plot.

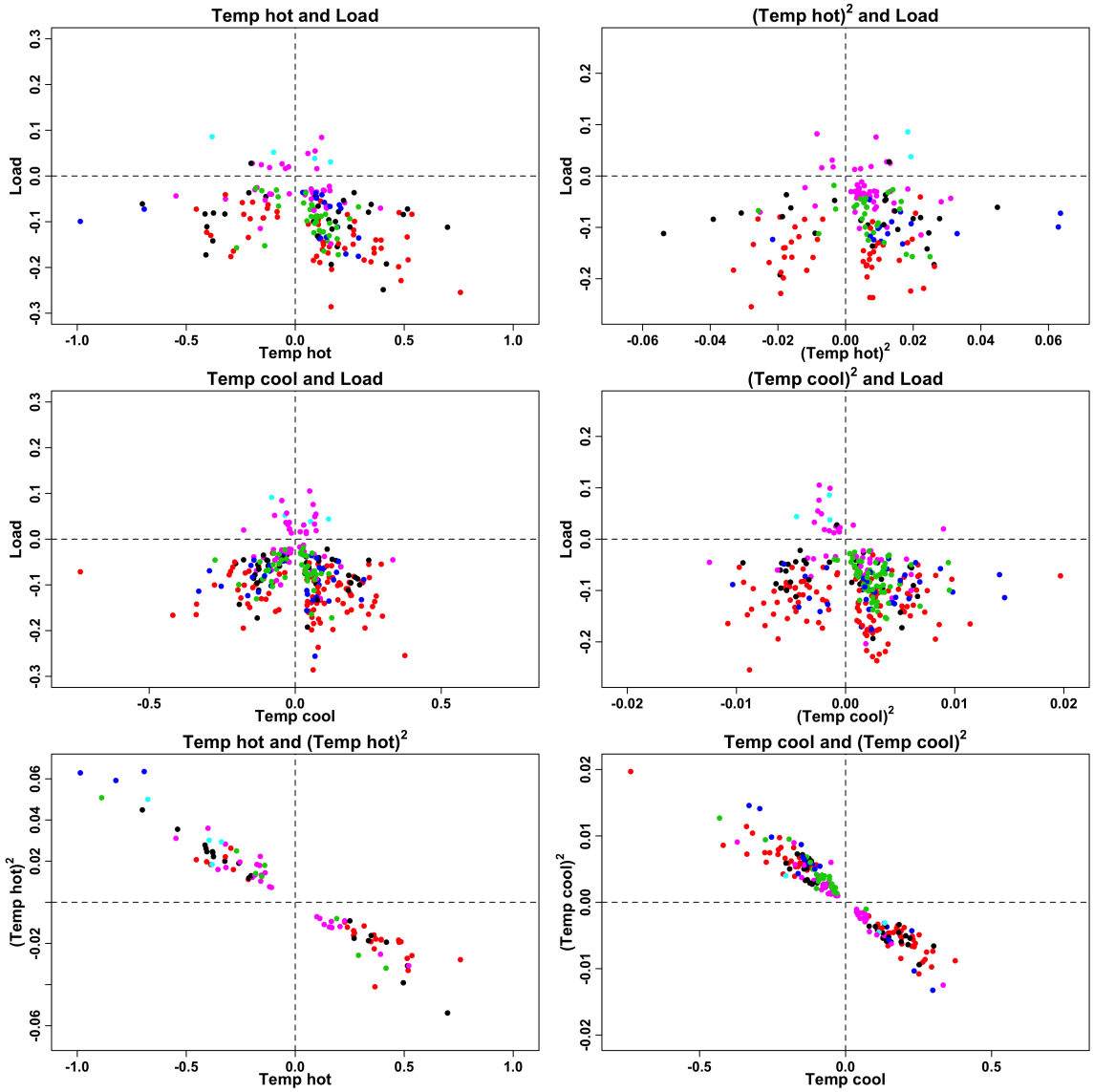

Figure B.11: Summarizing coefficient relationships for non-orthogonal covariate pairs for the recovery model (1995-2018 model fits). To be included in a plot in this figure, both relevant covariates must be present in a generator's final recovery model. Black is combined cycle, red is simple cycle, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine. Dashed lines indicate 0. Temperature units are degrees C. Load units are GW. Axis scales set independently in each plot.

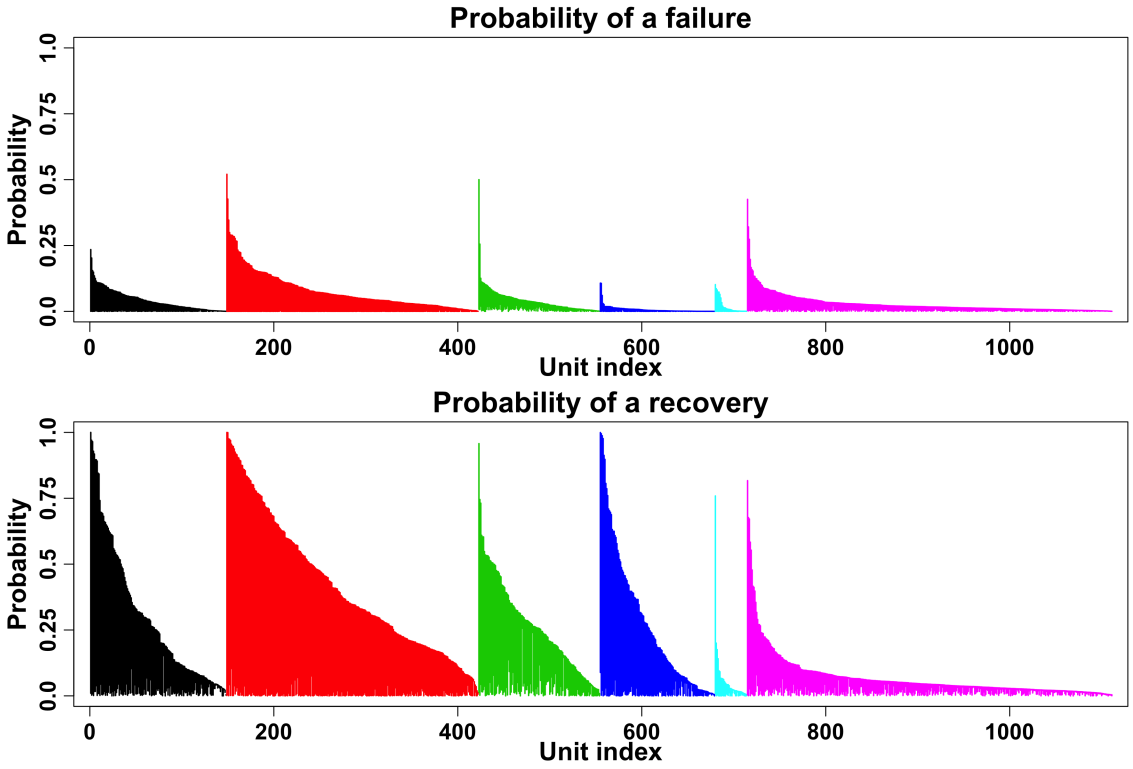

Figure B.12: Summarizing the empirical range of hourly transition probabilities (1995-2018 model fits). Plots include 1,111 generators with at least 10 failure and recovery events per statistically significant model parameter. Each generator is represented as a vertical line at an integer index (1 to 1,111). In each plot, generators are sorted by generator type and maximum experienced transition probability. Black is combined cycle, red is simple cycle, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine.

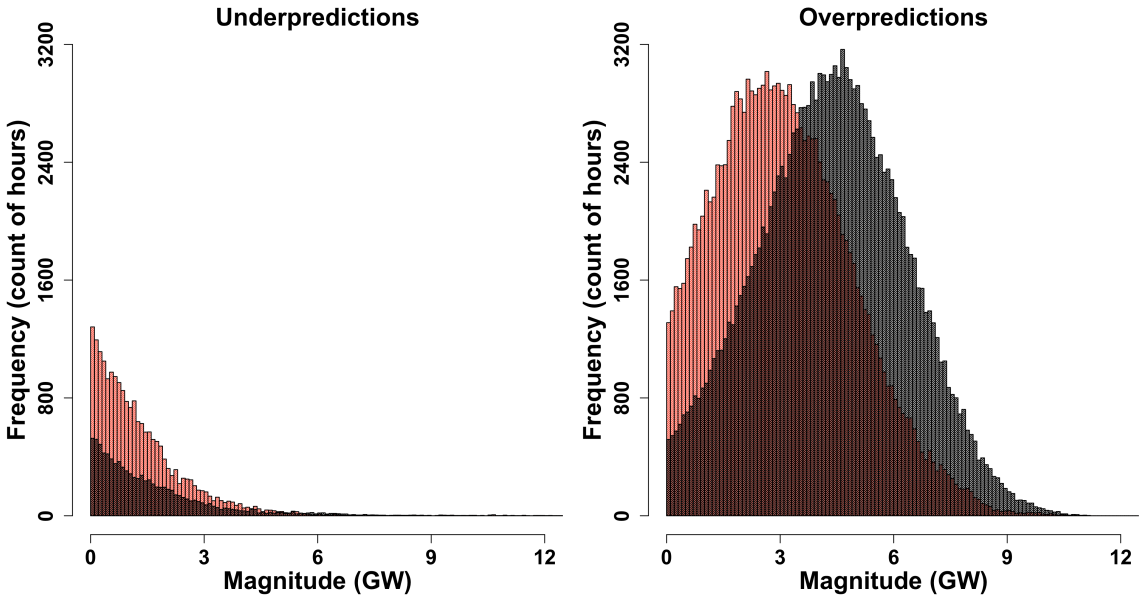

Figure B.13: Histograms of underpredictions and overpredictions of logistic regression (red) and current practice (black) models (1995-2018 model fits). The abscissa of the underpredictions histogram is truncated to improve presentation: the logistic regression model has a maximum underprediction of 16.2 GW, while the current practice model has a maximum underprediction of 20.8 GW; both models' maximum underprediction occurs during the Polar Vortex of January 2014.

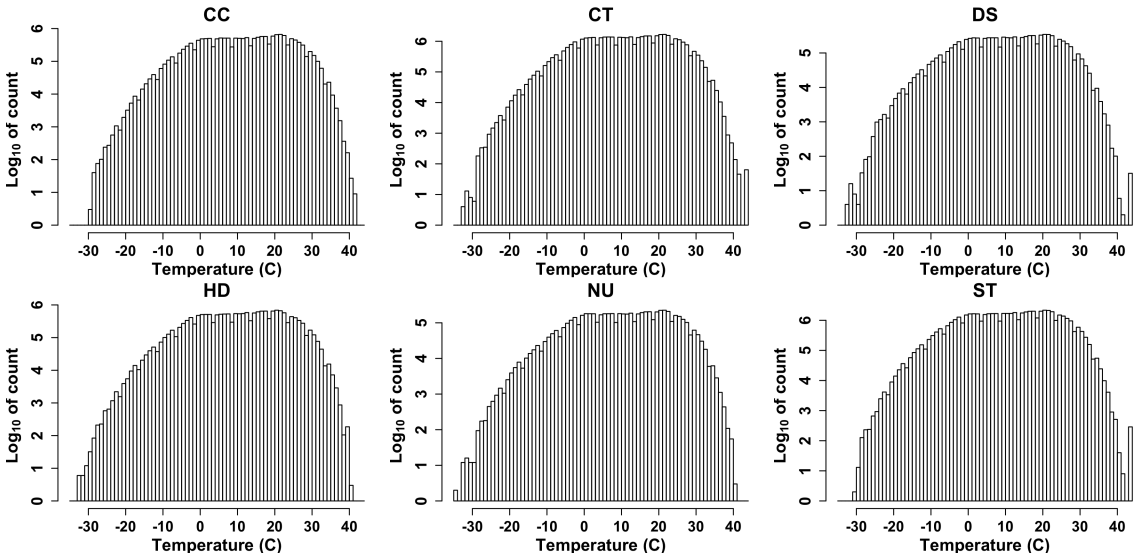

Figure B.14: Prevalence of temperatures experienced by 1,111 modeled generators by generator type (1995-2018 model fits). For use with Figure 3.6. Note the log scale on the ordinate. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

3.7 Appendix C: Figures and tables for 2004-2018 model fits

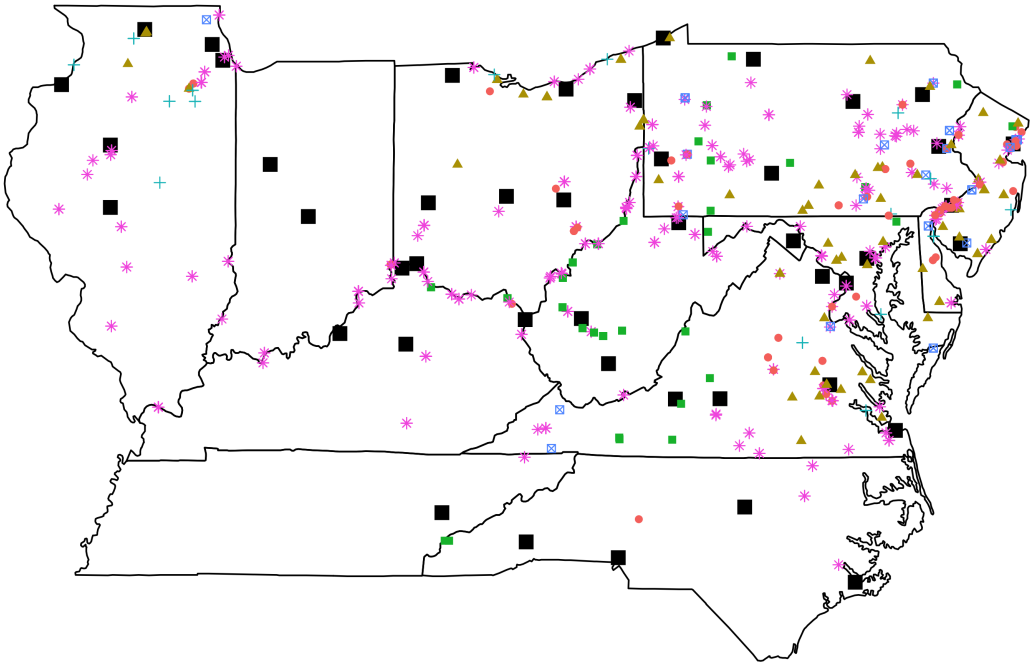

Figure C.1: Locations of 748 generators and linked weather stations, overlaid on corresponding U.S. states (2004-2018 model fits). Only generators with 10 transitions per statistically significant model parameter when fitting models using 2004-2018 data are retained. Note that all generators in multi-generator power plants have identical locations. Orange circles indicate combined cycle, blue squares with inset 'x' indicate simple cycle, yellow triangles indicate diesel, small green squares indicate hydroelectric and pumped storage, teal 'plus' signs indicate nuclear, purple asterisks indicate steam turbine, and large black squares indicate weather stations. A small number of retained generators are not shown for presentation considerations: Alabama (2), Louisiana (3), Michigan (12), Mississippi (3), South Carolina (1), Texas (7).

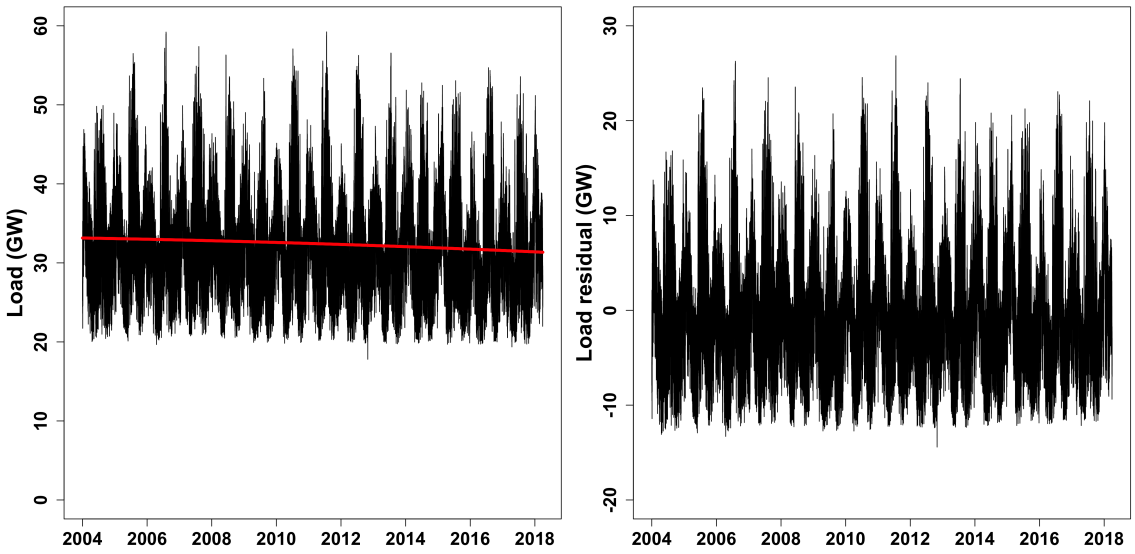

Figure C.2: Metered load time series (2004-2018 model fits). Left: hourly time series of metered system load for PJM transmission zones that have been part of PJM since database inception, with time trend (red curve) as given by Equation 3.6. Right: residuals from fitting time trend given by Equation 3.6 to the original series.

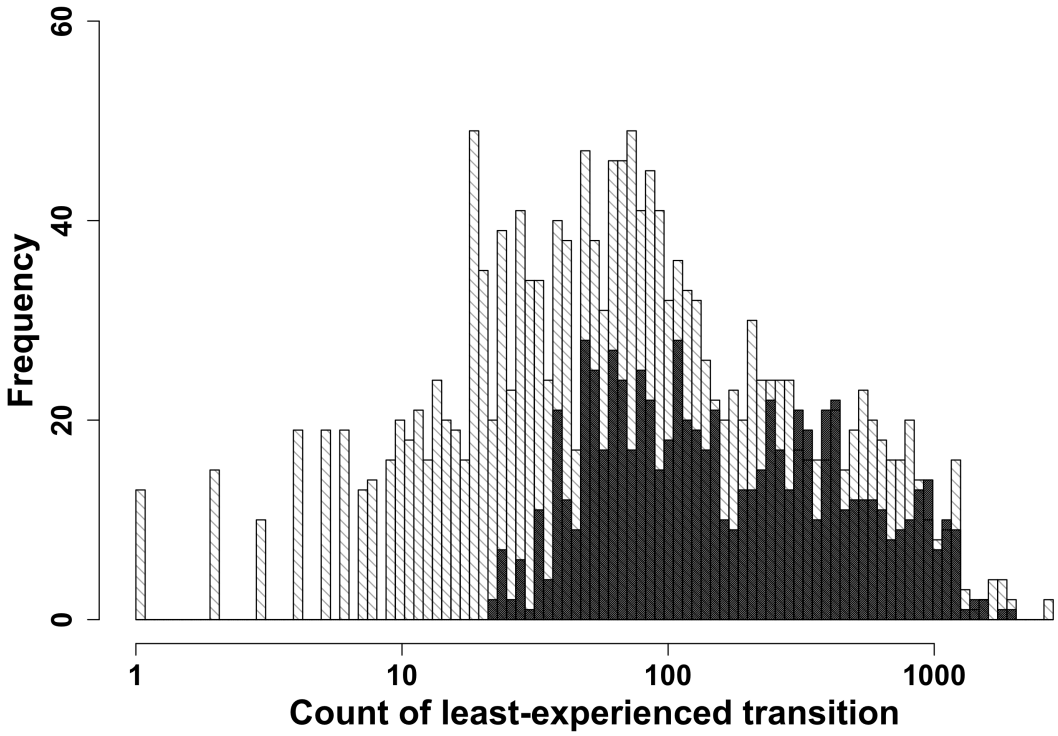

Figure C.3: Histogram of the count of each generator's least frequently experienced transition (2004-2018 model fits). The 1,748 generators with at least one full calendar year of data reporting and at least one unscheduled transition during 1995-2018 are plotted in light gray (consistent with Figure C.3). Of these, the 748 generators with at least 10 failure and recovery transitions per statistically significant parameter during 2004-2018 are then overplotted in dark gray. In contrast to Figure C.3, not all generators with at least 70 failure and recovery transitions are retained due to the more restrictive time period. Note the log scale on the abscissa.

Generator type	Total count	Retained count (%)	Total capacity	Retained capacity (%)
CC	224	126 (56)	53.4	31.2 (58)
CT	663	44 (7)	44.9	1.9 (4)
DS	236	120 (51)	0.8	0.4 (46)
HD	244	68 (28)	11.0	4.2 (38)
NU	35	32 (91)	37.2	34.2 (92)
ST	443	358 (81)	119.5	106.4 (89)
All	1,845	748 (41)	266.8	178.3 (67)

Table C.1: Summary of total and retained generator counts and capacity, by generator type (2004-2018 model fits). CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

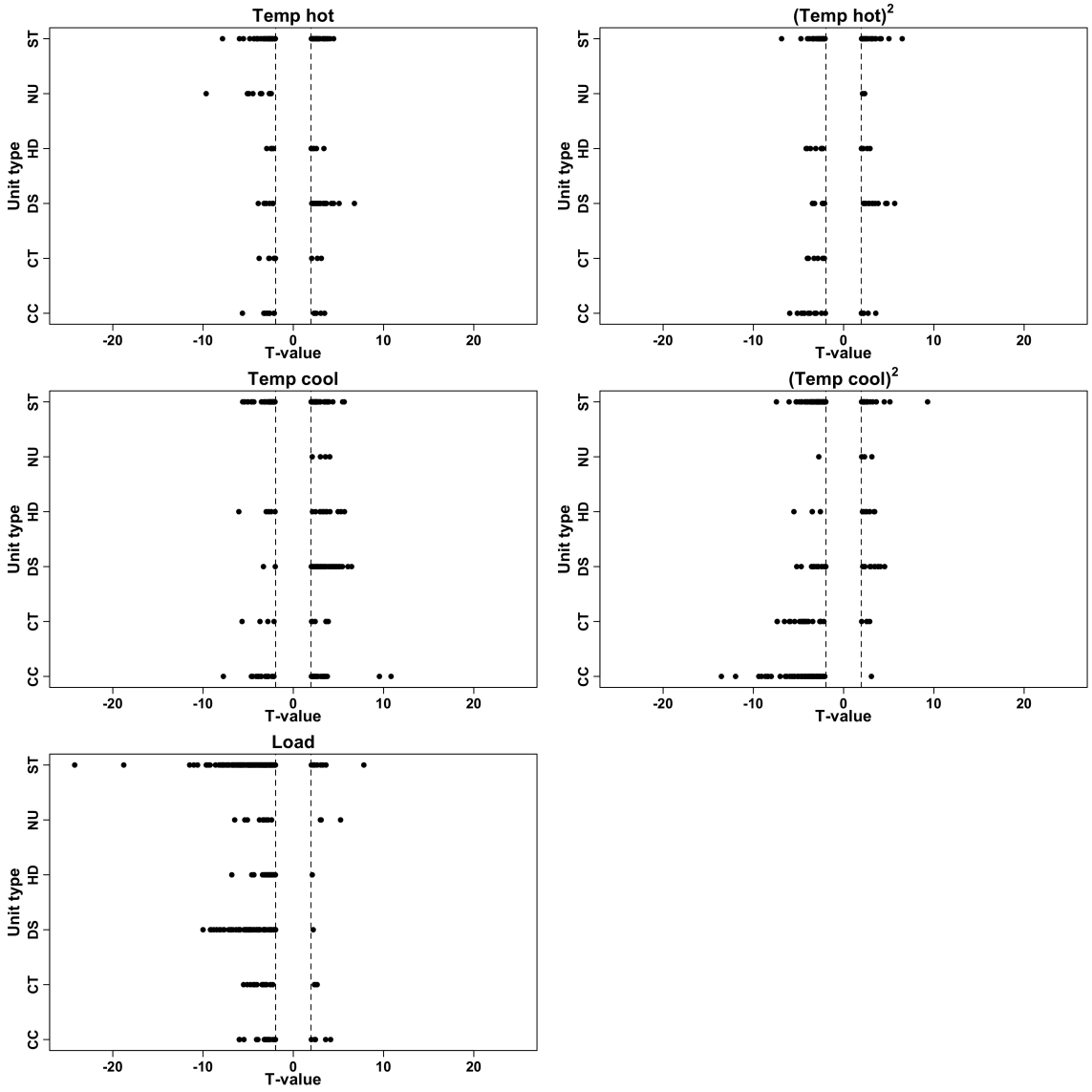

Figure C.4: Summarizing t-values for the failure model by covariate and generator type (2004-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Thresholds for significance (+/-1.96) indicated by dashed vertical lines. Constants are excluded. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

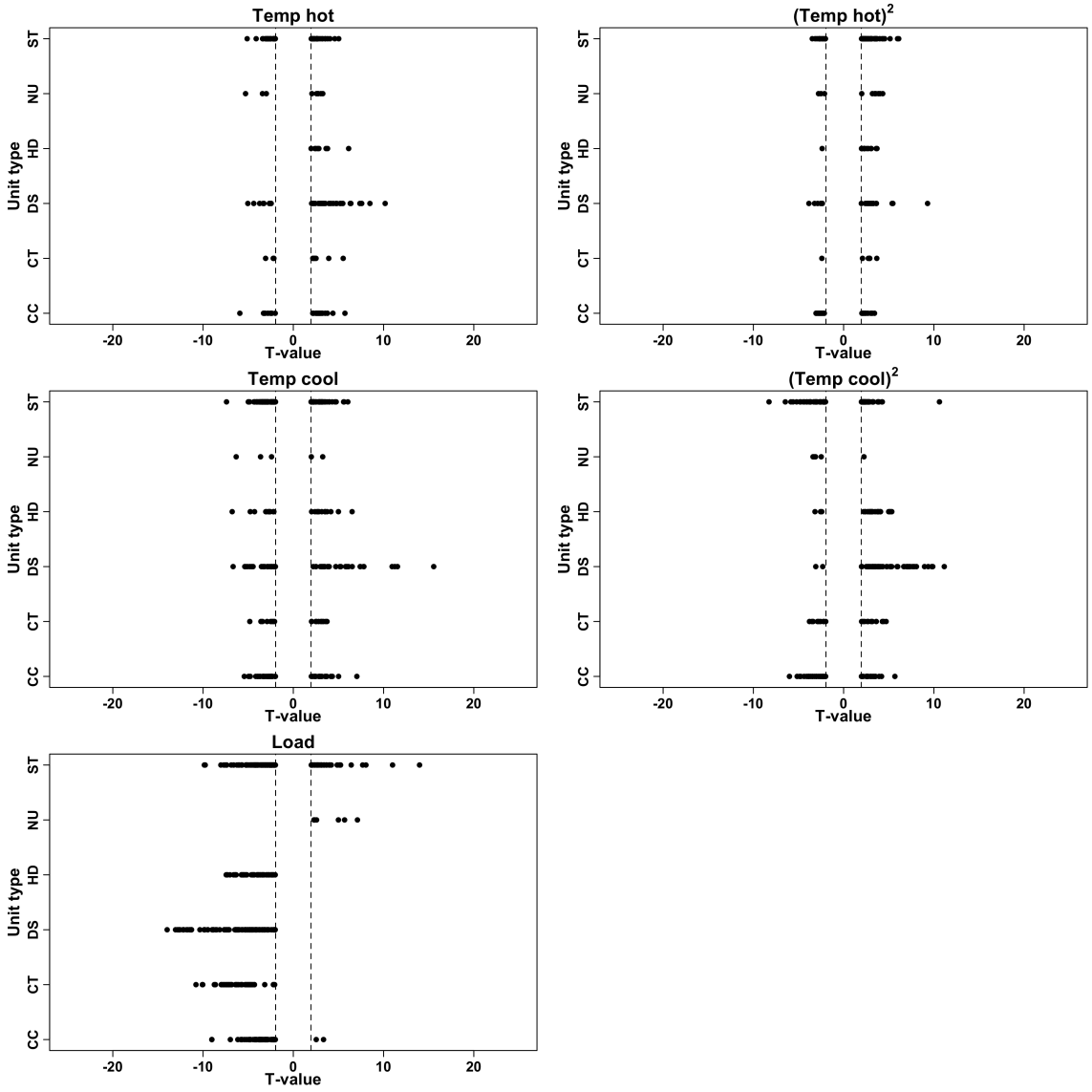

Figure C.5: Summarizing t-values for the recovery model by covariate and generator type (2004-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Thresholds for significance (+/-1.96) indicated by dashed vertical lines. Constants are excluded. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

Generator type	Count	Mean hot	Mean cool	Temp hot	Temp hot^2	Temp cool	Temp cool^2	Load
CC	126	126	126	17	21	43	80	21
CT	44	44	44	9	6	11	26	16
DS	120	120	120	36	29	57	27	93
HD	68	68	68	11	13	17	12	31
NU	32	32	32	9	2	4	4	12
ST	358	358	358	68	53	84	91	204
All	748	748	748	150	124	216	240	377

Table C.2: Number of times each model term is statistically significant at the 95% level for the failure model (2004-2018 model fits). Results reported for the 748 generating generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

Generator type	Count	Mean hot	Mean cool	Temp hot	Temp hot^2	Temp cool	Temp cool^2	Load
CC	126	124	126	26	27	52	47	69
CT	44	44	44	9	6	18	20	39
DS	120	120	119	41	24	55	62	100
HD	68	67	67	8	15	23	26	56
NU	32	32	32	9	10	5	5	6
ST	358	358	358	58	66	111	87	128
All	748	745	746	151	148	264	247	398

Table C.3: Number of times each model term is statistically significant at the 95% level for the recovery model (2004-2018 model fits). Results reported for the 748 generating generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

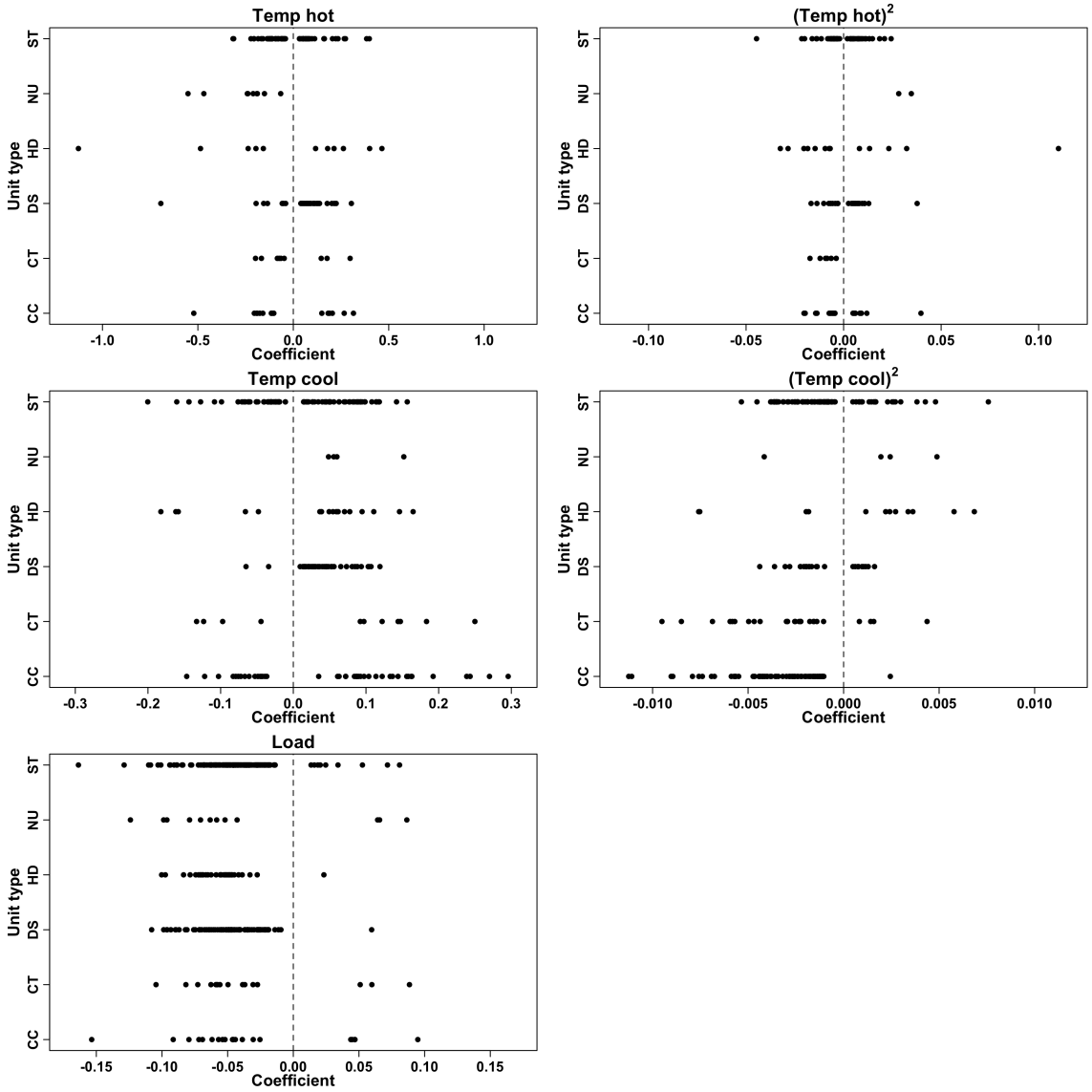

Figure C.6: Summarizing coefficients for the failure model by covariate and generator type (2004-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Dashed vertical lines indicate 0. Constants are excluded. Temperature units are degrees C. Load units are GW. Abscissa scales set independently. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

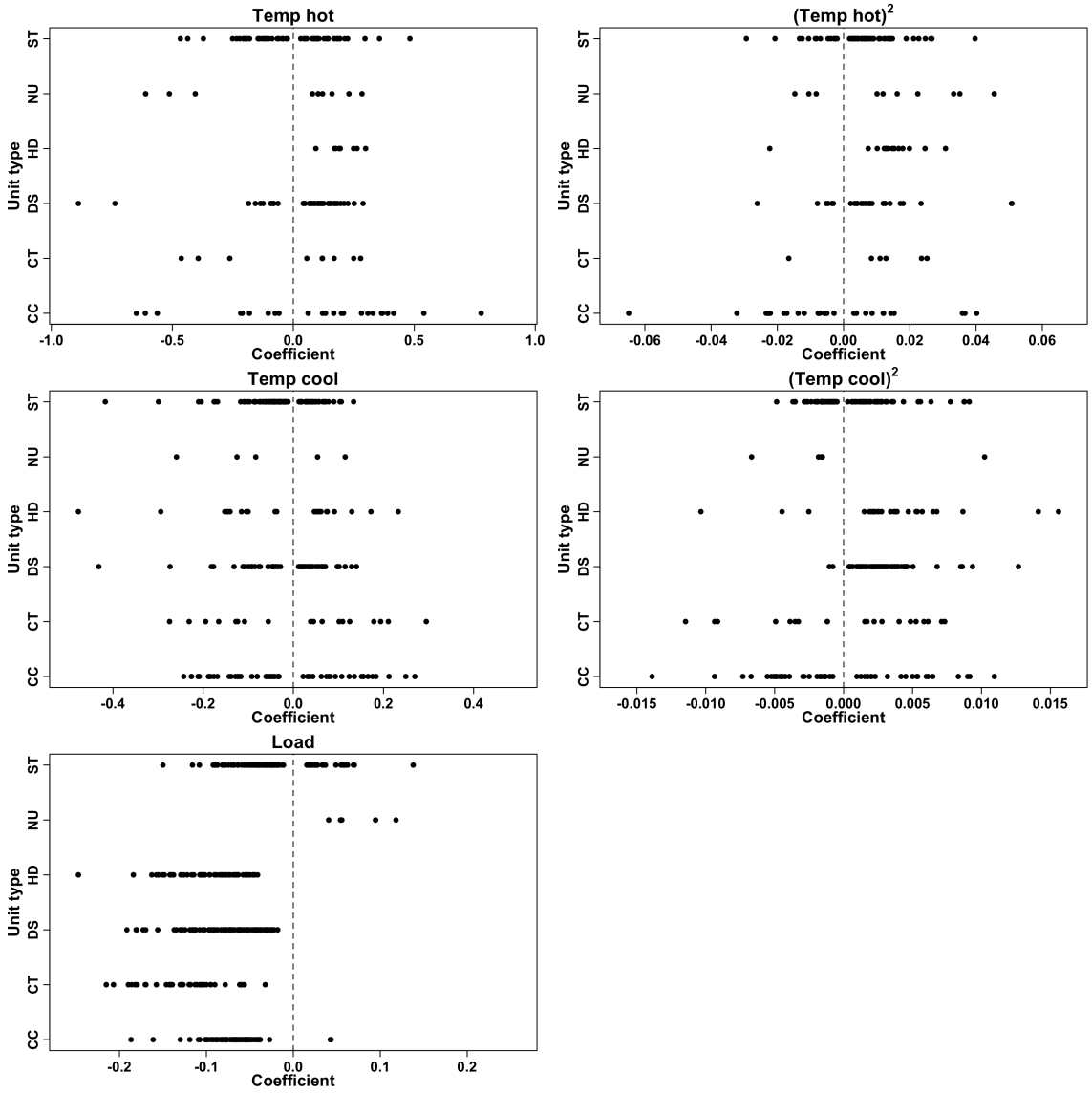

Figure C.7: Summarizing coefficients for the recovery model by covariate and generator type (2004-2018 model fits). Only generators for which the covariate is statistically significant at the 0.05 level are included. Dashed vertical lines indicate 0. Constants are excluded. Temperature units are degrees C. Load units are GW. Abscissa scales set independently. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

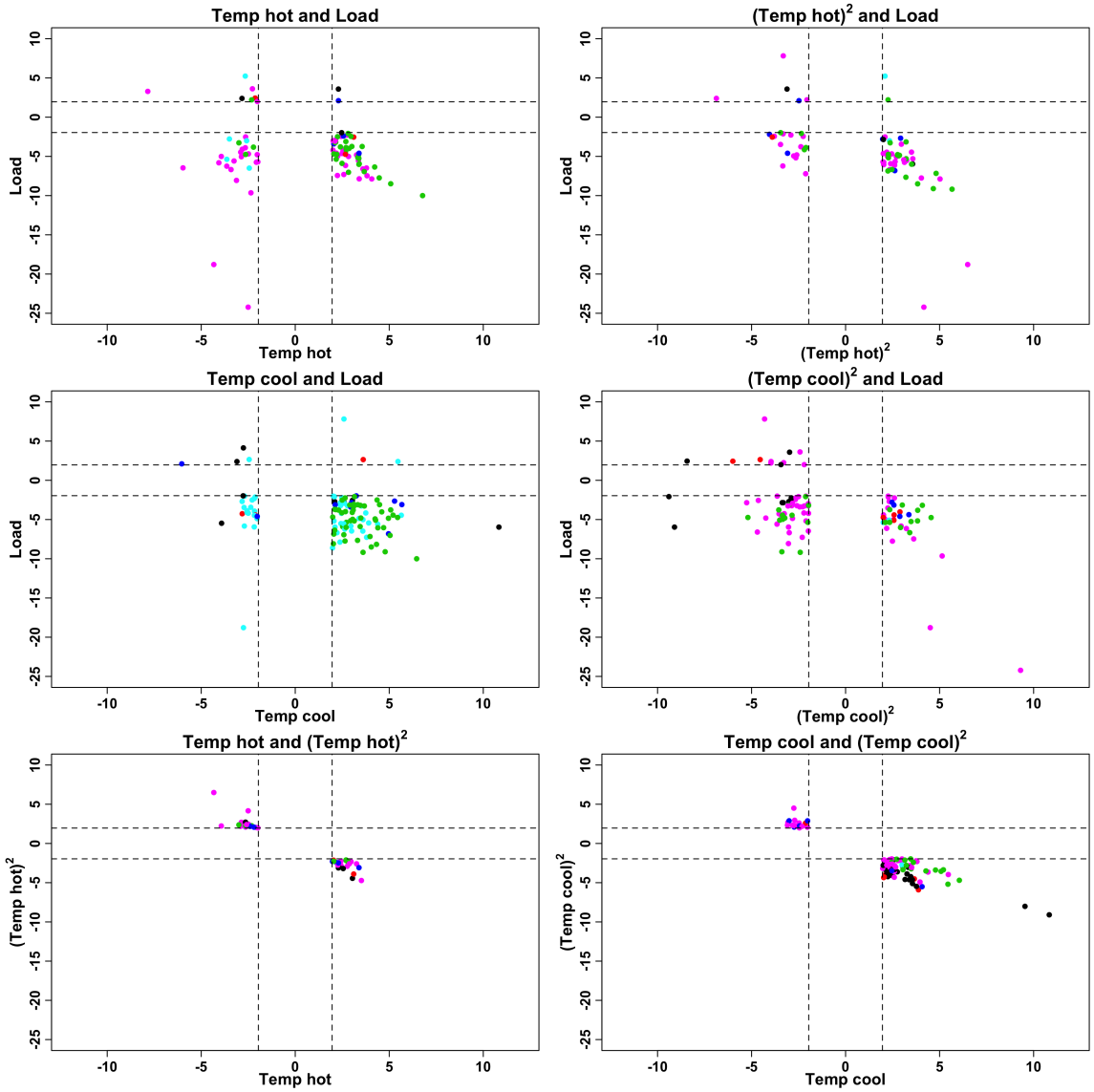

Figure C.8: Summarizing t-value relationships for non-orthogonal covariate pairs for the failure model (2004-2018 model fits). Thresholds for significance at 0.05 level (+/-1.96) indicated by dashed lines. To be included in a plot in this figure, both relevant covariates must be present in a generator's final failure model. Black is combined cycle gas, red is simple cycle gas, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine.

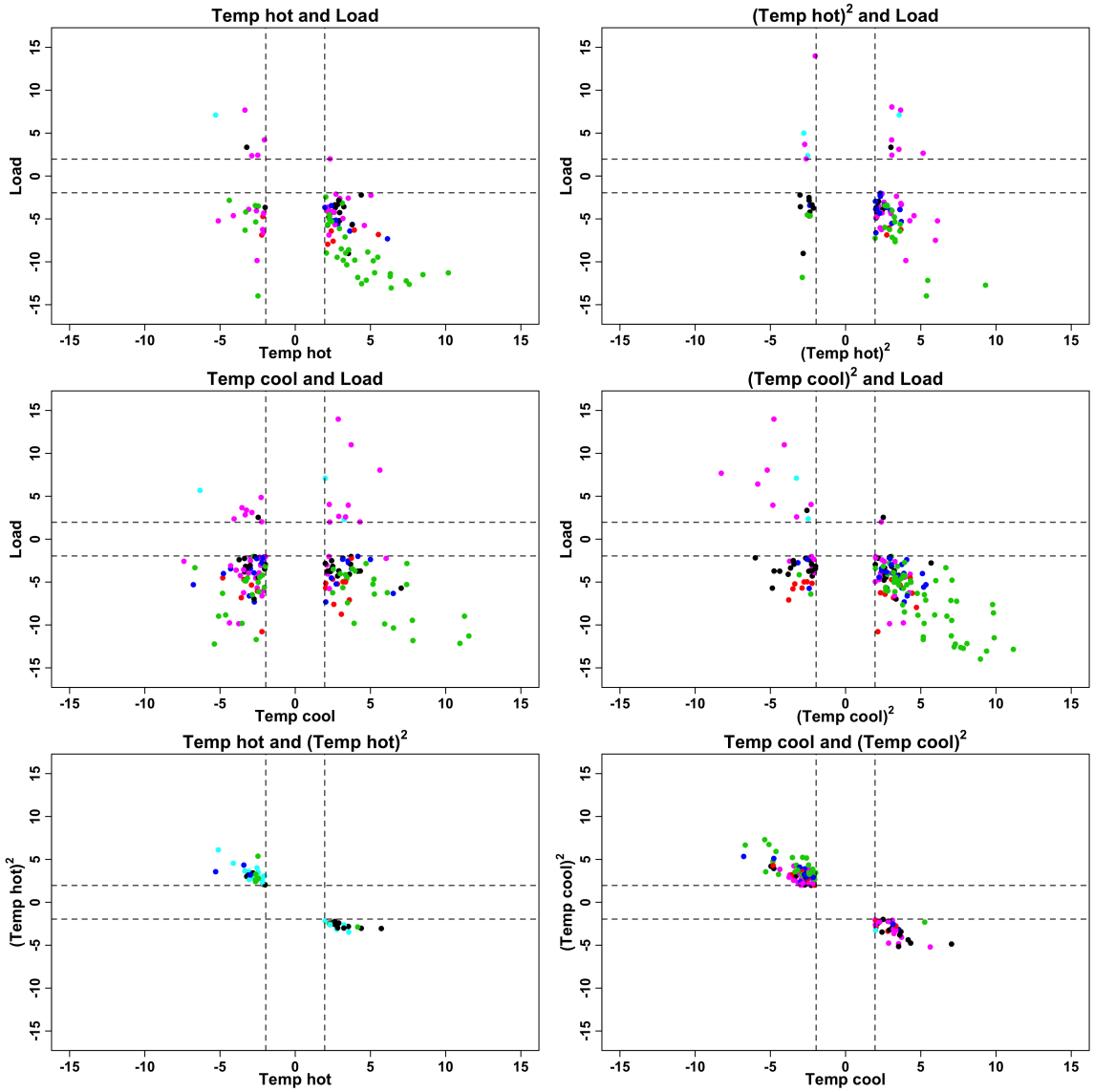

Figure C.9: Summarizing t-value relationships for non-orthogonal covariate pairs for the recovery model (2004-2018 model fits). Thresholds for significance at 0.05 level (+/- 1.96) indicated by dashed lines. To be included in a plot in this figure, both relevant covariates must be present in a generator's final recovery model. Black is combined cycle, red is simple cycle, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine.

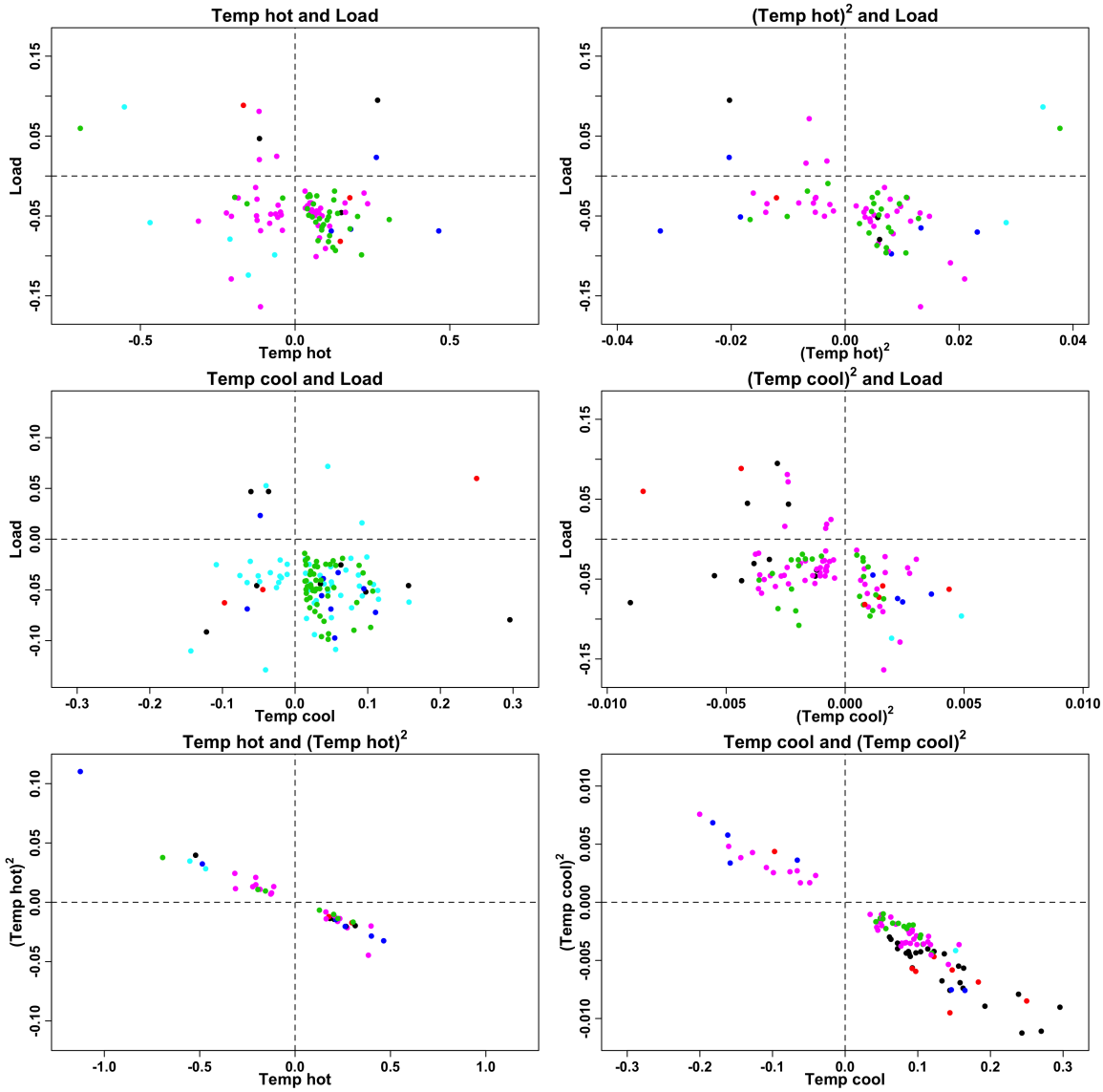

Figure C.10: Summarizing coefficient relationships for non-orthogonal covariate pairs for the failure model (2004-2018 model fits). To be included in a plot in this figure, both relevant covariates must be present in a generator's final failure model. Black is combined cycle gas, red is simple cycle gas, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine. Dashed lines indicate 0. Axis scales set independently.

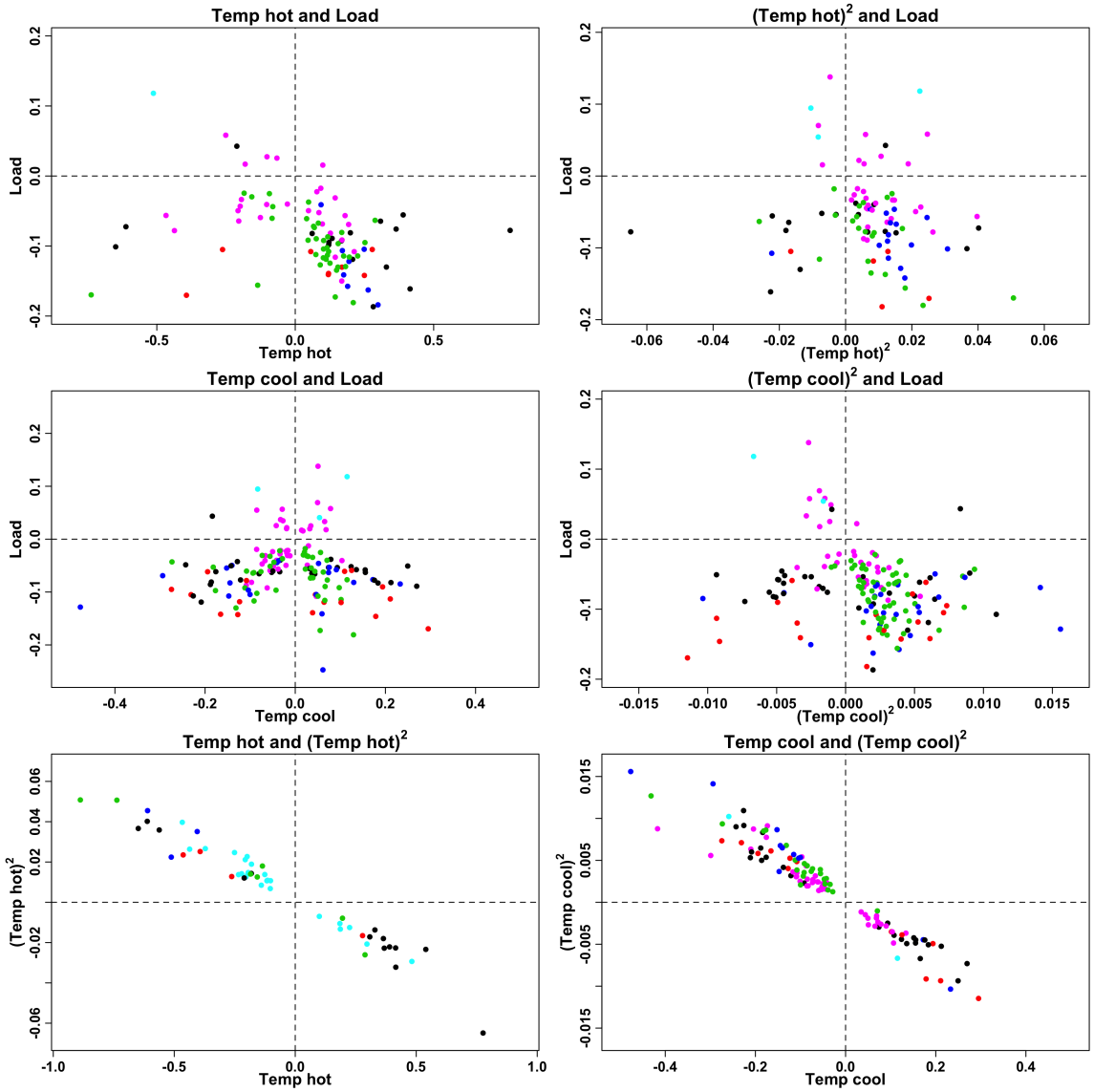

Figure C.11: Summarizing coefficient relationships for non-orthogonal covariate pairs for the recovery model (2004-2018 model fits). To be included in a plot in this figure, both relevant covariates must be present in a generator's final recovery model. Black is combined cycle gas, red is simple cycle gas, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine. Dashed lines indicate 0. Axis scales set independently.

Model	0	1	2	3	4	5	6	7
Failure	0	0	141	282	184	110	28	3
Recovery	0	0	150	237	169	145	42	5

Table C.4: Number of statistically significant parameters (including constants) for the 748 generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models (2004-2018 model fits).

Model	0	1	2	3	4
Failure	287	247	166	41	7
Recovery	263	233	187	57	8

Table C.5: Number of statistically significant temperature parameters (excluding constants) for the 748 generators with at least 10 instances of the less-common transition per parameter in both failure and recovery models (2004-2018 model fits).

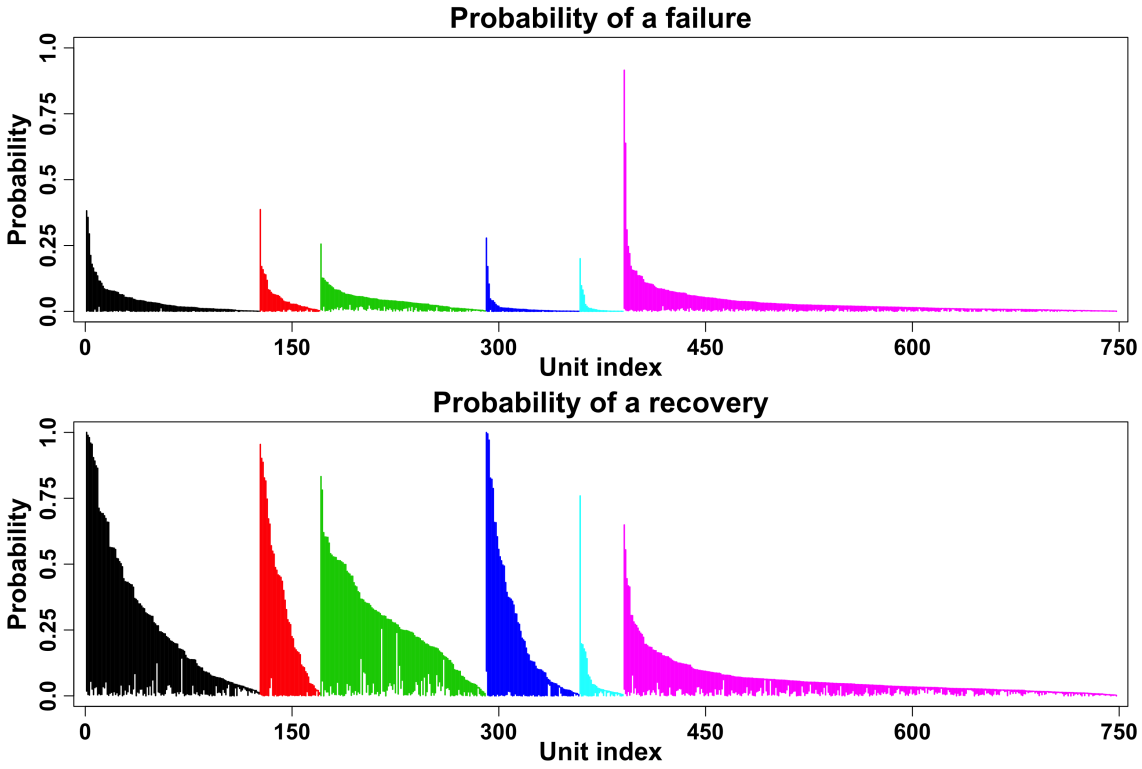

Figure C.12: Summarizing the empirical range of hourly transition probabilities (2004-2018 model fits). Plots include 748 generators with at least 10 failure and recovery events per statistically significant model parameter. Each generator is represented as a vertical line at an integer index (1 to 748). In each plot, generators are sorted by generator type and maximum experienced transition probability. Black is combined cycle gas, red is simple cycle gas, green is diesel, blue is hydroelectric, cyan is nuclear, magenta is steam turbine.

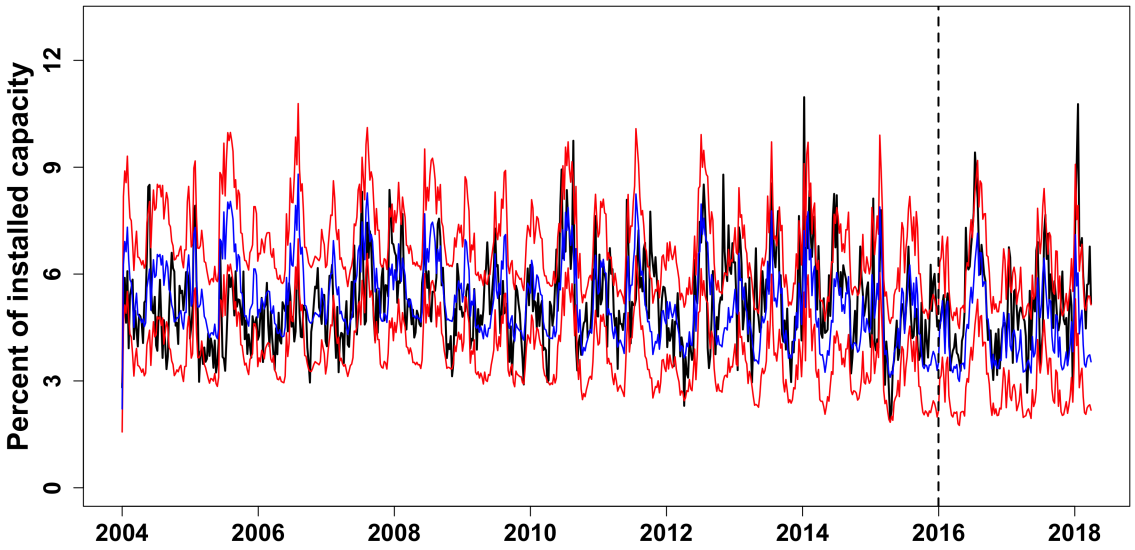

Figure C.13: Simulated time series from logistic regression model (2004-2015 model fits). Results presented for 703 generators with sufficient state transitions to support their failure and recovery models when fitting on 2004-2015; 2016-2018 used as test of model performance. The split between training and testing periods is denoted with a dashed vertical line. Weekly averages rather than hourly series. 5000 simulations conducted. Refer to Table C.6 for installed capacity by calendar year. Black trace is the empirical time series; blue trace is the concatenation of pointwise median simulation values; red traces are the concatenation of pointwise 2.5% and 97.5% simulation values.

Year	CC	CT	DS	HD	NU	ST	All
2004	14.8	1.6	0.1	2.9	31.8	79.3	130.5
2005	19.5	1.6	0.1	3.0	33.2	92.1	149.4
2006	20.4	1.6	0.1	3.0	33.2	92.3	150.5
2007	21.9	1.6	0.1	3.0	33.2	99.4	159.2
2008	23.1	1.6	0.1	3.0	33.2	98.8	159.9
2009	23.5	1.6	0.2	3.0	33.2	98.4	159.9
2010	23.3	1.8	0.2	3.0	33.2	98.7	160.1
2011	23.2	1.8	0.2	3.1	33.2	98.6	159.9
2012	25.1	1.8	0.2	3.1	33.2	92.6	155.9
2013	25.4	1.8	0.3	3.1	33.2	92.6	156.2
2014	25.4	1.8	0.3	3.1	33.2	91.0	154.6
2015	25.3	1.3	0.3	3.1	33.2	83.2	146.4
2016	25.3	1.3	0.3	3.1	33.2	82.2	145.4
2017	25.1	1.3	0.3	3.1	33.2	80.2	143.0
2018	25.1	1.3	0.3	3.1	33.2	80.2	143.0

Table C.6: Installed capacity (GW) of 703 retained generators by year and generator type (2004-2015 model fits). For use with Figure C.13. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

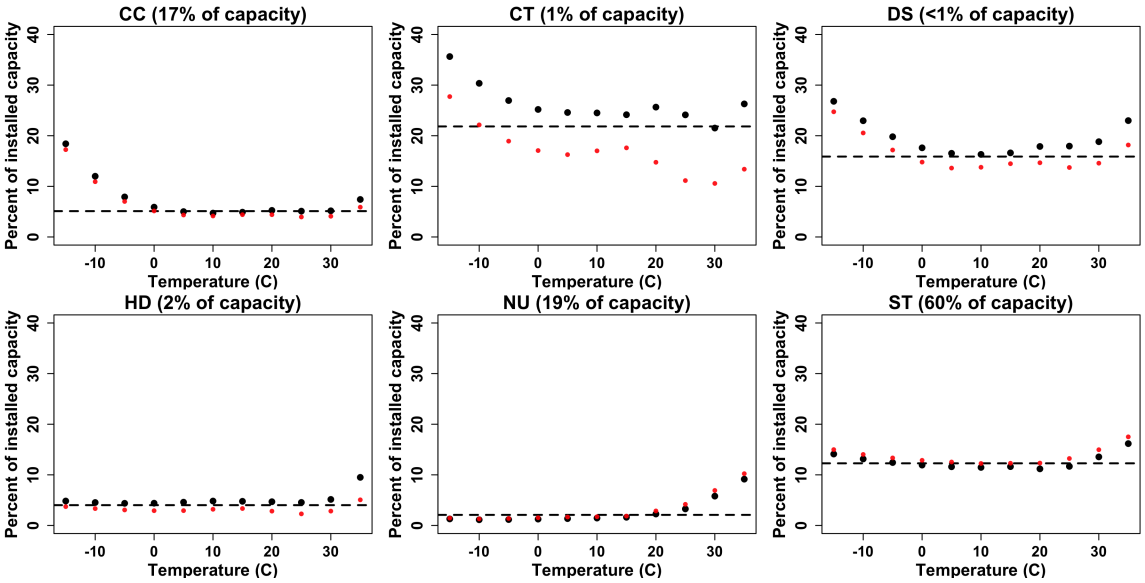

Figure C.14: Expected levels of unavailable capacity under logistic regression (dots) and current practice (dashed horizontal line), as a function of temperature (2004-2018 model fits). Black dots calculated using median load from temperature neighborhood, red dots calculated using 90^{th} percentile load from temperature neighborhood. Temperature neighborhood is defined as +/-10 degrees. Not all generators experience full temperature range; see Figure C.15 for prevalence of temperatures. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

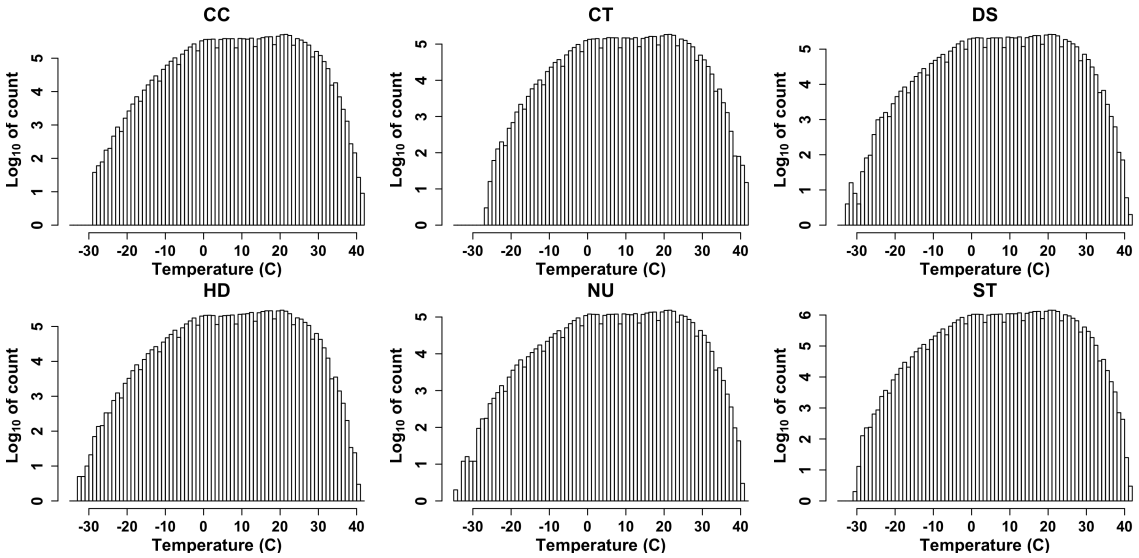

Figure C.15: Prevalence of temperatures experienced by 748 modeled generators by generator type (2004-2018 model fits). For use with Figure C.14. Note the log scale on the ordinate. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

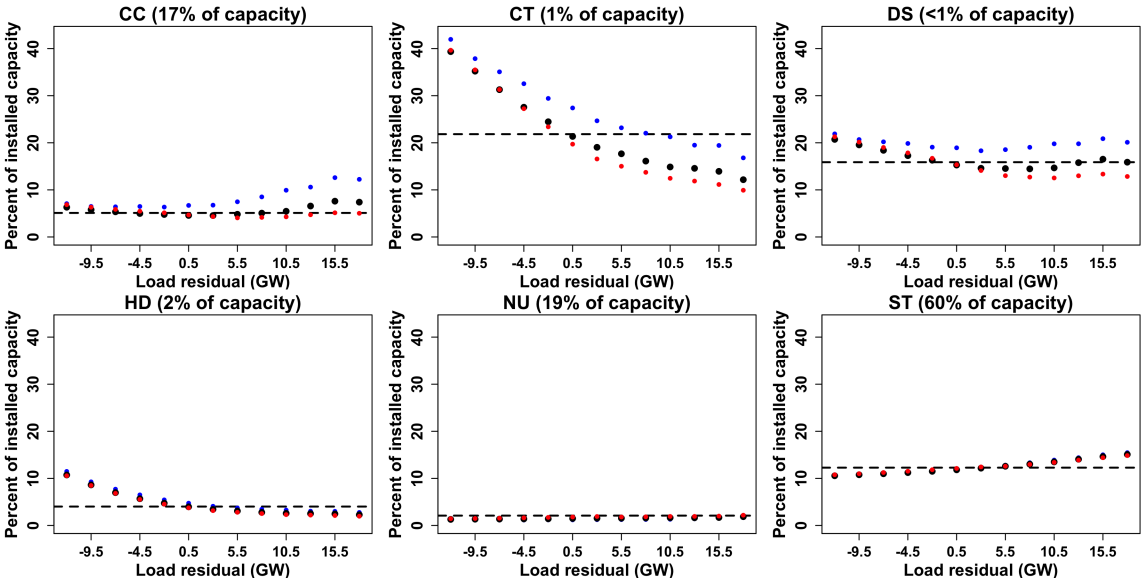

Figure C.16: Expected levels of unavailable capacity as a function of load under logistic regression (dots) and current practice (dashed horizontal line) when restricting to temperatures below 18.3 degrees Celsius (2004-2018 model fits). Black dots computed at median temperature from load neighborhood; blue and red dots correspond to 10^{th} and 90^{th} percentile temperatures from load neighborhood, respectively. Load neighborhood defined analogously to temperature neighborhood of Figure C.14. Current practice dashed line matches that of Figure C.14. Plot domain defined using only observations below 18.3 degrees. Abscissa spans different values than Figure 3.7 because load stationarizing procedure computed independently for 2004-2018 model fit. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

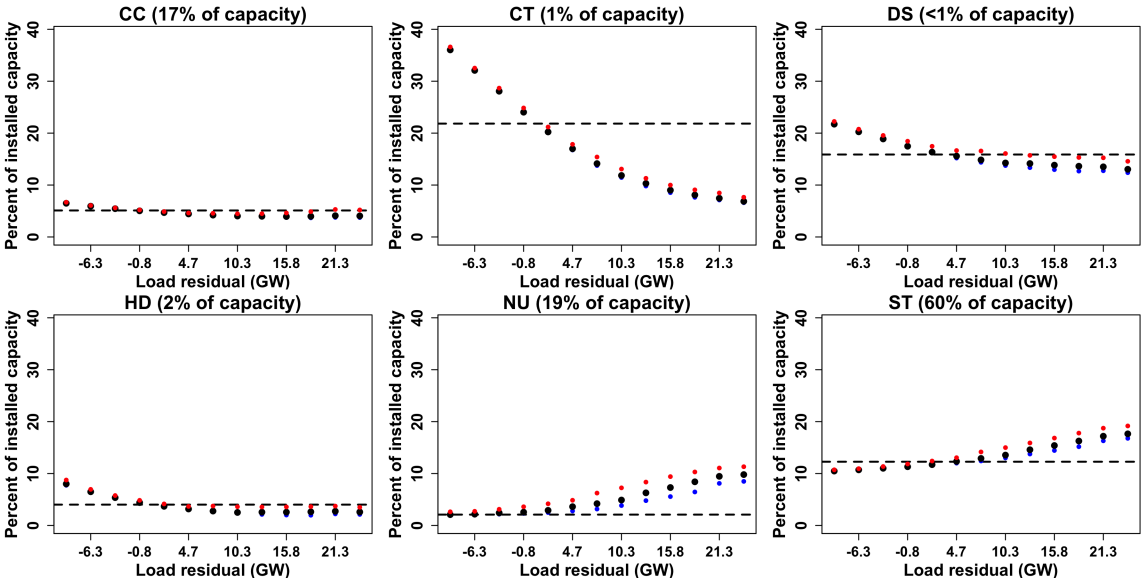

Figure C.17: Expected levels of unavailable capacity as a function of load under logistic regression (dots) and current practice (dashed horizontal line) when restricting to temperatures above 18.3 degrees Celsius (2004-2018 model fits). Black dots computed at median temperature from load neighborhood; blue and red dots correspond to 10^{th} and 90^{th} percentile temperatures from load neighborhood, respectively. Load neighborhood defined analogously to temperature neighborhood of Figure C.14. Current practice dashed line matches that of Figure C.14. Plot domain defined using only observations above 18.3 degrees. Abscissa spans different values than Figure 3.8 because load stationarizing procedure computed independently for 2004-2018 model fit. CC is combined cycle, CT is simple cycle, DS is diesel, HD is hydroelectric and pumped storage, NU is nuclear, ST is steam turbine.

Chapter 4 Resource adequacy implications of temperature-dependent generator availability*

Abstract

Standard resource adequacy modeling assumes generator failures are unconditionally independent and invariant to ambient conditions. We evaluate the resource adequacy implications of correlated failures in the PJM Interconnection by making use of recently developed temperature-dependent forced outage rates for the grid operator's conventional generator fleet. To carry out this analysis, we modify the Renewable Energy Capacity Planning Model, an open-source resource adequacy tool, and parameterize it for PJM's system. We demonstrate correlated failures pose substantial resource adequacy risk, increasing PJM's required reserve margin from 15.9% to 22.9% to achieve their target loss-of-load-expectation in the 2018/2019 delivery year. However, PJM actually procured a 26.6% reserve margin for this delivery year, translating to excess capacity payments of $315 million per year and an implied value of lost load of approximately $700k/MWh, a figure two orders of magnitude greater than typically used in operational contexts. This suggests substantial potential benefit in applying our method to aid grid operators in quantifying the capacity levels necessary to achieve a desired reliability target. Capacity requirements vary by month, with more than 95% of loss-of-load risk accruing in July, due to high temperatures increasing both demand for electricity and the probability of generator failures. We find that setting monthly capacity targets instead of an annual target could reduce procurement by approximately 16% in PJM over the delivery year. Given economic pressure being exerted on nuclear and coal generators in PJM, we examine the resource adequacy implications of replacing them with combined-cycle gas generators. We find moderate benefits of these resource transitions: approximately 1% and 3% reduction in capacity requirements, respectively, driven by lower forced outage rates for combined-cycle gas generators during peak-load hours. We identify modest resource adequacy risks from potential future climate scenarios, operationalized as temperature increases of 1 and 2 degrees Celsius relative to our study period. Holding loads fixed, these scenarios increase capacity requirements by approximately 0.5% and 1.5%, respectively. The resource transition scenarios and future climate scenarios offset each other to varying degrees, resulting in modest changes in capacity requirements in order to maintain the target reliability level.

* Joint work with Luke Lavin.

4.1 Introduction

Significant attention has been given to the question of how much generation capacity is required to ensure power system reliability [1–12]. In the U.S., system operators have formalized probabilistic approaches for determining capacity requirements in their balancing area. These methods are designed to satisfy mandatory reliability targets established by the North American Electric Reliability Corporation (NERC) and approved by the U.S. Federal Energy Regulatory Commission (FERC) [13]. In parts of the U.S. where planning and operation are mediated by an independent system operator (ISO) or Regional Transmission Organization (RTO), capacity procurement processes are also reviewed by planning leadership and stakeholder bodies. In the subset of ISOs with competitive, centralized capacity procurement, capacity obligations for load-serving entities in the balancing area may be used to parameterize the demand curve for the capacity market [14, 15]. Capacity markets are intended to help provide generators with revenue sufficiency and stability in support of long-term resource adequacy [16, 17].

Despite these quantitative procedures, there is a tendency in recent years toward capacity over-procurement. A recent NERC Summer Reliability Assessment reports that most assessment areas have anticipated reserve margins well in excess of their NERC reference summer target [18]. Similarly, the PJM Interconnection LLC (PJM), the largest system operator by installed generating capacity and load in North America, forecasts a planning reserve margin of approximately 25-29% over the period of 2019-2028, while their annual Reserve Requirement Study calls for no more than 16% over that period [19]. Higher procurement may be justified by a low cost of maintaining existing supply in an era of low load growth [20] or in response to systematic under-forecasting of high demand events. In PJM's case, several recent weather events have stressed the grid, including Hurricane Sandy (2012), the Polar Vortex of January 2014, and cold snaps in subsequent winters. This suggests that current approaches to resource adequacy modeling are not capturing important reliability risks, and that better methods may be able to reduce the need for heuristics in determining capacity procurement.

Recent research examined key assumptions made in resource adequacy modeling. Using four

years of Generating Availability Data System (GADS) data for the eight North American reliability regions, [21] demonstrates large generation outage events—correlated generator failures—occur with much greater frequency than is consistent with the assumptions of unconditional independence of generator failures and constant failure probabilities currently made [19]. Assuming unconditional independence and constant generator failure probabilities lowers the probability of large-outage events at the system level.

To identify a plausible mechanism to explain observed correlated failures, subsequent research explicitly modeled generator transition probabilities. Based on 23 years of GADS data for the PJM Interconnection, [22] used logistic regression to demonstrate significant correlations between transition probabilities and temperature. At both very cold and very hot temperatures, PJM's fleet is less available than on average. For example, nuclear generators are less available at high temperatures, likely due to cooling water constraints, while natural gas generators are less available at low temperatures, in part due to fuel unavailability. This temperature dependence was also shown to capture most of the observed correlated failures over the data period. Given that extreme temperatures tend to be associated with high loads and that temperatures are highly spatially correlated at the scale of an individual balancing area such as PJM, it follows that temperature dependence could pose significant resource adequacy risks for PJM.

Here we examine the resource adequacy implications of temperature-dependent generator availability. To do this we use the temperature dependence of generator resources identified in [22] and a modified version of the Renewable Energy Capacity Planning Model (RECAP) [23], an open-source resource adequacy tool originally developed by Energy+Environmental Economics (E3) in collaboration with the California Independent System Operator.[1] Like the resource adequacy planning tools used by PJM, RECAP computes loss of load expectation (LOLE) given a parameterized fleet of generators and a load forecast. We modify RECAP to allow each generator's forced outage rate to depend on hourly ambient temperature in PJM's footprint, rather than fixing it at an average value

[1] E3 includes the following disclaimer with the user guide to the public version of the model: "RECAP is open-source and not a commercial software package, thus users are welcome to modify or tailor any part of the RECAP model to their needs. RECAP has a complicated set of inputs and model settings and E3 takes no responsibility for the validity of results produced by third parties using this model, which requires some intuition and knowledge of the methodology to use successfully." We change the underlying RECAP Python source code to enable input of hourly generator forced outages dependent on ambient conditions.

in a given month or year, as is common industry practice. We use this modification to compute capacity requirements for PJM under two scenarios: current practice (representing assumptions of unconditional independence and constant failure probabilities) and when allowing generator availability to depend on temperature. By comparing these two sets of results, we can identify when violations of planning assumptions are most likely to result in loss-of-load events. This information can be used to incentivize appropriate risk mitigation investments. Given that extreme temperatures are seasonal, we also use RECAP to set monthly capacity targets to achieve the same reliability metric and quantify the implied reductions in capacity that could be obtained versus the current annual procurement approach. Finally, we explore how various bounding changes to the resource mix and future temperature increases under climate change scenarios may affect resource adequacy in PJM.

We find that relaxing the restrictive assumptions of unconditionally independent generator failures and constant generator failure probabilities uncovers substantial latent resource adequacy risk: a 6% increase in annual procurement (corresponding to a 7 percentage point increase in the reserve margin) is required to maintain the target 0.1 LOLE in the 2018/2019 delivery year. This is driven by the temperature dependence of generator availability, which can lead to substantially different distributions of available capacity at different hours of a given day, in conjunction with strong spatial dependence of temperature across the PJM footprint. Monthly capacity procurement targets could reduce annual average procurement by 16% in PJM with negligible impact on LOLE. An increase in temperature of 2 degrees Celsius from current levels would increase annual procurement requirements by 1.5%. Replacement of all nuclear and coal generation in PJM with equivalently sized combined-cycle gas generators having the same forced outage rates as PJM's existing combined-cycle generators leads to a reduction in annual procurement requirements of 1% and 3%, respectively, due to combined-cycle generators having lower forced outage rates during peak-load summer hours which have the highest LOLE. Finally, combining the temperature and retirement cases described above results in a 1% increase and 0.5% decrease in capacity requirements when replacing nuclear and coal generators, respectively.

The remainder of the paper is organized as follows. Section 4.2 reviews how resource adequacy

modeling is currently conducted in PJM. Section 4.3 describes the RECAP model. Section 4.4 describes data development, modeling, and RECAP parameterization. Section 4.5 presents results. Section 4.6 concludes.

4.2 Overview of resource adequacy modeling in PJM

Resource adequacy is concerned with ensuring sufficient forward procurement of generation capacity to serve firm load within the operational parameters of the grid. Each year PJM conducts a resource adequacy planning analysis that seeks to determine the level of capacity required to limit the frequency of loss-of-load events to once in 10 years, termed 0.1 loss of load expectation (0.1 LOLE) [13].[2] Each analysis forecasts capacity requirements for the next 11 delivery years.[3] The modeling process considers historical generator forced outage rates, generator maintenance requirements, load forecast error, peak load variability, interzonal transmission constraints, and emergency transfer capacity from neighboring synchronous balancing authorities [19, 25]. The installed reserve margin (IRM) is the quantity of capacity required to meet the reliability criterion, reported as a percentage above the forecast 1-in-2 (median) peak load [19, 25].

PJM forecasts the 1-in-2 peak load for the delivery year using an econometric model in conjunction with historical temperature profiles [19, 20, 25, 26]. To determine the peak demands for each week of the future delivery year, PJM takes the five most recent years of metered loads and normalizes each daily peak by that year's peak value. These normalized daily peak values are used to sort the 52 weeks of each historical load year in descending order by weekly peak by season. After magnitude-ordering the weeks, the 25 normalized peak-load observations for each calendar week are used to calculate a mean and standard deviation, defining a Gaussian distribution for that week, assumed to apply to each weekday of the week. This information, along with the forecast peak load and a forecast error factor, is used to calculate the most probable peak load for each week of the

[2]The LOLE metric used by PJM does not consider either the duration or magnitude of the load-shed event. Some U.S. system operators use a different interpretation where "once in ten years" instead means 24 total hours of expected load shed per 10 years; some international systems consider magnitude by using a quantity of expected unserved energy as the reliability metric [24].

[3]The delivery year begins June 1 and ends May 31 of the following calendar year.

future delivery year. The most probable peak for week i (MPP_i) is calculated as:

$$MPP_i = \mu_i + 1.163 * \psi_i \tag{4.1}$$

where μ_i is the mean of the Gaussian distribution for week i, $\psi_i = \sqrt{\sigma_i^2 + FEF^2}$ is the standard deviation after including the forecast error factor (FEF), σ_i is the standard deviation of the Gaussian distribution for week i, and 1.163 is a multiplier that relates the expected value of the maximum of a sample drawn from a Gaussian distribution to the mean and standard deviation of that distribution, based on the number of draws [27].[4] Uncertainty in each weekly peak load is assumed to follow a Gaussian distribution centered at MPP_i with standard deviation ψ_i. For purposes of calculating LOLE, PJM represents each week's Gaussian as a probability mass function using 21 equally spaced points spanning +/- 4.2 standard deviations. As will be described in detail below, the goal of PJM's procedure is to determine whether sufficient generation capacity will be available to meet each of the 21 points in each week of the delivery year. Points for which this is not true accrue LOLE.

With the 21 possible peak load values for each week of the delivery year established, PJM next forecasts the performance of its generator fleet for the delivery year. For each generator expected to serve the balancing area during the delivery year, PJM uses the most recent 5 years of GADS data to calculate the forced outage rate.[5,6] Each generator is modeled as a two-state homogeneous Markov model, with available and unavailable states, where the forced outage rate is assumed to represent the probability of the generator being unavailable when needed by the system. Note that the forced outage rate is assumed to apply in all hours of the delivery year—i.e., the probability of failure is assumed to be constant.

To determine the probability distribution of available capacity for the power system, all possible

[4]Here, 1.163 is the first Gaussian order statistic for n=5 draws, corresponding to five weekdays in each week.

[5]The precise forced outage rate used by PJM has changed over the years. Prior to the creation of the Capacity Performance capacity product, implemented through changes to PJM's capacity market in 2015 in response to poor generator performance during the 2014 Polar Vortex, PJM used XEFORd. XEFORd does not penalize generators for events considered "outside management control". For Capacity Performance resources, all failures are now considered the responsibility of the plant owner, and the availability statistic has switched to EFORd. Procurement of Capacity Performance resources began with the Base Residual Auction for the 2018/2019 delivery year (held in 2015); the Base Residual Auction for the 2020/2021 delivery year was the first that procured only Capacity Performance resources [28, 29].

[6]Generators with fewer than 5 years of historical data are supplemented by class-average data.

combinations of generator states are enumerated [19, 26]. The probability of a given state is computed as the product of the relevant generator-level probabilities, thereby assuming unconditional independence of generator failures. This forecast distribution of available capacity is then compared to the forecast peak load distribution in each weekday of the delivery year (approximately 260 days), as described previously, to compute the total LOLE for the delivery year [19]:

$$\begin{aligned} LOLE &= \sum_{i=1}^{260} LOLE_i \\ &= \sum_{i=1}^{260} \sum_{j=1}^{21} LOLP_j \\ &= \sum_{i=1}^{260} \sum_{j=1}^{21} P(D_j) * P(G < D_j) \end{aligned} \tag{4.2}$$

where $LOLE_i$ is the aggregate loss of load expectation for weekday i of the delivery year, $LOLP_j$ is the loss of load probability at the load value corresponding to the j^{th} position in the 21-point Gaussian representing weekday i's MPP, $P(D_j)$ is the probability of the j^{th} load value occurring, and $P(G < D_j)$ is the probability of available generation (i.e., generation not on forced, maintenance, or planned outage) being insufficient to meet the j^{th} load value.[7] The forecast peak load is then scaled until aggregate loss of load expectation (equation 4.2) precisely satisfies the reliability criterion (0.1 LOLE), and then the final IRM for the delivery year is calculated.

The IRM is used to define PJM's piecewise-linear demand curve for capacity for the corresponding delivery year, termed the variable resource requirement (VRR) curve. Capacity obligations for participating load-serving entities are procured through a set of forward auctions in the PJM capacity market, termed the reliability pricing model (RPM). The first auction is held three years prior to the start of the delivery year, with three subsequent auctions used to address any changes in capacity requirements that may emerge [25].

[7] G is shifted to account for ambient deratings during the summer and aggregate maintenance requirements (primarily scheduled during off-peak months). While it would therefore be more precise to use G_i to represent the distribution of available capacity, in broad strokes this distribution is the same across the delivery year due to the assumption of constant failure probabilities.

As will be discussed in detail below, our contribution is to evaluate the resource adequacy implications of relaxing the assumption of an invariant G, based on previous research demonstrating that the distribution of available capacity strongly depends on temperature [22].

4.3 RECAP

RECAP [23] was developed by Energy+Environmental Economics (E3) in collaboration with the California Independent System Operator to improve resource adequacy valuation of unconventional capacity resources like wind and solar. RECAP is similar in scope and capabilities to PJM's own resource adequacy tools and has been used by system operators, utilities, and in regulatory processes. It is implemented in Python with a Microsoft Excel front end for case development. We briefly describe the main functionality of the open-source version of the model used in this paper and our modifications.

RECAP divides the 8760 hours of the delivery year into time-slice bins for quantifying loss-of-load risk. By default, there are 576 time-slice bins: month (12) x hour (24) x weekday/weekend (2).[8] The user may define additional bins by load level within the month x hour x weekday/weekend time-slices to improve the model's ability to represent peak loads. Using historical input data, the contribution of wind and solar generation in each time-slice bin is subtracted from the load. RECAP then fits a user-selected probability distribution to the net-load data in each bin; we employ Gaussian distributions for consistency with PJM.

RECAP uses the forced outage rate for each conventional generator to compute a probability distribution of available capacity for the power system in each time-slice bin. We modify the RECAP source code, which by default can only consider generator outage probabilities at a monthly granularity, to consider hourly ambient temperature when generating this probability distribution. In this way we relax the assumption that available capacity of conventional generation is invariant to temperature.

[8] Recall from Equation 4.2 that PJM assumes that only non-holiday weekdays represent LOLE risk, while RECAP includes weekends and holidays. There is little discrepancy introduced by this difference: we find that weekends represent less than 0.1% of total LOLE under both the temperature-dependent and the unconditional forced outage rates scenarios.

RECAP computes the probability that net load exceeds available generation (termed the loss of load probability, LOLP) in each bin through convolution of the corresponding net load and generation availability probability density functions. Results are aggregated across any additional user-defined load bins, weighting appropriately.[9] By default, RECAP sums up resulting LOLP over all hours of the delivery year; however, to better match PJM's definition of LOLE, we modify RECAP to consider only the peak LOLP hour in each day.

4.4 Data development, modeling and parameterization

We parameterize RECAP with PJM's conventional generator fleet, hourly wind and solar generation profiles, hourly temperature data, and normalized historical PJM hourly load data. We discuss the development of each set of inputs below.

4.4.1 Conventional generator fleet

We use the PJM GADS database [30] to define the currently operating conventional generators serving the PJM balancing area, which includes all generators other than wind, solar, and behind-the-meter resources.[10] From PJM GADS we obtain full operating histories of each conventional generator serving the balancing area between 1995-2018Q1. In RECAP we parameterize conventional generators by their nameplate capacity and generator type, both of which are reported in PJM GADS, and their forced outage rate, which we calculate from PJM GADS.

We develop both temperature-dependent and unconditional forced outage rates for each generator. Temperature-dependent forced outage rates are calculated using the logistic regression approach described in [22]. In brief, the model specification includes linear and quadratic terms for both cold and hot temperatures in each hour, as well as a linear load term; the threshold delineating cold and hot temperatures is 18.3 degrees Celsius (65 degrees Fahrenheit). After fitting these models using all

[9] For example, if three load bins were used, with splitting occurring at the 80^{th} and 90^{th} percentiles, the three LOLP values for the current month x hour x weekday/weekend level would be weighted by 80%, 10%, and 10%.

[10] PJM GADS resource types include the following: combined-cycle gas (abbreviated as CC in figures and tables), simple-cycle gas (CT), diesel (DS), hydroelectric and pumped storage (HD), nuclear (NU), and steam turbine (ST). In 2017, the vast majority (95%) of ST generation in PJM was from coal, thus we use the two terms interchangeably [15].

available data, we evaluate them at a temperature of interest (along with the median load value at that temperature) to obtain a transition probability matrix governing failures and recoveries for each generator at that temperature. Eigendecomposition allows us to obtain the ergodic probability of a generator being unavailable at that temperature. We repeat this procedure over a temperature range of -30 degrees Celsius to 40 degrees Celsius for each generator and compute the average by generator type.

Repeating this process for each generator sweeping over temperatures allows us to obtain our final result: temperature-dependent forced outage rates.[11]

We calculate unconditional forced outage rates using the following equation:

$$EFOF_i = (FOH_i + EFDH_i)/PH_i \tag{4.3}$$

where $EFOF_i$ is the equivalent forced outage factor for generator i,[12] FOH_i (forced outage hours) is the sum of hours where generator i experiences a forced outage, $EFDH_i$ (equivalent forced derating hours) is the sum of hours where generator i experiences a forced derating, reported on a full-outage-equivalent basis, and PH_i (period hours) is the total number of hours of data reporting for generator i. While resource adequacy studies in PJM are conducted using the five most recent years of GADS data, here we use all historical data for each generator, consistent with our approach for computing temperature-dependent forced outage rates.[13] As with the temperature-dependent forced outage rates, we then compute averages by generator type. We use the unconditional forced outage rates to establish a baseline representing current PJM practice. This allows us to quantify latent resource adequacy risk from temperature-dependent generator availability.

Temperature-dependent distributions of available capacity for PJM's conventional generator fleet are shown in Figure 4.1. The distributions of available capacity at cold and hot temperatures are shifted to the left of the distributions at moderate temperatures, indicating less available capacity on

[11] See [22, Procedure 4] for more information.

[12] $EFOF$ is an approximation of $EFORd$, the availability statistic actually used by PJM, that allows us to circumvent data reporting limitations in PJM GADS prior to 2004. For further discussion see [22, Section 4].

[13] Given the infrequency of generator failures, the logistic regression approach requires many years of data to avoid introducing bias into the estimated transition probabilities [31]; see [22, Section 2] for more details.

average; they are also shorter and wider than the distributions at moderate temperatures, indicating greater risk of large deviations from the average.

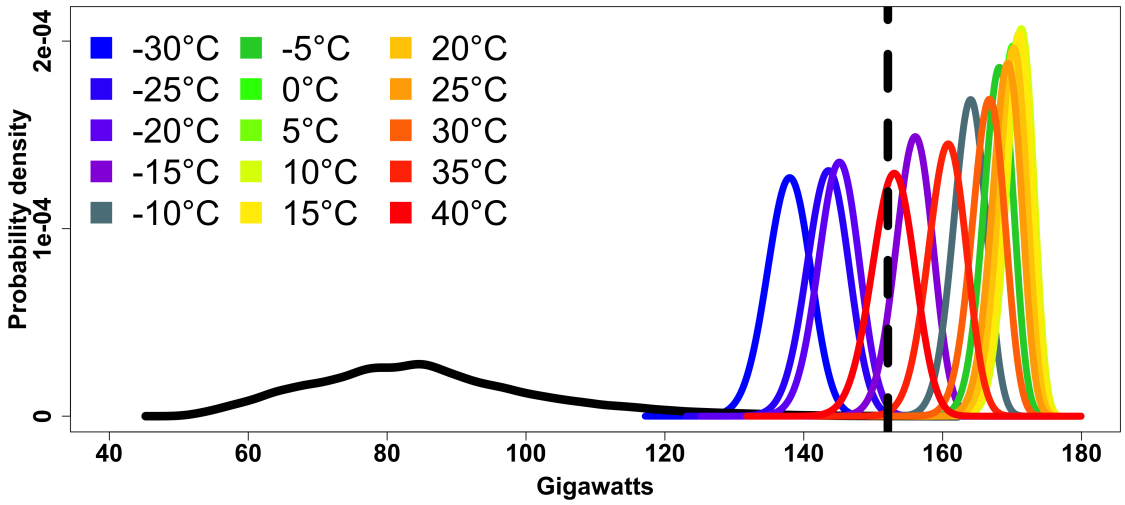

Figure 4.1: Distributions of available capacity as a function of temperature for PJM's conventional generator fleet (narrow curves, various colors). Note reduced capacity at both cold and hot temperatures. Wide black curve is the distribution of the 1-in-2 (median) load forecast. Dashed black line is the peak value from the 1-in-2 load forecast.

A comparison of available capacity for selected hours in July for the temperature-dependent and unconditional forced outage rate scenarios is shown in Figure 4.2. The temperature-dependent afternoon distribution is shifted to the left of, and has a higher variance than, the unconditional forced outage rate distribution due to the high temperatures that occur during afternoon hours in July. Current resource adequacy modeling procedures do not account for this shift in the available capacity distribution when determining capacity requirements.

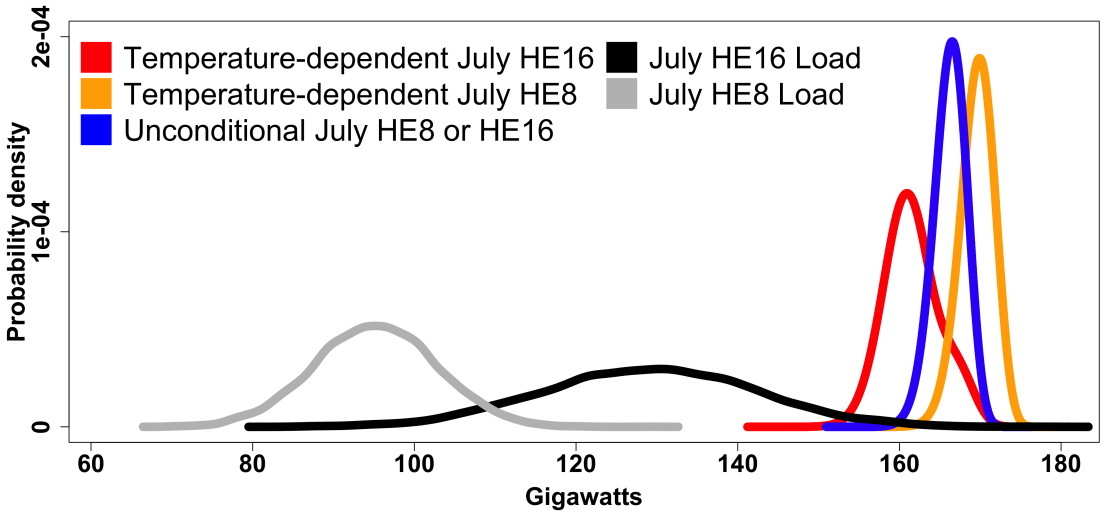

Figure 4.2: Illustration of how temperature-dependent forced outage rates can identify latent resource adequacy risk. Distributions of available capacity for selected July hours using temperature-dependent forced outage rates shown in red and orange; the distribution of available capacity when using unconditional forced outage rates shown in blue. Black and gray curves indicate load distributions for corresponding July hours. "HE" indicates hour-ending. Current practice does not recognize that the distributions of available capacity will differ between morning and afternoon due to the temperature differential, and consequently understates the risk of not being able to serve load.

4.4.2 Wind generation

To model wind generation output and its coincidence with peak load, RECAP requires an hourly normalized historical generation profile (i.e., as a fraction of maximum output) and quantity of installed wind capacity for the modeled delivery year. Because of the relatively recent vintage and sparse public hourly generation data for most wind generators in PJM, we choose to model wind generation rather than use empirical profiles.

To develop normalized wind generation profiles, we first identify online wind generators serving PJM as of December 2017 using Energy Information Administration (EIA) Form 923 [32]. Identified generators are then matched against the U.S. Wind Turbine Database [33] to obtain information about physical location, generator online date, turbine hub height and turbine technology, as well as to verify installed capacity. We exclude all installations with less than 4 MW capacity and those that could not be matched, leaving 89 wind generation locations in PJM totaling 8.3 GW of installed

capacity (more than 97% of total installed capacity).

To simulate hourly output from each of these wind sites, we then match their location data against the closest available 100-meter hub height wind speeds from the National Renewable Energy Laboratory's (NREL) Wind Integration National Dataset Toolkit [34], which contains 5-minute windspeeds for 2007-2012. Matched 100-meter windspeeds are scaled to individual generator hub heights using [35, Equation 2]. Wind speeds are then converted to power output using an appropriate power curve given the vintage and type of installation [34]. To account for degradation in wind turbine performance over time, we apply a historical degradation factor as a function of generator age in 2018 using Lawrence Berkeley National Laboratory's 2017 Wind Technologies Market Report [36, Figure 39].

Our resulting simulated wind generation output for 2007-2012 has a 39% capacity factor compared to a reported capacity factor of 32.6% for installations in the Great Lakes region and 44.4% for installations in the Interior region of the country [36, Figure 41]. Because most PJM wind capacity is located in Illinois, Indiana, and Ohio, closely mapping to the Great Lakes region, we derate our simulated hourly profiles to match the reported empirical 32.6% Great Lakes Region capacity factor.[14] Though our approach and RECAP's functionality enable calculation of the capacity value of each wind generation site in PJM individually, for study scope and model runtime reasons we aggregate generation into a single shape for input into RECAP.

4.4.3 Solar generation

As with wind generation, RECAP takes two key inputs for solar generation: hourly normalized generation profile(s) as a fraction of maximum output, and a total quantity of installed capacity corresponding to the profile(s) for the modeled delivery year. Given our focus on conventional resources and the relatively low level of utility-scale solar capacity in PJM, we model a single aggregate solar generation profile.

To develop the hourly normalized solar profile for PJM, we obtain solar radiation and weather data from NREL's National Solar Radiation Database [38, 39], and convert it to alternating current

[14] Reasons for simulated generation exceeding actual performance may include curtailment, turbine down time for maintenance, wake effects, and icing on blades [37].

(AC) generation output using Pacific Northwest National Laboratory's (PNNL) GridLab-D solar panel and inverter objects [40]. GridLab-D uses the same solar modeling as NREL's System Advisory Model [41], a widely used engineering-economic tool.

Key assumptions for our PJM solar shape include a fixed solar panel tilt of 30 degrees, a solar multiplier of 1.2, an inverter efficiency of 96%, a panel efficiency of 17% and a constant power factor of 1.0. We use a south-facing panel orientation, which maximizes capacity factor. Hourly modeled profiles are adjusted to account for daylight savings time and leap years. The resulting annual average AC capacity factor for this PJM shape is 19.3%.

Using EIA-923 [32] we identify approximately 2 GW_{AC} of installed utility-scale solar capacity online in PJM as of December 2017. This includes 1.3 GW_{AC} of capacity in Maryland, Virginia, Delaware, Ohio, Pennsylvania, Indiana, Illinois, and Kentucky, and an additional 0.7 GW_{AC} in North Carolina that may be deliverable to PJM. We assume behind-the-meter solar generation is accounted for in the input load profiles.

4.4.4 Temperature data

We obtain hourly historical temperature data from the National Oceanic and Atmospheric Administration [42]. Temperature data are used for two purposes: first, to specify a load regression and obtain a time trend for PJM loads; and second, to estimate the ambient temperature in PJM for the 2018 delivery year to model the probability of hourly generator forced outages under the temperature-dependent scenario.

We select weather stations corresponding to Chicago, Cleveland, Philadelphia and Washington, D.C. for 2006-2017, the same time period as our historical load data. We process each into hourly time series and then average them together to derive our input temperature series for PJM. Over this period there was nearly complete reporting for each weather station: in total only 25 hourly observations were missing across the four weather stations. Missing observations were interpolated using the nearest non-missing observation.

4.4.5 Load forecast

We obtain hourly historical metered load data from PJM by zone for 2006-2017 and aggregate to a PJM-wide load shape for input to RECAP. 2006 is chosen as a starting point for historical loads because in the years immediately prior there were significant changes to the PJM footprint with the addition of Duquesne Light (2005), Dominion (2005), Commonwealth Edison (2004), American Electric Power (2004), and Dayton Power and Light (2004). Since 2006, three zones have been added to PJM: American Transmission Systems Inc. (ATSI, 2011), Duke Energy Ohio and Kentucky (DEOK, 2012), and East Kentucky Power Cooperative (EKPC, 2013). To account for these missing zonal loads, we develop a correlation matrix between ATSI, DEOK, EKPC, and the zones that were present since 2006. We then match ATSI, DEOK, and EKPC to the zone with which each is most highly correlated. ATSI and DEOK most closely match Dayton Power and Light, while EKPC is matched to American Electric Power; the correlation of each of the three zones with their matched zone is always greater than 0.95. We then fill in the unobserved loads in ATSI, DEOK, and EKPC using the corresponding load in the matched zone, scaling by the ratio of the average loads.

With the unobserved loads added back in, we next seek to estimate the time trend in the historical PJM load data so that we can make the historical values comparable to loads in our future delivery year. We model daily loads using the following linear regression specification:

$$\begin{aligned} Load_t = {} & Chi\_HDD_t + Chi\_CDD_t + Cle\_HDD_t \\ & + Cle\_CDD_t + DC\_HDD_t + DC\_CDD_t \\ & + Phi\_HDD_t + Phi\_CDD_t + Weekday_t + T_t \end{aligned} \tag{4.4}$$

where $Load_t$ is the sum of the hourly loads in day t, Chi, Cle, DC, and Phi are abbreviations of the four cities in PJM's footprint for which we collect airport temperature data (Chicago, Cleveland, Washington, D.C., and Philadelphia), CDD_t is cooling degree days in day t, HDD_t is heating degree days in day t, $Weekday$ is a Boolean variable indicating whether day t is a non-holiday weekday, and T is a linear time trend.

We use the coefficient of T when modeled on the full 2006-2017 load data to de-trend historical loads to the 2018 delivery year for input to RECAP. In addition to the historical loads, RECAP takes in the forecast peak load and forecast total annual load for the delivery year. The 1-in-2 (median) unrestricted forecast peak load for our delivery year is 152.1 GW and the forecast annual load is 806.7 TWh (92.1 GW_{avg}) [43].

4.4.6 RECAP parameterization

Finally we parameterize RECAP with the developed generator fleet and load forecast for the modeled delivery year. We summarize our base parameterization in Table 4.1, matching PJM's parameterization reported in the 2018 Reserve Requirement Study (RRS) to the extent feasible [19].

We note that our parameterization does not account for all modeling details that are included in PJM's resource adequacy analysis. In particular, we do not include transmission constraints nor emergency imports from neighboring power systems. We do not include demand response (DR) due to the simplifications required to model duration-limited resources in RECAP; however, for this reason PJM also omits DR from the RRS.[15]

4.5 Results

We conduct several RECAP runs to examine the resource adequacy risks of temperature-dependent generator availability. First we compare capacity requirements to achieve 0.1 LOLE under unconditional and temperature-dependent forced outage rates. We next examine potential reductions in capacity requirements when setting monthly targets, rather than a single annual target. We then quantify the value of lost load (VOLL) implied by a range of reliability criteria by looking at the incremental change in expected unserved energy (EUE) associated with an incremental change in the quantity of capacity procured. Finally, we evaluate the resource adequacy implications of selected future generation and temperature scenarios.

[15] At minimum, the total number of calls per delivery year and the maximum duration of a call need to be specified to endogenously model DR's contribution to resource adequacy. This requires tracking system state over sequential hours. The version of RECAP modified for this study does not model sequential hours, though it does allow for ex-post analysis of DR's contribution by outputting hourly LOLP for the input historical hours.

Table 4.1: Parameterizing the RECAP model.

Metric	Value
Conventional generation[1, 2]	207.5 GW installed capacity
Solar generation	2.0 GW installed capacity
Wind generation	8.3 GW installed capacity
Demand response[3]	0 GW
Emergency imports[4]	0 GW
Footprint	All current PJM zones
Zonal disaggregation	None
Scheduled outages	Average historical requirements by generator type and calendar month
Peak load forecast[5]	152.1 GW
Load percentile bins[6]	0-80, 80-90, 90-100
Reliability metric[7]	0.1 LOLE per year
Capacity addition resource[8]	Combined-cycle gas generator

[1] We enforce a 2.5 GW reduction to thermal generator capacity July–August. This corresponds to the 2.5 GW of thermal capacity that PJM puts on planned outage during weeks 6-15 of the delivery year to account for thermal generator capability reductions driven by high temperature and humidity.

[2] This number includes all non-retired conventional resources reporting to GADS as of 2018 Q1. It includes resources in the PJM footprint as well as resources external to PJM which have obtained firm transmission to the PJM system and firm available transfer capability into PJM [19]. For comparison, 209 GW of installed capacity was eligible to be offered into the 2018/2019 Base Residual Auction, 189.6 GW of which was offered [28].

[3] We do not include demand response (DR) in the model due to the difficulty of parameterizing it in RECAP; PJM similarly omits it from their annual Reserve Requirement Study.

[4] Represents emergency capacity from neighboring power systems

[5] Corresponds to the peak load used in delivery year 2018/19 [19].

[6] Percentiles of the load in each hour x month x weekday/weekend "bin". Gaussian distributions are subsequently fit to each set of load observations.

[7] PJM does not consider weekends in calculating LOLE.

[8] This is the resource assumed to be built if the load forecast results in greater than 0.1 LOLE for the existing fleet.

4.5.1 Aggregate effect of accounting for temperature dependence

We begin by determining capacity requirements for achieving the 0.1 LOLE reliability target when treating generator forced outage rates as invariant to temperature. In this scenario, 174.5 GW of conventional generation capacity, corresponding to a 15.9% planning reserve margin, is required to cover the forecast 1-in-2 peak load.[16,17] For comparison, PJM reports that a 16.2% planning reserve margin achieves 0.1 LOLE for the 2018/2019 delivery year [19]. The close agreement between these two results gives us confidence that we have captured the primary drivers of LOLE in PJM despite not accounting for emergency imports or zonal transmission constraints, nor using PJM's internal modeling tools.

When using temperature-dependent forced outage rates, capacity requirements to achieve 0.1 LOLE increase by 10.6 GW to 185.1 GW, corresponding to a 22.9% planning reserve margin. This difference represents substantial latent resource adequacy risk from the simplifying assumption that forced outage rates are invariant to temperature. However, even the higher 185.1 GW figure is less than the 192.6 GW of total capacity (26.6% planning reserve margin) PJM procured for the 2018/2019 delivery year [28, 44].[18] Using a capacity market clearing price of \$184/MW-day (\$67.2/kW-year), this translates to \$315 million per year in additional capacity payments beyond what is required to achieve 0.1 LOLE when accounting for temperature dependence in generator availability.[19] To the extent that PJM has set the parameters of its capacity demand curve to hedge against unmodeled resource adequacy risk driven by correlated failures, the method presented here offers a means of more precisely quantifying capacity requirements for achieving a target LOLE. While greater reliability is desirable, a demand curve for forward capacity procurement in markets like

[16] We report installed capacity values rather than unforced capacity values throughout this section. Unforced capacity values deflate installed capacity by the generator's forced outage rate and are used by PJM to determine capacity payments for generators that clear the capacity market.

[17] We focus on conventional generation requirements (i.e., capacity requirements net of wind and solar) when reporting capacity quantity results. However we include the capacity value of the wind and solar resources when reporting the corresponding planning reserve margin.

[18] The majority of capacity required to serve the balancing area is centrally procured on behalf of load-serving entities via the Base Residual Auction (BRA); however, load-serving entities are allowed to procure capacity through a separate bilateral mechanism termed the Fixed Resource Requirement (FRR). In the 2018/2019 delivery year, 177.4 GW was procured through the BRA and 15.2 GW was procured through the FRR.

[19] \$184/MW-day is the average of 2018/2019 final zonal capacity prices, weighted by zonal unforced capacity obligations [45]. PJM capacity prices have fluctuated in recent years for multiple reasons but generally clear between \$100/MW-day–\$200/MW-day; other centralized capacity markets have experienced similar or more price volatility [46].

PJM's is meant to balance the incremental costs and benefits of capacity, and should be articulated as accurately as possible to achieve this aim.

As a sensitivity analysis, we determine the quantity of capacity required to achieve a range of reliability targets, from 0.02 LOLE to 0.18 LOLE (Figure 4.3). Achieving a lower LOLE target requires additional capacity procurement, implying a higher value of lost load (VOLL) for PJM and its stakeholders. The difference in capacity requirements between the unconditional and temperature-dependent forced outage rates scenarios is reasonably constant, ranging from 10.4 GW at 0.18 LOLE to 11.0 GW at 0.02 LOLE. PJM's procurement of 192.6 GW places it at 0.02 LOLE when accounting for temperature-dependent forced outage rates, the most stringent procurement scenario we consider.

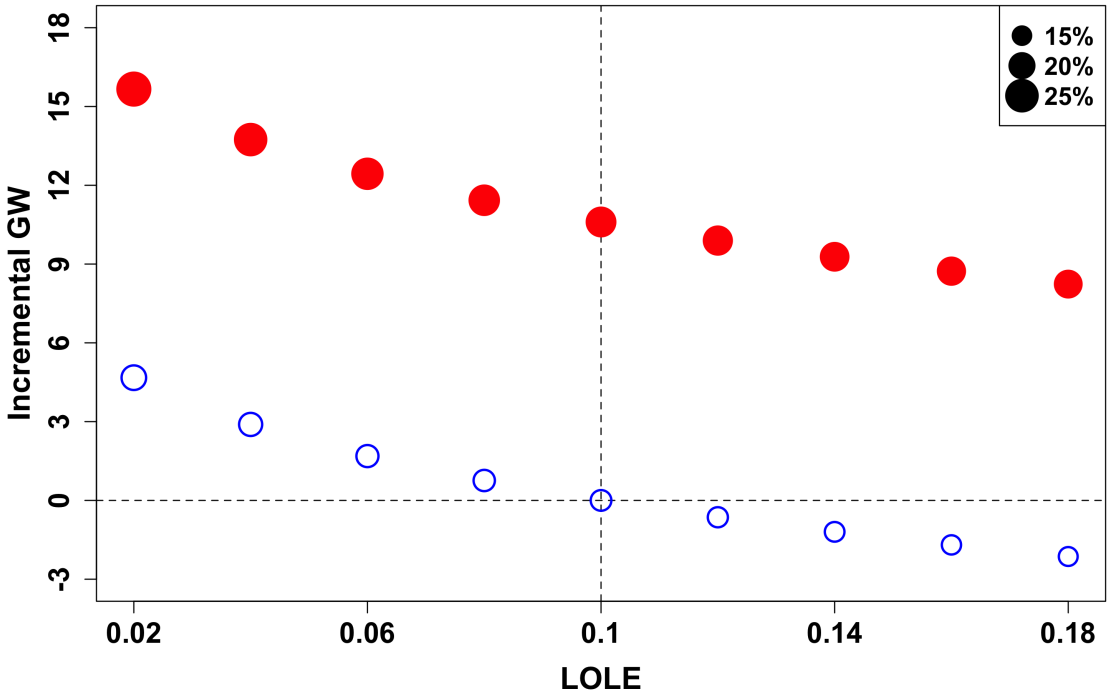

Figure 4.3: Incremental generator capacity required to achieve various LOLE targets under unconditional (hollow blue circles) and temperature-dependent (solid red circles) forced outage rates. Values are calculated with respect to the baseline reliability standard (0.1 LOLE) under unconditional forced outage rates. Circle size indicates corresponding planning reserve margin. For a fixed reliability level, the distance between the hollow blue circle and the red circle indicates the magnitude of latent resource adequacy risk in PJM's system when incorrectly treating forced outages as invariant with respect to temperature. Note that 0.02 LOLE in the unconditional forced outage rates scenario is a less stringent reliability target than 0.18 LOLE in the temperature-dependent forced outage rates scenario.

We then compute the incremental VOLL for each scenario. We do this by calculating the change in capacity procurement required to achieve the reliability target of interest, along with the corresponding change in EUE, reported in megawatt-hours (MWh). These values can be combined with an assumed cost of incremental capacity procurement to obtain an implied VOLL. We calculate the VOLL for both 0.08 LOLE and 0.12 LOLE using 0.10 LOLE as the baseline, and proceed "outward" from there (e.g., the VOLL for 0.06 LOLE is computed from a starting point of 0.08 LOLE, while the VOLL for 0.14 LOLE is computed from a starting point of 0.12 LOLE).[20] We again use a

[20]The VOLL corresponding to 0.1 LOLE is calculated as the arithmetic mean of the VOLLs for 0.08 LOLE and

capacity value of \$184/MW-day (\$67.2M/GW-year) to obtain our final result.[21] Implied VOLLs for this LOLE range are \$100k/MWh to \$700k/MWh, approximately two orders of magnitude higher than price caps at operational time scales in U.S. wholesale electricity markets [48, footnote 1]. As PJM's current procurement is at the lower end of our explored LOLE range, PJM's implicit VOLL corresponds to the upper end of our estimated VOLL range. These values add to previous studies that have argued that the 0.1 LOLE target may overvalue unserved energy and lead to a higher planning reserve margin than is economically optimal [47].

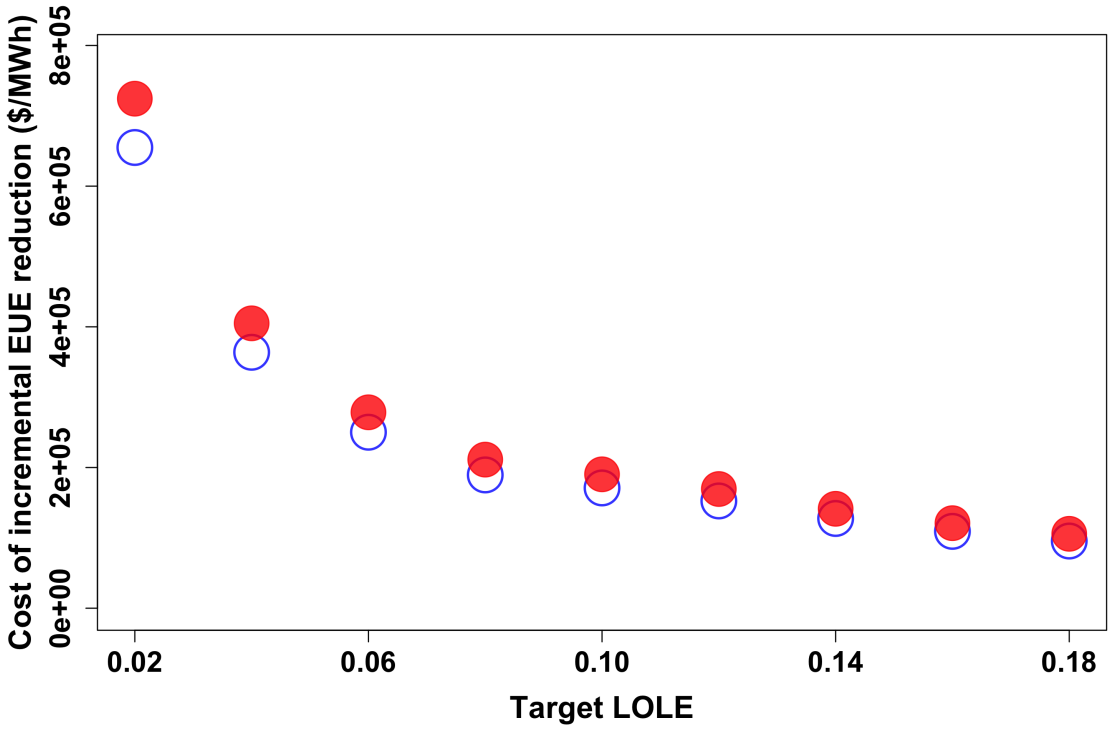

Figure 4.4: Incremental value of lost load (VOLL), in \$/MWh, as a function of the reliability target, under the temperature-dependent (solid red circles) and unconditional (hollow blue circles) forced outage rate scenarios. Values calculated using \$184/MW-day. With the exception of 0.1 LOLE, the VOLL at each reliability target is calculated with respect to the immediately adjacent reliability target in the direction of 0.1 LOLE; for example, the VOLL at 0.06 LOLE is calculated by computing the change in capacity procurement and expected unserved energy using 0.08 LOLE as the baseline. The VOLL corresponding to 0.1 LOLE is calculated as the arithmetic mean of the VOLLs for 0.08 LOLE and 0.12 LOLE.

0.12 LOLE.

[21] This likely represents an upper bound on the VOLL, given that PJM procured the maximum amount of capacity considered in this sensitivity analysis.

4.5.2 Resource adequacy implications of monthly capacity targets

Accounting for temperature dependence in generator failure rates significantly increases capacity procurement requirements over PJM's computed planning reserve margin (but as previously noted not above their procured reserves). Given the strong seasonality of extreme temperatures, we next consider whether and to what extent capacity requirements could be reduced through monthly procurement rather than the current PJM practice of annual procurement.

To conduct this analysis, we begin by looking at accumulated LOLE in each calendar month under an annual procurement approach, using both unconditional and temperature-dependent forced outage rates. In any month accounting for at least 0.01% of the total annual LOLE (i.e., 0.00001 LOLE given a 0.1 LOLE annual target), we retain the annual capacity procurement level as that month's requirement. For the remaining months—i.e., for those with zero, or effectively zero, LOLE—we allow RECAP to increase the LOLE to the monthly 0.01% contribution threshold by procuring a lower level of capacity.[22] In this way, we can quantify opportunities for reducing capacity procurement in low-load months without significantly increasing overall LOLE in the delivery year.

In the unconditional forced outage rates scenario, only July and August have more than 0.01% of annual LOLE, so we decrease capacity procurement until the LOLE threshold value is achieved in the other 10 months. This results in an average monthly procurement of 148.5 GW, a 15% reduction in capacity requirements on an annual basis. The result in the temperature-dependent forced outage rates scenario is similar, with again only July and August having more than 0.01% of annual LOLE. Here, decreasing procurement in the other 10 months to our threshold results in an average monthly procurement of 156.2 GW, a 16% reduction in capacity requirements on an annual basis. This capacity requirement is even 10% below annual base-case procurement in the unconditional forced outage rates scenario. A barplot showing monthly procurement requirements for achieving 0.1 LOLE under both forced outage rate scenarios is Figure 4.5.

[22] This is an arbitrary but reasonable target intended to illustrate the degree to which LOLE is concentrated in a small subset of months, effectively creating over-procurement in other months when procuring on an annual basis.

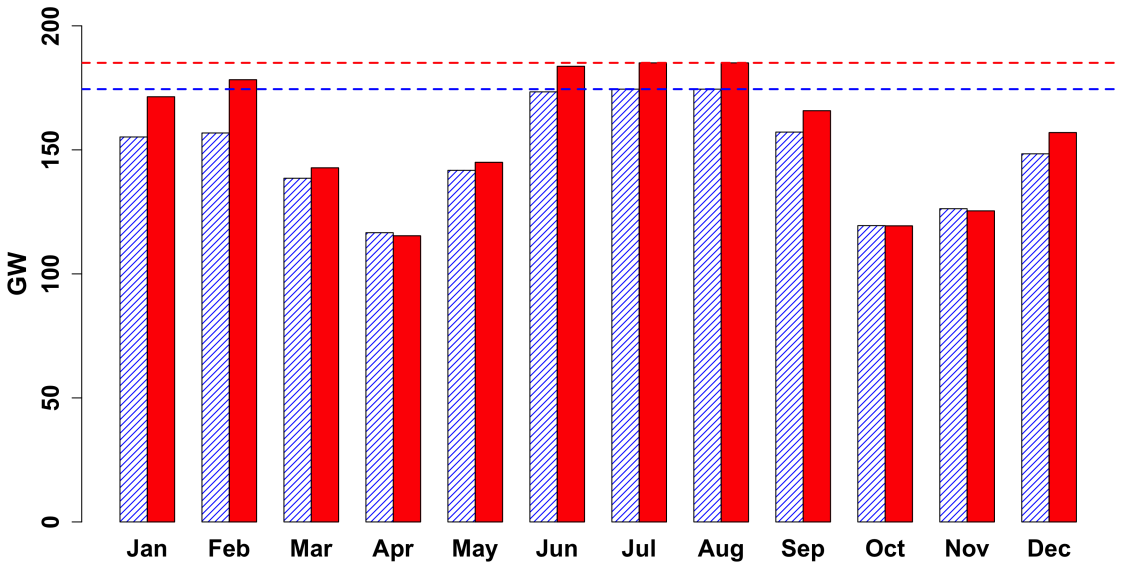

Figure 4.5: Monthly capacity requirements for achieving 0.1 LOLE under unconditional (blue hatching lines) and temperature-dependent (red) forced outage rates, with corresponding annual average procurement indicated by dashed horizontal lines. Annual capacity requirements in each scenario correspond to the height of the respective July bars. Monthly procurement levels for April, October and November are slightly lower in the temperature-dependent forced outage rates scenario than in the unconditional forced outage rates scenario.

We note that the 16% reduction obtained when procuring to achieve 0.1 LOLE on a monthly basis in the temperature-dependent scenario is a lower bound on the reduction that would be obtained by PJM if it opted to maintain its current reliability preference of 0.02 LOLE (implied by PJM's 26.6% realized reserve margin). This is because capacity requirements increase non-linearly in LOLE reductions, while LOLE is heavily concentrated in the summer. While we do not attempt to quantify the market implications of monthly procurement for capacity prices, we note that to the extent monthly procurement better enables inherently time-varying resources like wind, solar, and seasonally available demand-responsive loads to participate in the capacity market, monthly procurement should increase efficiency and reduce costs in the aggregate. However, we caveat that extreme temperatures are not the only cause of correlated failures. Further analysis on additional drivers of correlated failures should be conducted prior to proceeding with a seasonal or monthly procurement proposal.

4.5.3 Resource adequacy implications of future generation resource scenarios

The results presented thus far have considered capacity requirements for PJM's current conventional generator fleet. Given the economic pressure on nuclear and coal generators as a result of low natural gas prices [49, 59, 60], we next consider illustrative bounding scenarios on future fleet composition. In particular, we consider the resource adequacy implications of replacing the following with combined-cycle gas generators: 1). all existing nuclear generators; 2). all existing coal generators; and 3). all existing nuclear and coal generators. In each case, the replacement combined-cycle gas generator is the same size as the retired generator and has the same forced outage rates as PJM's existing combined-cycle gas generators.[23] Any of these fleet composition changes could increase or decrease capacity requirements, depending on how the change affects the system-level distribution of available capacity, particularly during peak-load hours.

In the unconditional forced outage rates scenario, replacing existing nuclear generators with combined-cycle gas generators results in a very slight increase in capacity requirements, while replacing existing coal generators reduces capacity requirements by approximately 5 GW. This is because combined-cycle generators' unconditional forced outage rates are slightly higher than those of nuclear generators but much lower than those of coal generators. In the temperature-dependent forced outage rates scenario, replacing existing nuclear generators reduces capacity requirements by approximately 2 GW, while replacing existing coal generators again reduces capacity requirements by approximately 5 GW. This is because combined-cycle generators' temperature-dependent forced outage rates are lower than those of nuclear and coal generators during the hot, high-load hours that contribute overwhelmingly to LOLE. Results are summarized in Table 4.2.

The inconsistent effect on resource adequacy of retiring nuclear generators under the two scenarios illustrates the importance of robust modeling of forced outage rates during high-load hours. We caution that if winter resource adequacy risks, particularly increased natural gas supply deliverability risks, are insufficiently captured by the logistic regression model of [22] then nuclear generators may

[23] This analysis makes a number of assumptions, notably that gas pipeline infrastructure would be expanded to achieve an equal level of deliverability for the replacement combined-cycle generators as for the current fleet. We also do not consider vintage effects, but note that the average age of existing coal and nuclear generators is higher than that of combined-cycle generators.

Table 4.2: Resource adequacy implications of fleet composition change scenarios under unconditional and temperature-dependent forced outage rates. Values reflect annual procurement.

Case	Unconditional	Temperature-dependent
Base case	174.5 GW	185.1 GW
Replace NU with CC	174.8 GW	183.2 GW
Replace ST with CC	169.2 GW	179.5 GW
Replace NU and ST with CC	169.5 GW	178.9 GW

provide reliability value not accounted for here. Increases in gas-fired generation capacity could exacerbate fuel supply constraints during cold weather events, particularly affecting the availability of gas under interruptible contracts [50]. In contrast, nuclear generators do not experience increased outage rates at cold temperatures [22].

4.5.4 Resource adequacy implications of future temperature scenarios

We next examine the resource adequacy risks of increased temperatures for PJM's existing fleet. Previous research examining the implications of climate change for power system reliability has focused on understanding changes in load patterns [51] as well as generator-level effects, such as increased deratings needed to comply with thermal pollution regulations [52, 53], with some work aggregating the implications for power systems planning [54]. Here we consider the resource adequacy implications of temperature increases as they translate to higher generator forced outages for PJM's conventional fleet. We examine two simplified scenarios: first, where the temperature in each hour increases by 1 degree Celsius, and second where the temperature in each hour increases by 2 degrees Celsius.[24] This approach is reasonable given that nearly all resource adequacy risk in PJM accrues in the summer, particularly during the hottest hours of July. These temperature scenarios increase capacity requirements by approximately 1 and 3 GW (0.5-1.5%) in the temperature-dependent forced outage rates scenario. By definition, increased temperatures have no effect on supply in the

[24] These temperature increases are with respect to the average temperatures occurring during 2006-2017, the period of our historical temperature data. Global temperatures during this time were already approximately 1 degree Celsius above the pre-industrial reference point [55]. Under business-as-usual emissions trajectories, an additional 1 and 2 degrees of temperature increase could be realized in approximately 2050 and 2100, respectively, though outcomes are highly sensitive to future emissions scenarios [56].

Table 4.3: Resource adequacy implications of temperature increases under unconditional and temperature-dependent forced outage rates. Values reflect annual procurement.

Case	Unconditional	Temperature-dependent
Base case	174.5 GW	185.1 GW
+1 degree Celsius	174.5 GW[1]	186.1 GW
+2 degrees Celsius	174.5 GW[1]	188.0 GW

[1] By definition there is no effect from temperature increases in the unconditional forced outage rates scenario, so these values are identical to the base case.

unconditional forced outage rates scenario.[25] Results are summarized in Table 4.3.

4.5.5 Resource adequacy implications of future temperature scenarios in conjunction with future generation resource scenarios

Finally, we consider capacity requirements under increased temperatures in conjunction with replacement of all nuclear and/or coal generators by combined-cycle gas generators. This simply combines the +2 degrees Celsius case with the three generator replacement scenarios considered previously. In the unconditional forced outage rates scenario, increased temperatures have no effect on capacity requirements, so each fleet replacement case matches the corresponding result from Table 4.2. In the temperature-dependent forced outage rates scenario, the capacity reductions from generator replacement offset the capacity increases required from increased temperatures to various extents; the net effect ranges from a reduction of 2 GW when replacing all coal and nuclear generators, to an increase of 1.5 GW when replacing only nuclear generators. Results are summarized in Table 4.4.

[25] We do not account for increases in the peak load as a result of temperature increases. This would increase capacity requirements in both scenarios, and creates an important interaction for consideration in power systems planning [54].

Table 4.4: Resource adequacy implications of temperature increases combined with generator replacement under unconditional and temperature-dependent forced outage rates. Values reflect annual procurement.

Case	Unconditional	Temperature-dependent
Base case (current temperature, current fleet)	174.5 GW	185.1 GW
+2 degrees Celsius with no fleet change	174.5 GW[1]	188.0 GW
+2 degrees Celsius with NU to CC	174.8 GW[1]	186.6 GW
+2 degrees Celsius with ST to CC	169.2 GW[1]	184.3 GW
+2 degrees Celsius with NU and ST to CC	169.5 GW[1]	183.0 GW

[1] By definition there is no effect from temperature increases in the unconditional forced outage rates scenario, so these values are identical to the corresponding fleet change element of Table 4.2.

4.6 Conclusions

We have examined the implications of temperature-dependent generator availability for resource adequacy in the PJM Interconnection. By combining PJM's GADS database with an open-source resource adequacy modeling tool and publicly available datasets, we demonstrate that the increased failure probabilities previously shown to affect PJM's conventional generator fleet at extreme temperatures [22] pose significant reliability risks not considered in standard resource adequacy modeling. Temperature dependence increases PJM's required reserve margin from 15.9% to 22.9% to achieve their reliability target of 0.1 LOLE in the 2018/2019 delivery year. However, PJM actually procured a 26.6% reserve margin for this delivery year, equivalent to 0.02 LOLE in our modeling framework, resulting in excess capacity payments of $315 million per year beyond what was required to achieve 0.1 LOLE.

While greater procurement of generation capacity reduces the probability of loss-of-load events, the incremental reliability benefits should be weighed against the incremental costs. We estimate that the VOLL implicit in PJM's selected level of capacity procurement during the 2018/2019 delivery year is approximately $700k/MWh, two orders of magnitude above common values used in operational contexts. This excess procurement helps explain why PJM has rarely triggered its operational scarcity pricing, even during recent extreme weather events [57], and creates doubt as to whether current initiatives on scarcity pricing reform [58] will have meaningful effects.

To the extent PJM understands that extreme temperatures increase generator failures and that

this effect cannot be accounted for in its current resource adequacy modeling approach, the decision to procure additional capacity is prudent. The key contribution of this work is to demonstrate a tractable method for incorporating correlated failures into resource adequacy modeling. In the case of PJM and other system operators with operationally similar conventional generator fleets and ambient conditions, we strongly recommend accounting for the correlation between temperature and generator failure probabilities. System operators in different contexts should adapt this approach by modeling key drivers of correlated failures in their balancing areas.

Given the seasonality of temperatures, we examine the potential benefits of monthly, rather than annual, capacity procurement for PJM. We find that monthly capacity targets could reduce total annual capacity procurement by approximately 16% without meaningfully increasing LOLE. While we do not estimate the economic benefits of a monthly capacity market, given how a monthly structure would likely alter generator bidding strategies, we note this potential reduction in capacity requirements is substantial. Further, a monthly or seasonal structure should allow for improved efficiency in the supply offers of truly seasonal resources, such as solar, wind, and some demand response.

Given the economic pressure being exerted on nuclear and coal generators, largely driven by inexpensive natural gas, we examine the resource adequacy implications of the retirement and replacement of all nuclear and/or all coal generators. When these resources are replaced with combined-cycle gas generators with forced outage rates equal to PJM's existing combined-cycle generators, the system experiences moderate resource adequacy benefits of approximately 1% and 3%, respectively. This result is driven by the lower forced outage rates of existing combined-cycle generators during the hot, high-load hours that contribute overwhelmingly to LOLE.

Given that all PJM generator types exhibit increased failure probabilities during hot temperatures, we examine the resource adequacy implications of future climate change scenarios corresponding to incremental temperature increases of 1 and 2 degrees Celsius. Holding other factors constant, these scenarios result in increased capacity requirements of approximately 0.5% and 1.5%, respectively, as a result of higher forced outage rates. This result suggests temperature-dependence of generator forced outages should be considered when planning for climate change. Combining the 2-degrees

Celsius temperature increase with the retirement cases results in a 1% increase and 0.5% decrease in capacity requirements when replacing nuclear and coal generators, respectively.

While we have endeavored to parameterize our model to be consistent with PJM, this analysis is not a substitute for PJM's own resource adequacy modeling process. For example, we do not account for transmission constraints nor emergency imports from neighboring systems. In addition, we make the following caveats about the temperature-dependent forced outage rate scenario. First, the regression specification employed may not capture the full extent of winter resource adequacy risk, particularly the risk of interruptions to non-firm gas transportation. That effect becomes even more important under the various retirement scenarios in which coal and/or nuclear generators are replaced by additional gas capacity. Similarly, we caution that extreme temperatures are not the only drivers of correlated failures. For example, states in PJM's territory have been affected by hurricanes since 2000, and the Atlantic hurricane season includes fall months which typically experience moderate temperatures. In addition, we note that our temperature-dependent forced outage rates are calculated on 23 years of PJM GADS data. PJM's Capacity Performance program—which did away with "outside management control" forced outage events and enacted penalties for resources that fail to perform when called—was implemented only after the Polar Vortex of 2014, and has led to significant investment in generator maintenance, dual-fuel capabilities, and firm gas supply contracts. Thus it is possible that the relationship between extreme temperatures and forced outages has weakened in recent years. Despite these caveats, this analysis demonstrates the importance of considering ambient conditions that significantly impact generator availability, particularly during high load hours.

Finally we note that planning models such as RECAP and PJM's own resource adequacy tools do not perform full sequential modeling of the system over time. Thus even when accounting for temperature-dependent generator availability, resource adequacy modeling may miss important reliability risks. Future work will consider temperature-dependent generator availability on operational timescales using a security-constrained unit commitment and economic dispatch model. Such an approach can better account for chronological risk of lost load from generator outages during sustained extreme weather events, incorporate demand response resources, quantify operational

flexibility needs, and allow system operators to better understand the value of procuring operating reserves, complementing this analysis.

References

[1] R. Billinton and R. Allan, *Reliability evaluation of power systems*. Plenum Press, second ed., 1996.

[2] R. Billinton and R. Allan, *Reliability evaluation of engineering systems*. Springer Science+Business Media, second ed., 1992.

[3] R. Billinton and R. Allan, *Reliability assessment of large electric power systems*. Kluwer Academic Publishers, first ed., 1988.

[4] R. Billinton and W. Li, *Reliability assessment of electric power systems using Monte Carlo methods*. Springer Science+Business Media, first ed., 1994.

[5] W. Li, *Risk assessment of power systems: Models, methods, and applications*. IEEE Press, first ed., 2005.

[6] S. Smith, "Spare capacity fixed by probabilities of outage," *Electrical World*, vol. 103, pp. 222–225, 1934.

[7] P. Benner, "The use of theory of probability to determine spare capacity," *General Electric Review*, vol. 37, no. 7, pp. 345–348, 1934.

[8] G. Calabrese, "Generating reserve capability determined by the probability method," *AIEE Transactions on Power Apparatus and Systems*, vol. 66, pp. 1439–1450, 1947.

[9] J. Hall, R. Ringlee, and A. Wood, "Frequency and duration methods for power system reliability calculations: Part I—generation system model," *IEEE Transactions on Power Apparatus and Systems*, vol. 87, no. 9, pp. 1787–1797, 1968.

[10] R. Ringlee and A. Wood, "Frequency and duration methods for power system reliability calculations: Part II—demand model and capacity reserve model," *IEEE Transactions on Power Apparatus and Systems*, vol. 88, no. 4, pp. 375–388, 1969.

[11] C. Galloway, L. Garver, R. Ringlee, and A. Wood, "Frequency and duration methods for power system reliability calculations: Part III—generation system planning," *IEEE Transactions on Power Apparatus and Systems*, vol. 88, no. 8, pp. 1224–1232, 1969.

[12] M. Bhavaraju, J. Hynde, and G. Nunan, "A method for estimating equivalent forced outage rates of multistate peaking units," *IEEE Transactions on Power Apparatus and Systems*, vol. 97, no. 8, pp. 1224–1232, 1978.

[13] ReliabilityFirst Corporation, "BAL-502-RFC-03: Planning resource adequacy analysis, assessment and documentation." `https://www.nerc.com/pa/Stand/Reliability%20Standards/BAL-502-RF-03.pdf`, 2017.

[14] S. Newell, D. Oates, J. Pfeifenberger, K. Spees, J. Hagerty, J. Pedtke, M. Witkin, and E. Shorin, "Fourth review of PJM's variable resource requirement curve." https://www.pjm.com/-/media/committees-groups/committees/mic/20180425-special/20180425-pjm-2018-variable-resource-requirement-curve-study.ashx, 2018.

[15] Monitoring Analytics, LLC, "State of the market report for PJM." https://www.monitoringanalytics.com/reports/PJM_State_of_the_Market/2017.shtml, 2017.

[16] P. Cramton, A. Ockenfels, and S. Stoft, "Capacity market fundamentals," *Economics of Energy & Environmental Policy*, vol. 2, no. 2, pp. 27–46, 2013.

[17] P. Cramton and S. Stoft, "Forward reliability markets: Less risk, less market power, more efficiency," *Utilities Policy*, vol. 16, no. 3, pp. 194–201, 2008.

[18] North American Electric Reliability Corporation, "2018 summer reliability assessment." https://www.nerc.com/pa/RAPA/ra/Reliability%20Assessments%20DL/NERC_SRA_05252018_Final.pdf, 2018.

[19] PJM Interconnection, "2018 PJM reserve requirement study." https://www.pjm.com/-/media/committees-groups/committees/pc/20181011/20181011-item-06b-2018-pjm-reserve-requirement-study-draft.ashx, 2018.

[20] PJM Interconnection, "PJM load forecasting model whitepaper." https://www.pjm.com/~/media/library/reports-notices/load-forecast/2016-load-forecast-whitepaper.ashx, 2016.

[21] S. Murphy, J. Apt, J. Moura, and F. Sowell, "Resource adequacy risks to the bulk power system," *Applied Energy*, vol. 212, pp. 1360–1376, 2018.

[22] S. Murphy, F. Sowell, and J. Apt, "A model of correlated power plant failures and recoveries." https://ceic.tepper.cmu.edu/publications/working-papers, 2018. Working paper CEIC-18-02.

[23] Energy+Environmental Economics Inc., "Renewable energy capacity planning model." https://www.ethree.com/tools/recap-renewable-energy-capacity-planning-model/, n.d.

[24] J. Pfeifenberger, K. Spees, K. Carden, and N. Wintermantel, "Resource adequacy requirements: Reliability and economic implications." https://www.ferc.gov/legal/staff-reports/2014/02-07-14-consultant-report.pdf, 2013.

[25] PJM Interconnection, "PJM manual 20: PJM resource adequacy analysis." https://www.pjm.com/~/media/documents/manuals/m20.ashx, 2019.

[26] PJM Interconnection, "PJM generating adequacy analysis: Technical methods." https://www.pjm.com/~/media/etools/oasis/references/whitepaper-sections-12.ashx, 2003.

[27] E. Pearson and H. Hartley, *Biometrika tables for statisticians*. Cambridge (United Kingdom): Cambridge University Press, 1966.

[28] PJM Interconnection, "2018/2019 RPM base residual auction results." https://www.pjm.com/-/media/markets-ops/rpm/rpm-auction-info/2018-2019-base-residual-auction-report.ashx?la=en, 2015.

[29] PJM Interconnection, "2020/2021 RPM base residual auction results." https://www.pjm.com/~/media/markets-ops/rpm/rpm-auction-info/2020-2021-base-residual-auction-report.ashx, 2017.

[30] Integ Enterprise Consulting, Inc., *PowerGADS 3.0 user manual*. Integ Enterprise Consulting, Inc., 2015.

[31] P. Peduzzi, J. Concato, E. Kemper, T. Holford, and A. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373–1379, 1996.

[32] U.S. Energy Information Administration, "Form 923, final 2017 data." = https://www.eia.gov/electricity/data/eia923/, 2018.

[33] B. Hoen, J. Diffendorfer, J. Rand, L. Kramer, C. Garrity, and H. Hunt, "United States Wind Turbine Database." https://eerscmap.usgs.gov/uswtdb, 2018.

[34] C. Draxl, A. Clifton, B. Hodge, and J. McCaa, "The Wind Integration National Dataset (WIND) toolkit," *Applied Energy*, vol. 151, pp. 355–366, 2015.

[35] W. Katzenstein, E. Fertig, and J. Apt, "The variability of interconnected wind plants," *Energy Policy*, vol. 38, no. 8, pp. 4400–4410, 2010.

[36] R. Wiser and M. Bolinger, "2017 wind technologies market report." https://emp.lbl.gov/wind-technologies-market-report, 2018.

[37] J. Rand. personal communication, Nov. 2018.

[38] National Renewable Energy Laboratory, "NSRDB Data Viewer." https://maps.nrel.gov/nsrdb-viewer/.

[39] A. Habte, M. Sengupta, and A. Lopez, "Evaluation of the National Solar Radiation Database (NSRDB): 1998-2015." https://www.nrel.gov/docs/fy17osti/67722.pdf, 2017.

[40] Pacific Northwest National Laboratory, "GridLAB-D." http://www.gridlabd.org/.

[41] National Renewable Energy Laboratory, "System Advisory Model." https://sam.nrel.gov/.

[42] NOAA National Centers for Environmental Information, "Global surface hourly (DS3505)," 2001. www7.ncdc.noaa.gov/CDO/cdopoemain.cmd?datasetabbv=DS3505&countryabbv=&georegionabbv=&resolution=40.

[43] PJM Interconnection, "PJM load forecast report." https://www.pjm.com/-/media/library/reports-notices/load-forecast/2018-load-forecast-report.ashx, 2018.

[44] PJM Interconnection, "2019/2020 RPM base residual auction planning period parameters." https://www.pjm.com/-/media/markets-ops/rpm/rpm-auction-info/2019-2020-rpm-bra-planning-parameters-report.ashx?la=en, 2016.

[45] PJM Interconnection, "Final zonal UCAP obligations, capacity prices and CTR credit rates." https://www.pjm.com/-/media/markets-ops/rpm/rpm-auction-info/2018-2019-final-zonal-ucap-obligations-capacity-prices-ctr-credit-rates.ashx?la=en, 2018.

[46] C. Byers, T. Levin, and A. Botterud, "Capacity market design and renewable energy: Performance incentives, qualifying capacity, and demand curves," *The Electricity Journal*, vol. 31, no. 1, pp. 65–74, 2018.

[47] S. Newell, R. Carroll, A. Kaluzhny, K. Spees, K. Carden, N. Wintermantel, and A. Krasny, "Estimation of the market equilibrium and economically optimal reserve margins for the ERCOT region." http://www.ercot.com/content/wcm/lists/143980/10.12.2018_ERCOT_MERM_Report_Final_Draft.pdf, 2018.

[48] D. Patton, "Comments of David B. Patton, Ph.D., regarding state policies affecting eastern RTOs," tech. rep., United States Federal Energy Regulatory Commission, 2017. Docket No. AD17-11-000: State policies and wholesale markets operated by ISO New England Inc., New York Independent System Operator, Inc., and PJM Interconnection, LLC.

[49] United States Department of Energy, "Staff report to the secretary on electricity markets and reliability." https://energy.gov/sites/prod/files/2017/08/f36/Staff%20Report%20on%20Electricity%20Markets%20and%20Reliability_0.pdf, 2017.

[50] PJM Interconnection, "Fuel security analysis: A PJM resilience initiative." https://www.pjm.com/-/media/library/reports-notices/fuel-security/2018-fuel-security-analysis.ashx, 2018.

[51] F. Fonseca, P. Jaramillo, M. Bergés, and E. Severnini, "Seasonal effects of climate change on intra-day electricity demand patterns," *Climatic Change*, 2019.

[52] S. Chandramowli and F. Felder, "Impact of climate change on electricity systems and markets—A review of models and forecasts," *Sustainable Energy Technologies and Assessments*, vol. 5, pp. 62–74, 2014.

[53] M. Cook, C. King, F. Davidson, and M. Webber, "Assessing the impacts of droughts and heat waves at thermoelectric power plants in the United States using integrated regression, thermodynamic, and climate models," *Energy Reports*, vol. 1, pp. 193–203, 2015.

[54] M. T. Craig, S. Cohen, J. Macknick, C. Draxl, O. J. Guerra, M. Sengupta, S. E. Haupt, B.-M. Hodge, and C. Brancucci, "A review of the potential impacts of climate change on bulk power system planning and operations in the United States," *Renewable and Sustainable Energy Reviews*, vol. 98, pp. 255–267, 2018.

[55] Intergovernmental Panel on Climate Change, "Climate change 2014 synthesis report summary for policymakers." https://www.ipcc.ch/site/assets/uploads/2018/02/AR5_SYR_FINAL_SPM.pdf, 2018.

[56] Intergovernmental Panel on Climate Change, "Global warming of 1.5°C: An IPCC special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty." https://www.ipcc.ch/sr15/, 2018.

[57] M. Watson, "PJM's winter 2019 experience reinforces need for reserve pricing reform." https://www.spglobal.com/platts/en/market-insights/latest-news/electric-power/031819-pjms-winter-2019-experience-reinforces-need-for-reserve-pricing-reform. Accessed: 2019-04-10.

[58] PJM Energy Price Formation Senior Task Force, "Price formation." https://www.pjm.com/-/media/committees-groups/task-forces/epfstf/20181214/20181214-item-04-price-formation-paper.ashx. Accessed: 2019-04-10.

[59] M. Roth and P. Jaramillo, "Going nuclear for climate mitigation: An analysis of the cost effectiveness of preserving existing US nuclear power plants as a carbon avoidance strategy," *Energy*, vol. 131, pp. 67–77, 2017.

[60] J. Jenkins, "What's killing nuclear power in US electricity markets? Drivers of wholesale price declines at nuclear generators in the PJM Interconnection." http://ceepr.mit.edu/files/papers/2018-001.pdf, 2018.