Resource adequacy risks to the bulk power system in North America

Sinnott Murphy[†], Jay Apt[†,‡,*], John Moura[§], Fallaw Sowell[‡]

[†]Department of Engineering & Public Policy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, U.S.A.
[‡]Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, U.S.A.
[§]North American Electric Reliability Corporation, 3353 Peachtree Road Suite 600, Atlanta, Georgia 30326, U.S.A.
[*]Corresponding author's email address: apt@cmu.edu

PLEASE DO NOT CITE OR QUOTE WITHOUT THE PERMISSION OF THE AUTHORS

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 (hourly time series) 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. The distributions of unscheduled unavailable capacity in each region are reasonably well modeled by Weibull and lognormal distributions. 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.

Keywords

Generating Availability Data System, Correlated failures, Resource adequacy, Reserve margins.

Highlights

- Correlated failures of NERC electric power generators occurred in 2012-2015
- Correlated failures happen in most NERC regions even when major storms are removed
- Correlated outages should be considered in defining resource adequacy requirements

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 [2,3]. 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 [4–7].

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 [8]. 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 [9].[3]

The availability statistic approach to RAM distills multiple years of availability history for each generating unit to a single number. Because all temporal information is discarded, it implicitly assumes failures are independent among generating units [10]. 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 [11]. For lack of a tractable alternative, the assumption of independent failures is also made by scholars working outside RAM: [12] 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

[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 [21,32]. 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 [33]; various reliability regions have other interpretations [14].

[3] Current resource adequacy planning procedures for several control areas in the United States may be found in the following sources: [34–39].

demonstrate our method using the Generating Availability Data System (GADS), a proprietary database maintained by the North American Electric Reliability Corporation (NERC) [13]. 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 [10,14].

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 introduces the GADS data. Section 3 describes the steps we take to clean the data and generate time series of unavailable capacity. Section 4 presents our results. Section 5 concludes.

The novel results discussed in Section 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 S-1 in the supplementary materials[4]). 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.

[4] Tables and Figures presented in the online supplementary materials are denoted with the S- prefix.

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 [15]. 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 [16,17].

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 [18]. 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.[5] 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 [19]. 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 failures (the GADS term for these is SF), the three unscheduled outages (U1, U2, and

[5] Reporting failures that represent less than 2% of a unit's capacity and last less than 30 minutes is voluntary. Hydro 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 [19].

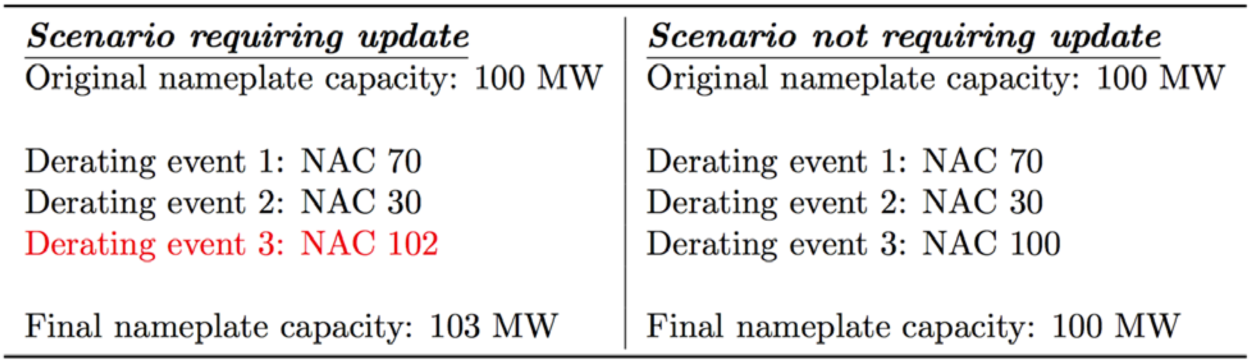
U3), and the three unscheduled deratings (D1, D2, and D3).[6] These are the primary event types considered in RAM. We next describe our methods for processing the raw GADS events data into time series of unavailable capacity.

3. Methods

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 3.2), we check that each unit's nameplate capacity is greater than its largest reported NAC. An example is shown in Figure 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.



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 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 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 [20]. For units whose time zone was updated by this process, we adjust the start and end times of its events accordingly.[7]

[6] 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.

[7] 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.

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.

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:

$$Magnitude\ of\ event = Nameplate\ capacity\ of\ unit - NAC\ of\ event \qquad (1)$$

However, deratings can overlap and usually the magnitude of the succeeding derating must be calculated as a function of the derating(s) already underway [19]. 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:

$$Magnitude\ of\ event = NAC\ of\ previous\ event - NAC\ of\ event \qquad (2)$$

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.

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.[8] 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.[9] An example time series is shown in Figure 2.

[8] 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 S-6 and Figure S-7.

[9] 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.

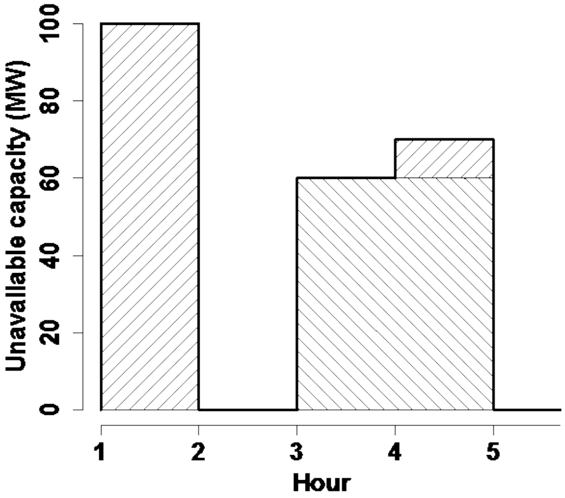


Figure 2: Illustrative unavailable capacity time series for one generating unit. In hour 1 the unit experiences an outage and is fully unavailable. In hour 2 the outage has been repaired and the unit is fully available. In hour 3 the unit experiences a 60-MW derating. In hour 4 the unit experiences a 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.

4. Results

We begin by presenting a brief descriptive summary of the GADS data by NERC region (Section 4.1). We then test for correlated failures (Section 4.2) and for seasonal patterns in unscheduled unavailable capacity (Section 4.3). Finally, we present a set of analyses with direct application for reliability analysis (Section 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.

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 1. We show time series of unscheduled unavailable capacity for each NERC region as a percentage of installed capacity in Figure 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 S-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 S-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 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	60,100	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: ABB Velocity Suite.
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: ABB Velocity Suite.
3. The number of units experiencing an unscheduled event during the study period.

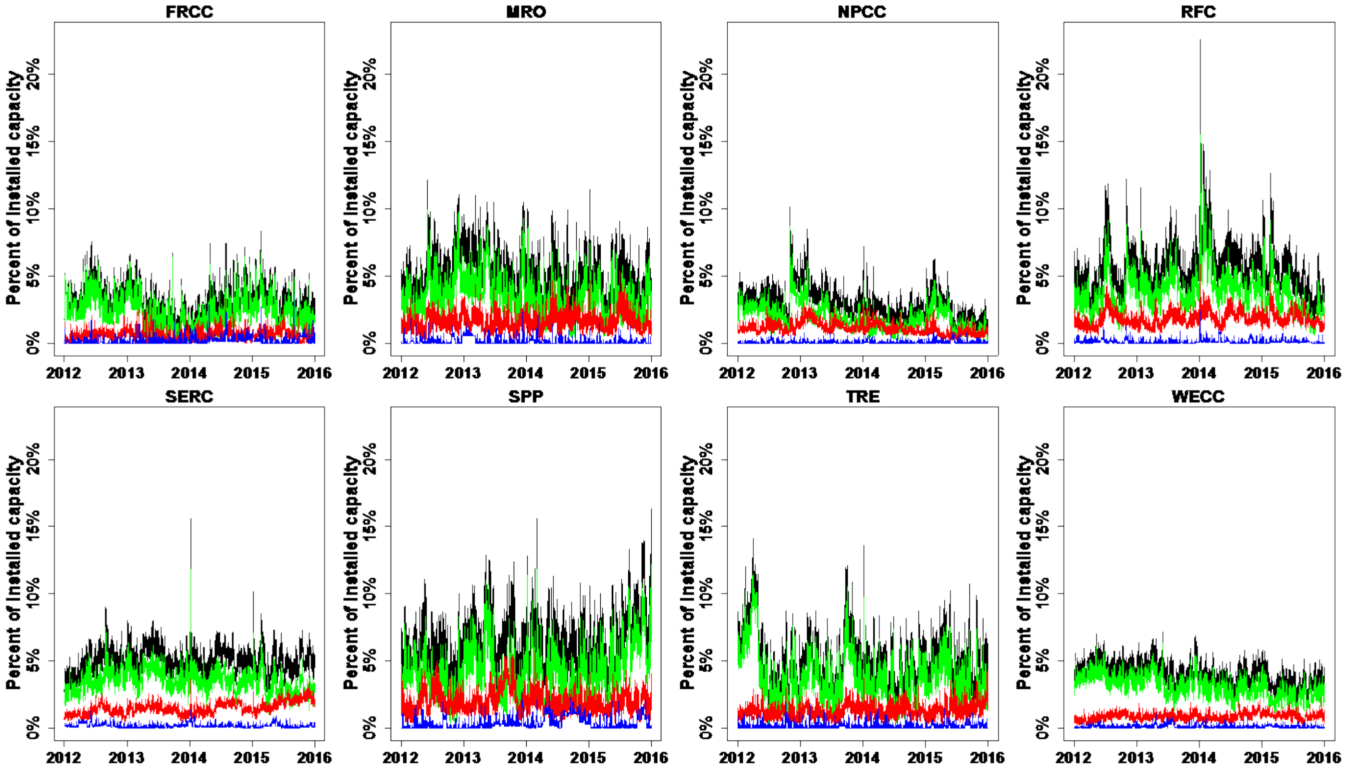


Figure 3: Unscheduled unavailable capacity as a percent of installed capacity, by unscheduled event type. Green: unscheduled outages only; red: unscheduled deratings only; blue: start-up failures only; black: all unscheduled events (the sum of green, red, and blue curves).

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 S-3. We present breakdowns of the count of unscheduled events by event type in Figure S-4.

4.2 Testing for independence among generators

We wish to test the assumption that generator failures are independent. From Figure 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 (Section 4.2.1). 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 (Section 4.2.2). 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.[10]

4.2.1 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.[11] 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.[12] 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 S-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 S-4.

[10] 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.

[11] The block length is a function of the autocovariance sequence, the spectral density function, and the length of the time series [40]. We compute each unit's block length using the "np" library in R [41,42]. Subsampling is carried out using the "boot" library in R [43,44].

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

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 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 [21] 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 S-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 S-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 S-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.

4.2.2 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:

$$P(arrival_i) = \frac{C\left(events_{D1:D3,SF,U1:U3_i}\right)}{C(hours_{1:T}) - C(hours_{SF,U1:U3_i})} \qquad (3)$$

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.[13] When calculating this probability, we subtract the number of hours in

[13] 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.

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 (~6 weeks) during the study period; this removes nine units from the analysis. Histograms of the estimated probabilities are reported in Figure S-14. Parameters from lognormal fits to the estimated probabilities are reported in Table S-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 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.

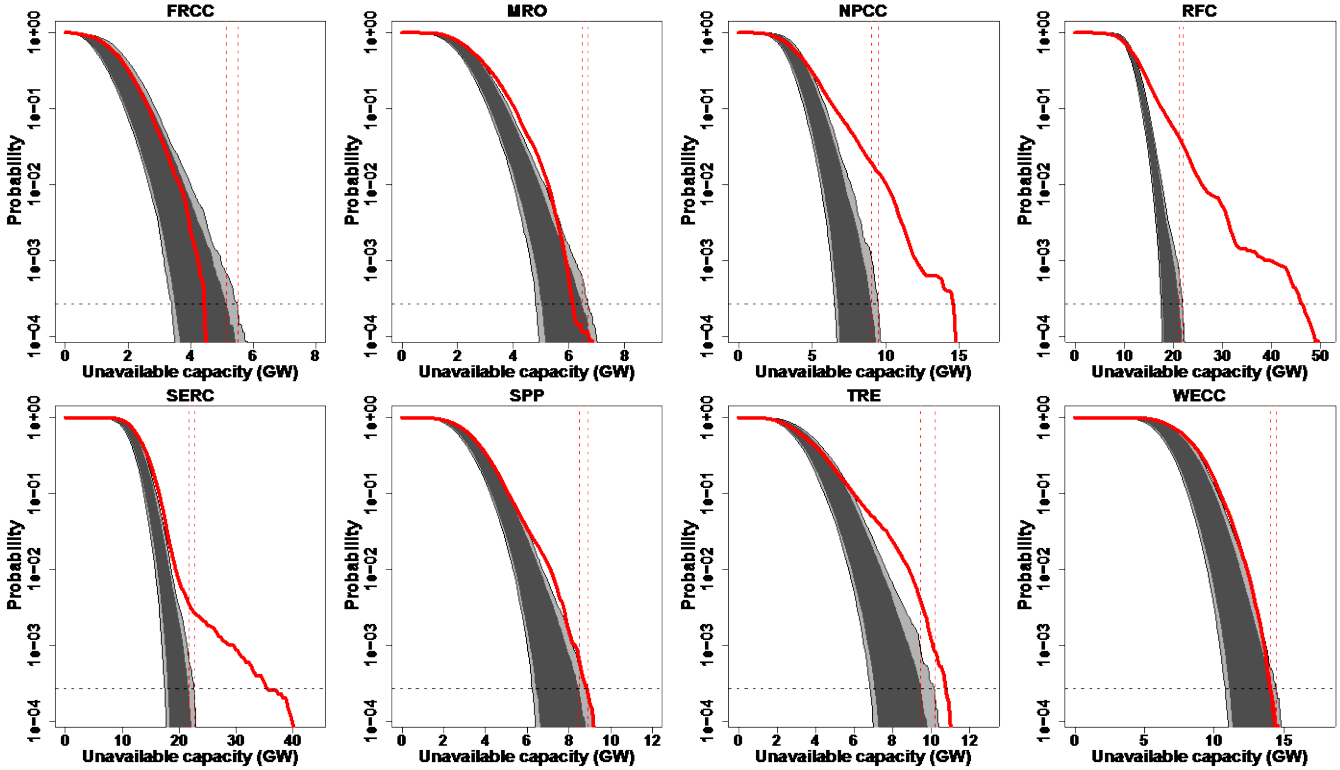


Figure 4: 95% and 99% confidence bands from 1000 binomial simulation runs shown in dark and light gray, respectively; empirical distributions from full study period shown in red. Dashed

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 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.

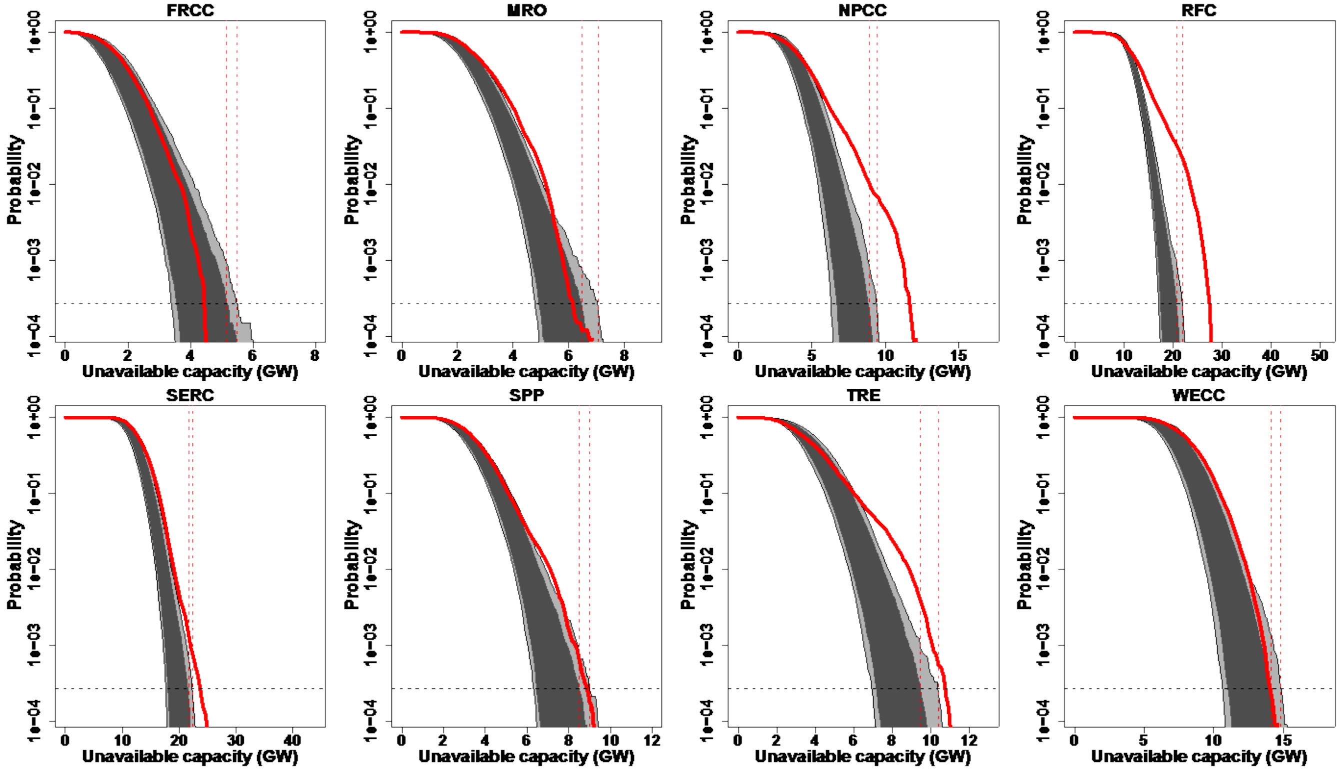


Figure 5: 95% and 99% confidence bands from 1000 binomial simulation runs shown in dark and light gray, respectively; empirical distributions from removing Hurricane Sandy and January 2014 shown in red. Dashed horizontal line indicates 2.4 LOLH threshold; dashed vertical lines indicate intersection of 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.

Table 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 50th 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 the 2.4 LOLH, with greater incidence than allowed under 2.4 LOLH resource adequacy requirement). The definitions are nested such that a region cannot satisfy the second definition without also satisfying the first.**

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

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 event for our study period, larger than any event we observed in the region during the four years we studied [22].

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.

4.3.1 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 [23]. 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.[14] From this, we compute 95% and 99% confidence bands and plot them as exceedance curves along with the corresponding empirical distributions (Figure S-15 through Figure S-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.

4.3.2 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 S-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).[15] 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 S-22). From these results we conclude that generally we cannot support the hypothesis of seasonality in average unavailable capacity from unscheduled events.

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 S-23).[16] 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.

4.3.3 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 S-25). The residuals resemble white noise and appear to be homoskedastic. Autocorrelation functions of the residuals show no significant remaining structure (Figure S-26). Values and t-statistics for the

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

[15] A weakly stationary AR(1) model can be written $x_t = \mu + \rho x_{t-1} + \varepsilon_t$ 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.

[16] 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).

AR(1) parameters are reported in Table S-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.

4.3.4 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 from Section 4.3.1 and plot the results as exceedance curves for both the full study period (Figure S-28) and when excluding Hurricane Sandy and the Polar Vortex (Figure S-29). Consistent with the results from our previous tests of seasonality, we find minimal benefits from calculating availability statistics separately for each season.

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.

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 $\mu = 1/MTTR$ [24].[17] 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 [8]. 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

[17] 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 [24].

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 [25,26].

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 [27], MTTR and MTTF values for a single coal-fired generating unit modeled with 10 availability states [28], MTBF values for seven gas turbine units in India [29], MTBF and MTTR values for 11 gas turbine units in Nigeria [30], and MTTF values for a single combined-cycle unit modeled with 8 availability states in Israel [31].

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.

4.4.1 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 S-30) and with January 2014 and Hurricane Sandy removed (Figure S-31). We report the parameters of each fit in Table S-9.

4.4.2 Parametric fits to distributions of normalized derating magnitudes

We fit Weibull distributions to each unit type's distribution of normalized derating magnitudes (Figure S-32). We report the parameters of each fit in Table S-10.

4.4.3 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 S-6 through S-8. We report the fit parameters in Table S-2.

4.4.4 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 3. We present histograms of the results in Figure S-14. We report the parameters of each fit in Table S-5.

4.4.5 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.[18] 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

[18] 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.

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 6 through Figure 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.[19] 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 S-14 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 3 to facilitate comparison. Larger MTBF values indicate greater reliability.

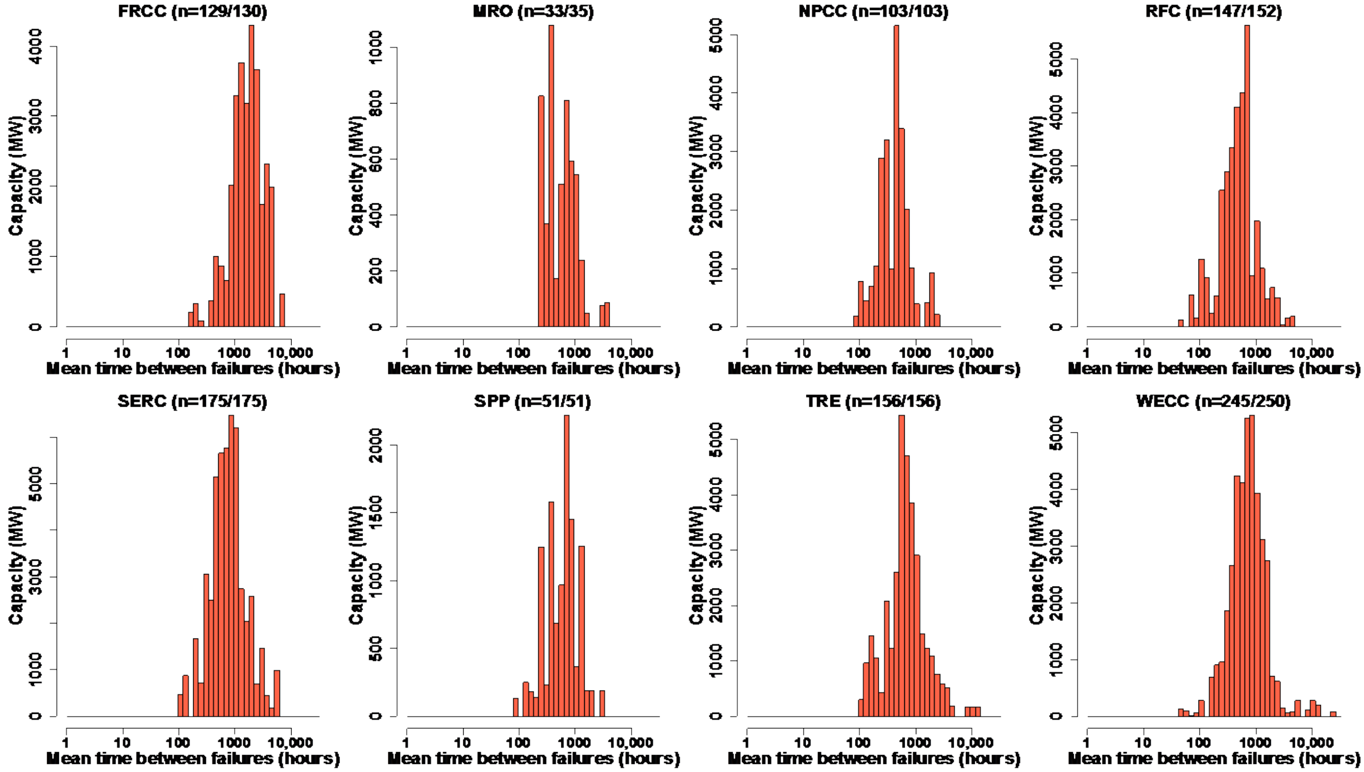


Figure 6: Capacity weighted mean time between failure (MTBF) values for combined cycle gas units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. 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 S-14).

[19] 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.

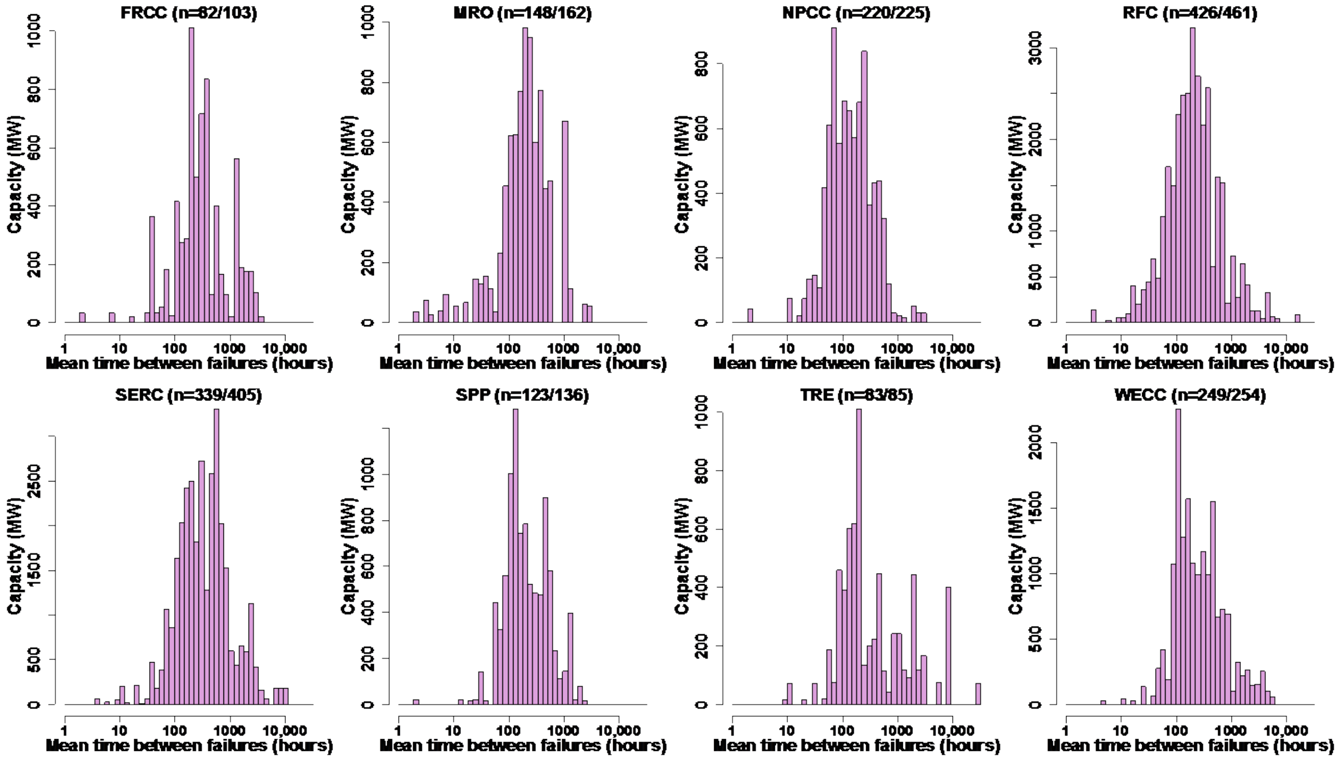


Figure 7: Capacity weighted mean time between failure (MTBF) values for simple cycle gas units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. 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 S-14).

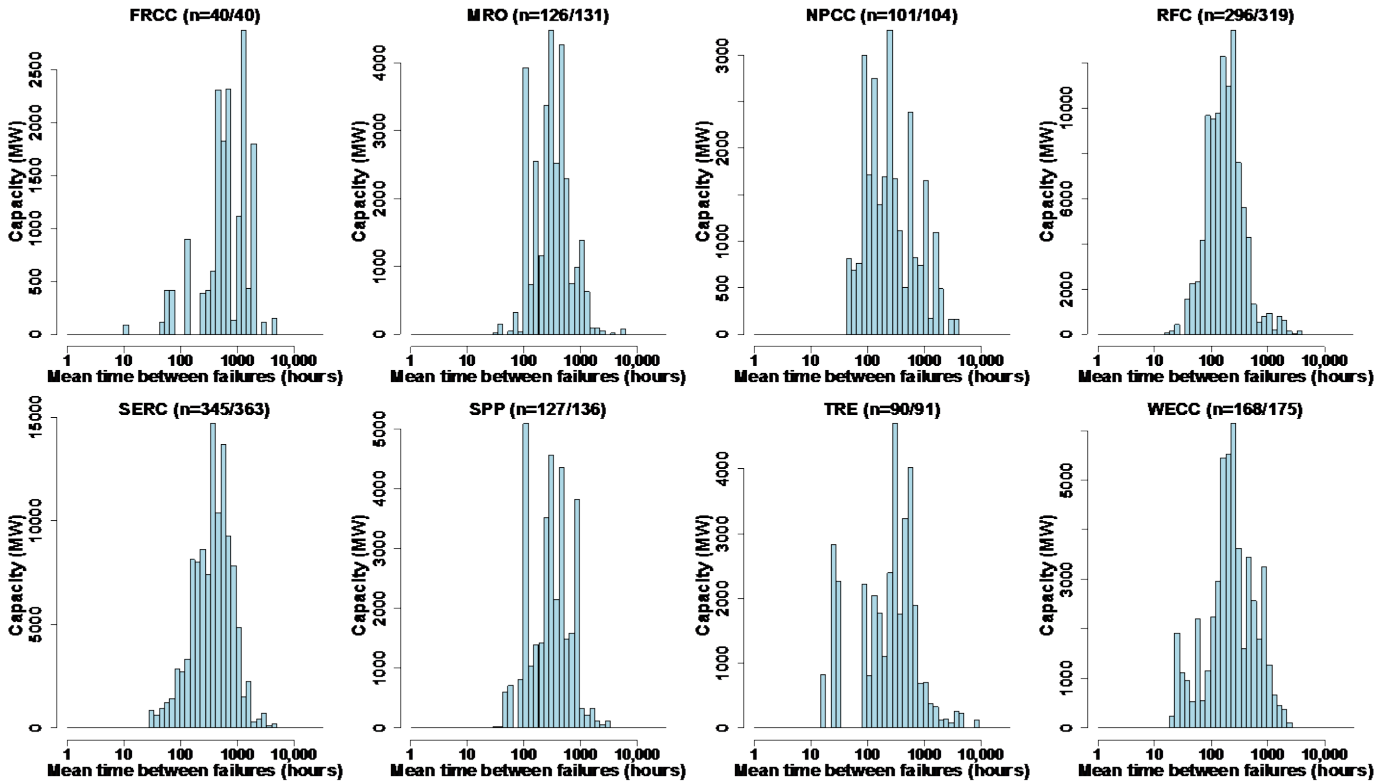


Figure 8: Capacity weighted mean time between failure (MTBF) values for fossil steam and fluidized bed units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could 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 S-14).

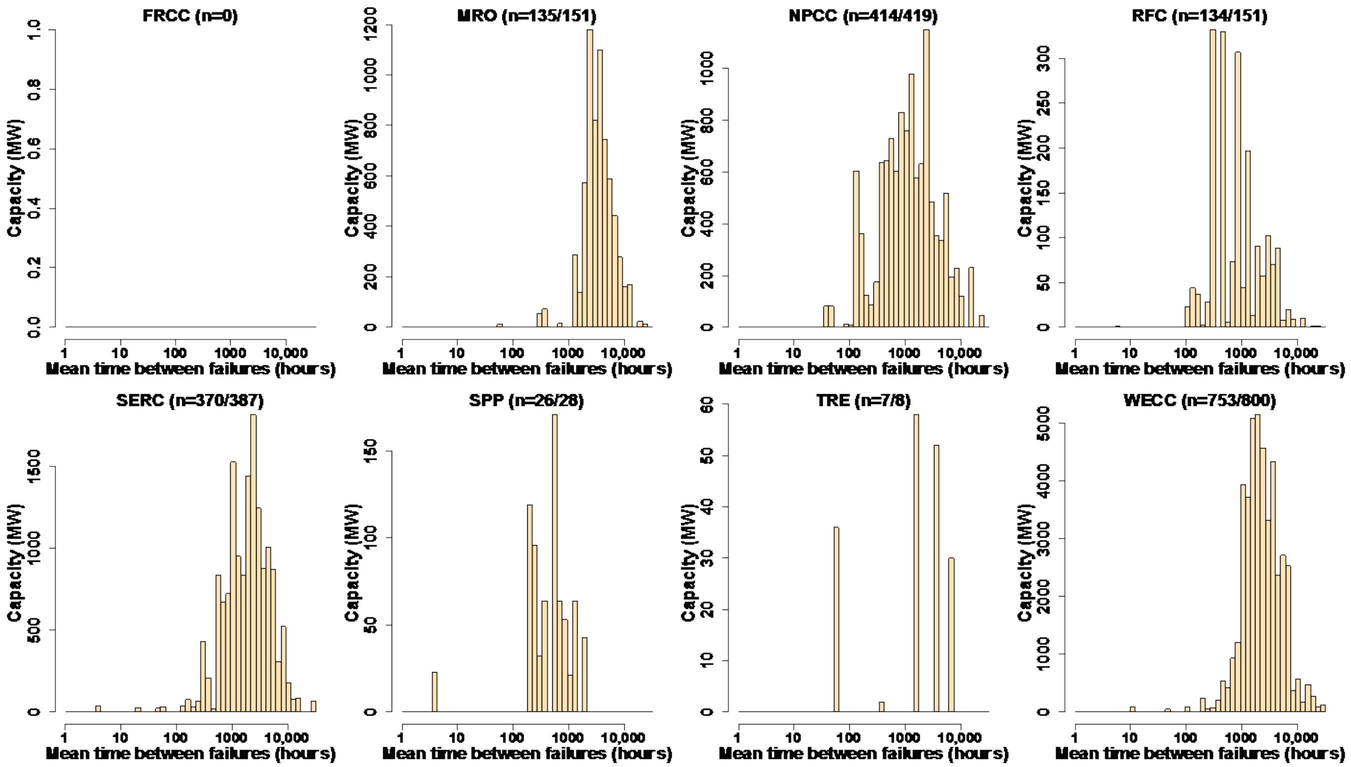


Figure 9: Capacity weighted mean time between failure (MTBF) values for hydroelectric units. There are no such units in FRCC. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for 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 S-14).

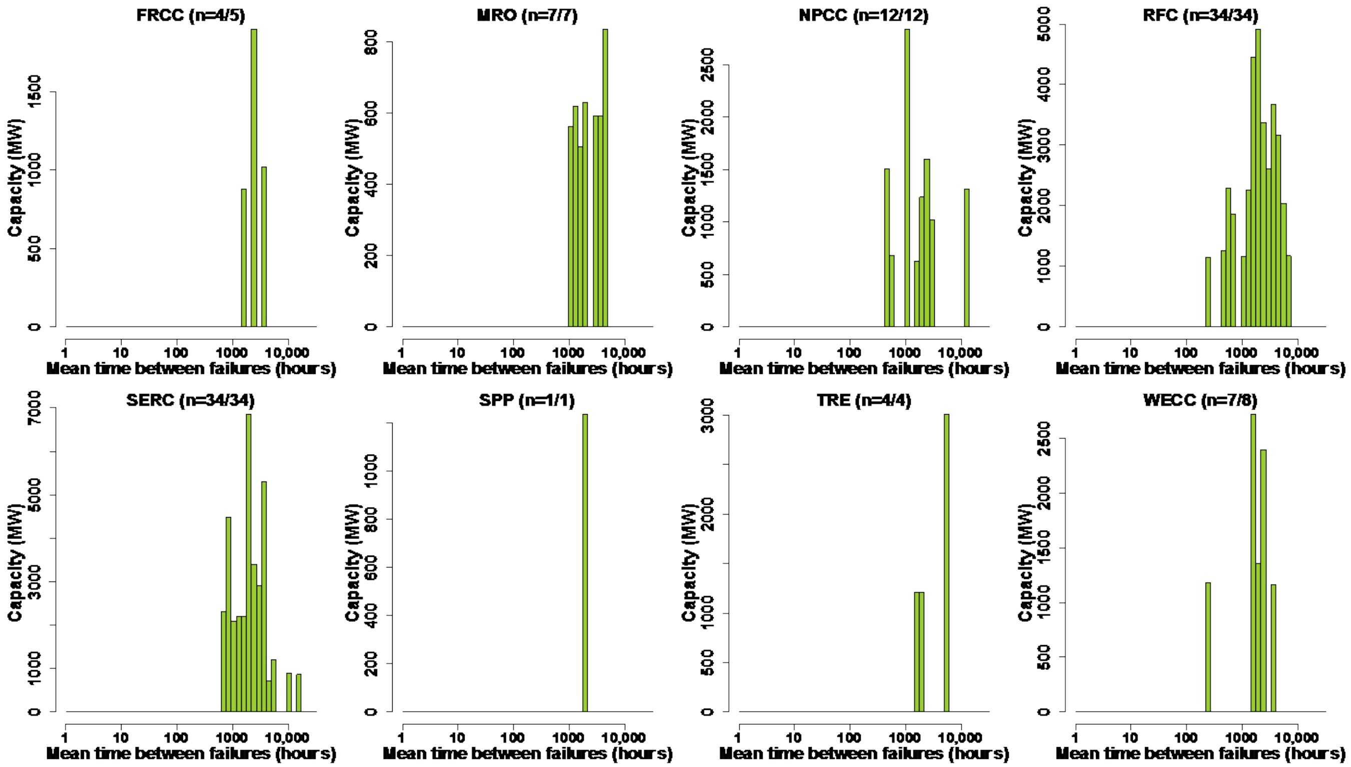


Figure 10: Capacity weighted mean time between failure (MTBF) values for nuclear units. Note the log scale for MTBF. Values are calculated with all reserve shutdown hours removed so as to restrict attention to service hours. Numerator indicates count of units for which an MTBF could be calculated. Denominator 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 S-14).

Table 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 S-14). Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

	10th	*20th*	*30th*	*40th*	*50th*	*60th*	*70th*	*80th*	*90th*
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

We conclude that nuclear, hydro, and combined cycle units tend to run longer before failing than simple cycle and fossil steam units.[20] We fit Weibull and gamma distributions to the MTBF values by unit type and report the parameters in Table S-11 and Table S-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 S-38). 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 S-33 through Figure S-37. 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 S-39 through Figure S-43). 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 4. Smaller MTTR values indicate shorter average repair durations.

Table 4: Selected percentiles of capacity-weighted mean time to recovery values by unit type (hours). Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

	10th	*20th*	*30th*	*40th*	*50th*	*60th*	*70th*	*80th*	*90th*
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

We conclude that combined cycle units typically have among the lowest MTTR values while nuclear units have among the highest.[21] We fit Weibull and gamma distributions to the MTTR values by unit type and report the parameters in Table S-15 and Table S-16. We present histograms of the number of failure periods used to calculate each unit's MTTR in Figure S-44 through Figure S-48. We note that some units' MTTRs are calculated based on only a single

[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 MTBF.

[21] 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.

failure period. With a longer time series, the proportion of units with MTTRs based on very few failures periods would decrease, increasing confidence in the robustness of these results.

4.4.6 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 5. We also conduct the test when aggregating over regions. Table 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 S-49 through Figure S-53.

Table 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 S-14). Abbreviations: CC combined cycle units, CT simple cycle gas units, FSFB fossil steam and fluidized bed units, HY hydroelectric, NU nuclear.

	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 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.

	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

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, 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 S-13.

4.4.7 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.[22] We present time series of these events, overlaid with our time series of unscheduled events for context, by unit type (Figure 11 through Figure 15). We also publish these data so 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.

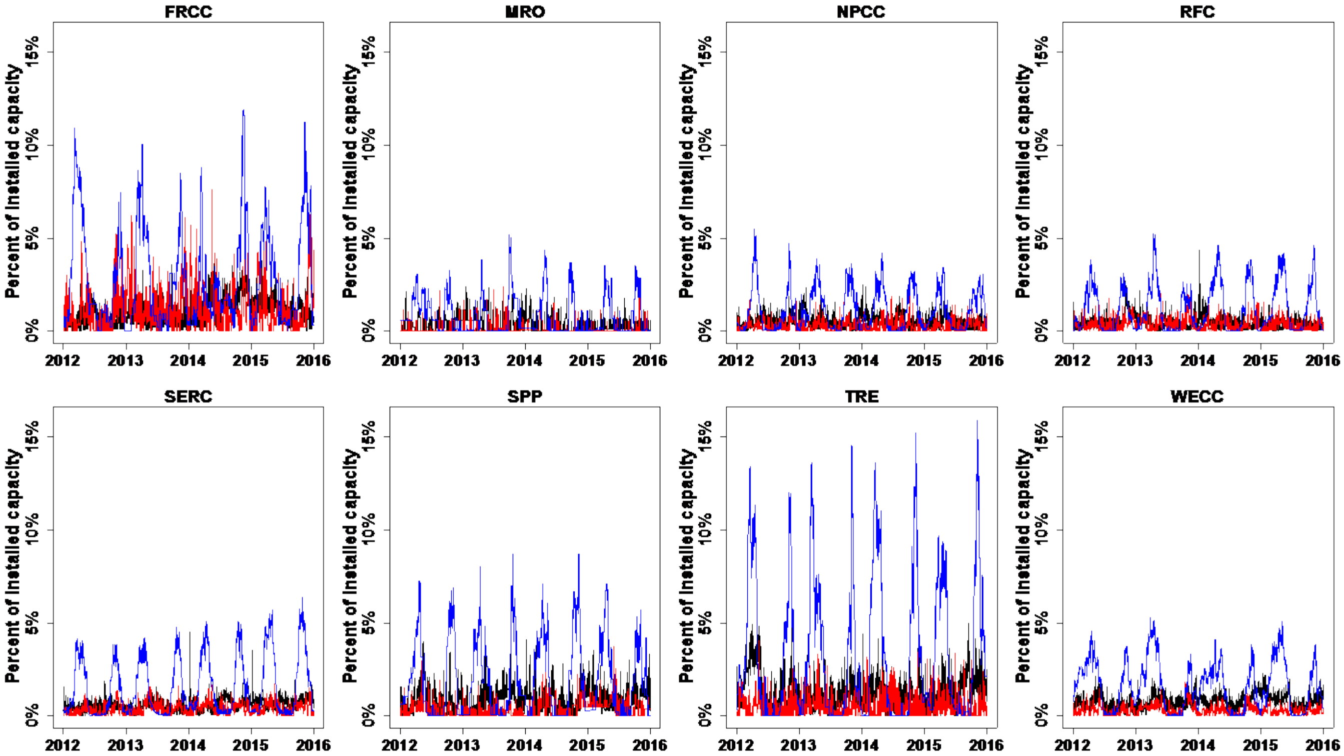


Figure 11: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for combined cycle units.

[22] 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 (NERC, 2017). 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 (PJM, 2017b).

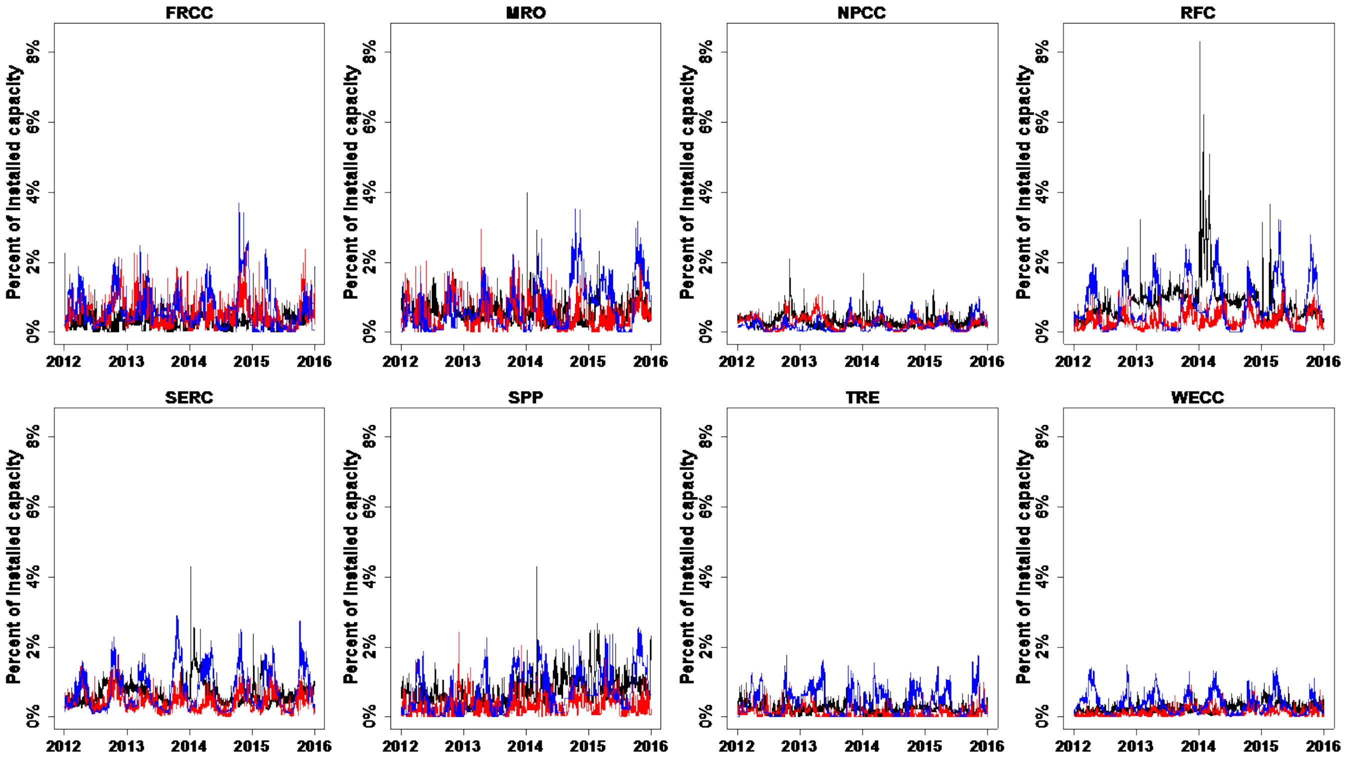


Figure 12: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for simple cycle gas units.

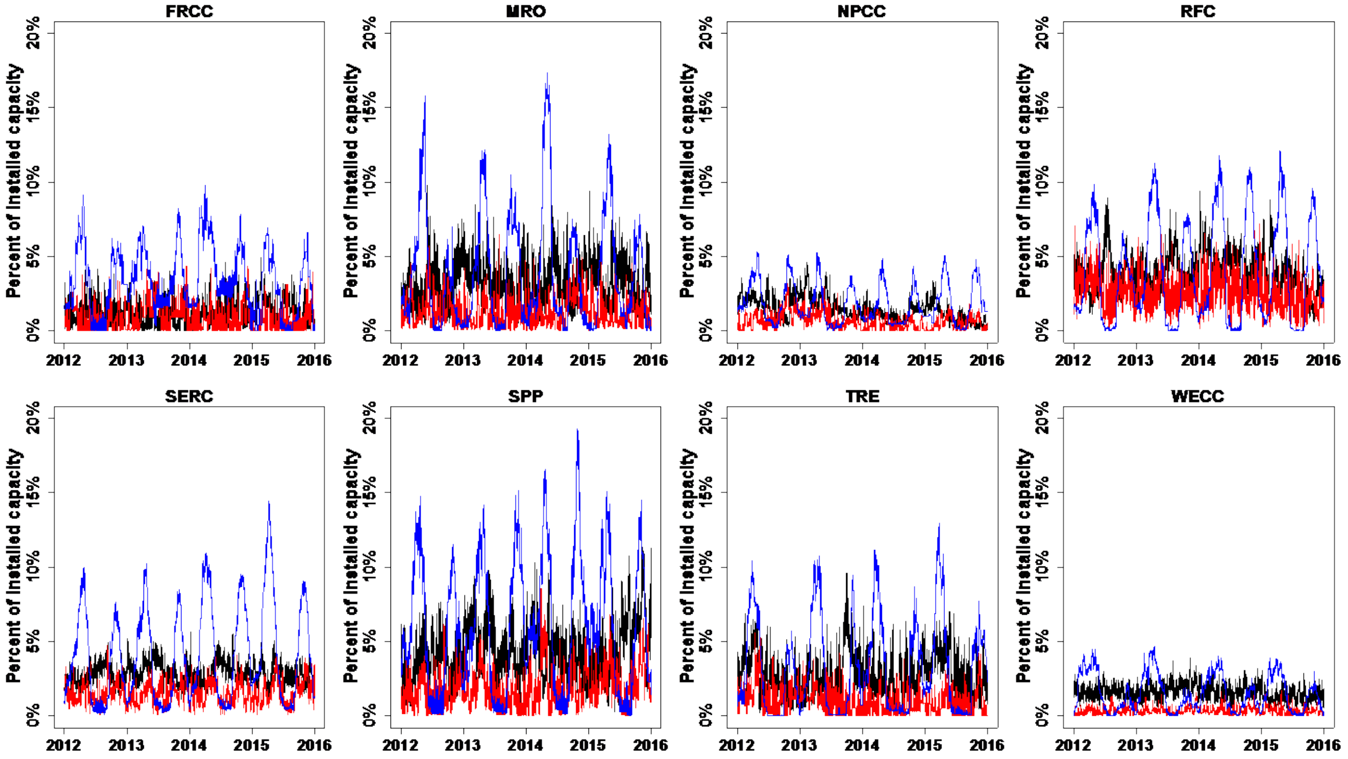


Figure 13: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for fossil steam and fluidized bed units.

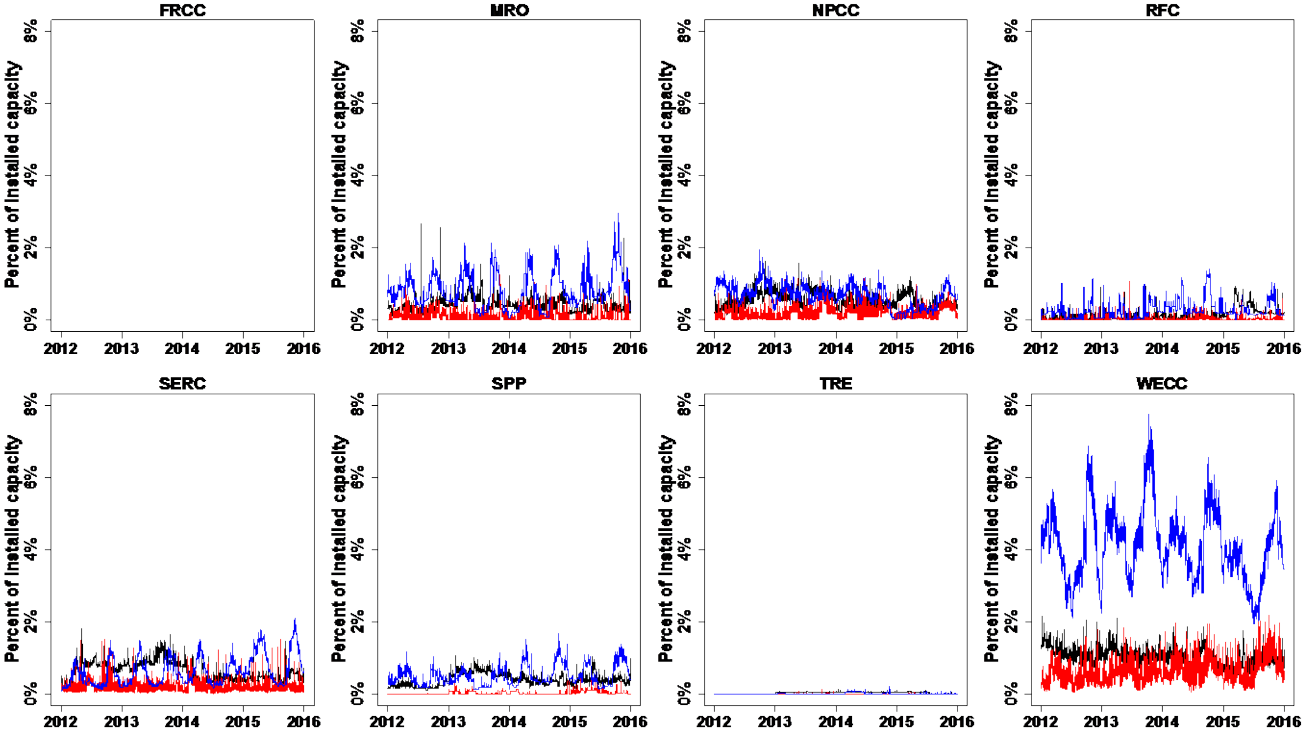


Figure 14: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for hydroelectric units.

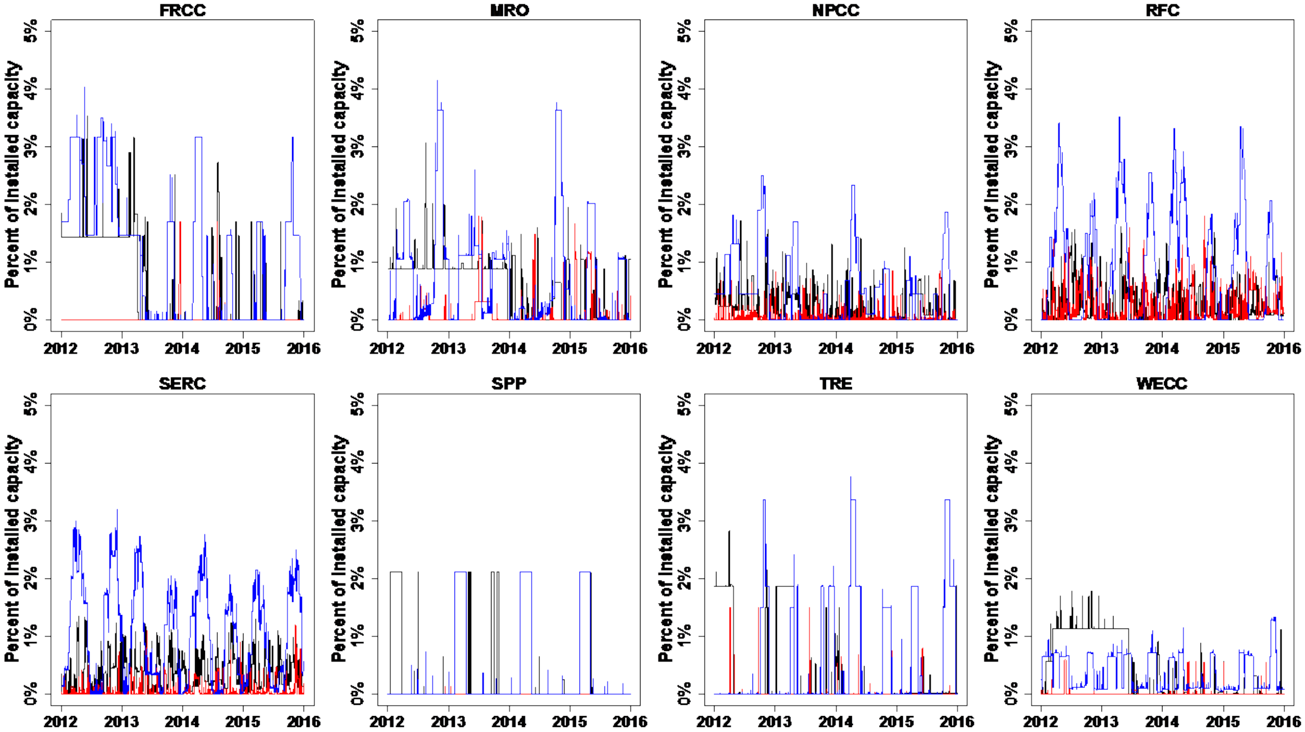


Figure 15: Unscheduled (black), maintenance (red), and scheduled (blue) unavailable capacity for nuclear units.

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[23], 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.

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

[23] 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.

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.

Acknowledgments

We thank Noha Abdel-Karim, David Calderon, Sean Donovan, Elliott Nethercutt, Donna Pratt, Lee Thaubald, and Matthew Varghese of NERC for help accessing and working with the GADS database. We also thank Ron Fluegge and Cosma Shalizi for helpful discussions.

S.M. was supported by the Electric Power Research Institute, the U.S. Department of Energy, the North American Electric Reliability Corporation (NERC), and the Richard King Mellon Foundation. J.A. received partial support from the Carnegie Mellon Climate and Energy Decision Making Center (CEDM), formed through a cooperative agreement between the National Science Foundation and CMU (SES-0949710).

With the exception of J.M. of NERC, the funders/sponsors had no role in the study design, collection, analysis, and interpretation of the data, in the writing of the report, or in the decision to submit the article for publication.

The authors declare no conflicts of interest.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://

S.M.’s ORCID: 0000-0002-4572-8295
J.A.’s ORCID: 0000-0002-6195-0355
F.S.’s ORCID: 0000-0001-5042-7617

References

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

[2] Smith SA. Service reliability measured by probabilities of outage. Electr World 1934;103:371–4.

[3] Benner PE. The use of the theory of probability to determine spare capacity. Gen Electr Rev 1934;37:345–8.

[4] Calabrese G. Generating reserve capacity determined by the probability method. Trans Am Inst Electr Eng 1947;66:1439–50.

[5] Lyman WJ. Calculating probability of generating capacity outages. Trans Am Inst Electr Eng 1947;66:1471–7.

[6] Loane ES, Watchorn CW. Probability methods applied to generating capacity problems of a combined hydro and steam system. Trans Am Inst Electr Eng 1947;66:1645–57.

[7] Seelye HP. Outage expectancy as a basis for generator reserve. Trans Am Inst Electr Eng 1947;66:1483–8.

[8] Bhavaraju MP, Hynds JA, Nunan GA. A method for estimating equivalent forced outage rates of multistate peaking units. IEEE Trans Power Appar Syst 1978:2067–75.

[9] Li W. Reliability assessment of electric power systems using Monte Carlo methods. Springer US; 1994.

[10] Felder FA. Incorporating resource dynamics to determine generation adequacy levels in restructured bulk power systems. KIEE Int Trans Power Eng 2004;4:100–5.

[11] Fegan GR. Reliability calculations for interdependent plant outages. EPRI report EL-3669. 1984.

[12] Borenstein S, Bushnell JB, Wolak FA. Measuring market inefficiencies in California’s restructured wholesale electricity market. Am Econ Rev 2002;92:1376–405.

[dataset] [13] NERC. The Generating Availability Data System. 2016. <http://www.nerc.com/pa/RAPA/gads/Pages/default.aspx>.

[14] Lueken R, Apt J, Sowell F. Robust resource adequacy planning in the face of coal retirements. Energy Policy 2016;88:371–88.

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

[16] NERC. The Generating Availability Data System (GADS): Applications and benefits. 1995. <http://www.nerc.com/pa/RAPA/gads/Publications/GADS-Applications-and-Benefits.pdf>.

[17] NERC. NERC operating manual. 2016. <http://www.nerc.com/comm/OC/Pages/Operating-Manual.aspx>.

[18] NERC. Generating availability data system: Mandatory reporting of conventional generation performance data. 2011. <http://www.nerc.com/pa/RAPA/gads/MandatoryGADS/Revised_Final_Draft_GADSTF_Recommendation_Report.pdf>.

[19] NERC. Generating Availability Data System data reporting instructions. 2017.

<http://www.nerc.com/pa/RAPA/gads/Pages/Data Reporting Instructions.aspx>.

[dataset] [20] ABB. Velocity Suite. 2016. <http://new.abb.com/enterprise-software/energy-portfolio-management/market-intelligence-services/velocity-suite>.

[21] Pfeifenberger JP, Spees K, Carden K, Wintermantel N. Resource adequacy requirements: Reliability and economic implications. Brattle Gr 2013.

[22] FERC, NERC. Arizona-Southern California outages on September 8, 2011: Causes and recommendations. 2012. <http://www.nerc.com/pa/rrm/ea/Pages/September-2011-Southwest-Blackout-Event.aspx>.

[23] NERC. 2016 summer reliability assessment. 2016. <http://www.nerc.com/pa/RAPA/ra/Reliability Assessments DL/2016 SRA Report_Final.pdf>.

[24] Billinton R, Allan RN. Reliability evaluation of engineering systems. Springer; 1992.

[25] Li W. Risk assessment of power systems: Models, methods, and applications. John Wiley & Sons; 2014.

[26] Papic M, Agarwal S, Allan RN, Billinton R, Dent CJ, Ekisheva S, et al. Research on common-mode and dependent (CMD) outage events in power systems: A review. IEEE Trans Power Syst 2017;32:1528–36.

[27] Sugianto LF, Mielczarski W. A fuzzy logic approach to optimise inventory. In: Forsyth G, Ali M, editors. Ind. Eng. Appl. Artif. Intell. Expert Syst. Proc. Eighth Int. Conf., 1995, p. 419–24.

[28] Elmakias D. New computational methods in power system reliability. vol. 111. Springer Science & Business Media; 2008.

[29] Sarkar A, Kumar D, Kumar S, Singha M. Reliability assessment of Rukhia gas turbine power plant in Tripura 2012;2:184–95.

[30] Oyedepo SO, Fagbenle RO, Adefila SS. Assessment of performance indices of selected gas turbine power plants in Nigeria. Energy Sci Eng 2015;3:239–56.

[31] Lisnianski A, Laredo D, Benhaim H. Multi-state Markov model for reliability analysis of a combined cycle gas turbine power plant. 2016 Second Int. Symp. Stoch. Model. Reliab. Eng. Life Sci. Oper. Manag., 2016, p. 131–5.

[32] NERC. 2014 probabilistic assessment. 2015. <http://www.nerc.com/pa/RAPA/ra/Reliability Assessments DL/2014ProbA April Report Final_Final.pdf>.

[33] Spees K, Newell SA, Pfeifenberger JP. Capacity markets—Lessons learned from the first decade. Econ Energy Environ Policy 2013;2:1–26.

[34] 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. 2016. <https://www.iso-ne.com/static-assets/documents/2016/01/icr_values_2019_2020_report_final.pdf>.

[35] MISO. Business practices manual 11: Resource adequacy. 2016. <https://www.misoenergy.org/Library/BusinessPracticesManuals/Pages/BusinessPracticesManuals.aspx>.

[36] NYISO. Installed capacity manual. 2017. <http://www.nyiso.com/public/webdocs/markets_operations/documents/Manuals_and_Guides/Manuals/Operations/icap_mnl.pdf>.

[37] New York State Reliability Council. New York control area installed capacity requirement. 2016. <http://www.nysrc.org/pdf/Reports/2017 IRM Study Report Final 12-

2-16 (002).pdf>.

[38] PJM Interconnection. PJM manual 20: PJM resource adequacy analysis. 2016. <http://www.pjm.com/~/media/documents/manuals/m20.ashx>.

[39] PJM Interconnection. PJM manual 18: PJM capacity market. 2017. <https://www.pjm.com/~/media/documents/manuals/m18.ashx>.

[40] Politis DN, White H. Automatic block-length selection for the dependent bootstrap. Econom Rev 2004;23:53–70.

[41] Hayfield T, Racine JS. Nonparametric econometrics: The np package. J Stat Softw 2008;27:1–32.

[42] Patton A, Politis DN, White H. Correction to “Automatic block-length selection for the dependent bootstrap” by D. Politis and H. White. Econom Rev 2009;28:372–5.

[43] Canty A, Ripley B. “boot: Bootstrap R (S-Plus) Functions.” R package version 1.3-16 2015.

[44] Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge University Press; 1997.

Glossary

CC: Combined cycle units

CT: Simple cycle units

D1, D2, D3: Unscheduled deratings (partial failures), event types in GADS

FERC: Federal Energy Regulatory Commission

FRCC: Florida Reliability Coordinating Council, a NERC Regional Entity

FSFB: Fossil steam and fluidized bed combustion units

GADS: Generating Availability Data System, a database of historical generator availability

HY: Hydroelectric and pumped storage units

MRO: Midwest Reliability Organization, a NERC Regional Entity

NAC: Net available capacity, a variable reported in a GADS event record indicating the amount of available capacity remaining on the unit at the start of the corresponding event

NERC: North American Electric Reliability Corporation, the FERC-designated Electric Reliability Organization for the United States

NPCC: Northeast Power Coordinating Council, a NERC Regional Entity

NU: Nuclear units

ORAP: Operational Reliability Analysis Program, a database of historical generator availability

PJM: PJM Interconnection LLC, a regional transmission operator in the eastern United States

RAM: Resource adequacy modeling

RFC: ReliabilityFirst Corporation, a NERC Regional Entity

RS: Reserve shutdown event type

RTO: Regional transmission operator

SERC: Southeast Electric Reliability Council, a NERC Regional Entity

SF: Startup failure, an unscheduled event type in GADS

SPP: Southwest Power Pool, a NERC Regional Entity

TRE: Texas Reliability Entity, a NERC Regional Entity

U1, U2, U3: Unscheduled full outages, event types in GADS

WECC: Western Electricity Coordinating Council, a NERC Regional Entity

Online supplementary material: Resource adequacy risks to the bulk power system in North America

Sinnott Murphy[†], Jay Apt[†,‡,*], John Moura[§], Fallaw Sowell[‡]

[†]Department of Engineering & Public Policy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, U.S.A.
[‡]Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, U.S.A.
[§]North American Electric Reliability Corporation, 3353 Peachtree Road Suite 600, Atlanta, Georgia 30326, U.S.A.
[*]Corresponding author's email address: apt@cmu.edu

69 pages, 51 figures, 16 tables

Contents

List of Figures

List of Tables

1 Characteristics of the generator availability data

A map of the eight NERC regions in the conterminous United States and the Canadian provinces is shown below in Figure S-1.

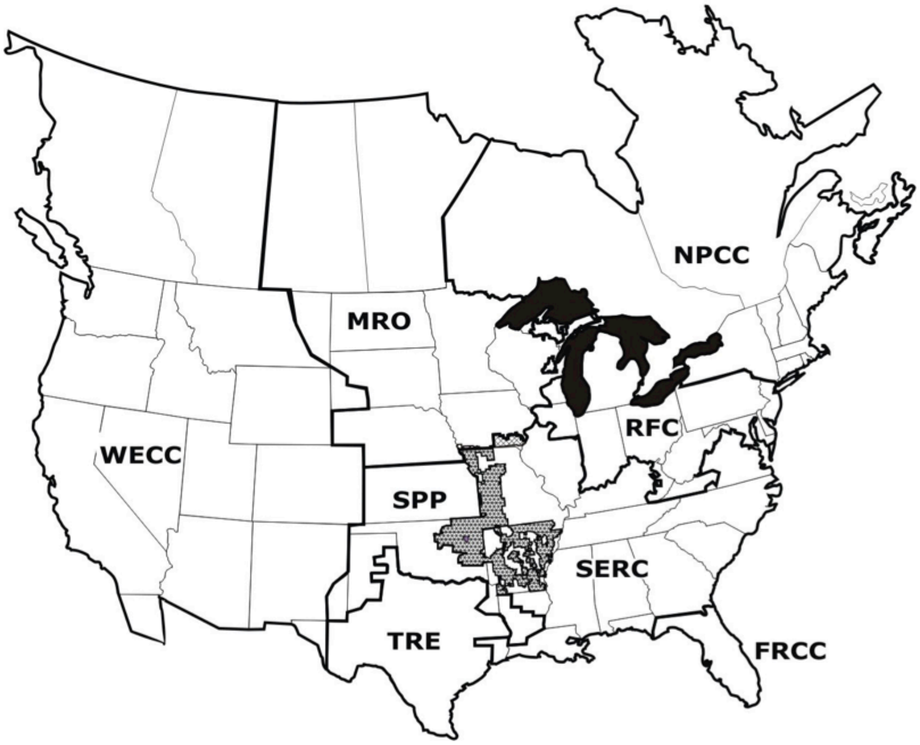


Figure S-1: The NERC regions. Source: NERC.

Overall outage statistics for each region's time series are below in Table S-1.

Table S-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.

We present hourly time series of the percent of generating units in each region affected by an unscheduled event in Figure S-2. In contrast to Figure 3 in the main text, Figure S-2 shows correlated failures without weighting by capacity. Higher values in this plot denote instances where a large number of units are simultaneously affected by an unscheduled event, without regard to the nameplate capacity of the unit or, in the case of partial failures, the magnitude of the failure itself. Where this figure differs from Figure 3 in the main text can help identify when a large unavailable capacity event is due to a few large units failing together.

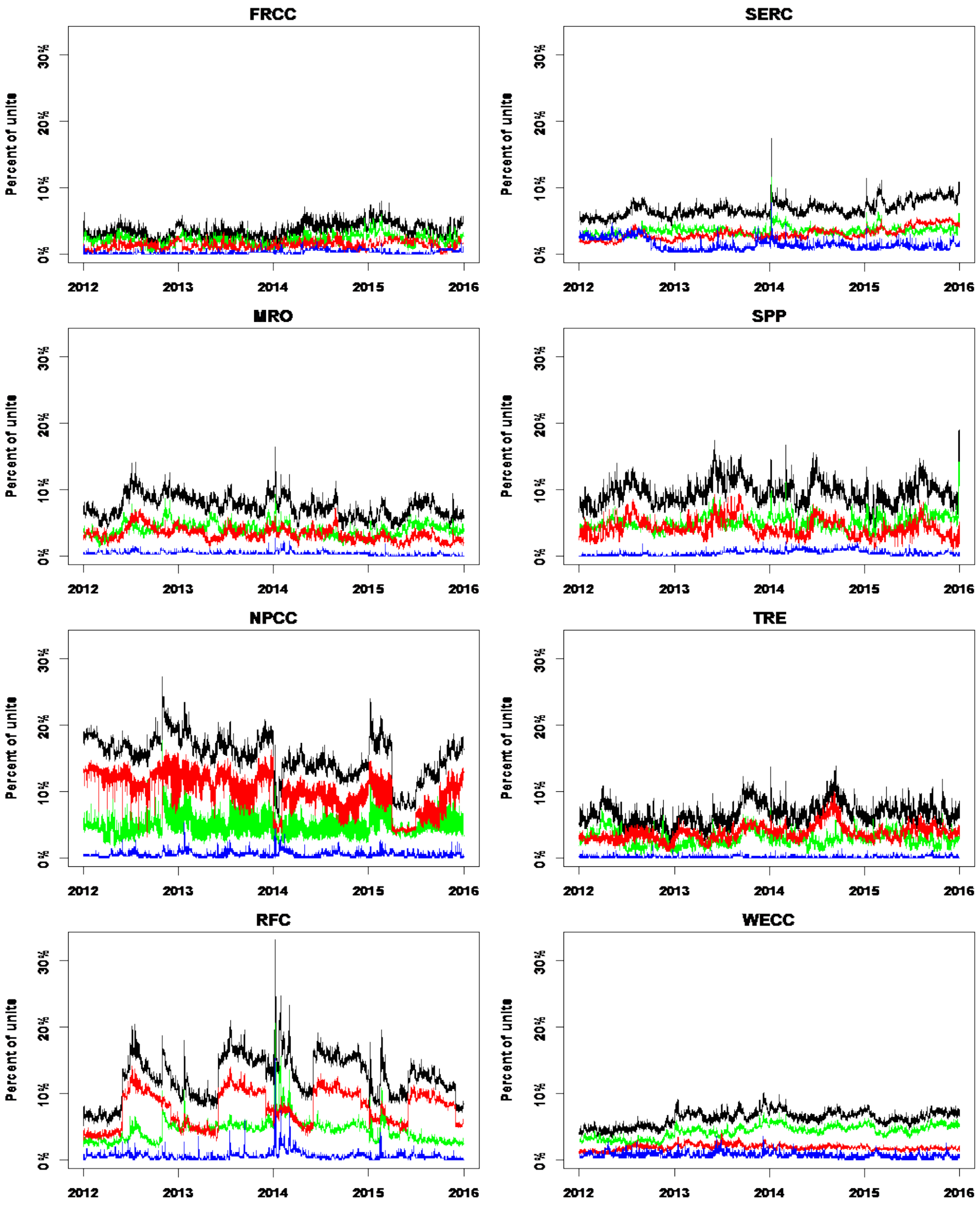


Figure S-2: Percent of units in each region reporting an unscheduled event in each hour of time series Green: unscheduled outages only; red: unscheduled deratings only; blue: start-up failures only; black: all unscheduled events (the sum of green, red, and blue curves).

The percent of unavailable capacity from each of the unscheduled event types (startup failures, unscheduled outages, and unscheduled deratings) is shown as a pie chart for each region in Figure S-3. This weights the prevalence of unscheduled events by their magnitude. We complement this with the percent of event counts by unscheduled event type in Figure S-4. Deratings are the most common unscheduled event type in every region but full outages always represent more unscheduled unavailable MWh.

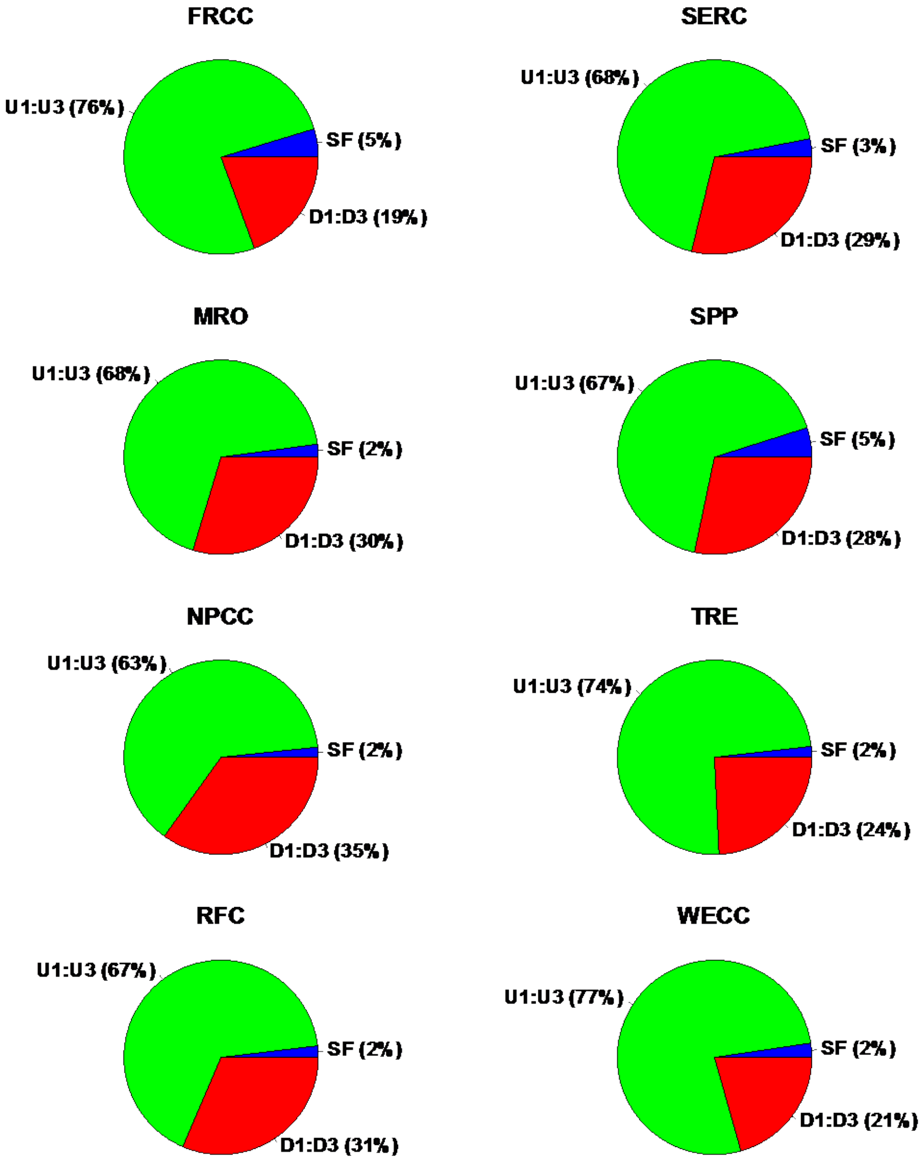


Figure S-3: Proportion of unscheduled unavailable MWh by event type category.

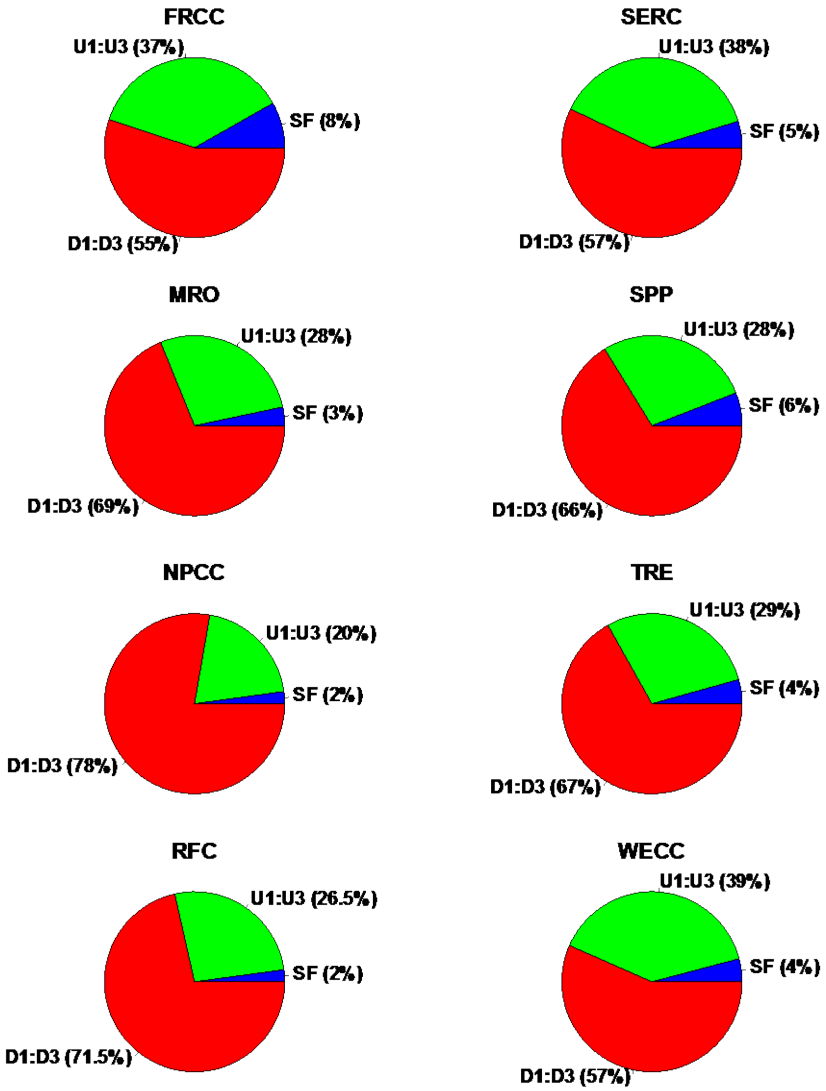


Figure S-4: Proportion of unscheduled event counts by event type category.

A histogram of the number of months each unit was in operation during the four years of our analysis is shown in Figure S-5. This supports the MTBF, MTTR, and correlated failures analyses, which implicitly assume that all units are in operation for the full study period. More than 60% of units were active the entire study period and 80% were active for at least three years.

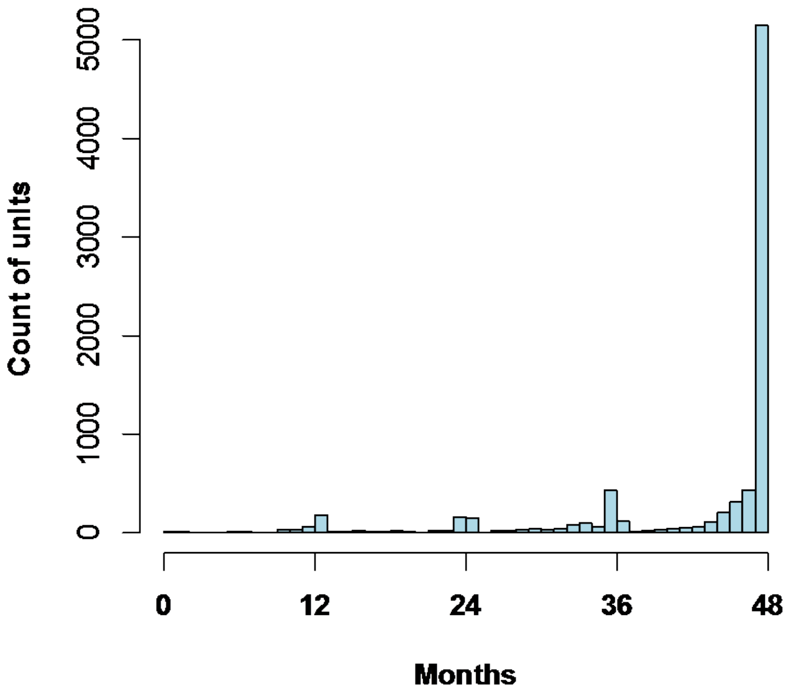


Figure S-5: Histogram of months each unit was in operation (2012-2015).

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 S-6, S-7, and S-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 [1]. 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.

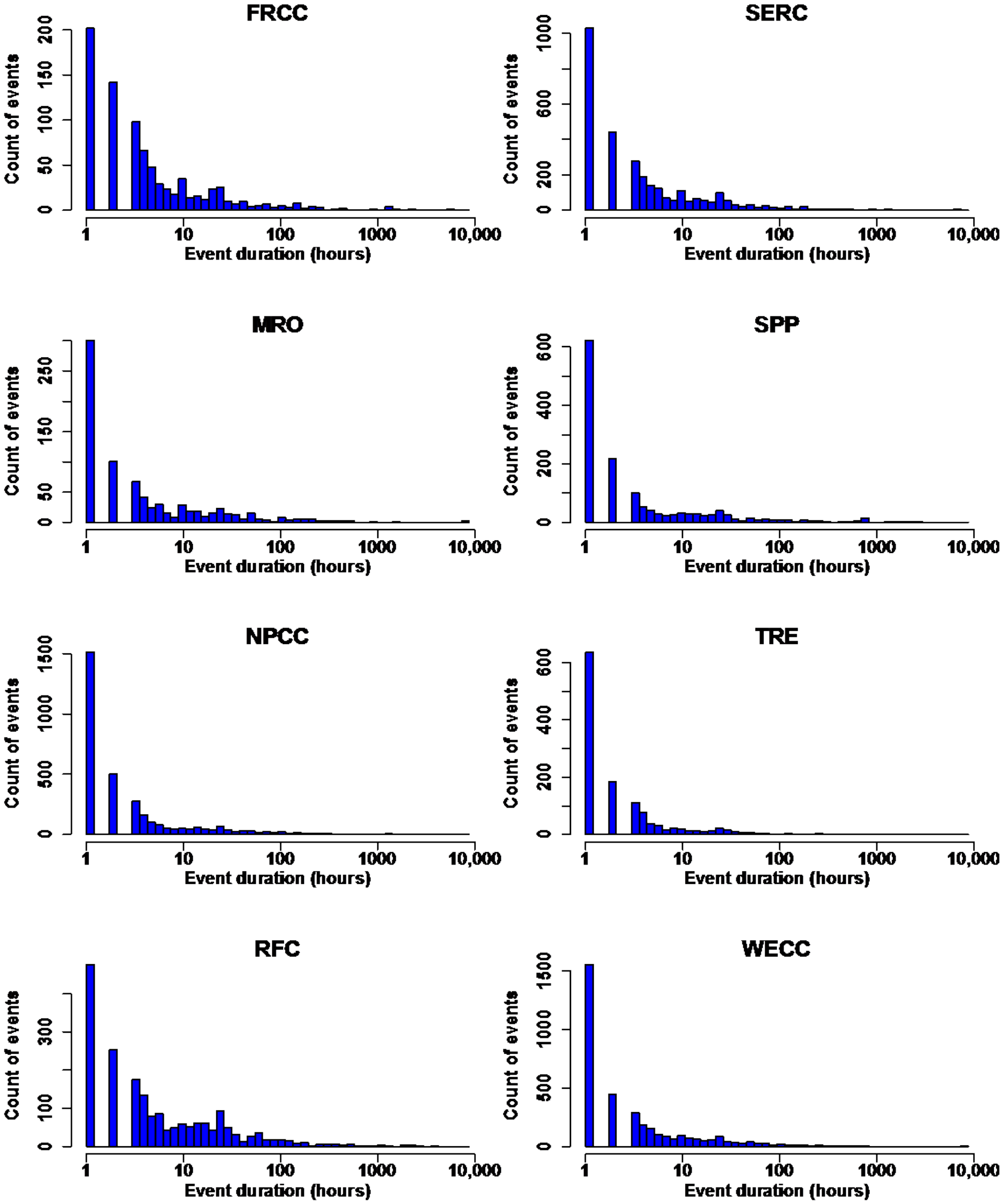


Figure S-6: Histogram of durations for startup failures. Full study period. Note the log scale for event durations.

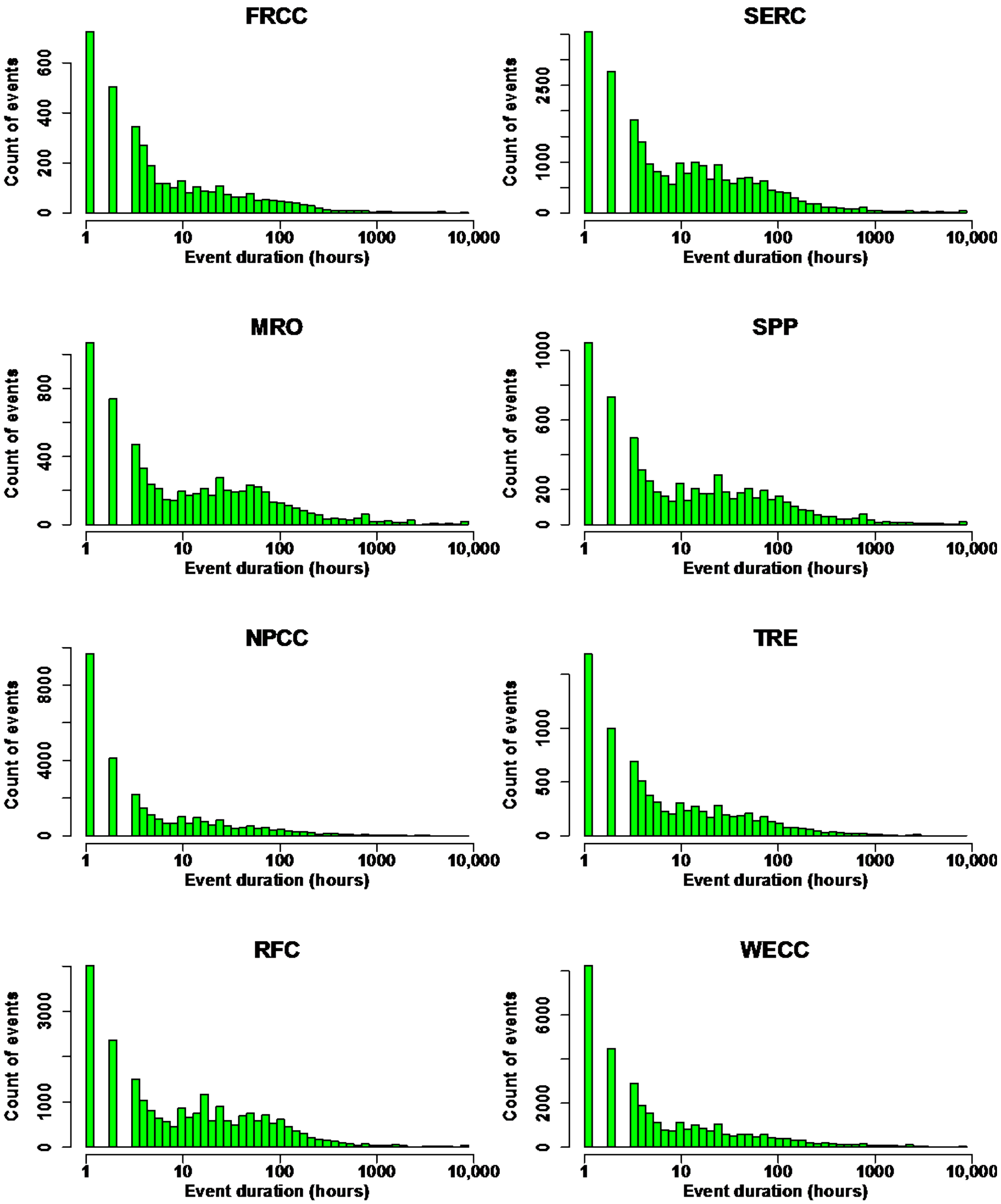


Figure S-7: Histogram of durations for unscheduled outages. Full study period. Note the log scale for event durations.

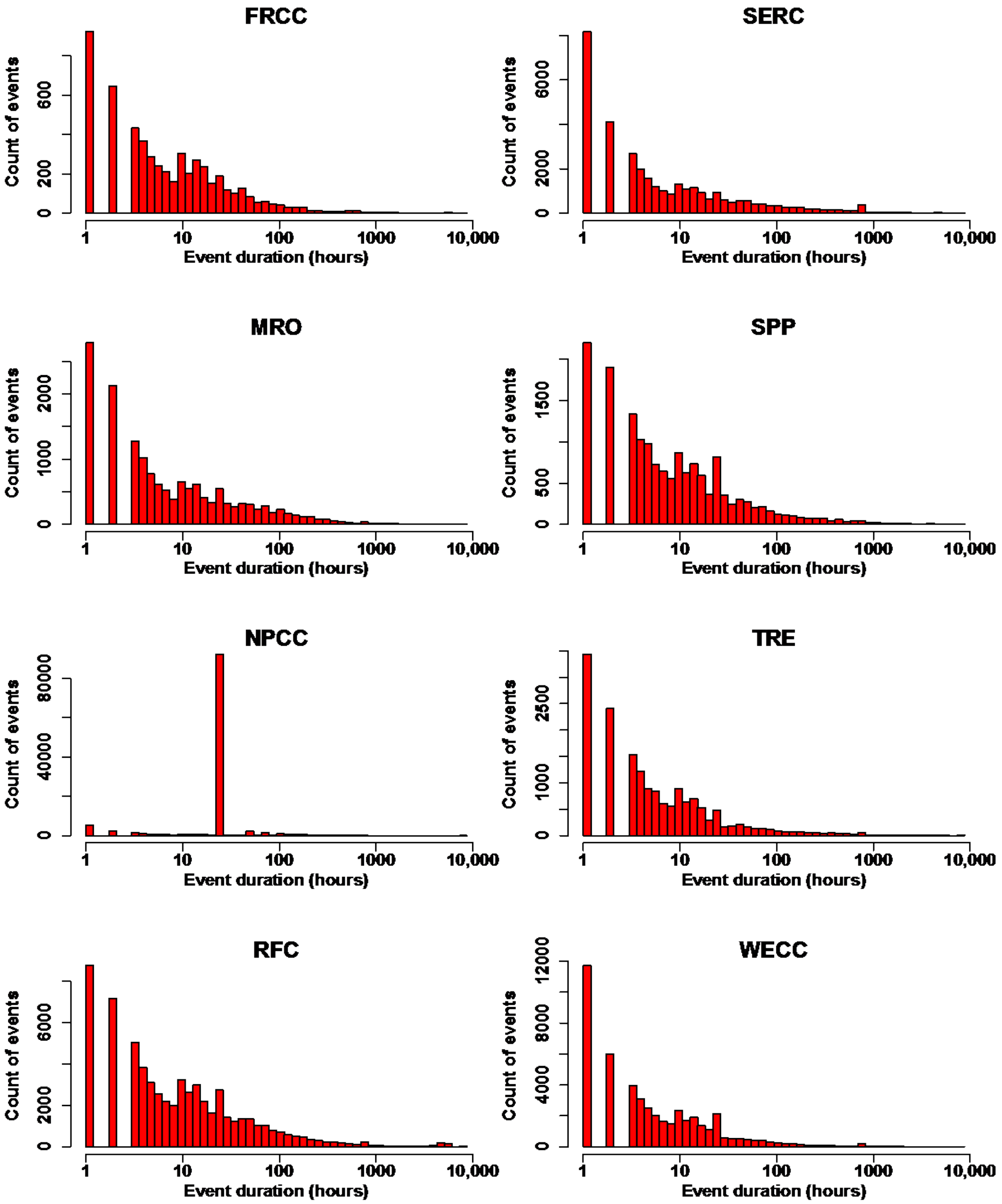


Figure S-8: Histogram of durations for unscheduled deratings. Full study period. Note the log scale for event durations.

Parameters of lognormal distributions fit to each region’s unscheduled event durations are reported by event type in Table S-2.

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

	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 (0.0080)	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. Histograms of the start and end minute of each unscheduled event are shown in Figures S-9 and S-10.

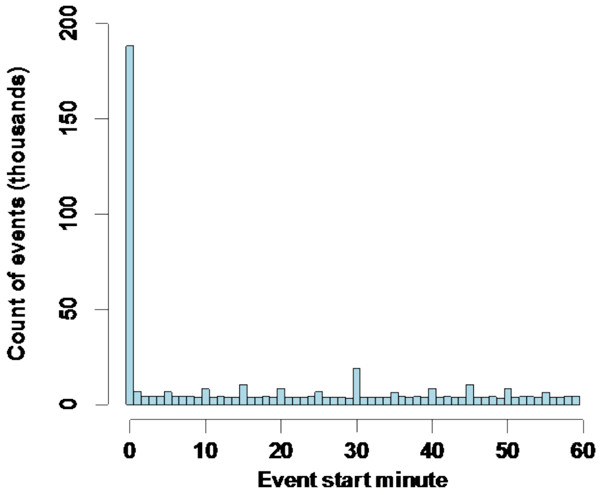


Figure S-9: Histogram of recorded start minute of unscheduled events.

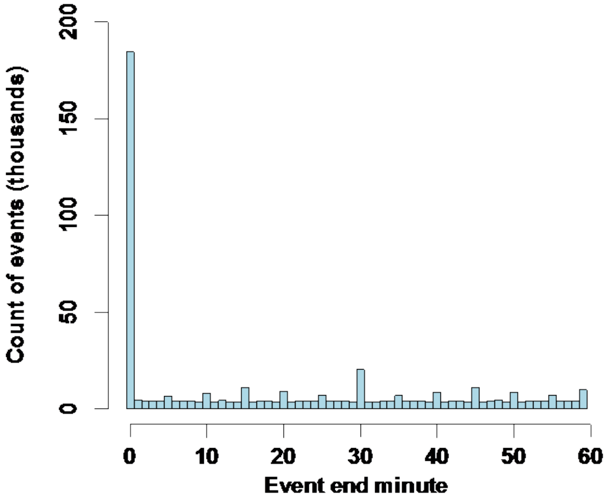


Figure S-10: Histogram of recorded end minute of unscheduled events.

Summary statistics of the number of unscheduled events experienced by each generating unit are shown in Table S-3.

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

Region	*10th*	*20th*	*30th*	*40th*	*50th*	*60th*	*70th*	*80th*	*90th*
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	41	65

2 Tests for independence among generators

2.1 Block subsampling supplementary data

Figure S-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 S-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 S-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 S-11 and Figure S-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.

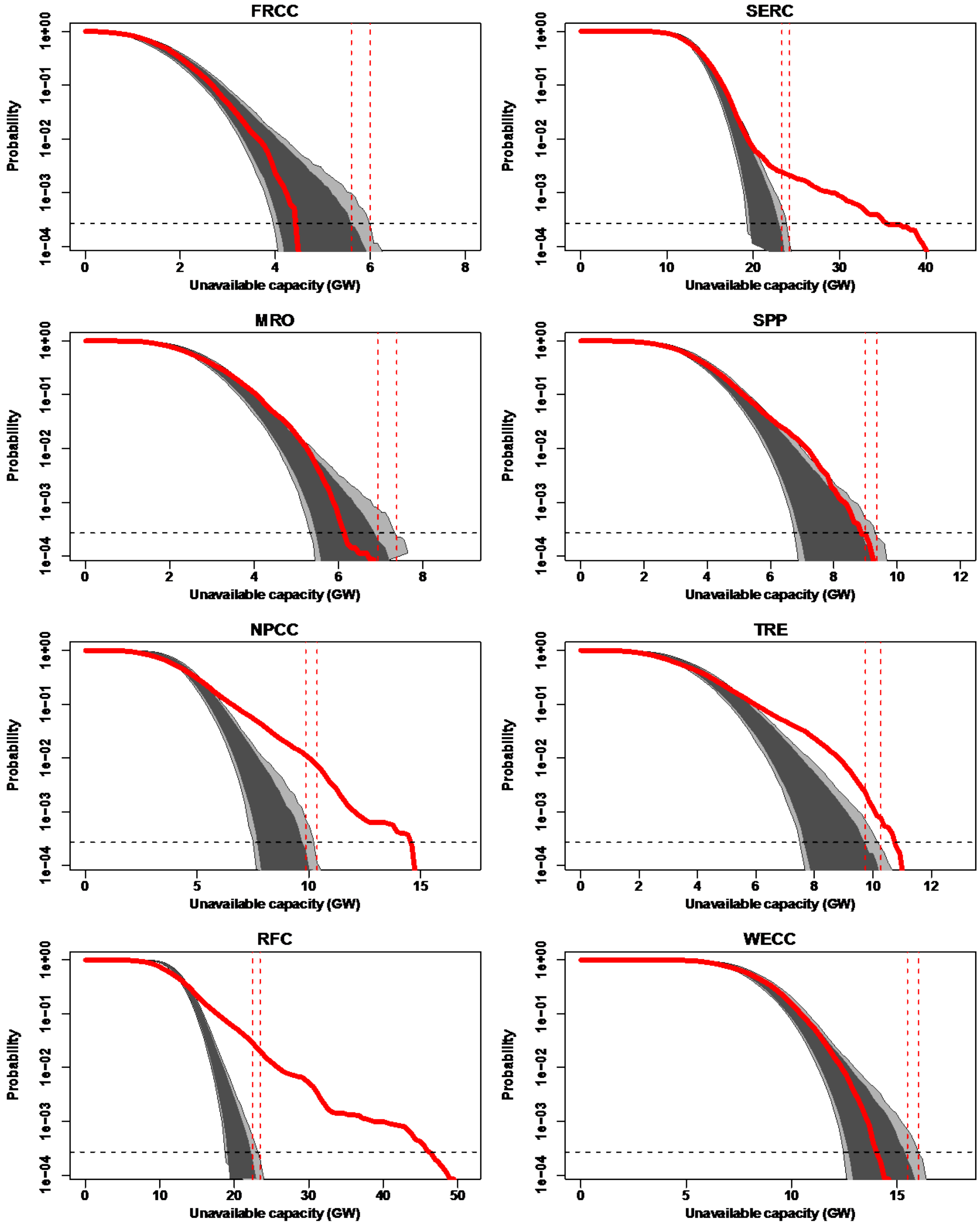


Figure S-11: 95% and 99% confidence bands from 1000 block subsampling runs (non-seasonal subsampling, full 2012-2015 period) in dark and light gray, respectively; full empirical distributions in red. Dashed horizontal line indicates 2.4 LOLH threshold; dashed vertical lines indicate intersection of 2.4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

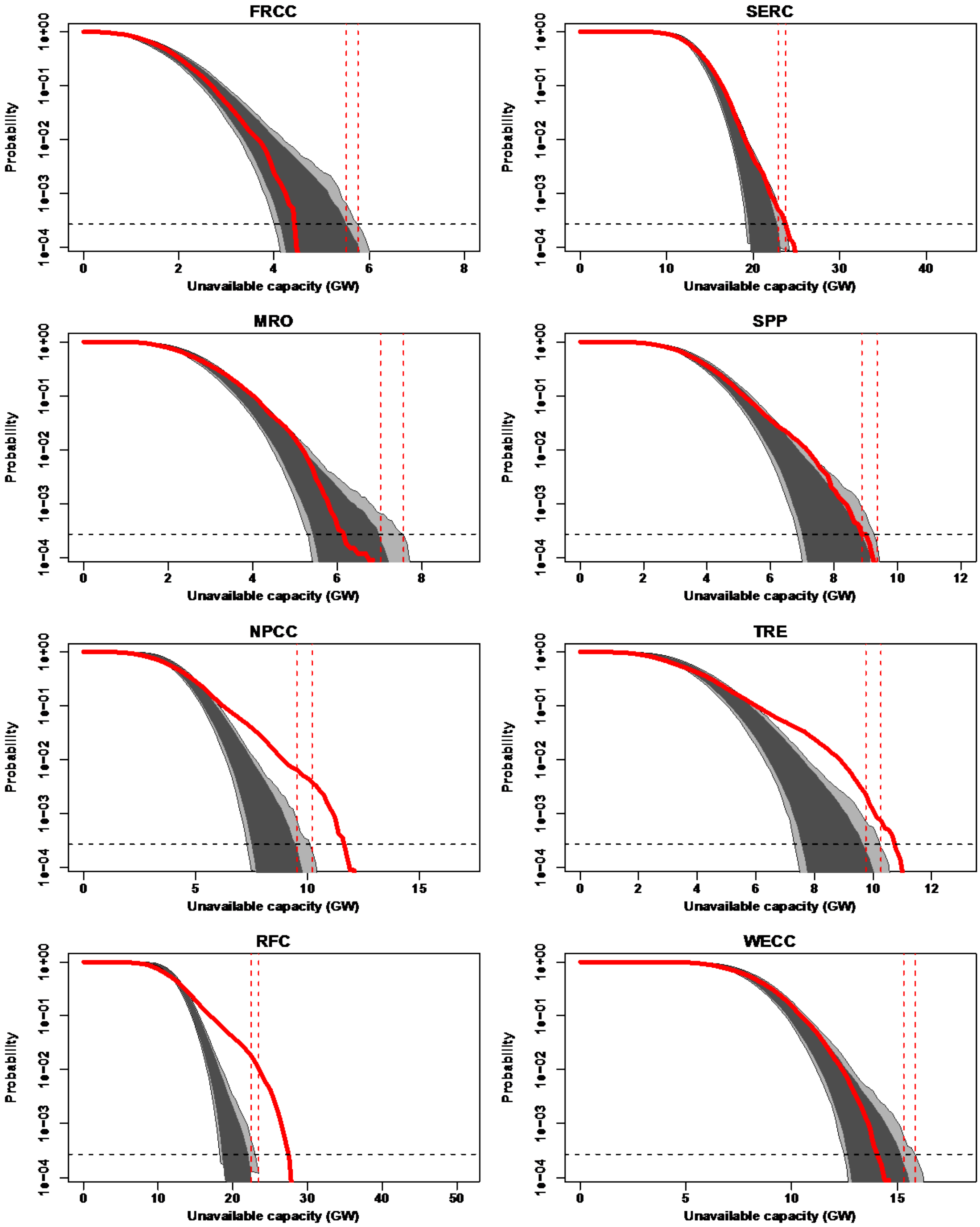


Figure S-12: 95% and 99% confidence bands from 1000 block subsampling runs (non-seasonal subsampling, Hurricane Sandy and January 2014 removed) in dark and light gray, respectively; empirical distributions (excluding Hurricane Sandy and January 2014) in red. Dashed horizontal line indicates 2.4 LOLH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

Table S-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).

	Full period		*Removing Sandy and January 2014*	
Region	*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	--	--	--	--

We present histograms of the block length (in hours) used for each generating unit for the block subsampling analysis (Figure S-13). Block lengths range from a single hour to a maximum of 562 hours (~23 days).

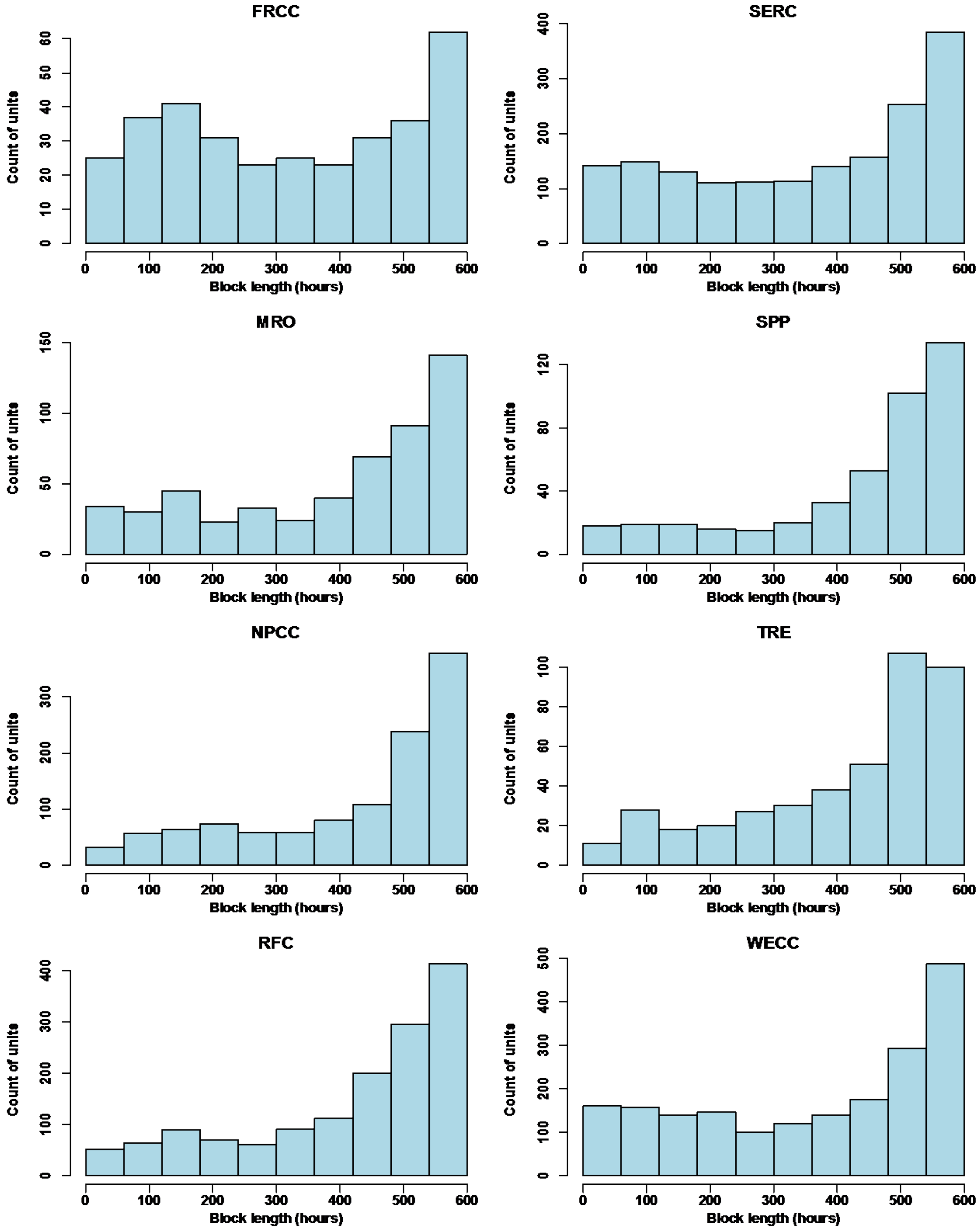


Figure S-13: Histograms of optimal block length used in the block subsampling analysis.

2.2 Binomial event arrivals supplementary data

We present histograms of each unit’s hourly probability of experiencing an unscheduled event arrival for the full study period, overlaid with a fitted lognormal distribution, in Figure S-14. Each unit’s event arrival probability is estimated according to Equation 3 in the main text.

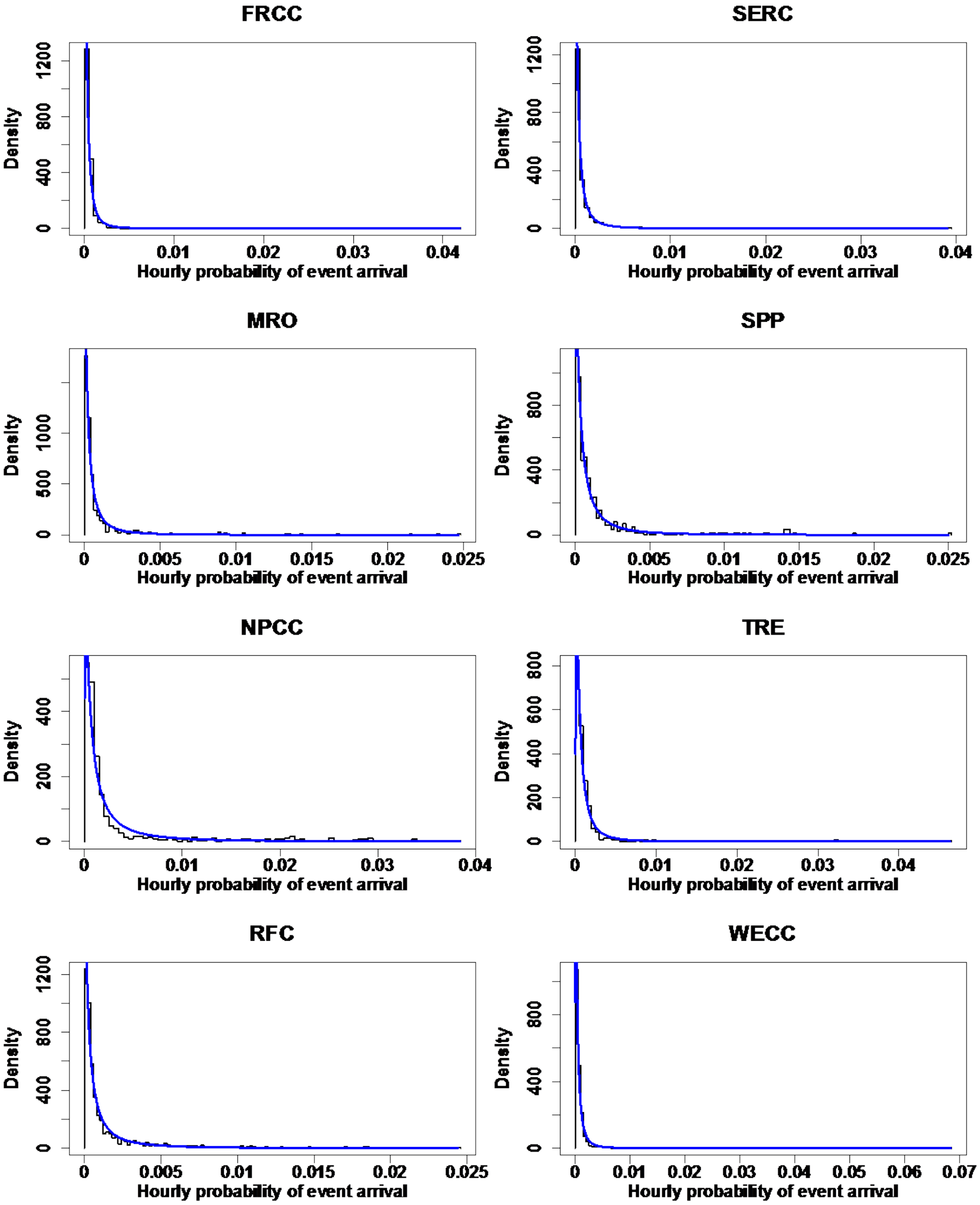


Figure S-14: Event arrival probabilities by region with fitted lognormal distribution overlaid. Probabilities estimated via Equation 3 in the main text. Full study period. Excludes the nine units that are not at least partially available for 1000 or more hours during the study period.

We summarize the parameters of lognormal fits to each region's distribution of unscheduled event arrival probabilities in Table S-5.

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

	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 (0.230)
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)

2.3 A supplemental statistical test for failure correlation

We also test for positive correlation for every possible pair of units in each region using a t-test as described in [3]. 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 S-6 and when removing Hurricane Sandy and January 2014 in Table S-7.

Table S-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 S-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%

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 4.2 in the main text, where we found meaningful violations of the independence assumption only in four regions: NPCC, RFC, SERC, and TRE.

3 Seasonality

3.1 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 S-15 through S-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.

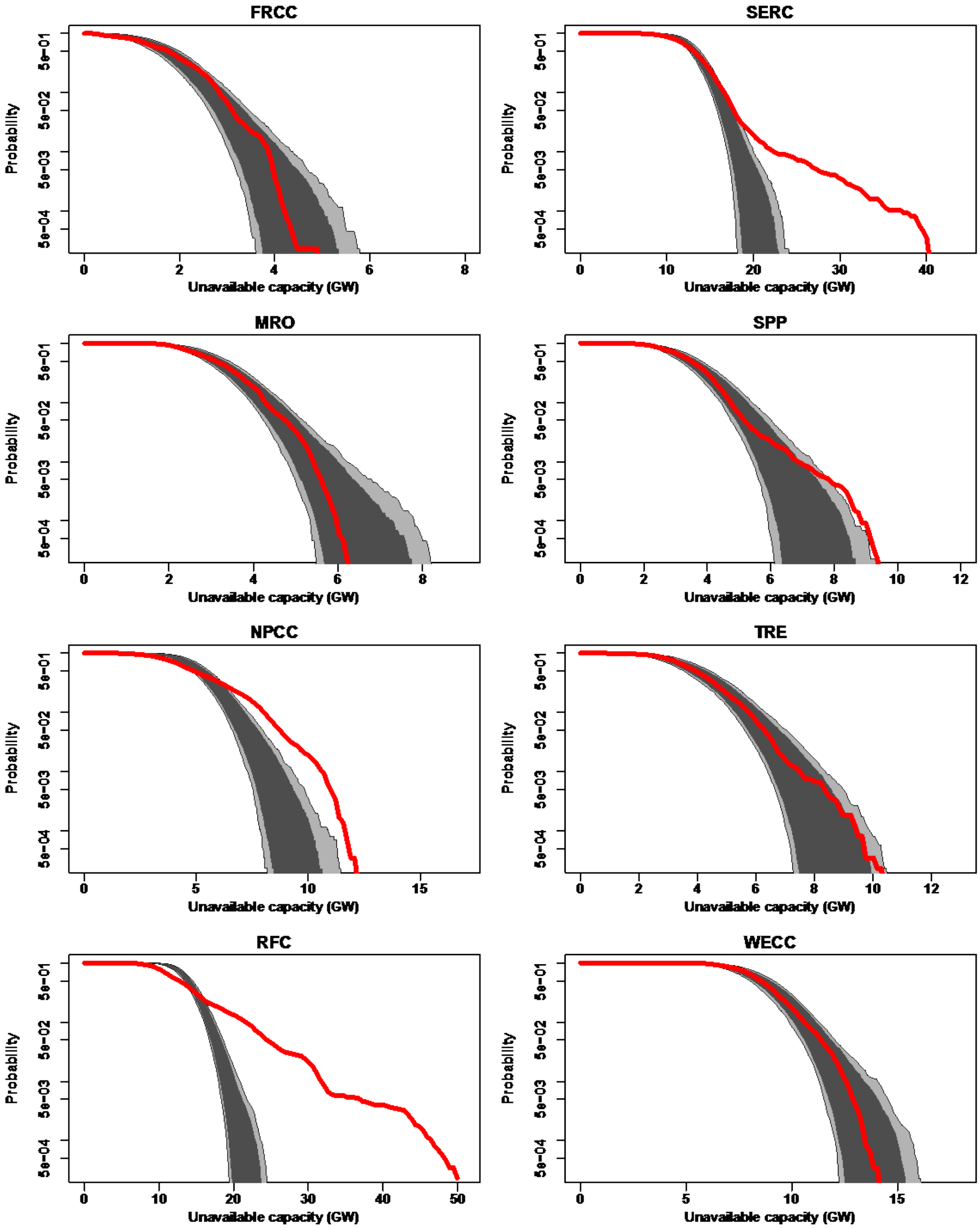


Figure S-15: 95% and 99% confidence bands from 1000 block subsampling runs for the full winter period (December, January, and February 2012-2015) in black, empirical distributions for only winter months in red.

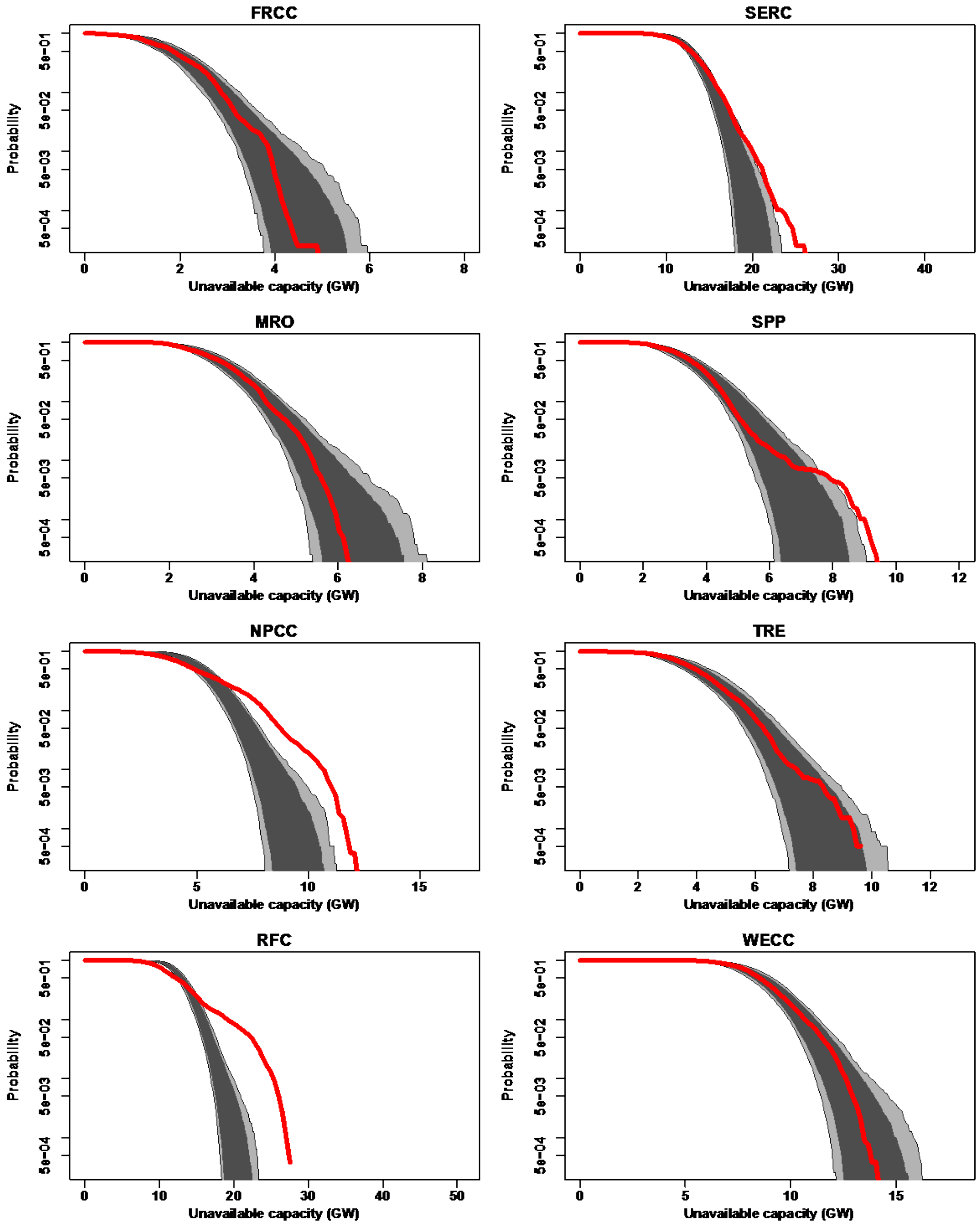


Figure S-16: 95% and 99% confidence bands from 1000 block subsampling runs for the winter period except January 2014 in black, empirical distributions for corresponding winter months in red.

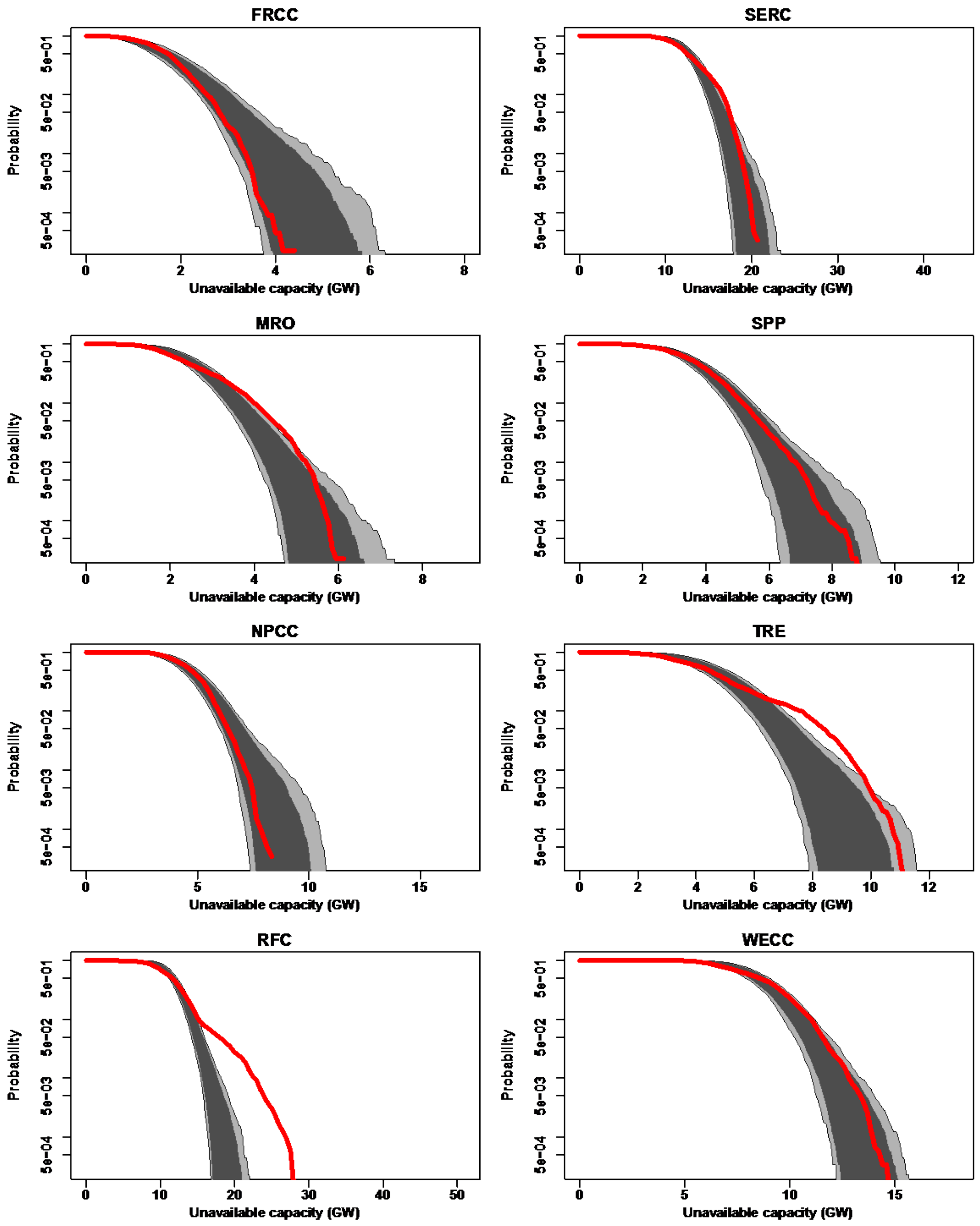


Figure S-17: 95% and 99% confidence bands from 1000 block subsampling runs for the spring period (March, April, and May 2012-2015) in black, empirical distributions for corresponding months in red.

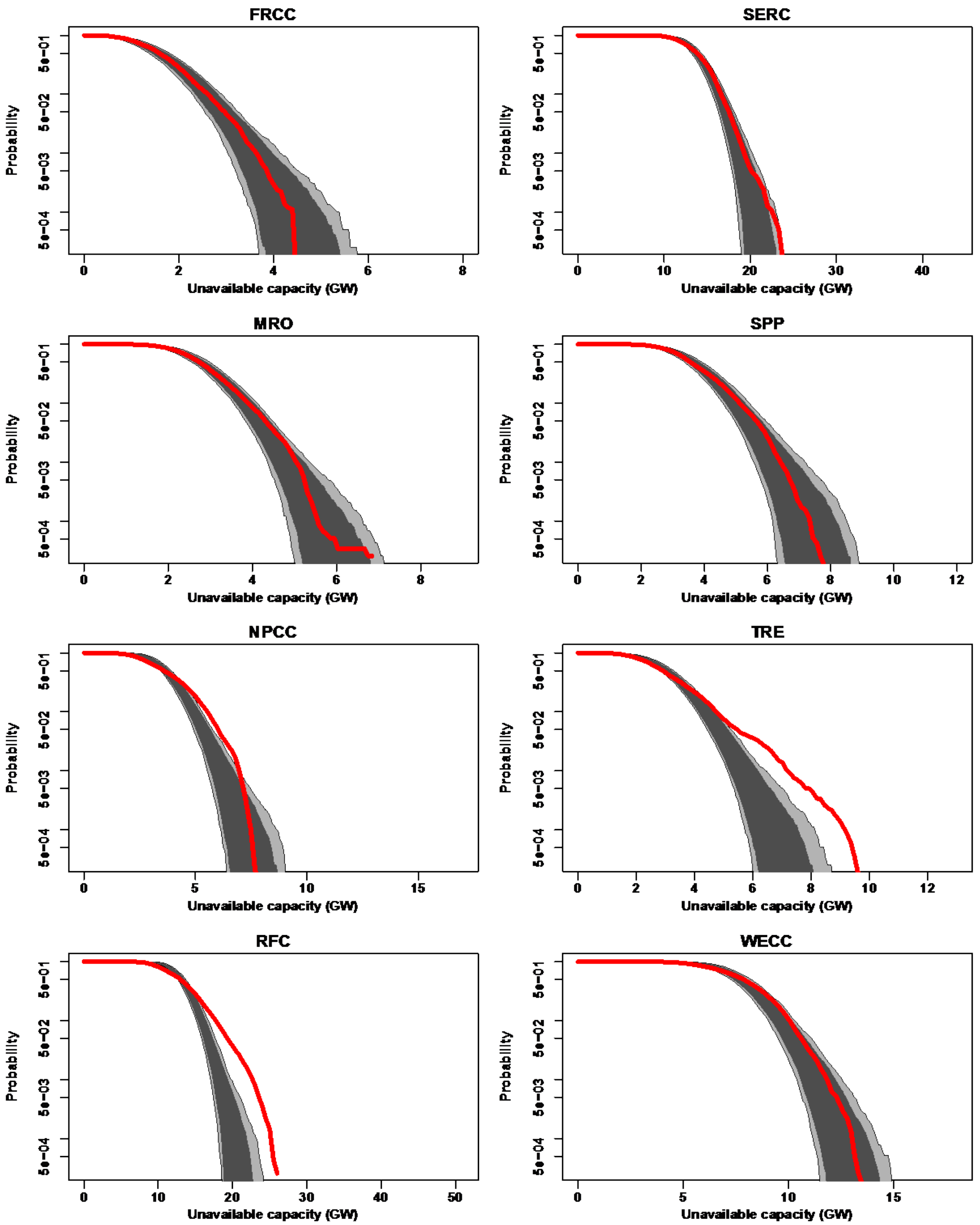


Figure S-18: 95% and 99% confidence bands from 1000 block subsampling runs for the summer period (June, July, August, and September 2012-2015) in black, empirical distributions for corresponding months in red.

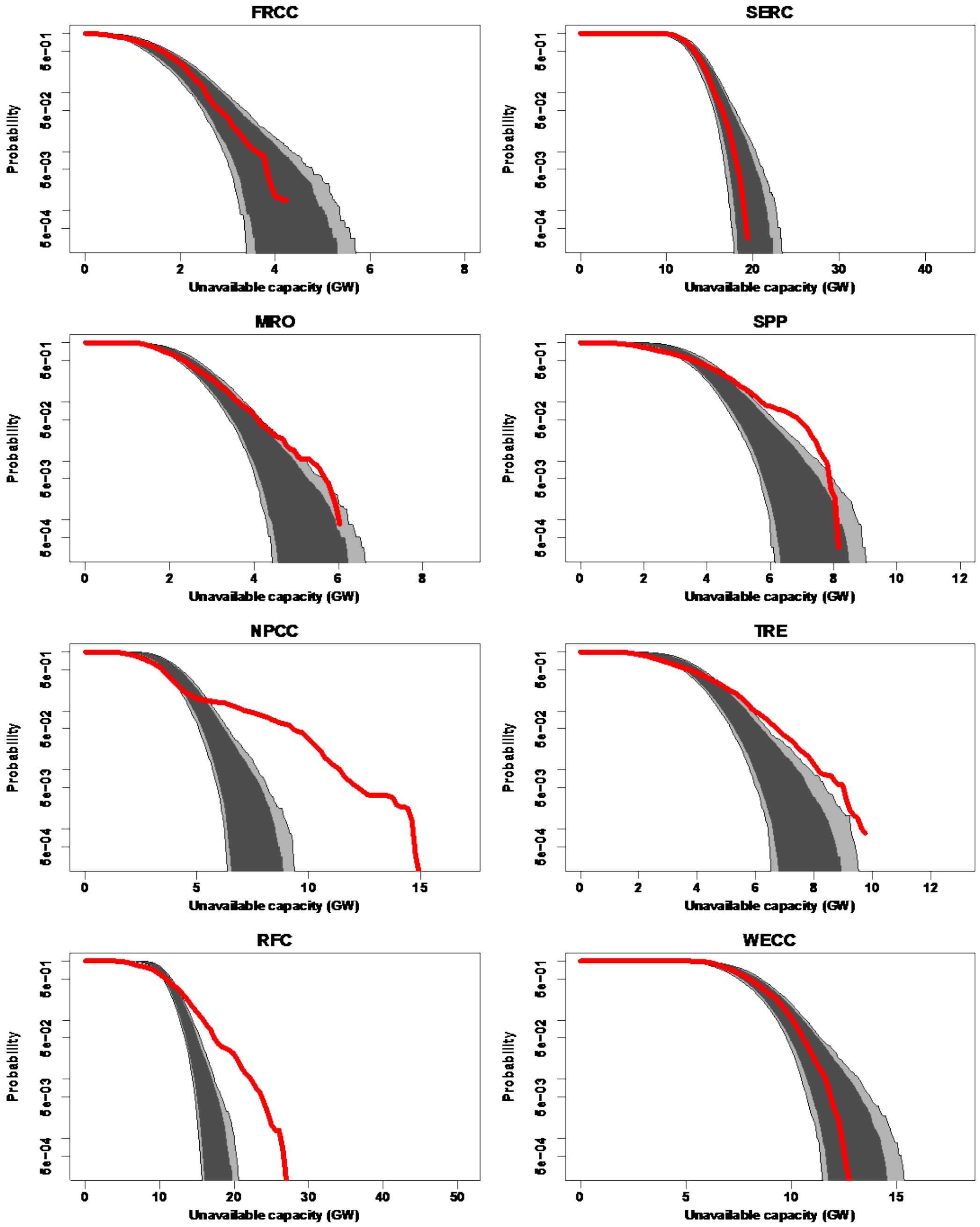


Figure S-19: 95% and 99% confidence bands from 1000 block subsampling runs for the fall period (October and November 2012-2015) in black, empirical distributions for corresponding months in red.

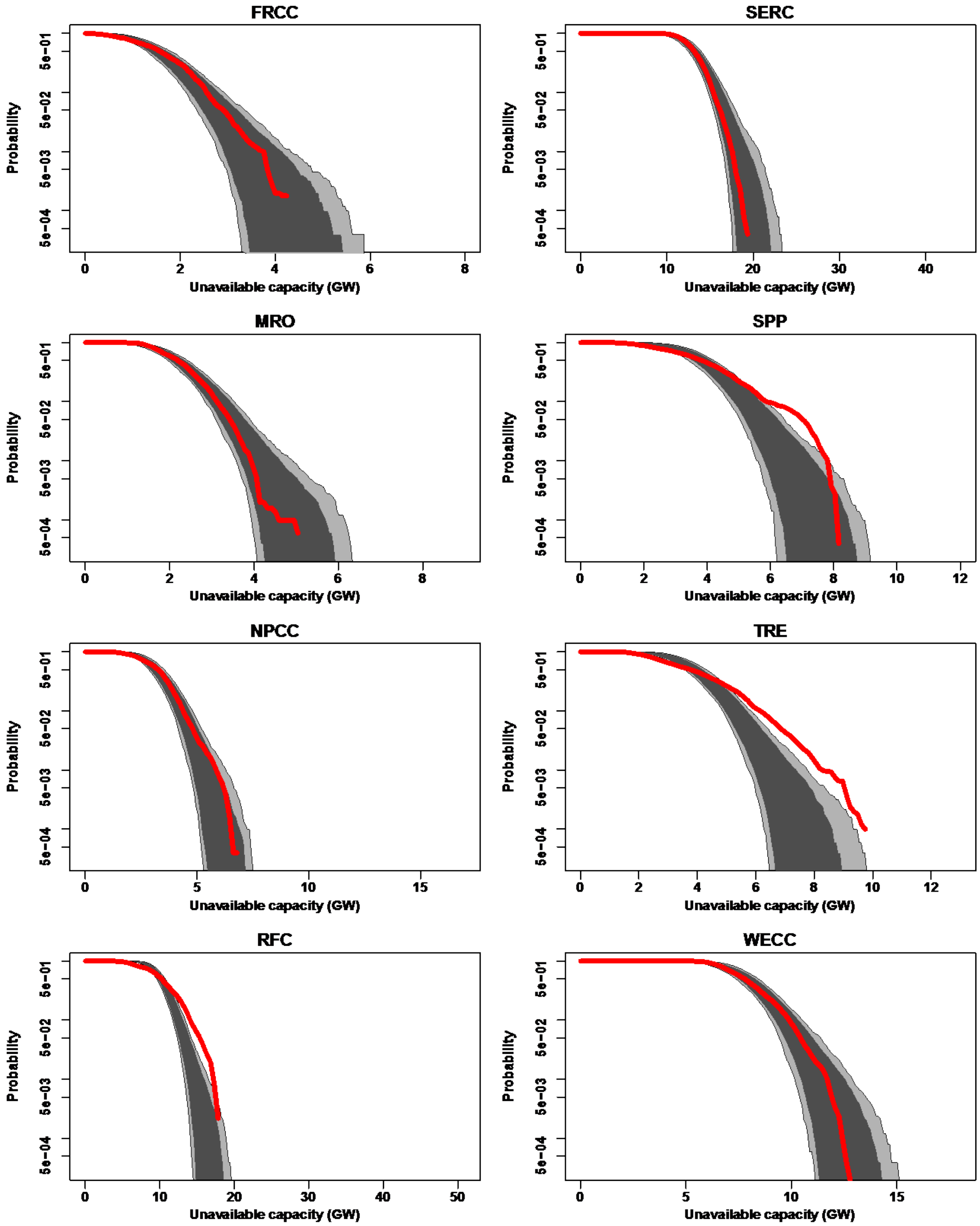


Figure S-20: 95% and 99% confidence bands from 1000 block subsampling runs for the fall period except Hurricane Sandy (October and November 2012-2015 except for October 29-November 30, 2012) in black, empirical distributions for corresponding months in red.

3.2 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 S-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 S-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.

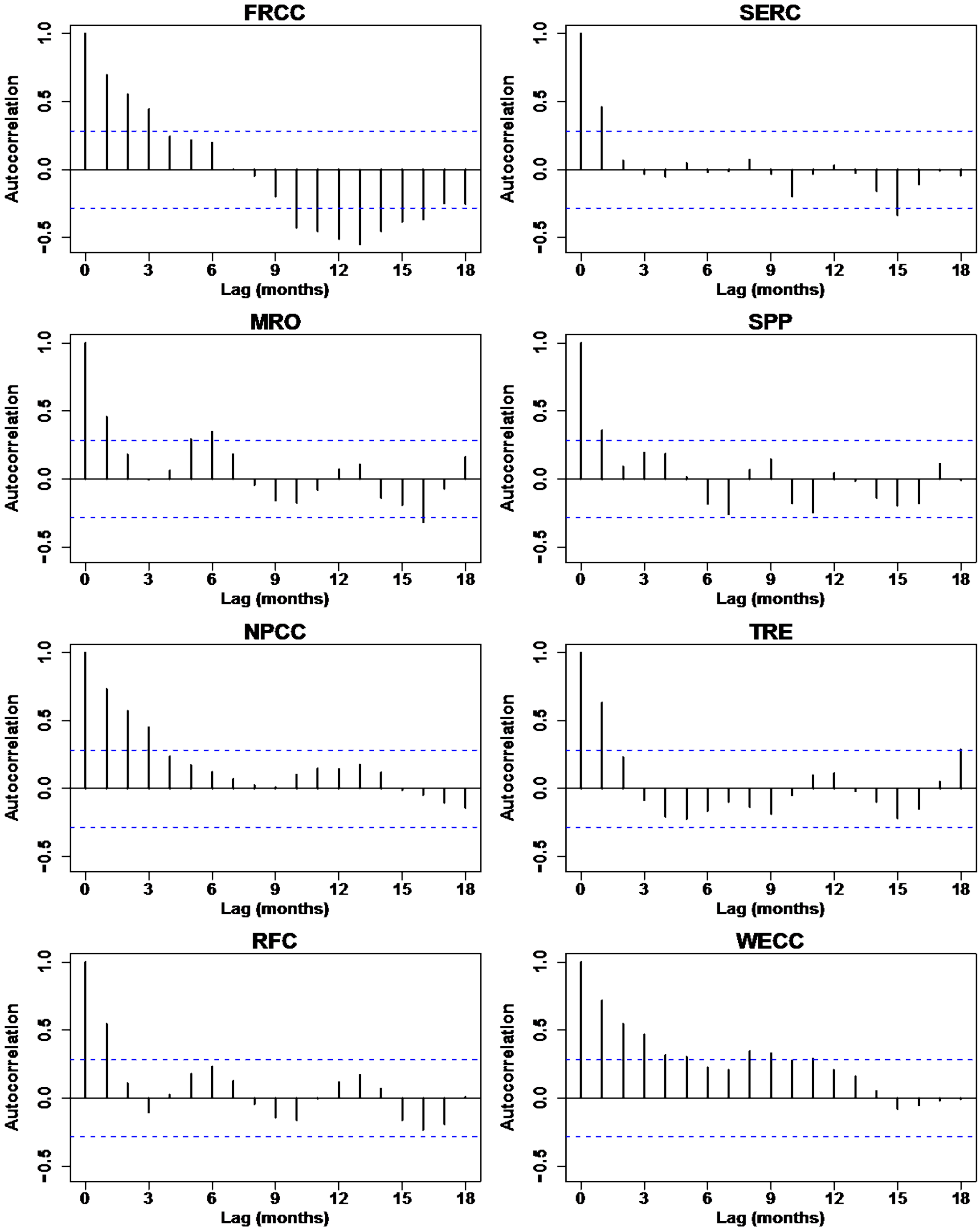


Figure S-21: Autocorrelation functions of monthly average unscheduled unavailable capacity by region with 95% confidence interval for assessing statistical significance of lags (dashed blue lines).

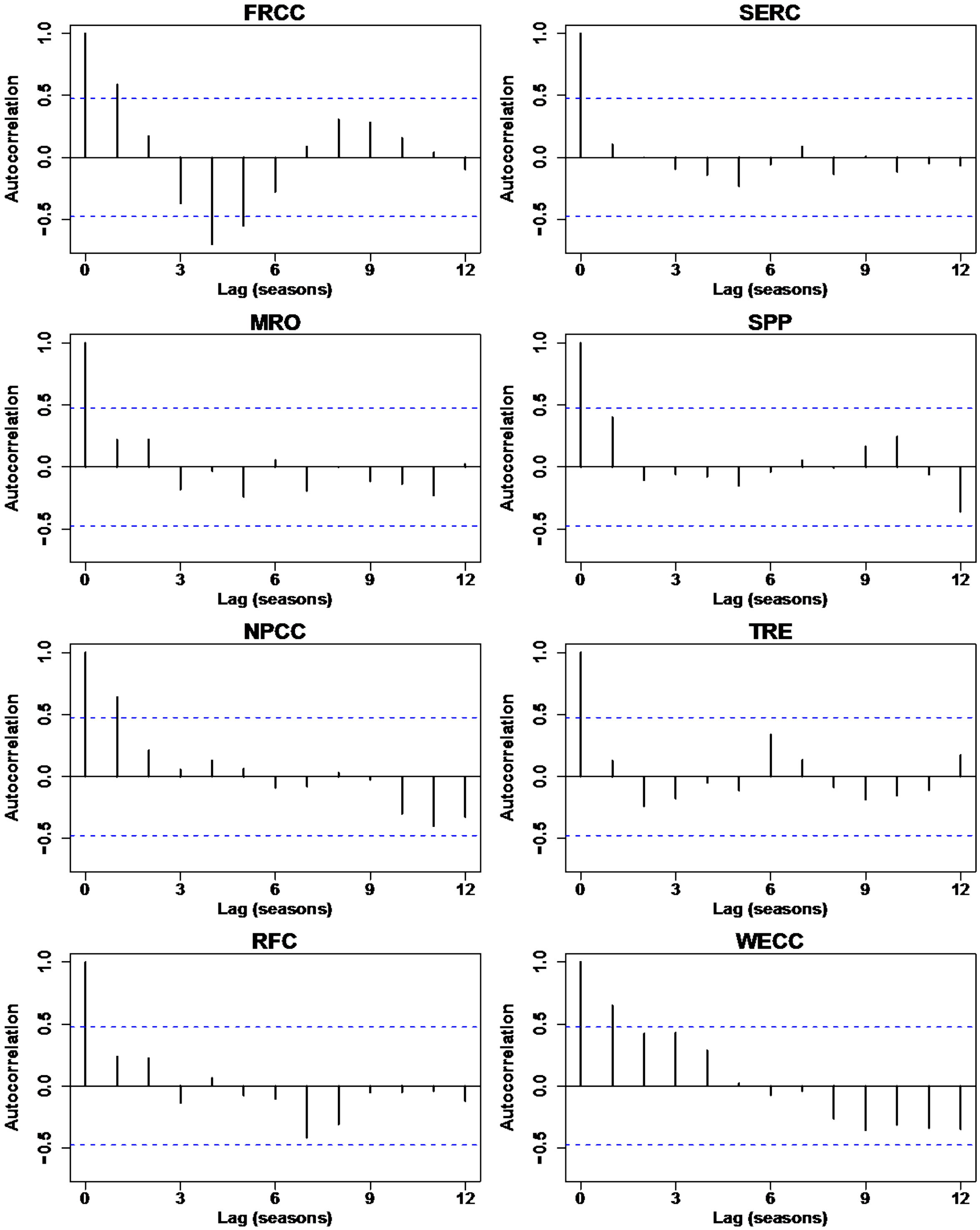


Figure S-22: Autocorrelation functions of seasonal average unscheduled unavailable capacity by region with 95% confidence interval for assessing statistical significance of lags (dashed blue lines).

We next plot each season's distribution of unscheduled unavailable capacity as exceedance curves (Figure S-23). The seasonal exceedance curves generally overlap, again suggesting little evidence of consistent seasonality.

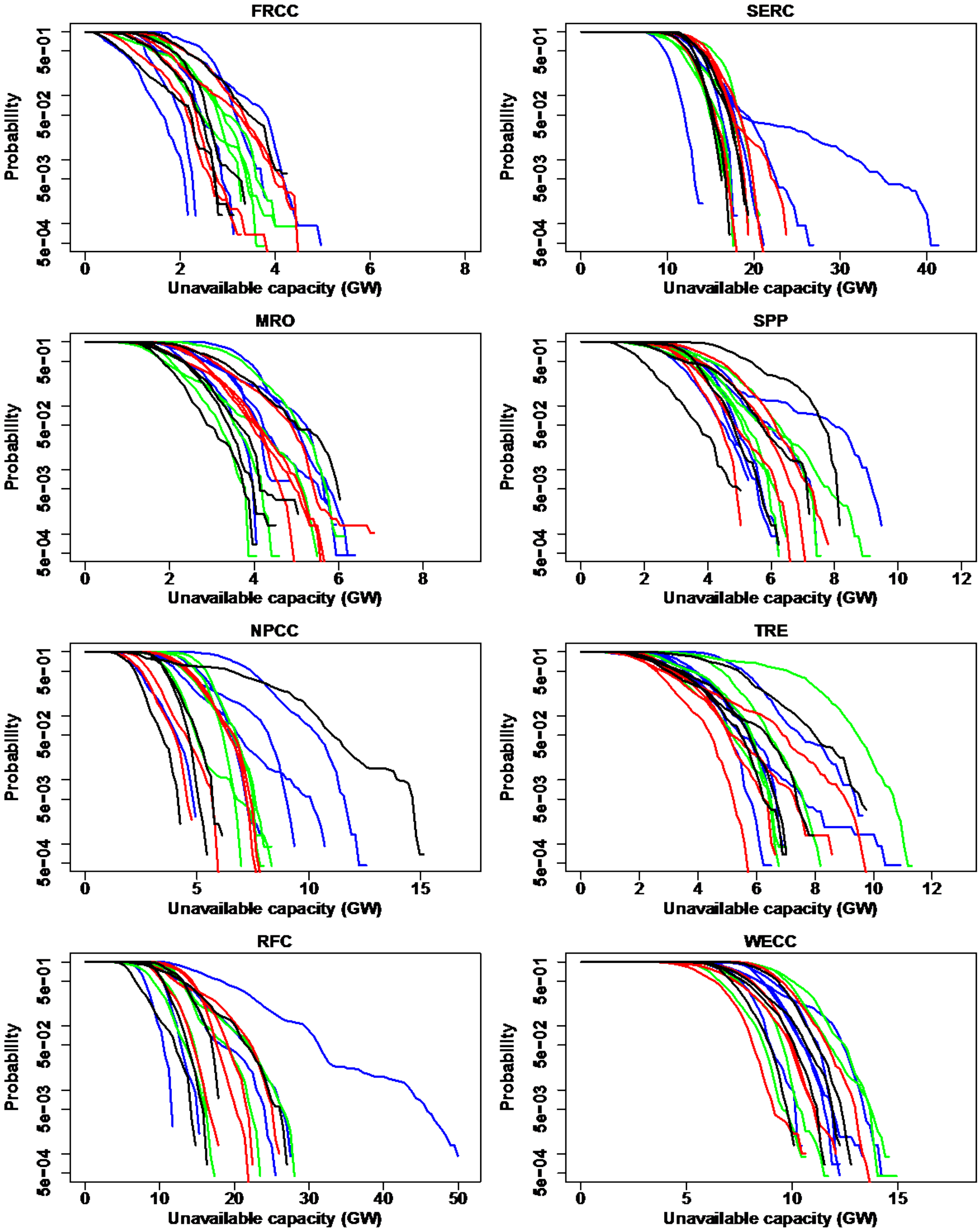


Figure S-23: Seasonal distributions of unscheduled unavailable capacity. Winter is shown in blue, spring in green, summer in red, and fall in black for each of the four years analyzed.

3.3 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 S-24). The residuals resemble white noise and appear to be homoskedastic. Autocorrelation functions of the residuals show no significant remaining structure (Figure S-25). We report AR(1) coefficients, standard errors, and t-statistics in Table S-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.

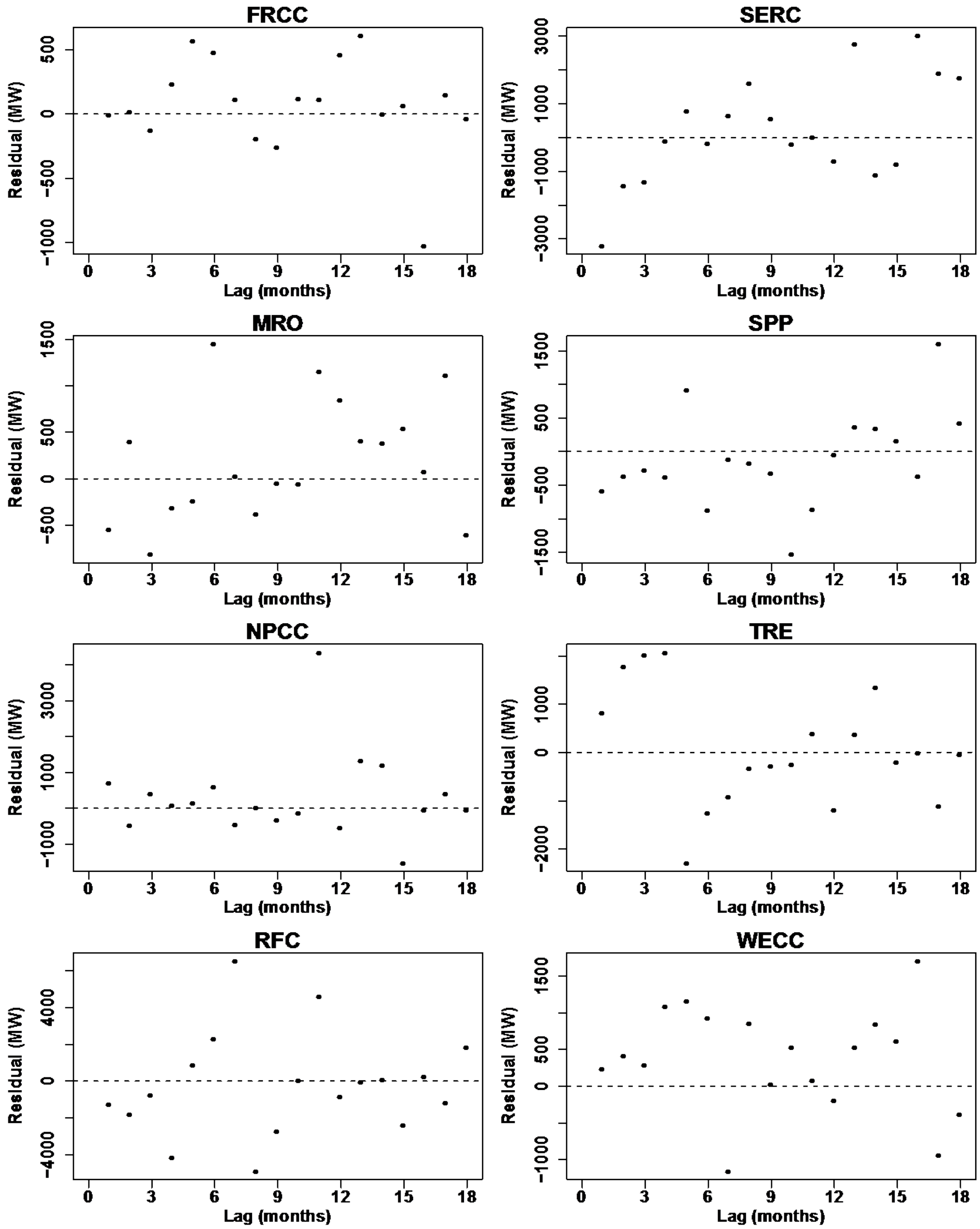


Figure S-24: AR(1) residuals for each NERC region.

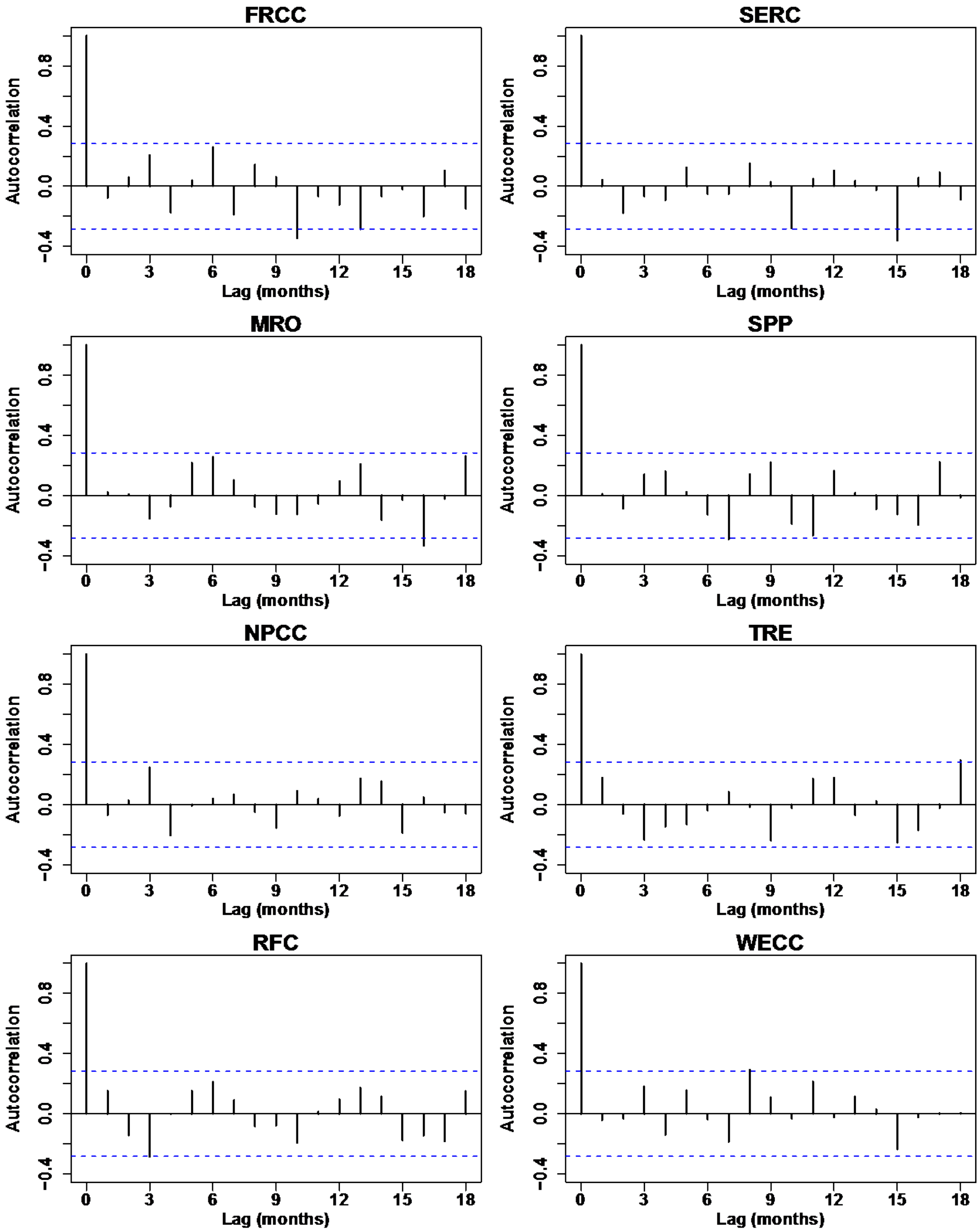


Figure S-25: Autocorrelation functions of the AR(1) residuals for each NERC region with 95% confidence interval for assessing statistical significance of lags (dashed blue lines).

Table S-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

3.4 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 4.3.1 in the main text and plot the results as exceedance curves for both the full study period (Figure S-26) and when excluding Hurricane Sandy and the Polar Vortex (Figure S-27). We find minimal benefits from using seasonal availability statistics for resource adequacy modeling.

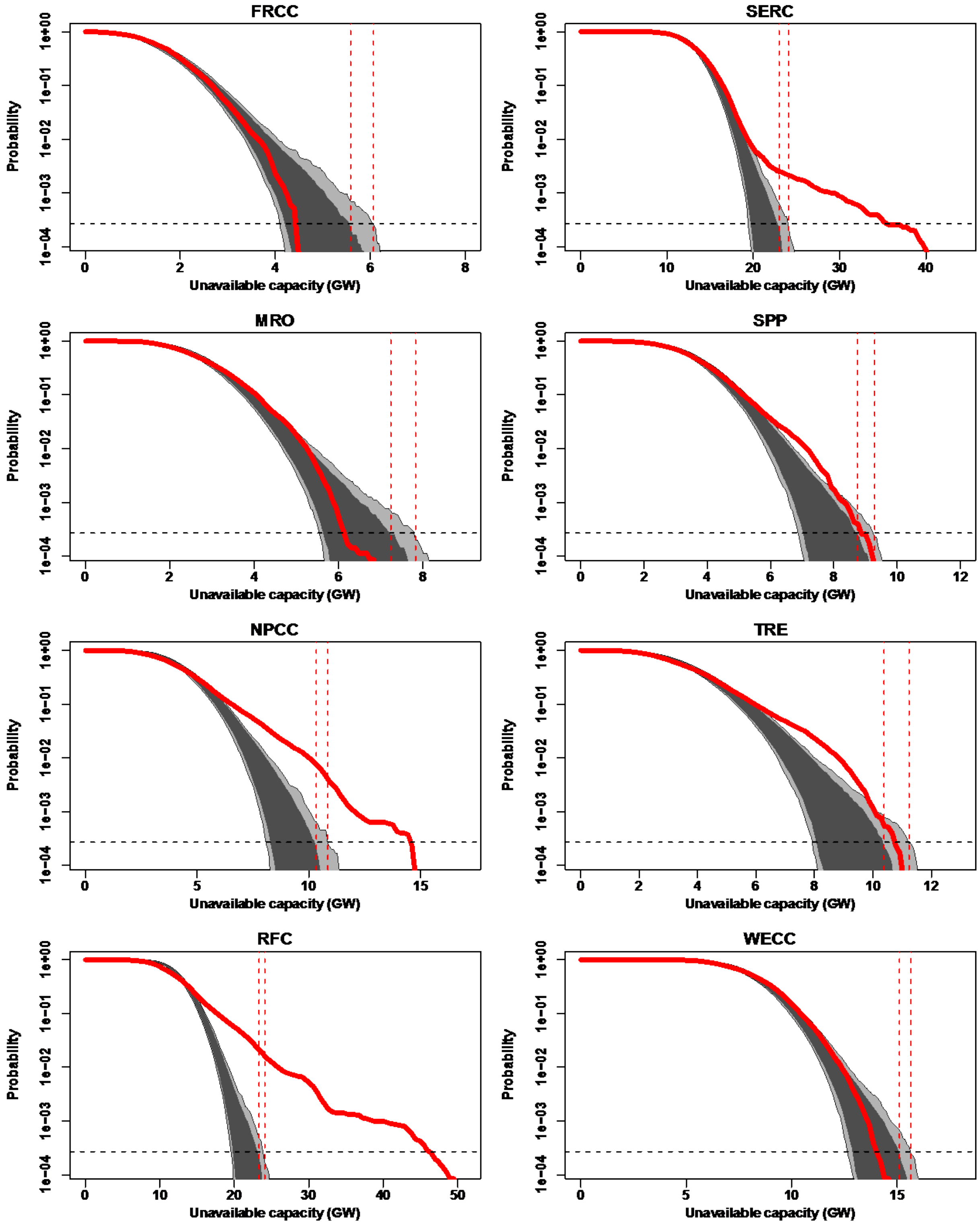


Figure S-26: 1000 seasonal block subsampling runs for the full period 2012-2015.

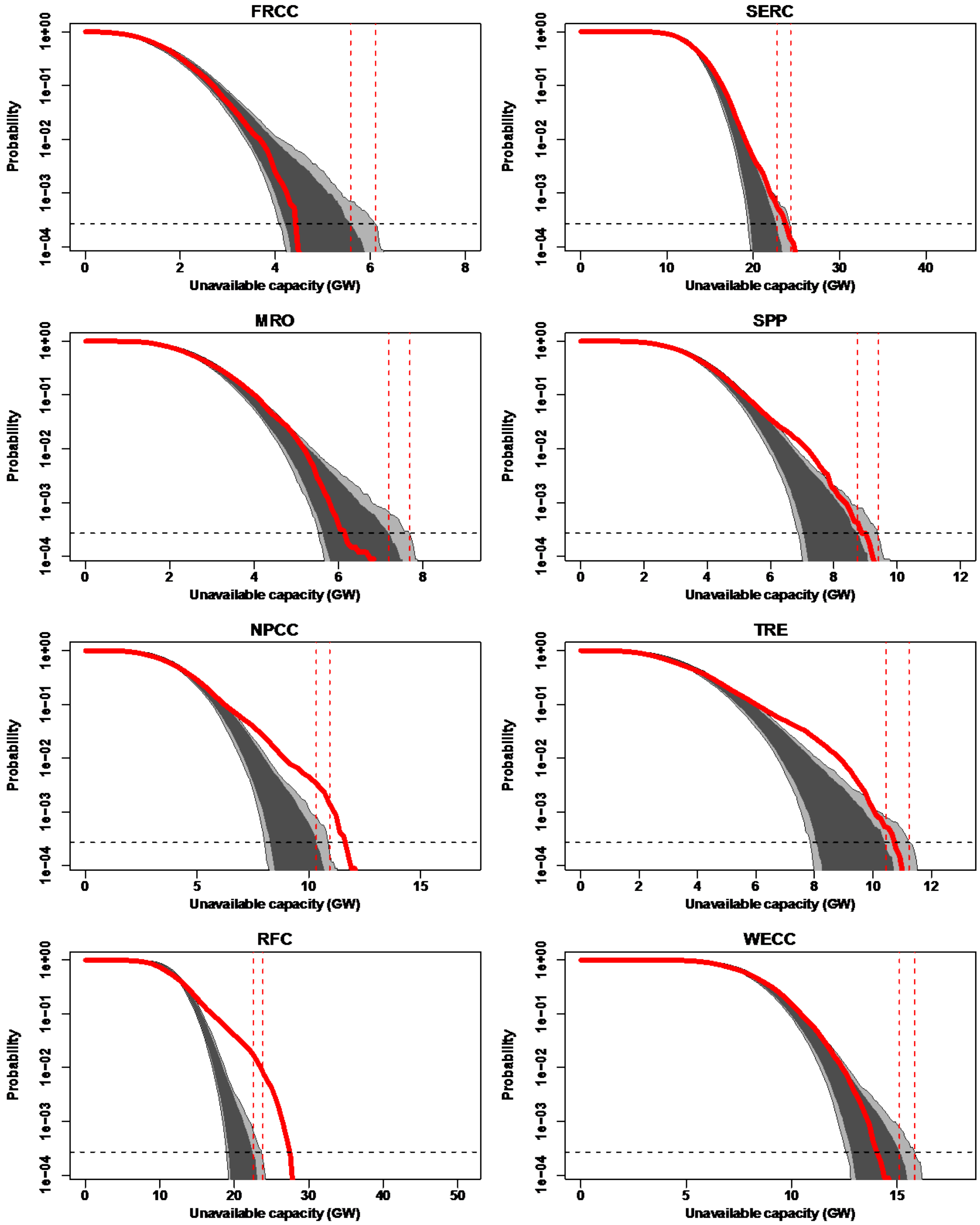


Figure S-27: 1000 seasonal block subsampling runs for the period 2012-2015 excluding Hurricane Sandy and January 2014.

4 Reliability applications

4.1 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 S-28) and with January 2014 and Hurricane Sandy removed (Figure S-29). We report the parameters of each fit in Table S-9.

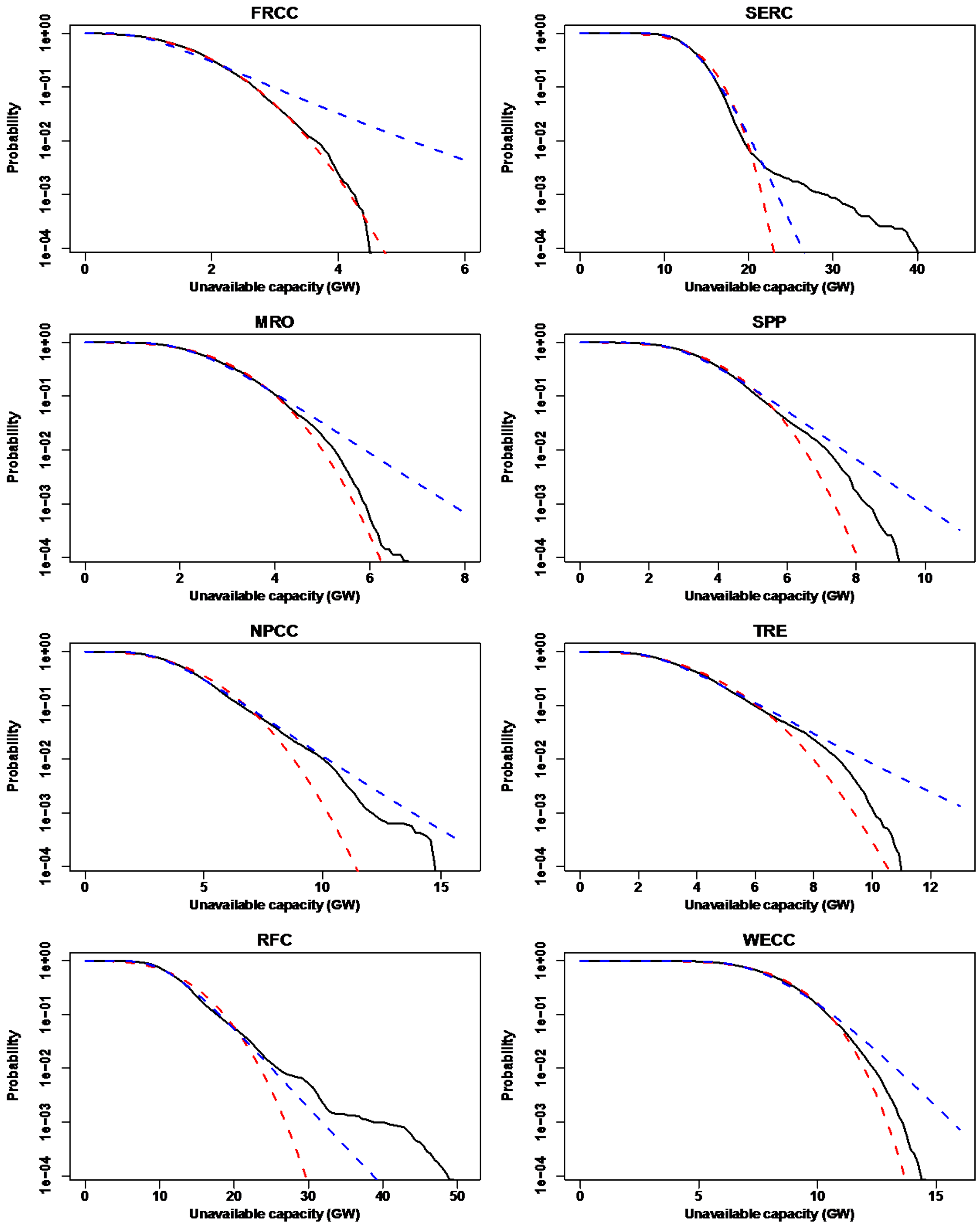


Figure S-28: Overlay of Weibull (red) and lognormal (blue) fits to observed exceedance curves (black) for the 2012-2015 period. Weibull and lognormal parameters are given in Table S-9.

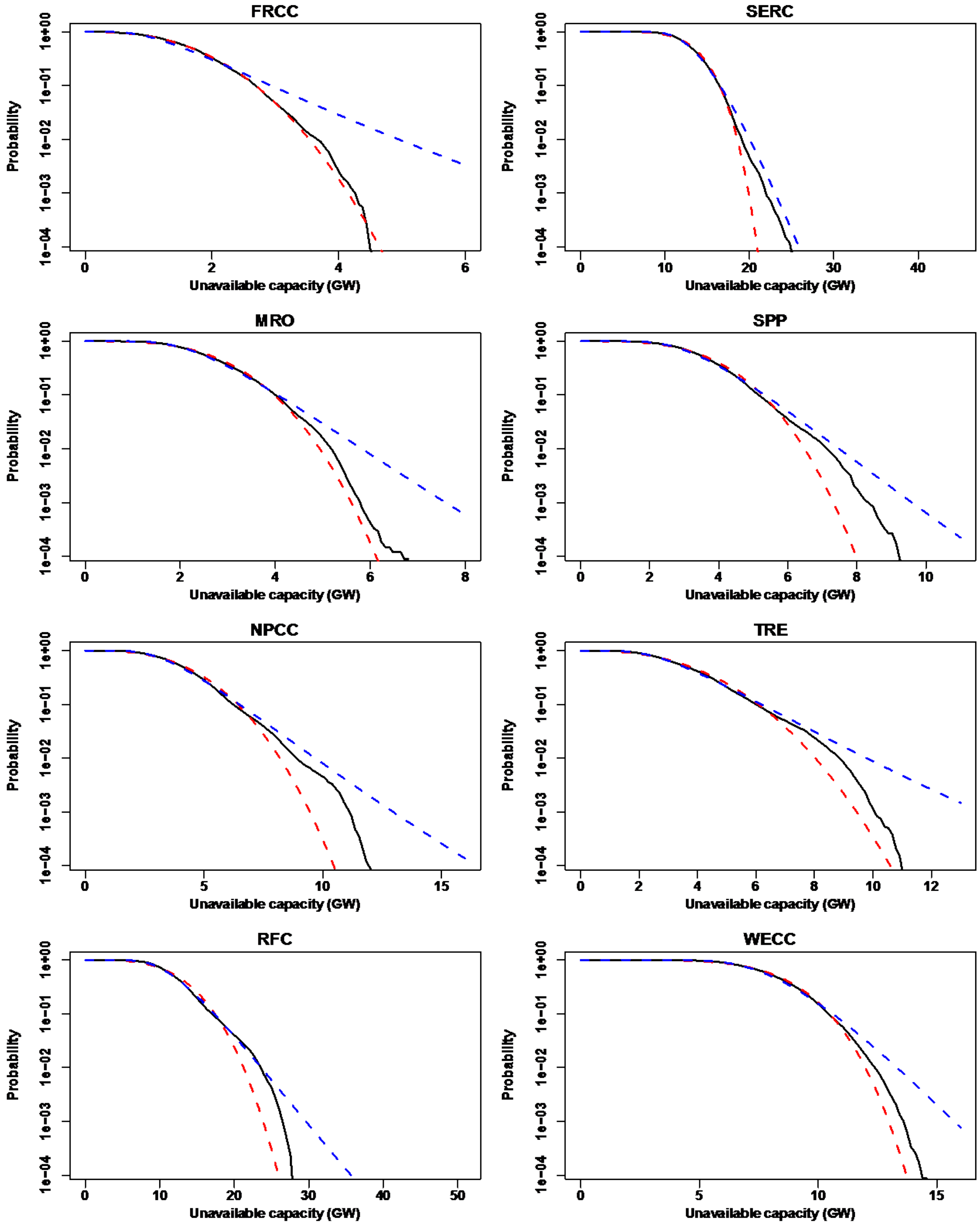


Figure S-29: Overlay of Weibull (red) and lognormal (blue) fits to observed exceedance curves (black), removing Hurricane Sandy and January 2014. Weibull and lognormal parameters are given in Table S-9.

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

	Full period				*Removing Hurricane 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)

4.2 Parametric fits to distributions of normalized derating magnitudes

We fit Weibull distributions to each unit type's normalized derating magnitudes (Figure S-30). We report the parameters of each fit in Table S-10. This information can be used in conjunction with the other results presented in Section 4.4 of the main text in Markov modeling of power systems.

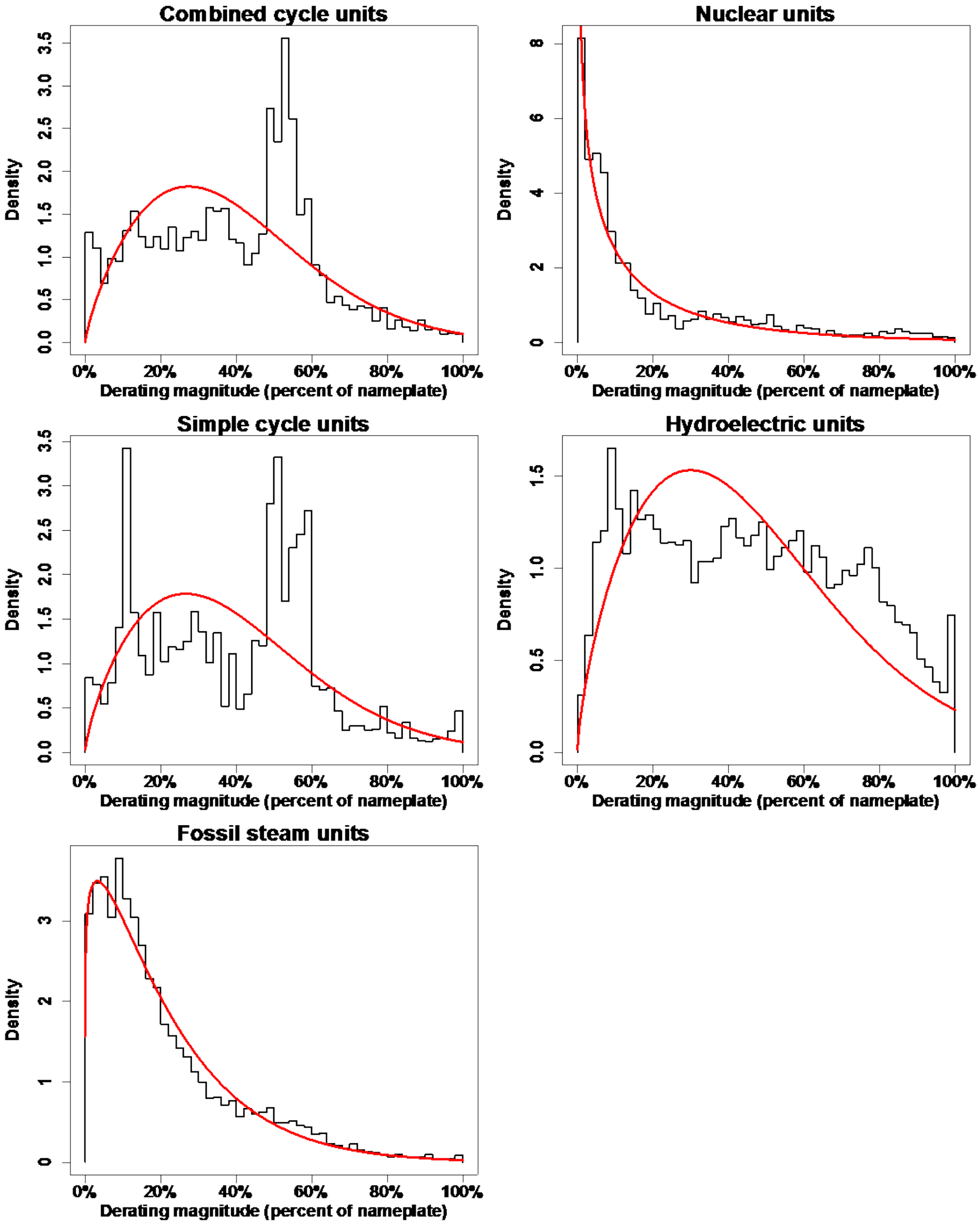


Figure S-30: Distribution of derating magnitudes as a percentage of nameplate capacity, with fitted Weibull distributions.

Table S-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)

4.3 Mean time between failure and mean time to recovery

4.3.1 Supplementary MTBF results

The parameters for Weibull and gamma distributions fit to each unit type's capacity-weighted MTBF values excluding reserve shutdown (RS) hours are given in Tables S-11 and S-12. We do not report parameters at the region-by-unit-type level due to small sample sizes in several instances.

Table S-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	1,060	0.78	440	1.08	420	1.13	3,120	1.34	2,880
(0.0016)	(2.12)	(0.0015)	(1.71)	(0.0012)	(0.65)	(0.0029)	(10.35)	(0.0029)	(7.03)

Table S-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	680	0.75	710	1.30	310	1.34	2,220	1.94	1,350
(0.004)	(2.19)	(0.0026)	(3.41)	(0.0026)	(0.76)	(0.006)	(12.04)	(0.0079)	(6.20)

Histograms of the number of between-failure periods used to calculate each unit's MTBF are shown in Figure S-31 through Figure S-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.

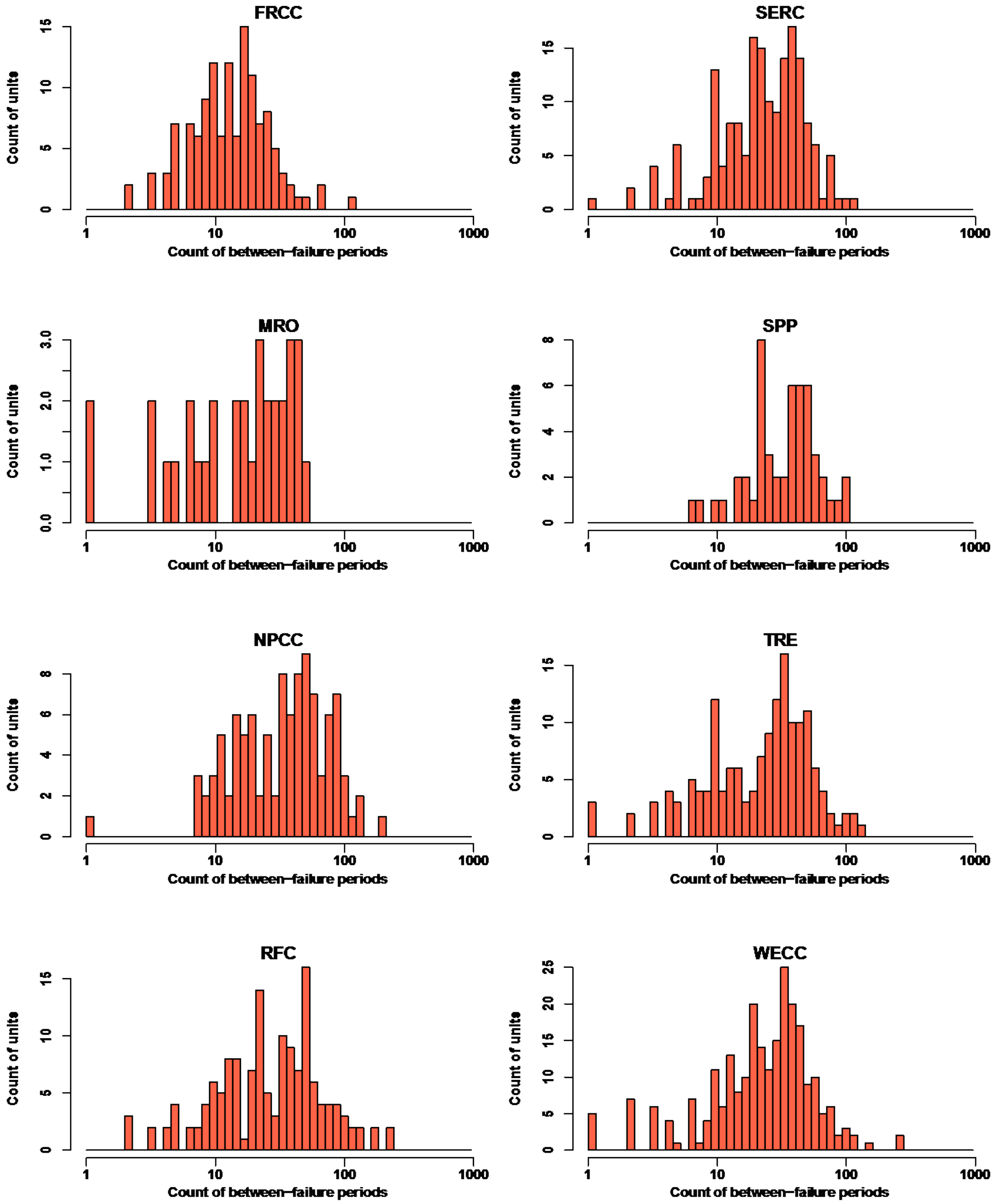


Figure S-31: Count of between-failure periods used to calculate mean time between failure for combined cycle units. Units with significant reserve shutdown reporting discrepancies are excluded.

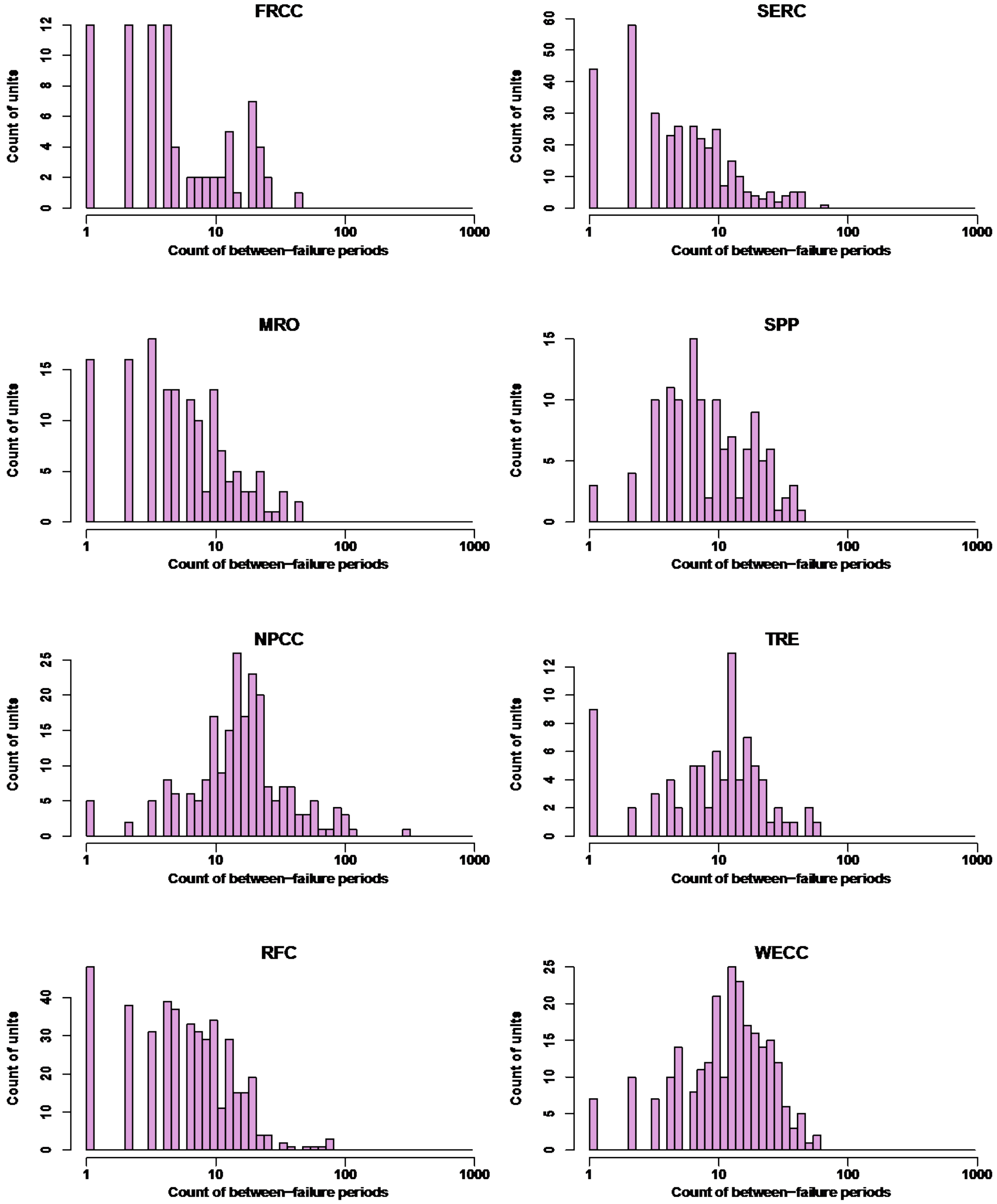


Figure S-32: Count of between-failure periods used to calculate mean time between failure for simple cycle units. Units with significant reserve shutdown reporting discrepancies are excluded.

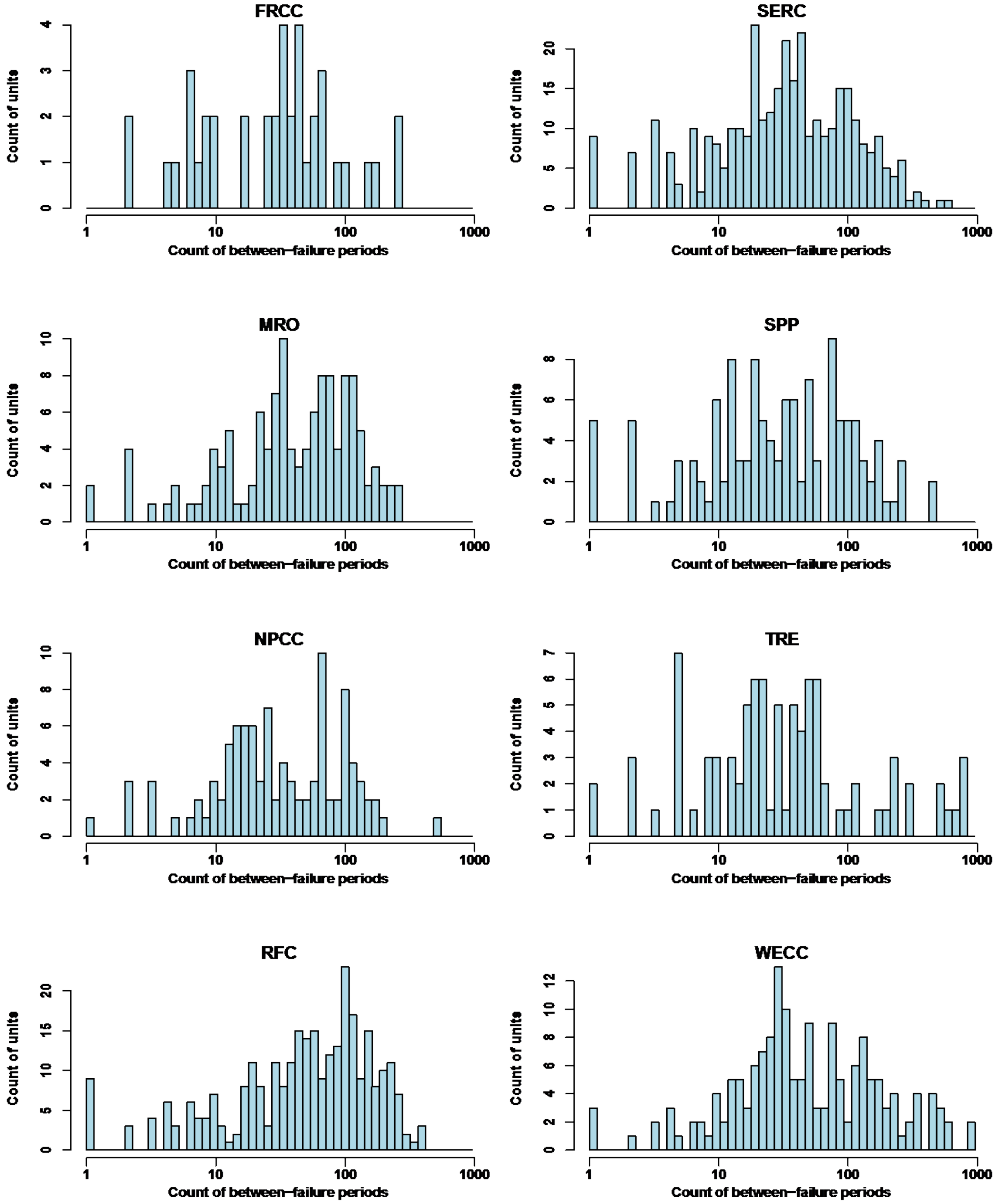


Figure S-33: Count of between-failure periods used to calculate mean time between failure for fossil steam and fluidized bed units. Units with significant reserve shutdown reporting discrepancies are excluded.

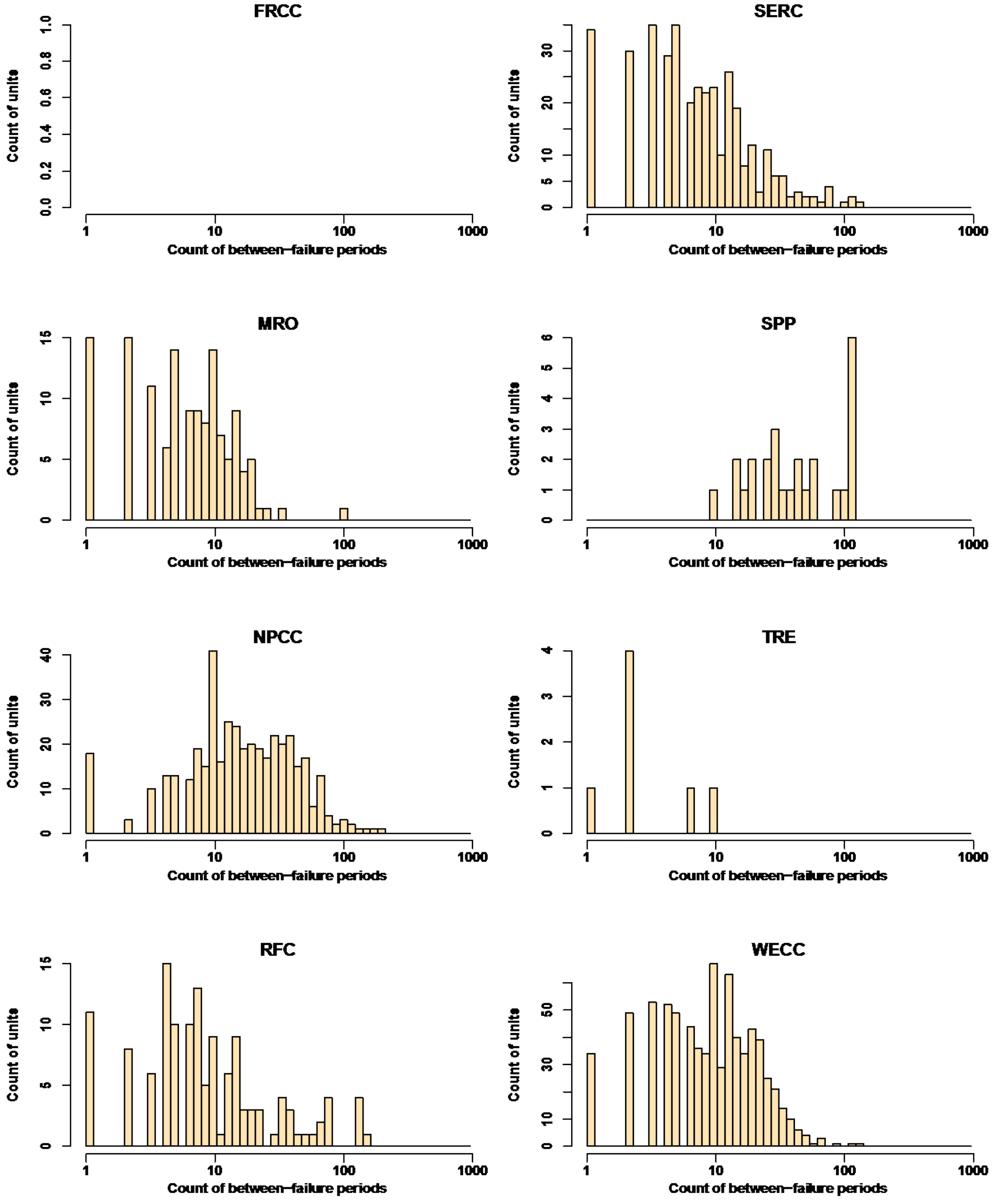


Figure S-34: Count of between-failure periods used to calculate mean time between failure for hydroelectric units. FRCC has no such units. Units with significant reserve shutdown reporting discrepancies are excluded.

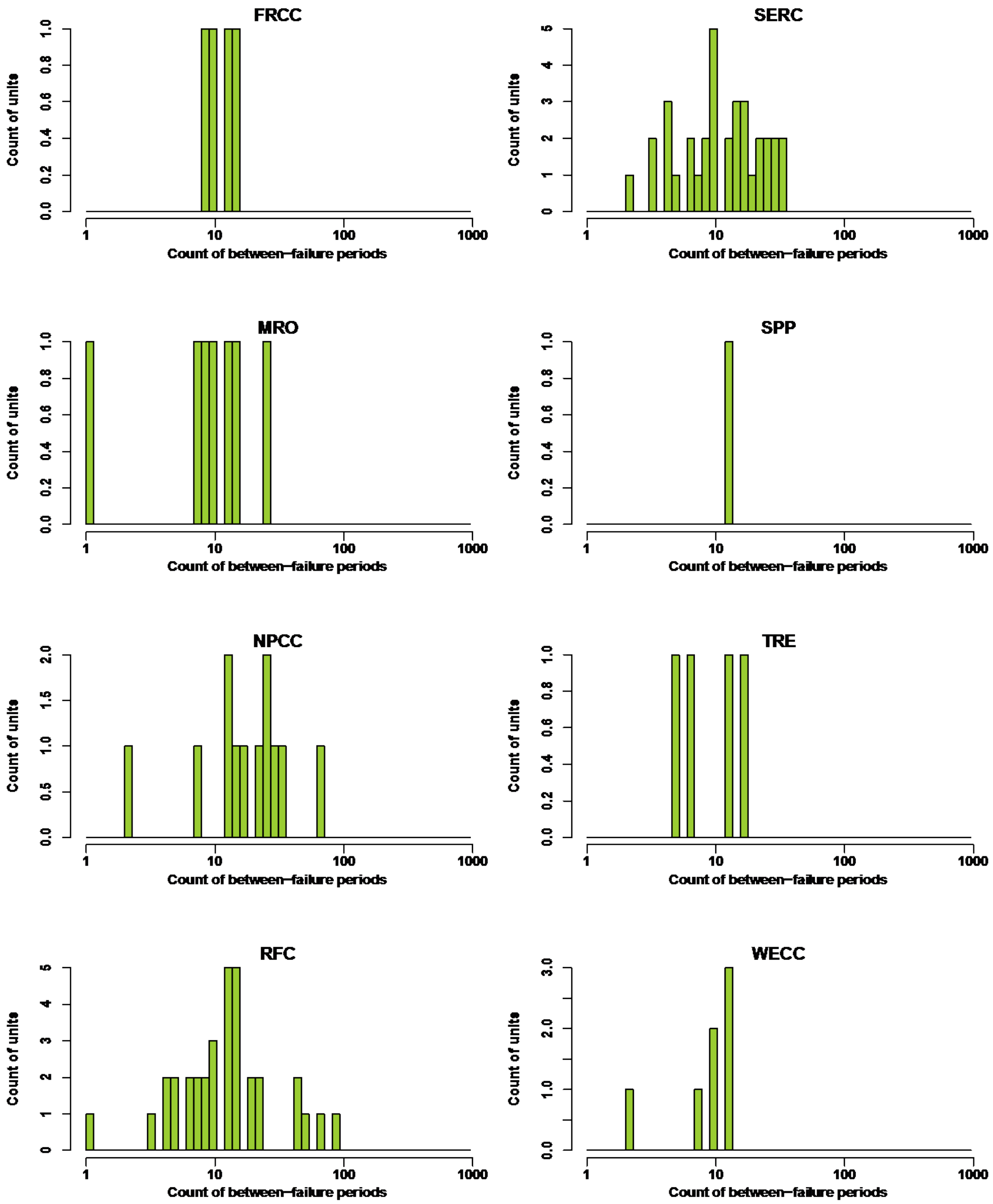


Figure S-35: Count of between-failure periods used to calculate mean time between failure for nuclear units. Units with significant reserve shutdown reporting discrepancies are excluded.

4.3.2 MTBF calculation information

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 S-36). We conclude that the calculation of MTBF is not very sensitive to the selection of the minimum percent of unit's nameplate capacity that must be unavailable to constitute a "failure".

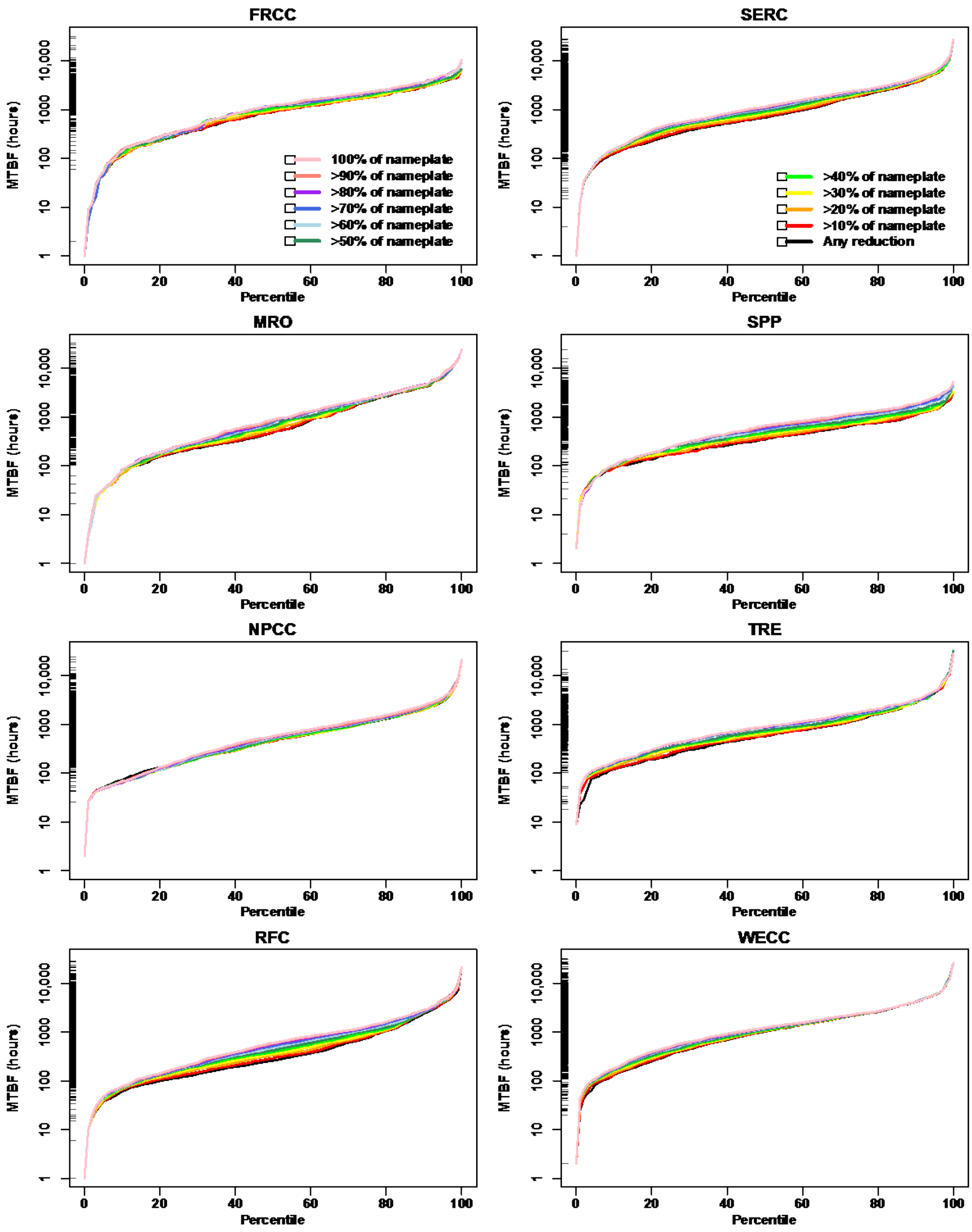


Figure S-36: Sensitivity of MTBF to the minimum percent of unit's nameplate capacity that must be unavailable to constitute a "failure". The legend has been split between the top two panels for readability. The rug along the y-axis indicates each region's MTBF values when a failure is defined as any reduction in availability. Values have been calculated with all reserve shutdown hours removed. Units with significant reserve shutdown reporting discrepancies are excluded (see Table S-14).

4.3.3 Supplementary MTTR results

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 S-37 through S-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 S-42 through Figure S-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 S-15 and S-16. We do not report parameters at the region-by-unit-type level due to small sample sizes in several instances.

Table S-13: Parameters for Weibull fits to capacity-weighted MTTR distributions 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	45	0.53	95	0.60	106	0.47	77	0.66	386
(0.00085)	(0.14)	(0.00090)	(0.48)	(0.00056)	(0.28)	(0.00092)	(0.50)	(0.0013)	(1.90)

Table S-14: Parameters for gamma fits to capacity-weighted MTTR distributions 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	125	0.37	644	0.44	528	0.30	914	0.55	1,071
(0.0014)	(0.45)	(0.0010)	(3.21)	(0.00076)	(1.51)	(0.00098)	(5.62)	(0.0020)	(5.84)

4.3.4 MTTR calculation information

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.

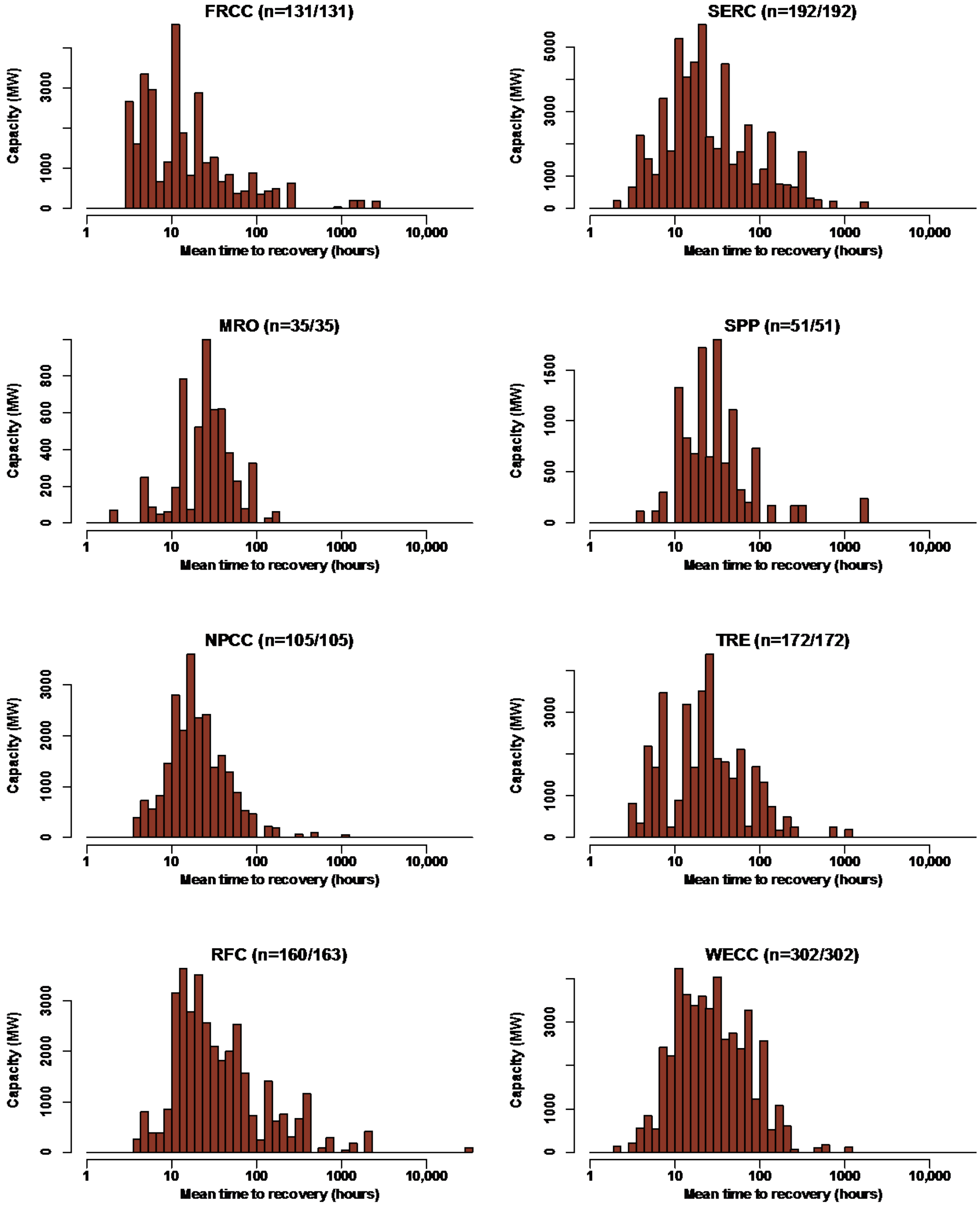


Figure S-37: Capacity-weighted mean time to recovery (MTTR) values for combined cycle gas units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units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

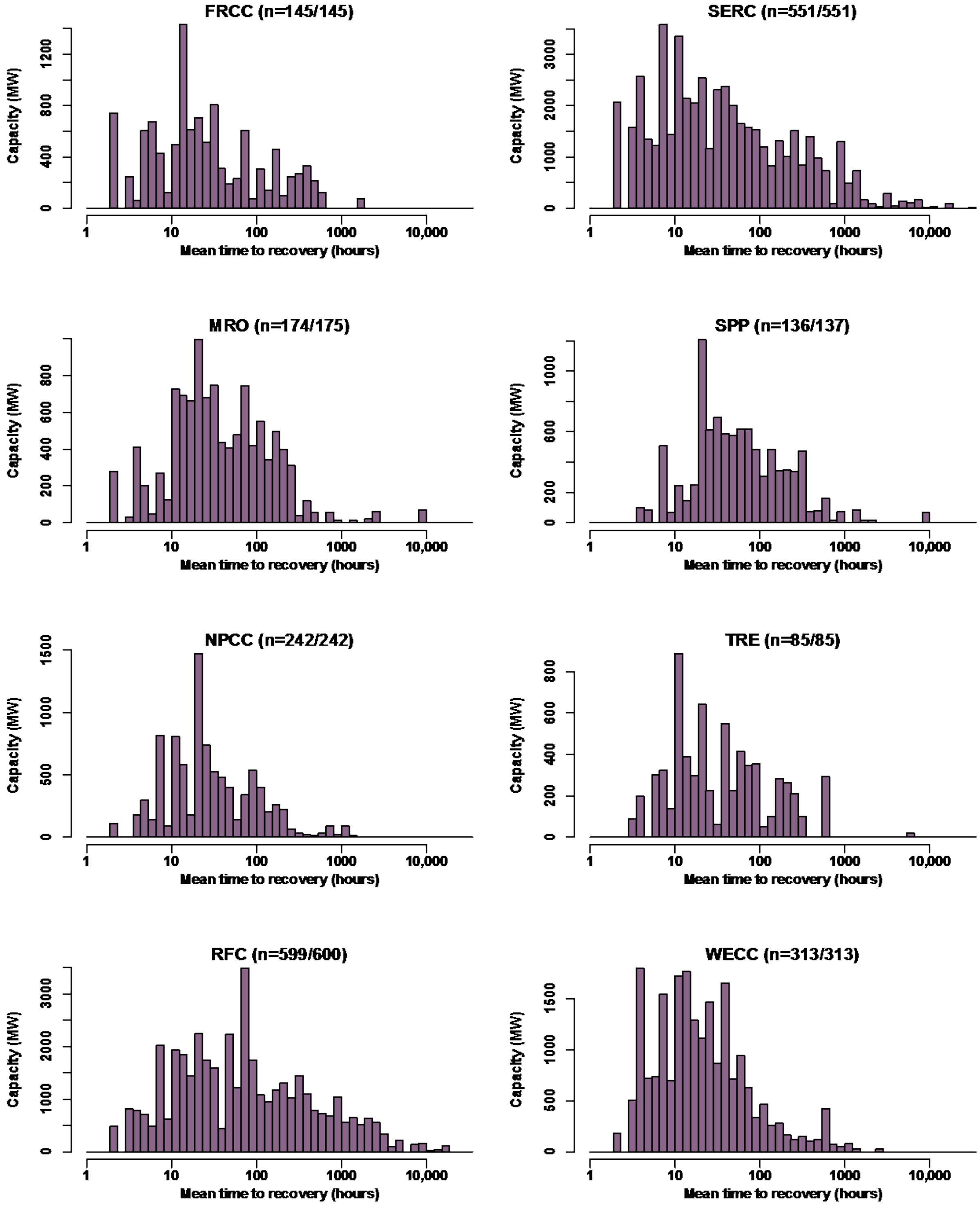


Figure S-38: Capacity-weighted mean time to recovery (MTTR) values for simple cycle gas units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units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

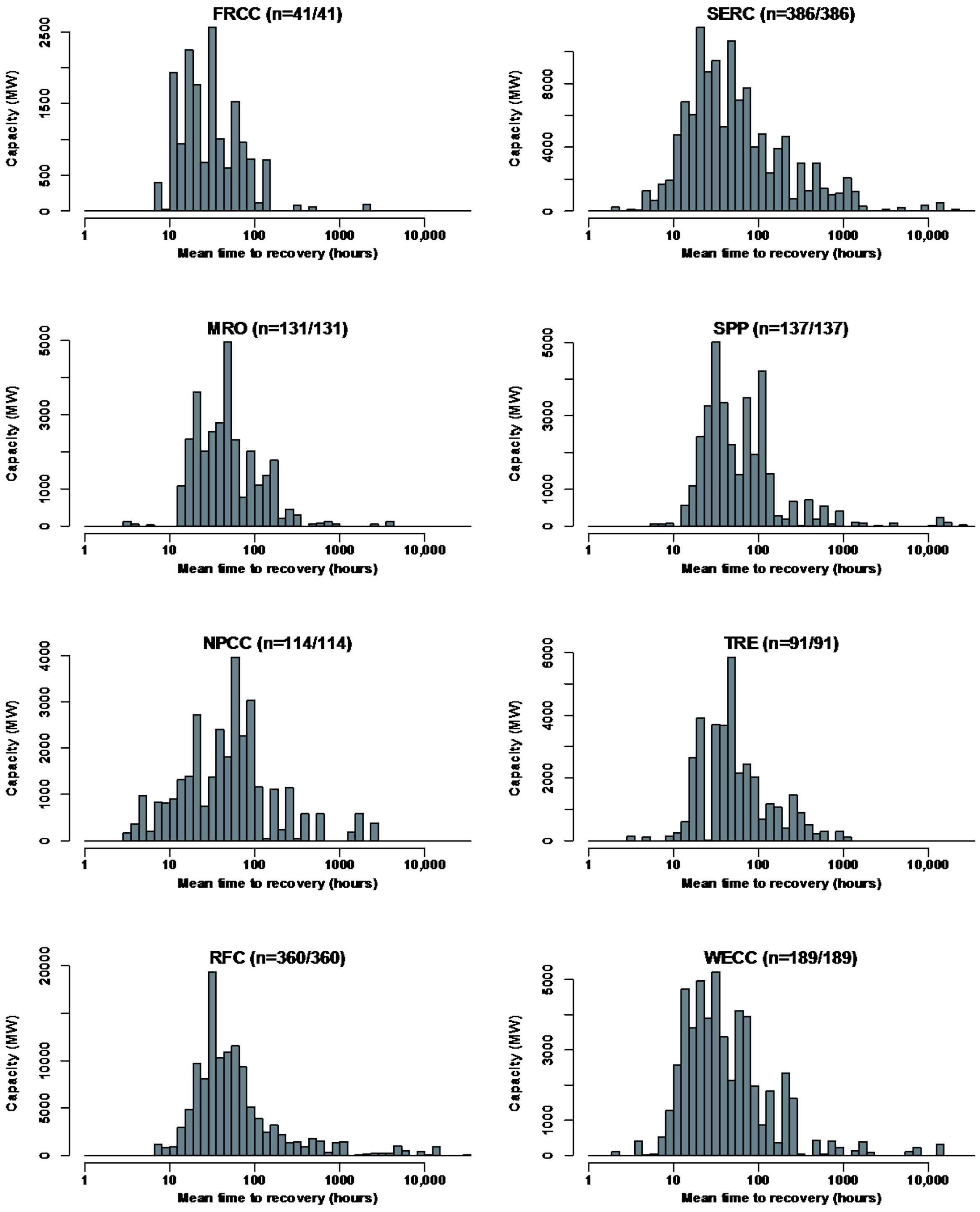


Figure S-39: Capacity-weighted mean time to recovery (MTTR) values for fossil steam and fluidized bed units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the 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).

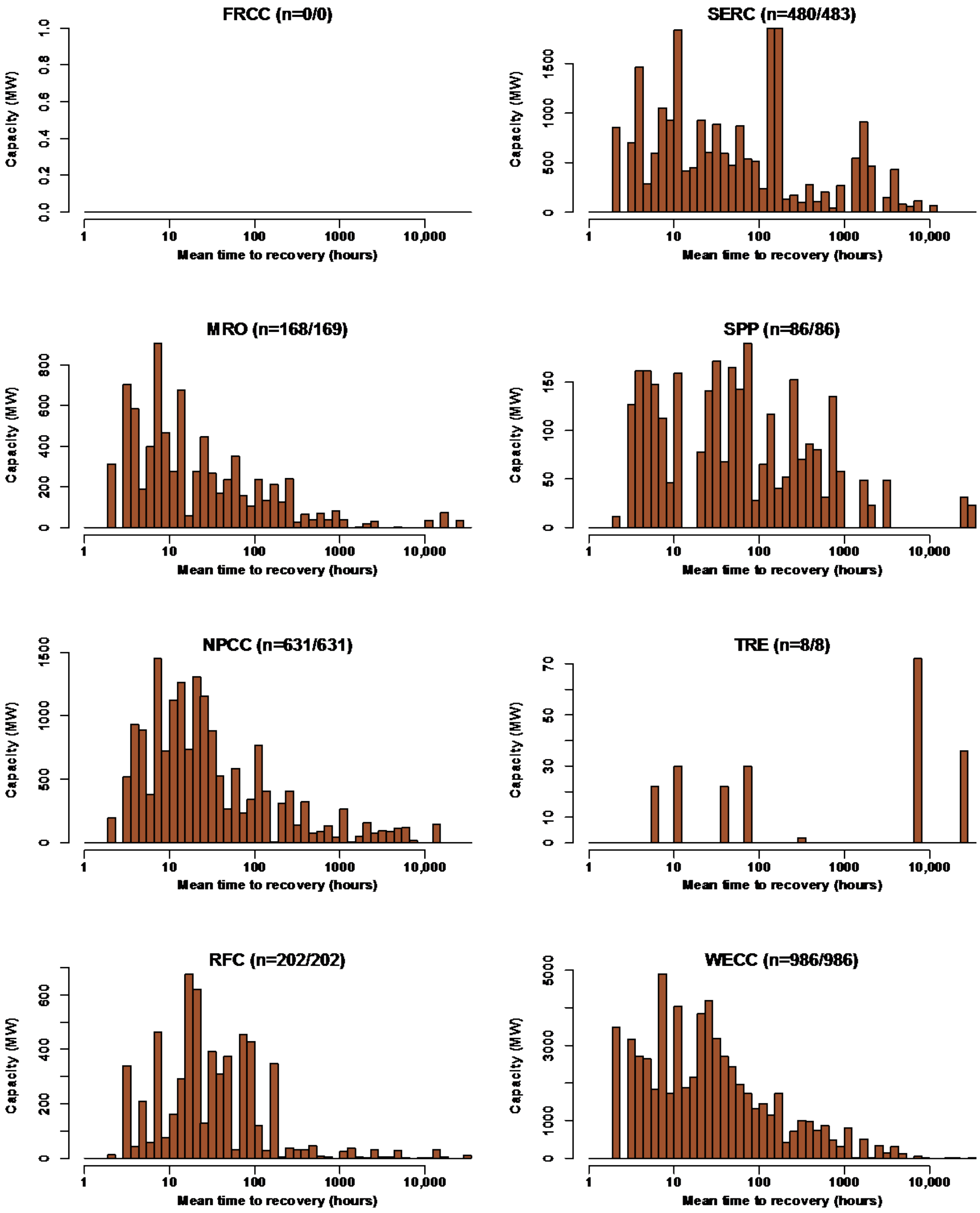


Figure S-40: Capacity weighted mean time to recovery (MTTR) values for hydroelectric units. FRCC has no such units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates 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).

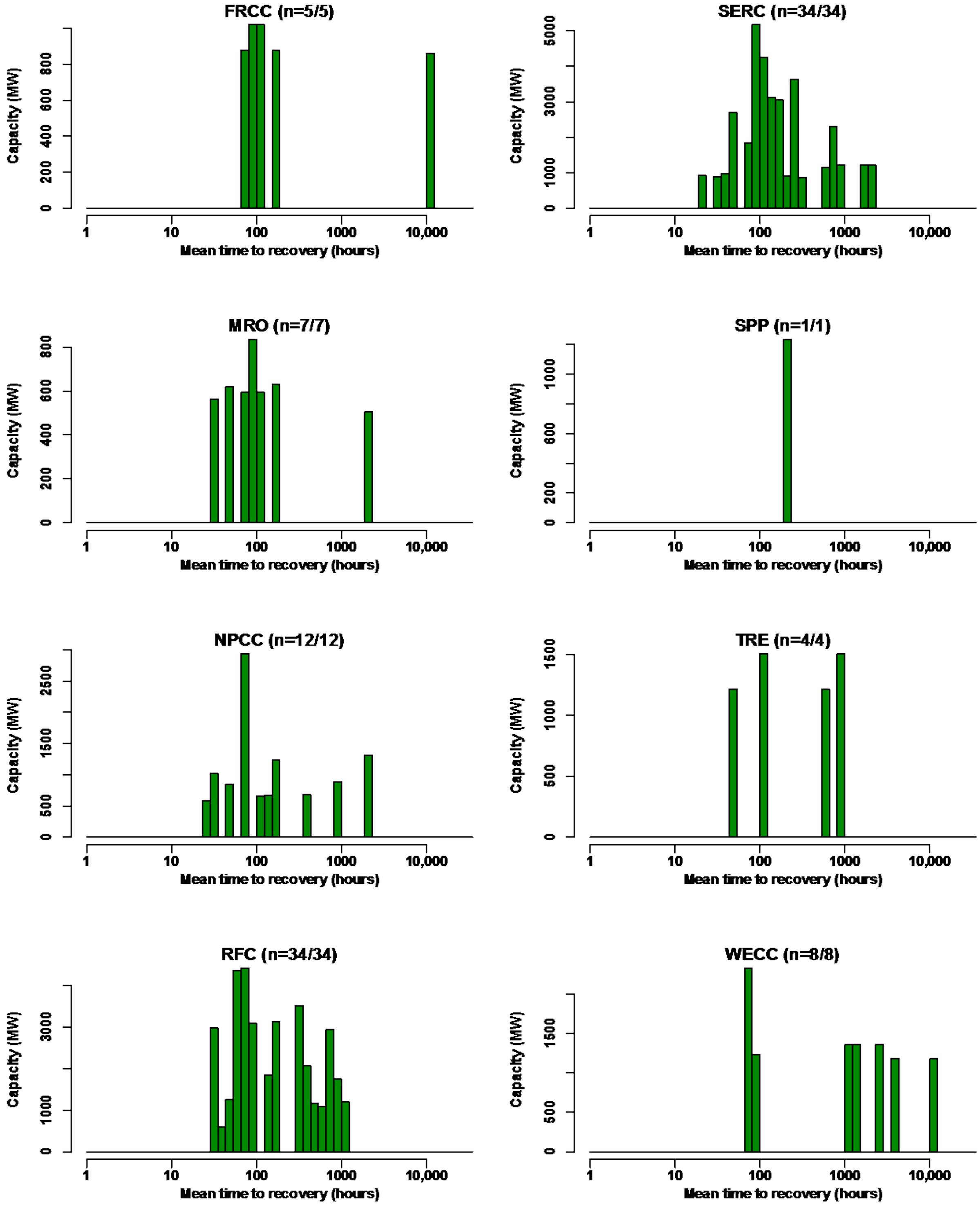


Figure S-41: Capacity-weighted mean time to recovery (MTTR) values for nuclear units. Note the log scale for MTTR. In the parenthetical notation after the region name, the numerator indicates the count of units for which an MTTR could be calculated. The denominator indicates the count of units experiencing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

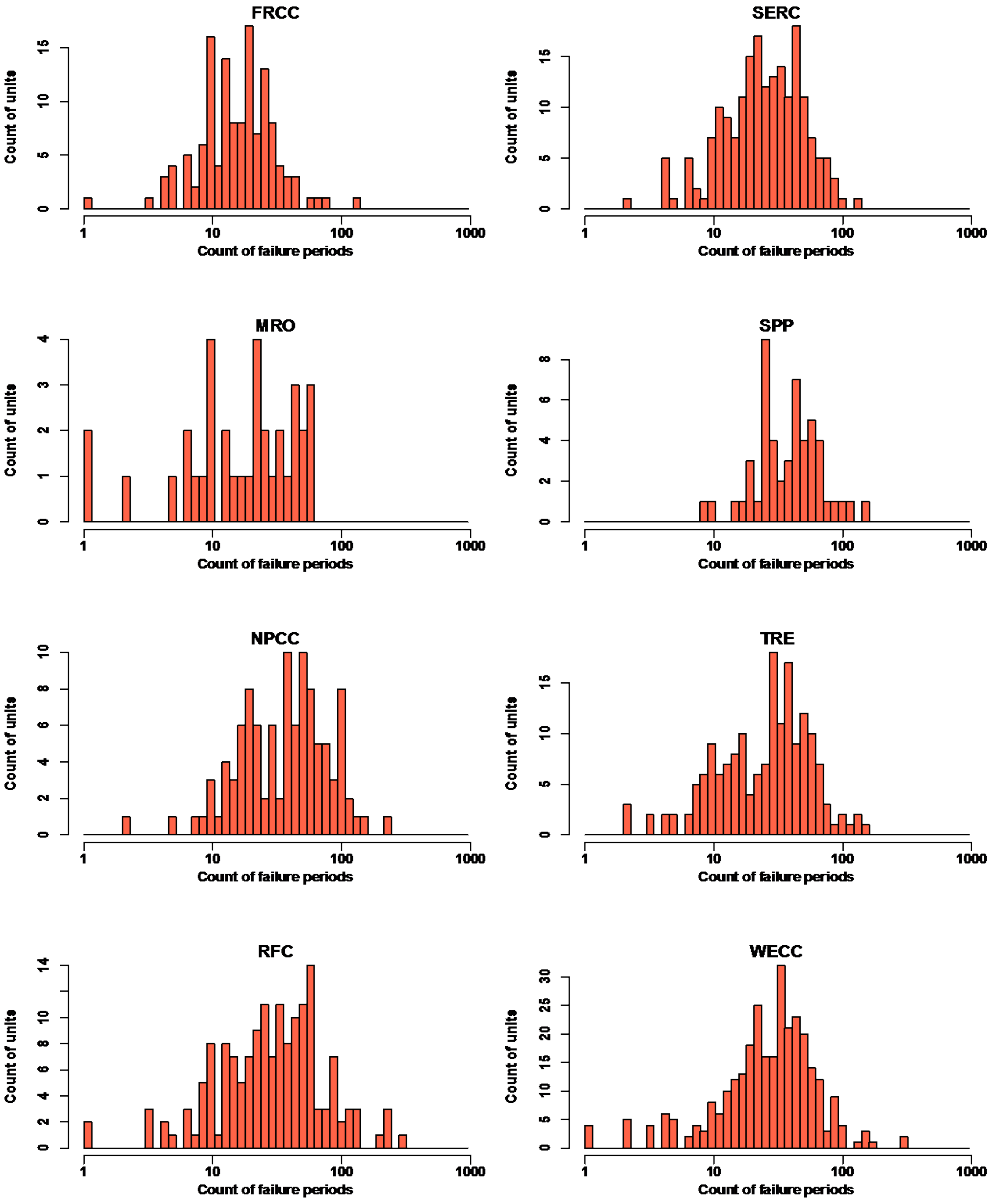


Figure S-42: Count of failure periods used to calculate mean time to recovery for combined cycle units.

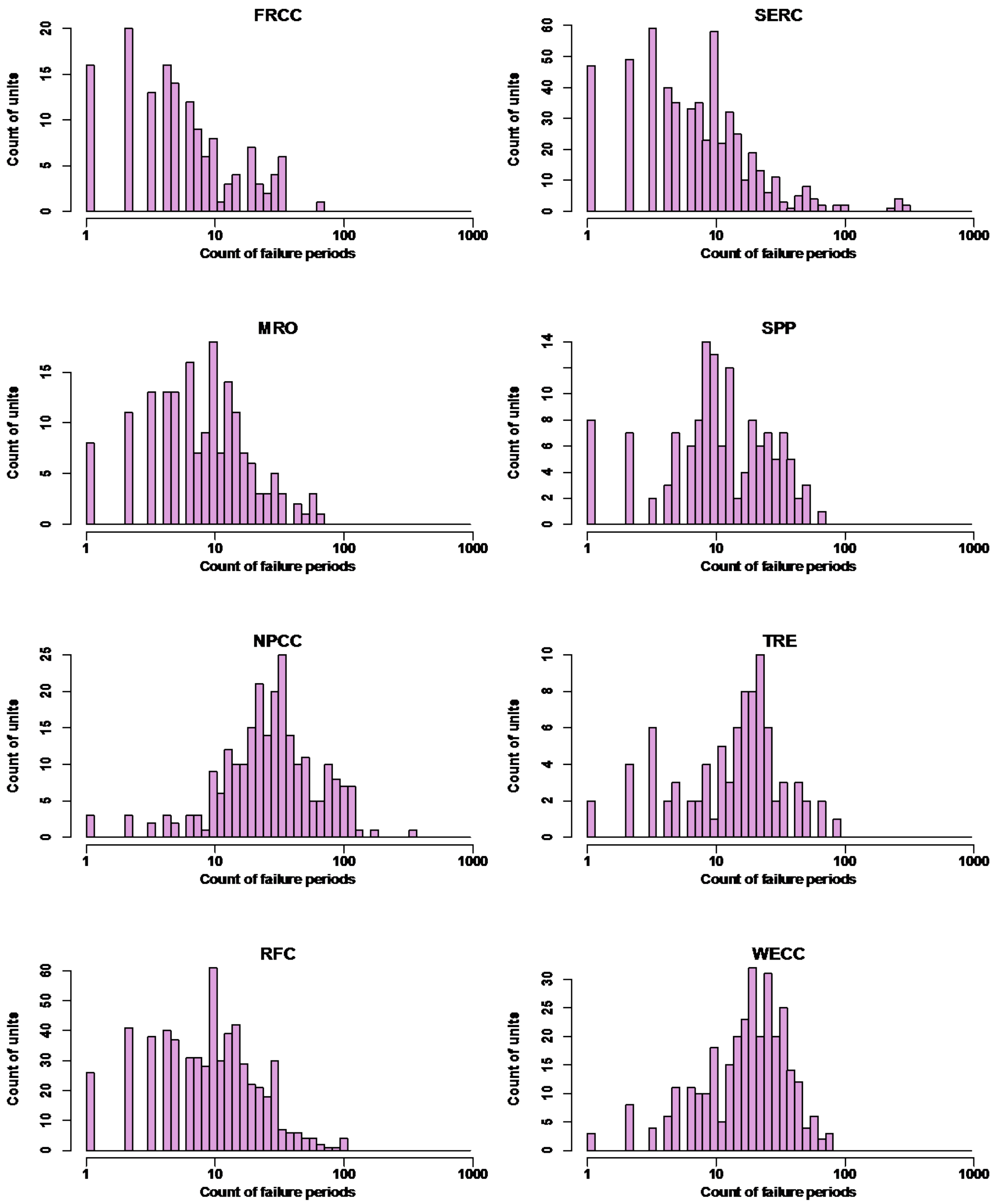


Figure S-43: Count of failure periods used to calculate mean time to recovery for simple cycle units.

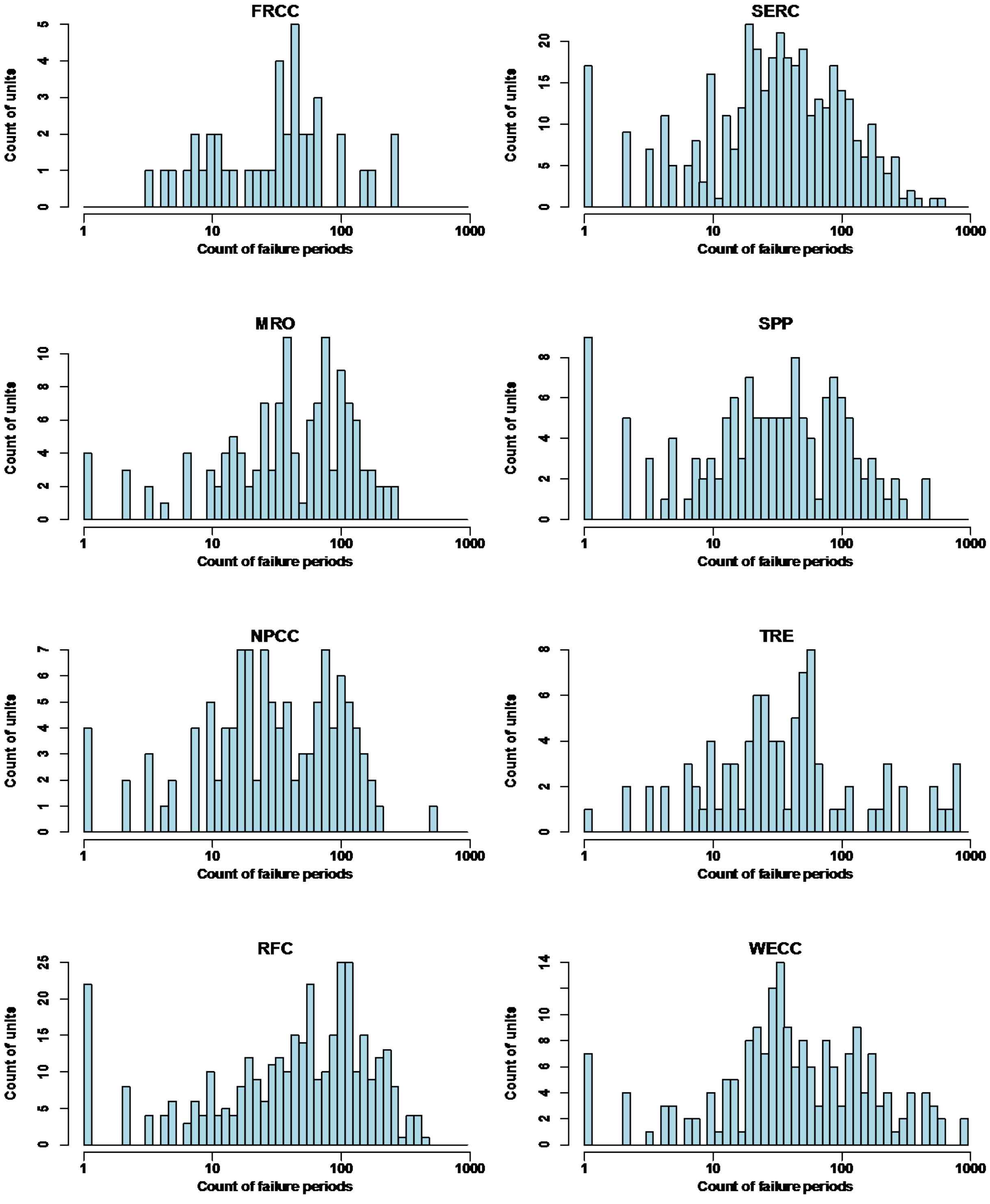


Figure S-44: Count of failure periods used to calculate mean time to recovery for fossil steam and fluidized bed units.

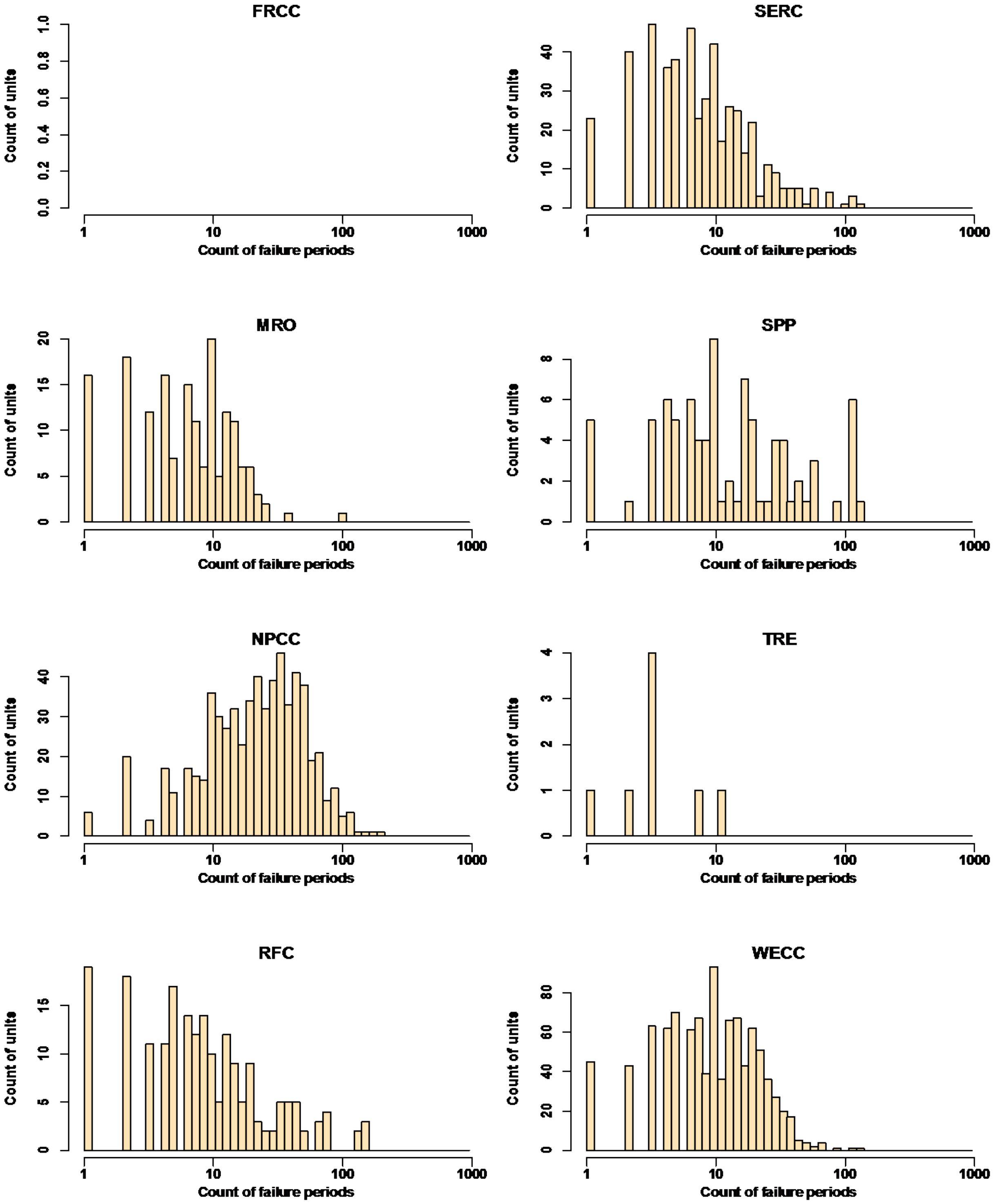


Figure S-45: Count of failure periods used to calculate mean time to recovery for hydroelectric units. FRCC has no such units.

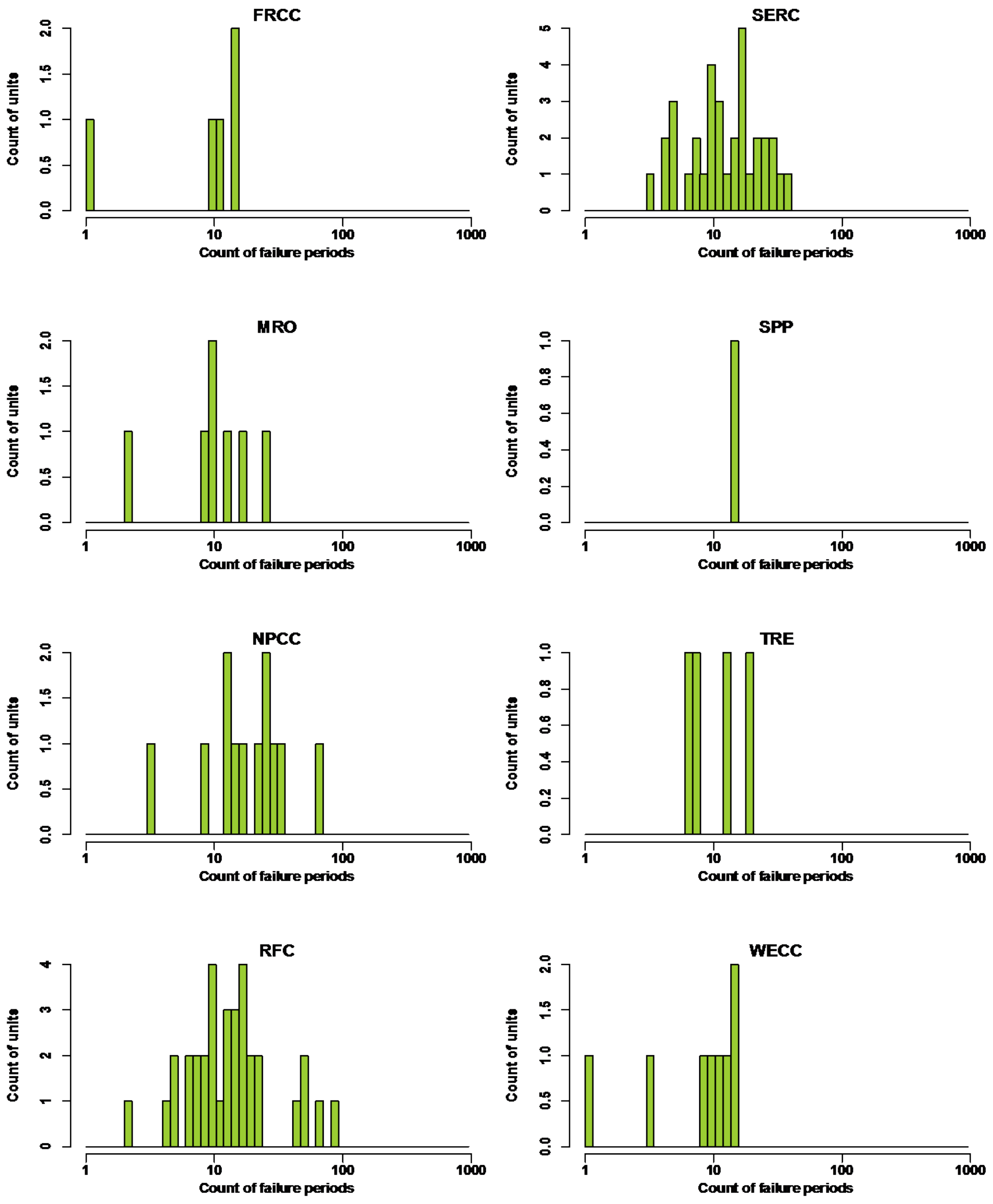


Figure S-46: Count of failure periods used to calculate mean time to recovery for nuclear units.

4.3.5 Do the MTBFs of small and large units differ?

In section 4.4.6 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 6 in the main text. Here we present the two-sample Kolmogorov-Smirnov test results (Table S-13). Results are highly consistent with the Mann-Whitney U test results (Table 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 S-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.

	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 S-47 through Figure S-51. We again exclude units with significant discrepancies in RS reporting between the GADS Events and GADS Performance tables.

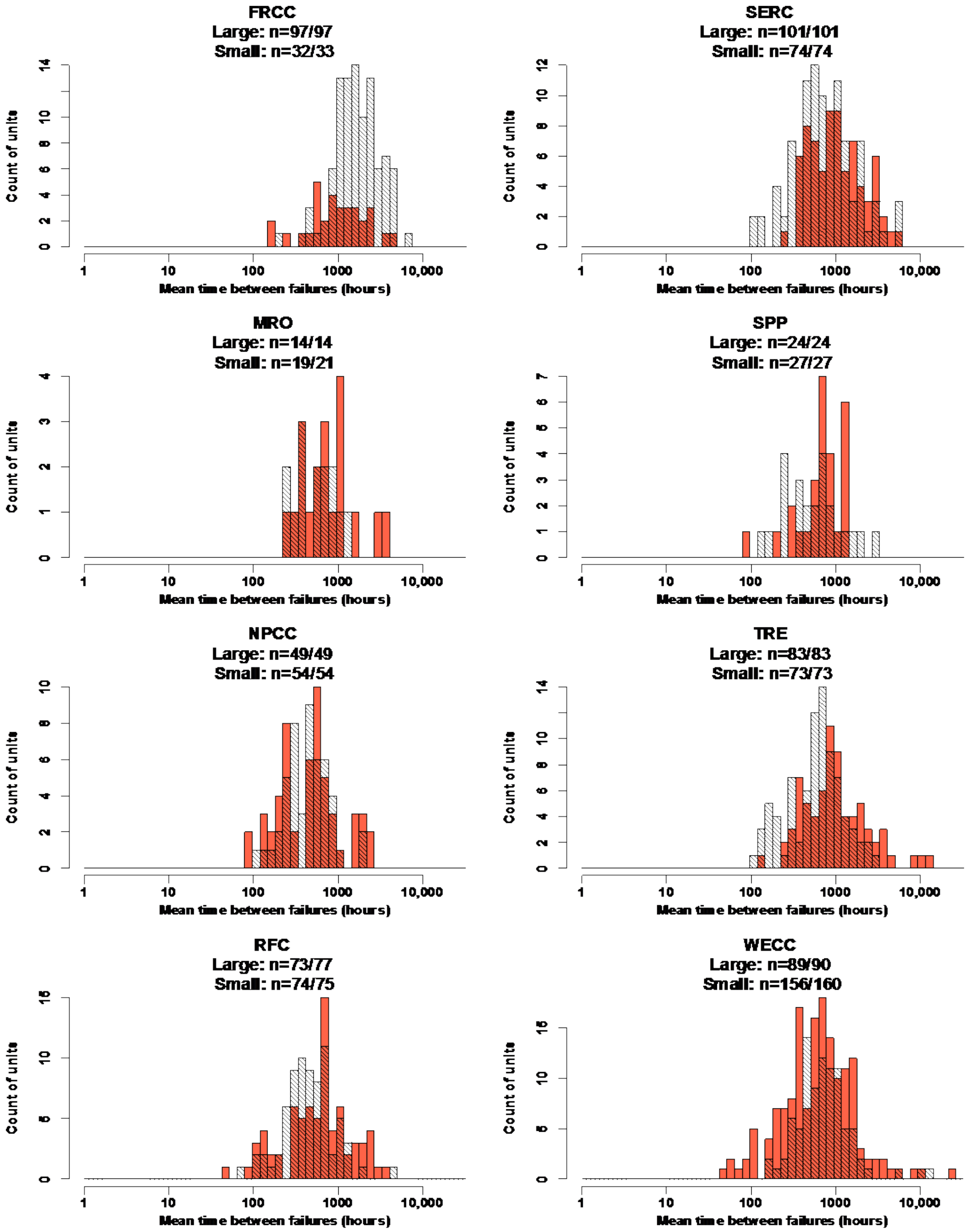


Figure S-47: Mean time between failure values for small (orange) versus large (black) combined cycle gas units; threshold is 185 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled 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 S-14).

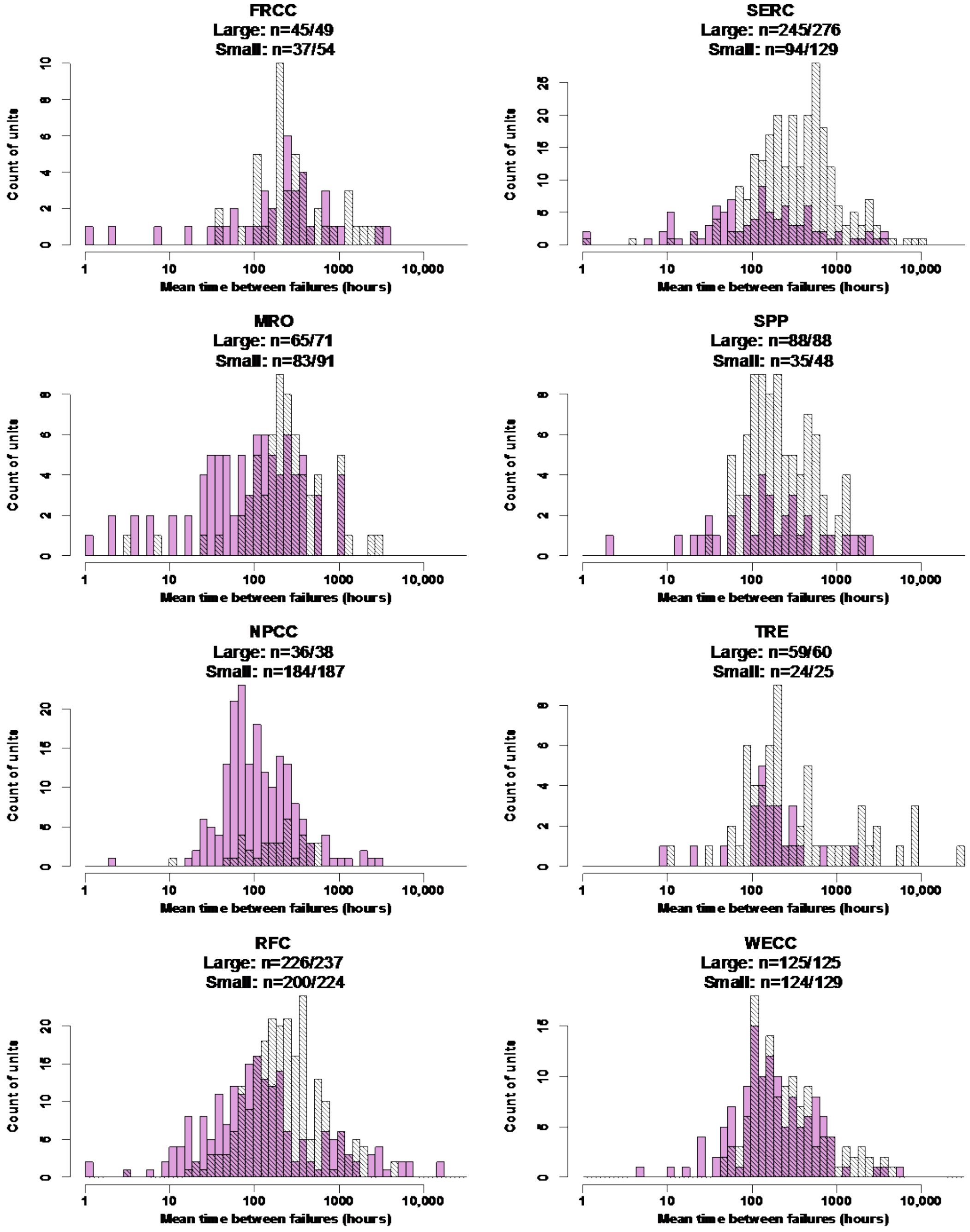


Figure S-48: Mean time between failure values for small (purple) versus large (black) simple cycle gas units; threshold is 57 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled 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 S-14).

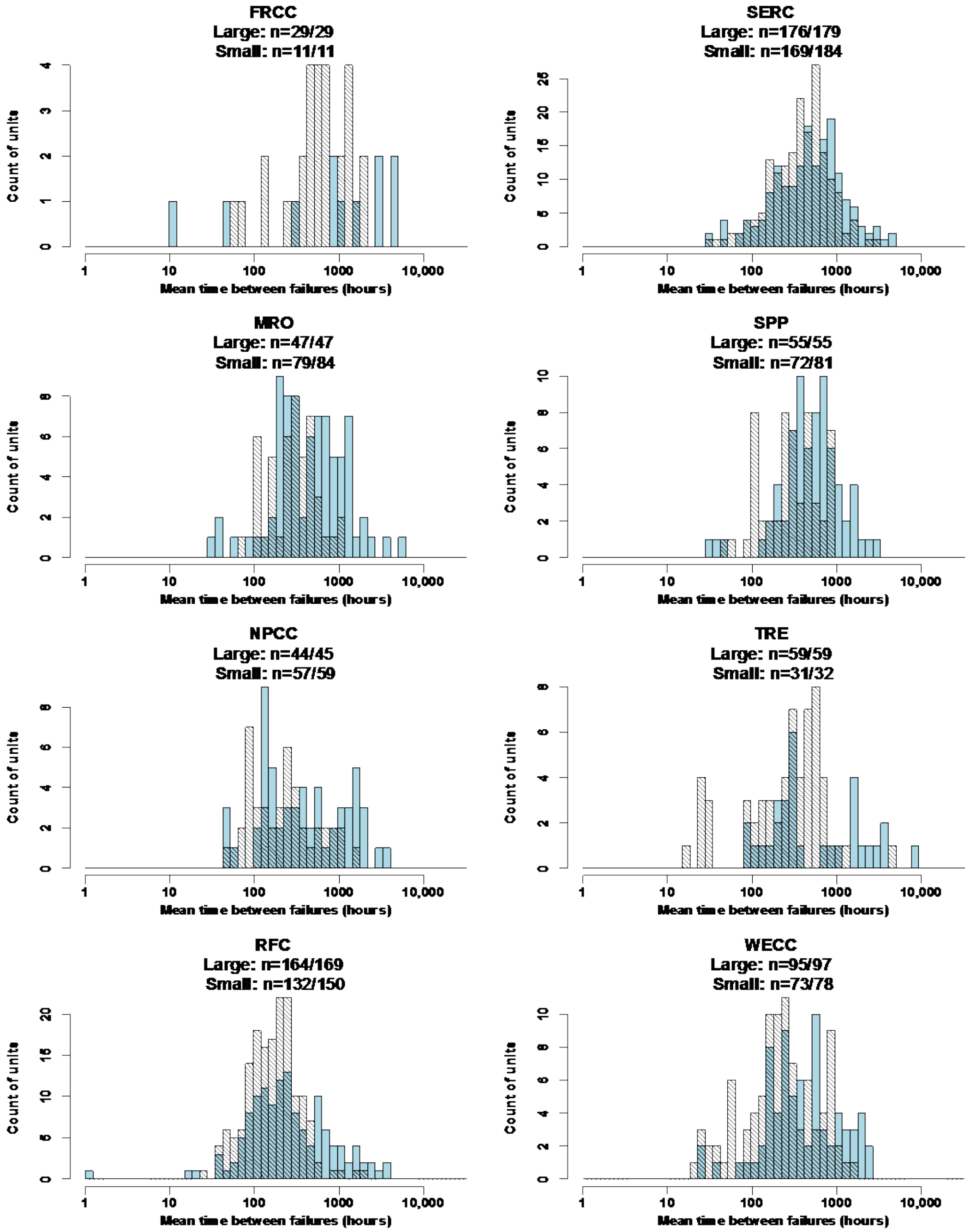


Figure S-49: Mean time between failure values for small (blue) versus large (black) fossil steam and fluidized bed units; threshold is 212 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least 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 S-14).

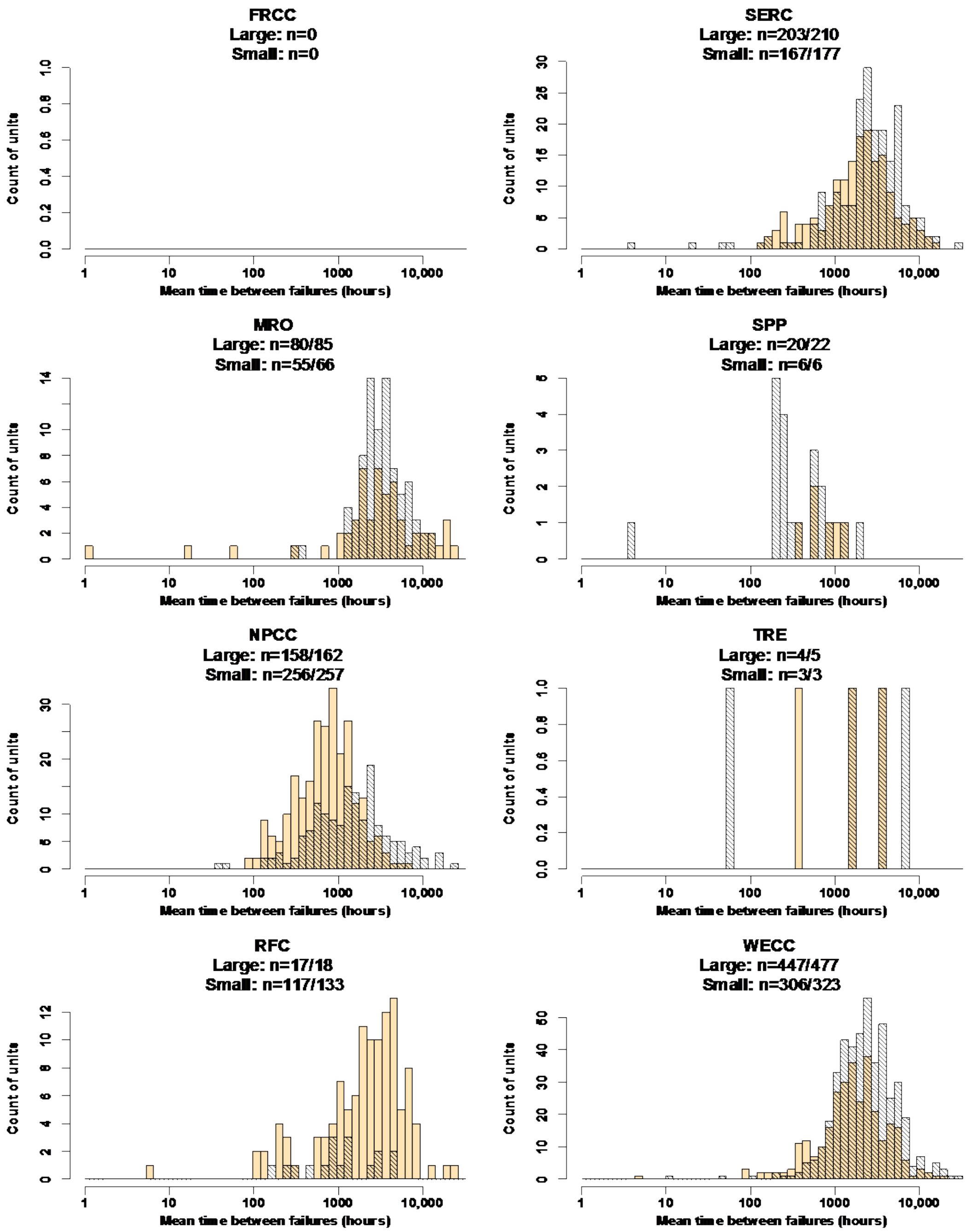


Figure S-50: Mean time between failure values for small (tan) versus large (black) hydroelectric units; threshold is 23 MW. FRCC has no such units. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing 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 S-14).

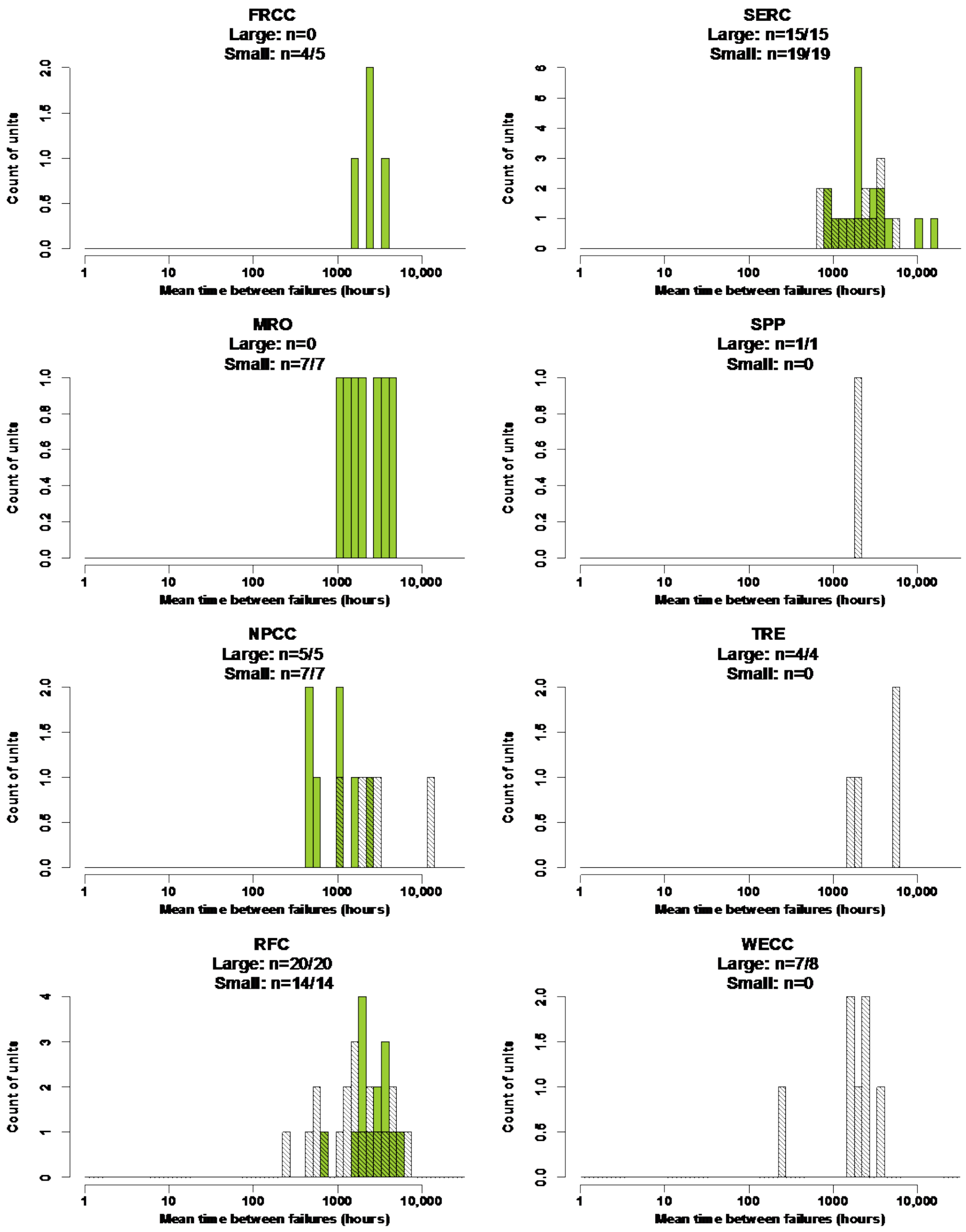


Figure S-51: Mean time between failure values for small (green) versus large (black) nuclear units; threshold is 1022 MW. Note the log scale for MTBF. Numerator indicates count of units for which an MTBF could be calculated. Denominator indicates count of units experiencing at least one unscheduled event 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 S-14).

4.3.6 Removing reserve shutdown hours from MTBF calculations

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 [1]. 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 S-14.

Table S-16: Summary of capacity (MW) falling into each of three reserve shutdown reporting categories, by unit type. All capacity represented in column 2 and column 4 is included in the MTBF analysis.

	MW not reporting RS events	*MW with RS events/performance discrepancy >100 hours*	*MW with RS events/performance discrepancy ≤100 hours*	*Percent of MW included in MTBF analysis*
Combined cycle	11,660	16,060	215,690	93.4%
Simple cycle	4,840	29,970	129,480	81.8%
Fossil steam	64,670	30,080	348,530	93.2%
Hydroelectric	32,740	36,790	51,440	69.6%
Nuclear	103,710	0	3,630	100%
Total	217,620	112,900	748,770	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.

References

[1] NERC. Generating Availability Data System data reporting instructions. 2017. <http://www.nerc.com/pa/RAPA/gads/Pages/Data Reporting Instructions.aspx>.

[2] Pfeifenberger JP, Spees K, Carden K, Wintermantel N. Resource adequacy requirements: Reliability and economic implications. Brattle Gr 2013.

[3] Paul SR, Ho NI. Estimation in the bivariate Poisson distribution and hypothesis testing concerning independence. Commun Stat Methods 1989;18:1123–33.