Enhanced Representations of Probabilistic Resource Adequacy Risk in Power System Capacity Expansion Modeling

Gordon Stephen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Daniel S. Kirschen, Chair

Baosen Zhang

Miguel A Ortega Vazquez

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Enhanced Representations of Probabilistic Resource Adequacy Risk
in Power System Capacity Expansion Modeling

Gordon Stephen

Chair of the Supervisory Committee:
Daniel S. Kirschen
Electrical and Computer Engineering

Power system infrastructure investment optimizations traditionally rely on capacity-based reserve margin heuristics to produce least-cost solutions that also meet probabilistic resource adequacy criteria. While these conventional strategies have always had shortcomings, they are becoming increasingly untenable as the grid's supply-demand balance comes to depend more and more on variable renewable generation, energy storage technologies, and interregional transmission. This work develops several novel iterative mathematical formulations to better represent probabilistic risk and the potential contributions of these new resources inside optimization-based capacity expansion models, without relying on capacity accreditation or stochastic optimization. Strategies for capturing resource adequacy impacts from supply uncertainty, weather-driven variability, temporal energy shifting via storage, and spatial energy shifting via transmission are developed independently before being integrated into a unified mathematical framework. The final iterative framework is then applied to demonstrate the methods on multiple test systems representing diverse climate regions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my dissertation advisor, Prof. Daniel Kirschen, for his insights, guidance, and flexibility over the course of my somewhat unconventional doctoral studies; as well as the other members of my dissertation committee, Profs. Baosen Zhang, Miguel Ortega-Vazquez, and Chaoyue Zhao, for their feedback and support throughout this process.

I was fortunate enough to retain my research position at NREL while completing this dissertation, and I am grateful to all my past and present NREL colleagues for providing a stimulating workplace and encouraging my efforts to balance work and school. Particular thanks go to Bethany Frew, who provided a supportive environment for my research interests and numerous opportunities to align efforts throughout my studies; Aaron Bloom, who first introduced me to probabilistic resource adequacy assessment and the methodological challenges that would become the topic of this dissertation; and Jaquelin Cochran, whose early support and encouragement was instrumental in allowing me to stay at NREL while performing this work.

I am also grateful to all the industry practitioners I had the opportunity to engage with throughout this research, including members of the IEEE Resource Adequacy Working Group and ESIG Resource Adequacy Task Force. They provided key perspectives into the challenges facing today's power system planners, resource adequacy practitioners, and software tool developers, and helped me understand the practical role this work could play in addressing those challenges.

I would also like to thank all of the members of the Renewable Energy Analysis Laboratory and broader University of Washington community who were a part of this journey with me. I am especially grateful to classmates, colleagues, and friends Lane Smith, Mareldi Ahumada Paras, Daniel Tabas, Nina Vincent, and Diego Pena, and to the dedicated teaching of Professors Daniel Kirschen, Maryam Fazel, Archis Ghate, and Lillian Ratliff, whose

courses challenged me to grow and helped lay the core theoretical foundations for this work.

Finally, I am deeply indebted to my family for their assistance, encouragement, and sacrifice, not merely through my doctoral studies, but my entire life. Their steadfast commitment to providing me with opportunities is what has enabled me to get to this point, and I am appreciative of and humbled by that fact. Despite our geographic distance over the past decade, I have never failed to sense their presence and support.

Chapter 1
INTRODUCTION

Global energy systems, and power systems in particular, are experiencing rapid transformations driven by the urgent need for decarbonization and continued cost declines in disruptive energy technologies. These unprecedented transformations impact both the supply and demand sides of electric power systems. On the supply side, system planners need to integrate variable renewable generation without jeopardizing the grid's ability to meet requirements for energy, capacity, and flexibility. Meanwhile, the task is complicated by the fact that those very requirements are changing under planners' feet, given increased electrification in the heating and transportation sectors, and the potential significant role of electricity in producing carbon-free liquid fuels. The efficient integration of all these resources also requires supporting investments in energy storage and transmission infrastructure to diversify electrical energy availability across space and time. Planning power system infrastructure investments that effectively facilitate this transition while balancing wide-ranging and uncertain technical, economic, policy, and social considerations therefore becomes a high-dimensional decision problem subject to significant uncertainty.

Power system planning has always been a challenging task, even in periods characterized by greater technology and policy stability. Given the capital-intensive nature and long lead times associated with grid infrastructure projects, decisions must often be made well before the exact needs of the future system are known with certainty. This planning process has only become more challenging as factors such as disruptive technological innovations in the supply and demand of electricity, policy uncertainty, and the growing prevalence of extreme weather events drive rapid changes in the operating context of current and future power systems, and introduce new questions about long-term system needs and capabilities.

While the challenge of planning modern power systems may be significant, the cost of failing to do so is even greater. Modern society is highly dependent on a reliable supply

of affordable electrical energy, a dependence that will only increase under decarbonization efforts. Planning to ensure the continued reliability of power systems through the energy transition is therefore paramount, and requires addressing not only the traditional challenges around uncertainty of supply and demand, but also the increasing prominence of weather-driven resource variability and the the spatial and temporal coupling introduced by increased reliance on long-distance transmission and storage.

Maintaining reliability through the energy transition requires modern planning tools capable of capturing these crucial emerging dynamics. Policy makers and power system planners often turn to capacity expansion models (CEMs) to identify economically responsible infrastructure investment strategies and assess the potential long-run impacts of policy decisions; these models typically formulate mathematical optimization problems to identify least-cost system designs that satisfy technical and policy constraints [1] in order to quantify the potential implications of emerging opportunities and risks, and help inform here-and-now decisions that must be taken with incomplete information about the future [2].

As a cost minimization, arguably the most fundamental aspect of this analysis is knowing the minimum level of investment that is still sufficient to maintain resource adequacy, which means maintaining an acceptably low probability of failing to balance electrical supply and demand [3]. Since such reliability investments are subject to diminishing returns and the risk of unserved electrical load can never be eliminated entirely, power system planners, regulators, and society at large must decide what level of risk they're willing to tolerate, taking into consideration the cost of reliability.

Tools that support this decision process require some internalized understanding of the level of reliability that different potential system designs would be able to deliver. The non-linear nature of quantifying probabilistic shortfall risk has required traditional CEMs to rely on deterministic, capacity-centric heuristics such as planning reserve margins to approximate this critical outcome, while more sophisticated probabilistic analyses [4] are performed separately. Probabilistic assessments, which are based on many alternate years of weather data and stochastic discrete generator outages [5], produce probabilistic metrics such as loss-of-load probability (LOLP), loss-of-load expectation (LOLE), and expected unserved energy (EUE) [6, 7]. Insights from these studies are mostly decoupled from the planning

optimization process, typically only interacting through the one-off choice of the CEM's planning reserve margin. This process has always left a disconnect between optimization-based planning decisions and probabilistic adequacy outcomes, even in traditional contexts dominated by thermal generating resources; however, the growing sophistication required of modern adequacy assessments has made these shortcomings ever more apparent and increasingly challenged the viability of this approach.

There are at least two options available to bridge the gap between these traditionally distinct domains of power system planning. The first is to fully merge portfolio optimization and probabilistic assessment through the use of stochastic optimization techniques. Chapter 2 considers two approaches for doing this, one parametric and one based on empirical sampling. The sample-based approach in particular is very flexible, conceptually simple, and aligned with traditional probabilistic adequacy assessment methods. However, it is also highly computationally demanding. With appropriate decomposition strategies and sufficient parallel computing resources, this class of solution strategies is capable of producing optimized solutions for large stochastic planning problems.

However, as a socio-technical exercise, power system planning is also characterized by many constraints and performance measures that are challenging to precisely represent in an optimization formulation. The process generally happens through a public, multi-stakeholder process and involves many "messy" and unquantifiable sociopolitical considerations that defy unambiguous mathematical representation. A single least-cost investment plan produced by even the largest, most sophisticated optimization model may be not be optimal or even feasible when judged on these softer and potentially inconsistent criteria. In other cases, cost parameters or causal mechanisms may be subject to deep uncertainties that erode the meaningfulness of a "least-cost" solution. Given these inevitable parametric and structural inaccuracies, a set of many "near-optimal" solutions that are reasonably cost-effective, satisfy the problem's core constraint set, and identify a robust range of acceptable alternative solutions may be more useful in practice than a single prescribed "master" strategy. This suggests that a more productive use of finite computing resources may be to solve many related-but-different planning problems quickly, rather than formulating and solving a single large, highly-detailed, but potentially misleading stochastic optimization problem.

A second approach to bridging capacity expansion optimization and probabilistic adequacy assessment is to iterate between the two classes of models, passing information back and forth to incrementally move towards a cost-effective system design that meets all probabilistic adequacy criteria. Chapter 2 considers a simple version of this strategy for planning thermal-dominated power systems and finds that, while less sophisticated than the stochastic optimization alternatives, it provides a pragmatic and computationally-efficient means to achieve desired planning outcomes. Chapter 3 investigates more sophisticated elaborations of this iterative approach and identifies adaptive stress period planning and sparse storage chronology as compelling strategies for efficiently valuing contributions from variable and energy-limited resources.

Despite the compelling advantages of the iterative techniques discussed above, they are still unable to compete with stochastic optimization in providing an endogenous understanding of probabilistic shortfall risk and how that risk might change under different infrastructure portfolios. Chapter 4 considers this challenge and develops a new iterative approach that uses full probability distributions from the resource adequacy assessment step to parametrize deterministic capacity expansion optimization problems, providing an approximated endogenous understanding of probabilistic risks in a much more tractable CEM formulation. Chapter 5 then combines the different iterative strategies from previous chapters into a single unified mathematical framework, demonstrating the complementary nature of the constituent approaches on three new test systems. Finally, Chapter 6 provides concluding remarks and identifies new research questions arising from this work which could be addressed in future efforts.

Chapter 2

REPRESENTING ADEQUACY CONTRIBUTIONS OF THERMAL GENERATING RESOURCES

Power system resource adequacy assessment has historically been dominated by the risk of mechanical component failures in thermal generating units. While other sources of uncertainty are becoming increasingly relevant in modern power systems, endogenizing an awareness of thermal outage risk in a planning optimization problem remains challenging despite the maturity of traditional adequacy assessment methods.

The fundamental challenge arises from the fact that, while thermal unit failures may be correlated due to shared causal relationships between outage likelihood and common exogenous drivers such as ambient temperature, they remain statistically independent after conditioning on the environmental context in which the units are operating. A mechanical failure and resulting forced outage at one generating facility will not induce a failure in another plant.

Thermal outages and repairs should therefore be modeled as stochastic transitions between unit-specific discrete states. Generally, only two states per unit are considered: available, with some deterministic but potentially time-varying level of generating capacity, and forced offline, with no abililty to generate [4].

When modelled non-sequentially (assessing each time-period independently of all others), such outages are assumed to be distributed as independent Bernoulli trials, with total system available capacity represented as the sum of these unit-specific random variables. In a sequential simulation, unit-level outages are most often considered serially dependent according to a two-state Markov random process, fully parametrized by the probabilities of transitioning from an operational state to a forced outage and vice versa [5].

Whether considered sequentially or non-sequentially, this discrete, independent nature of thermal generating units and their availability states results in a nonconvex set of fea-

sible expansion plans, and nonlinear relationships between build decisions and generating capacity available to the system at a given point in time. Grouping capacity from similar generators into a homogenous, aggregate "technology class", a common approach to reducing problem size in large capacity expansion models [8], is no longer possible when discrete unit outages are being considered.

The conventional means of accommodating supply uncertainty in a CEM is to define a planning reserve margin (PRM) above the expected peak system load [1]. A constraint can then be added to the optimization to require that the total installed capacity (or some derated "capacity credit" or "capacity value" [9]) for each resource in the system add together to meet or exceed the peak demand plus the PRM. PRMs vary between systems, but for larger interconnected systems are typically in the range of 10-20% of peak load [10].

While simple to implement in a mathematical program, the PRM approach to ensuring RA in the planned system fails to consider the operating characteristics of individual resources, leading to multiple issues. The most relevant here is that, given independent unit outages, PRM analysis fails to capture the substantially-increased risks of relying on a small number of large generators, where a single unit outage could eliminate a large fraction of the system's avaialable generating capacity. The precise PRM parameter necessary to cost-effectively satisfy a given probabilistic risk criterion is therefore a function of the model's build decisions, and so cannot be known a priori.

This chapter develops and tests three alternate adequacy constraint formulations intended to address this challenge, with Section 2.1 providing a basic capacity expansion model formulation to which each of these methods can be applied.

The first constraint formulation (Section 2.2) uses the traditional reserve margin-based planning approach but iterates between optimization solves and a full resource adequacy model. This enables direct assessment of the risk implications of discrete unit outages, with automatic tuning of the optimization problem's planning reserve parameter to find the lowest reserve margin requirement that meets the system's resource adequacy criteria.

The second method (Section 2.3) directly endogenizes the Monte Carlo methods used in stand-alone RA assessments into a stochastic optimization framework (sample average approximation). This method is conceptually simple and allows for optimizing to a desired

level of probabilistic system adequacy, as well as studying the economic tradeoffs between system reliability and cost. Unfortunately, its representation of thermal units introduces non-convexities which impact the method's computational practicality.

The third method (Section 2.4) applies a series of distributional approximations to capture the different risk profiles associated with thermal generator unit size in a convex optimization framework. While this method allows for certain probabilistic RA metrics to be imposed as direct system constraints, the approximations and relaxations required are unlikely to be accurate enough for satisfactory application in a wide range of practical settings.

Finally, Section 2.5 develops empirical tests and compares the computational performance, accuracy, and generating portfolios selected by each of these approaches.

2.1 Core Capacity Expansion Model Formulation

The basic goal of a planning optimization is to design a cost-minimizing (or, if demand elasticity is being modeled, welfare-maximizing) portfolio of generating resource investments. For a simple single-region model, this can be represented as follows:

$$\min_{n,p} \sum_{g \in G} n_g P_g C_g^c + \sum_{g \in G, t \in T} p_{gt} C_g^o \theta_t \tag{2.1}$$

Where n_g is the number of units to build in particular generator class $g \in G$, P_g is the capacity of a single unit in that class, and C_g^c is the per-MW capital cost of a unit in that class. Since we are minimizing total system cost, we must also consider the variable cost of operating those units C_g^o at a particular level of dispatch p_{gt} for each time period $t \in T$. As it is typically impractical to model operations for every time period of the operating horizon, we use representative periods with weighting factors θ_t.

Since we cannot invest in negative generators, n_g must be positive and provides an upper bound on the level of generation available from each generator class:

$$0 \leq n_g \ \forall \ g \in G \tag{2.2}$$

$$0 \leq p_{gt} \leq n_g P_g a_{gt} \ \forall \ g \in G, t \in T \tag{2.3}$$

Here a_{gt} represents the fraction of the class' capacity that is available to generate at time t. For wind or solar generators this may be time varying, while for thermal units it may be 100% or represent a fixed derate based on the unit class' average historical reliability.

Generation is dispatched to balance electricity supply and demand in each time period:

$$\sum_{g \in G} p_{gt} = L_t \ \forall \ t \in T \tag{2.4}$$

Here, L_t represents the electrical load in a given period. The traditional planning reserve margin constraint is then given as:

$$\sum_{g \in G} n_g P_g \mathrm{CC}_g \geq L_{\max}(1 + \mathrm{PRM}) \tag{2.5}$$

Where $L_{\max}$ is the peak system load, PRM is the planning reserve margin, and CC_g is the fractional capacity credit attributed to each resource based on its expected ability to contribute to serving the peak load. Since this section focuses on risk from independently-occuring thermal outages, these capacity credits can be taken as fixed based on the average reliability of a given technology. Calculating capacity credit for variable or energy-limited resources is more complex and will be discussed further in Chapter 3.

2.2 Reserve Margin Iteration

The traditional approach to enforcing adequacy criteria in planning models is to parametrize the model with technology-specific capacity credits (CC_g) and preselect an installed capacity surplus (PRM) that will result in a system buildout that delivers the desired level of resource adequacy. If capacity credits are accurately assigned based on average availability during the period of peak demand, a PRM of zero would imply that the system, in expectation, just manages to serve peak demand, with any above-average generator outages during the peak period resulting in load dropping.

Most power systems operate to much higher adequacy standards that this, and so such a system would not be considered resource adequate. By increasing the PRM, we can enforce

a higher adequacy constraint. The correct parameter value to choose depends not just on the desired reliability level for the system, but also on the variance of total generator availability around the expected level.

Rather than pre-specify a PRM level that may or may not yield an adequate system, the Reserve Margin Iteration approach [11] iteratively solves the planning problem for different choices of PRM, bisecting on possible values to find the one that yields a system sufficiently close to the desired level of adequacy, as assessed by a dedicated external resource adequacy tool. The optimization re-solves can be performed quickly since only a single constraint parameter is being varied, allowing the previous solution to effectively hot-start the next iteration's optimization.

2.3 Endogenous Monte Carlo Sampling

Given the impracticality of evaluating an exponential number of combinations of discrete component availability states, RA models typically apply Monte Carlo sampling [5] to estimate expected values of uncertain properties of interest, such as the total energy unserved over a simulation horizon, or whether the system is able to balance supply and demand in a given time period. These expected values can then be used as reliability criteria for determining the adequacy of the system [12].

Since this kind of Monte Carlo sampling is how dedicated adequacy models assess system reliability, one logical course of inquiry would be to embed those same calculations directly inside the planning problem. Embedding Monte Carlo sampling inside an optimization problem amounts to sample average approximation [13], in which the expected value of some outcome distribution is incorporated into the problem formulation by considering a finite number of randomly drawn input samples. By the Law of Large Numbers, the average outcome across the samples will tend towards the true expected value of the distribution.

In the case of an RA-aware CEM, the samples used correspond to stochastically-generated availability profiles for thermal generators, and the sample average to an adequacy risk metric. Once formulated, the metric may be incorporated into the model as either an engineering constraint (e.g. expected unserved energy must not exceed 0.001% of total demand) or an economic signal (e.g. considering the expected value of lost load as a system cost to be

co-optimized against capital expenditures).

As discussed previously, the discrete nature of thermal unit availabililty means that directly endogenizing RA considerations inside a CEM requires explicit consideration of unit-level thermal investments. The rest of this section outlines two possible approaches to this explicit representation that support incorporating discrete outages directly into the model's operational representation.

2.3.1 Independent binary variable approach

A conceptually simple method of considering discrete unit buildouts would be to explicitly model each unit j in a given class of generator g. A set of random outage profiles can then be created for each individual generating unit and, taken together with the generator class' rated capacity P_g, provides a time- and scenario-dependent maximum generation P_{gjst} to be included in the specific unit's operating constraints. For the jth generating unit of class g:

b_{gj} denotes the binary decision to build the jth unit

p_{gjst} denotes the unit's dispatch decision (MW) in timestep t of scenario $s \in S$

With these values defined, we can redefine the objective function given in Section 2.1 as follows:

$$\min_{n,p} \sum_{gj} b_{gj} P_g C_g^c + |S|^{-1} \sum_{gjst} p_{gjst} C_g^o \theta_t \tag{2.6}$$

For the jth generating unit of class g, the dispatch limit is given as:

$$0 \leq p_{gjst} \leq b_{gj} P_{gjst} \tag{2.7}$$

However, since the finite set of sampled future random outage profiles for each unit is known to the solver when making build decisions, it would be able to selectively build units with convenient random outage behaviour, biasing internal estimates of probabilistic resource adequacy metrics to be overly optimistic.

2.3.2 One-hot binary variable approach

An alternative approach is to use a one-hot binary encoding in the optimization problem to require units to be built in a particular order. For a class of generator with n possible units to build, this reduces the number of build combinations from 2^n to just n, reducing the solution's ability to overfit its build decisions to the finite samples provided and reducing (but not eliminating) potential bias in assessed adequacy. Under this formulation, an available capacity parameter P_{gjst} represents the cumulative available capacity of all units up to j in class g, considering randomly-generated forced outages for timestep t in sample s.

By using a similar clustered dispatch approach to that of Palmintier and Webster [14], units within a class of generator can be modeled with a shared dispatch variable, reducing the problem size. Unlike that work, however, integers cannot be used to track the number of units built, as available capacity at different times no longer scales linearly with unit count when considering the potential for random outages.

Now, for a given class g of generating units:

b_{gj} denotes the binary decision to build j units of class g, with a one-hot constraint $\sum_j b_{gj} \leq 1$

p_{gst} denotes the class' total dispatch in timestep t of scenario s

The problem's objective function then becomes:

$$\min_{n,p} \sum_{gj} j b_{gj} P_g C_g^c + |S|^{-1} \sum_{gst} p_{gst} C_g^o \theta_t \tag{2.8}$$

while the dispatch constraint is:

$$0 \leq p_{gst} \leq \sum_j b_{gj} P_{gjst} \tag{2.9}$$

By using one-hot binary variables instead of integers, the model is able to represent the nonlinear nature of cumulative capacity under random draws of independent outage

profiles. Given that thermal units within a given class of generator are assumed to have identical capital costs and operating characteristics, the only impact to the model over the independent device representation is the reduced ability to cherry-pick units with favorable outage profiles.

Variable generating resources or temporal patterns in unit rating or reliability are easily represented by changing the values of the availability time series P_{gjst}. Since each probabilistic scenario is represented explicitly, more complicated considerations such as transmission and storage constraints can be modeled directly as well. This approach can also easily calculate and apply many alternative probabilistic risk metrics.

Unfortunately, this approach is unlikely to be viable for larger systems. Sample error and selection bias are introduced through the sample average approximation process, and reducing these to acceptable levels requires many Monte Carlo replications [13]. Furthermore, each additional generating unit to be considered adds a new binary variable to the problem. As a result, obtaining an accurate probabilistic representation of a large power system necessarily requires solving a very large non-convex optimization problem, which may be computationally impractical.

2.4 Convex Relaxation with Normal Approximation

Given the computational challenges of Monte Carlo sampling, one may wish to simplify the system representation such that direct analytical assessments of supply shortfall probabilities become tractable. Assuming identical generating capacities c_g and availability rates p_g for a grouping g of n_g thermal units, the total count of available units in the group is distributed as a binomial distribution $B(n_g, p_g)$ [4]. The total available capacity is therefore distributed as $c_g B(n_g, p_g)$. For simplicity, we assume here that c_g and p_g are the same in all time periods: resources with time-varying availability characteristics (such as variable renewables) could be modeled by generalizing c_g or p_g to vary across time (for example, fully correlated renewables would use a time-varying c_{gt} with $p_g = 1.0$).

While this representation of available capacity is conducive to analytical calculations of probabilistic system risk metrics (such as the probability of shortfall at a given load level), its discrete and nonlinear nature is clearly not ideal for adding capacity across different

types of units and embedding in an optimization problem. We can instead approximate this capacity distribution by another with more convenient properties.

Under certain conditions, the Binomial distribution $B(n_g, p_g)$ can be approximated as a continuous Normal distribution with mean $n_g p_g$ and variance $n_g p_g(1-p_g)$ [15]. Scaling by unit size c_g, we can also represent the available capacity as a Normal distribution with $\mu_g(n_g) = n_g c_g p_g$ and $\sigma_g^2(n_g) = n_g c_g^2 p_g(1-p_g)$. Both of these parameters are linear functions of n_g, the number of units of g built. Since Normal distributions are closed under addition and their parameters combine linearly, the total available generating capacity (across all generator classes g) can be represented as a single Normal distribution with parameters that are also linear functions of investment levels n.

$$\mu(n) = \sum_g \mu_g(n_g), \quad \sigma^2(n) = \sum_g \sigma_g^2(n_g) \tag{2.10}$$

$$x \sim \sum_g N(\mu_g(n_g), \sigma_g^2(n_g)) = N(\mu(n), \sigma^2(n)) \tag{2.11}$$

Subtracting x from system demand L_t gives s_t, the distribution of system capacity shortfall at time t:

$$L_t - x = s_t \sim N(L_t - \mu(n), \sigma^2(n)) \tag{2.12}$$

Integrating this distribution over negative values (surplus conditions) gives p_t, the probability that the system has sufficient capacity to meet demand in period t.

$$p_t = P(s_t \leq 0) = P\left(\frac{s_t - L_t + \mu(n)}{\sigma(n)} \leq \frac{-L_t + \mu(n)}{\sigma(n)}\right) \tag{2.13}$$

Note that by the definition of s_t, $\frac{s_t - L_t + \mu(n)}{\sigma(n)}$ is Normally distributed with mean 0 and variance 1, meaning adequacy probability can be calculated in terms of the cumulative distribution function Φ of the standard Normal distribution:

$$p_t = P(s_t \leq 0) = \Phi\left(\frac{\mu(n) - L_t}{\sigma(n)}\right) \tag{2.14}$$

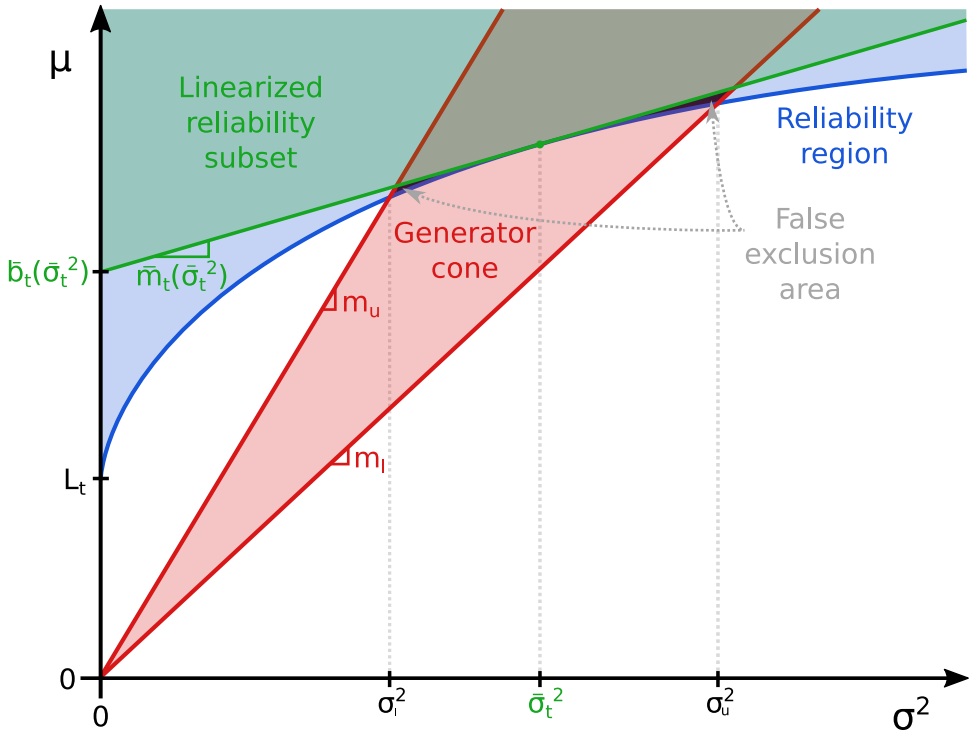

Figure 2.1: Each period's LOLP-driven reliability constraint is a non-convex set in terms of the mean (μ) and variance (σ^2) of total available generating capacity. A convex half-space approximation of the reliability region can be chosen based on the properties of the thermal generator options to be considered.

Choosing some maximum allowable LOLP, which corresponds to a minimum adequacy probability p_t, we can express a probabilistic reliability constraint in terms of build decisions n. After some rearranging, we obtain:

$$\mu(n) - \Phi^{-1}\left(1 - \text{LOLP}_{\max}\right)\sigma(n) \geq L_t \tag{2.15}$$

Unfortunately, while $\mu(n)$ and $\sigma^2(n)$ are linear functions of n, $\sigma(n)$ is not, and the set of n satisfying this reliability constraint is non-convex for typical values of $\text{LOLP}_{\max} < 0.5$.

Figure 2.1 illustrates that the non-convex set of investment options satisfying the reliability contraint can be approximated by defining a halfspace in (μ, σ^2). The set of candidate dividing lines for the halfspace are the tangent lines to the original set boundary. While there are infinite such choices, the selection can be informed by the characteristics of the generator

build options available. From the definitions of μ_g and σ_g^2 we see that $\sigma_g^2 = c_g(1-p_g)\mu_g$ defines a ray for each generator class, limiting the 'accessible' (μ, σ^2) space to that contained within the conic combination of generator rays. We can then choose a subsetting halfspace to minimize the 'false exclusion error' area outside the halfspace but within the intersection of the generator cone and the reliabilty region.

With the constraint expressed as $\mu(n) \geq L_t + \alpha\sqrt{\sigma^2(n)}$, $\alpha = \Phi^{-1}(1 - \text{LOLP}_{\max}) \geq 0$, we can express the boundary tangent line at a given $\bar{\sigma}_t^2$ as $\mu = \bar{m}_t(\bar{\sigma}_t^2)\sigma^2 + \bar{b}_t(\bar{\sigma}_t^2)$ with $\bar{m}_t$ and $\bar{b}_t$ given as:

$$\bar{m}_t(\bar{\sigma}_t^2) = \frac{\alpha}{2\sqrt{\bar{\sigma}_t^2}}, \quad \bar{b}_t(\bar{\sigma}_t^2) = L_t + \frac{\alpha}{2}\sqrt{\bar{\sigma}_t^2} \tag{2.16}$$

At this point it is convenient to define m_u and m_l, the slopes of the two rays bounding the accessible generation cone in (μ, σ^2) space. m_u is the maximum of $\frac{1}{c_g(1-p_g)}$ over g, corresponding to the 'least variable' class of generator and therefore the slope of the steepest generator ray. m_l is the minimum of the same quantity and therefore corresponds to the 'most variable' class of generator and the slope of the shallowest generator ray.

Requiring that $\bar{m}_t({\bar{\sigma}_t}^2) < m_l \leq m_u$ so that the false exclusion error area is finite, we can calculate this area (shown in black in Figure 2.1) via integration and minimize as a function of the tangent point $\bar{\sigma}_t$:

$$\min_{\bar{\sigma}_t} \frac{m_u - m_l}{2}\bar{\sigma}_t^2 \frac{\frac{\alpha^2}{4}\bar{\sigma}_t^2 + \alpha L_t \bar{\sigma}_t + L_t^2}{m_u m_l \bar{\sigma}_t^2 + \frac{\alpha}{2}(m_u + m_l)\bar{\sigma}_t + \frac{\alpha^2}{4}} + F(m_u, m_l, \alpha, L_t) \tag{2.17}$$

This minimization can be solved through one-dimensional line search over $\max(\frac{\alpha}{2m_l}, \sigma_l) \leq \bar{\sigma}_t \leq \sigma_u$, where $m_u\sigma_l^2 = L + \alpha\sigma_l$ and $m_l\sigma_u^2 = L + \alpha\sigma_u$.

The false exclusion area can also be approximated by integrating the μ-distance between the linear and non-convex constraint bounds over $[\sigma_l^2, \sigma_u^2]$. This value can be shown to be locally convex in $\bar{\sigma}_t^2$, with a minimizer at $\bar{\sigma}_t^{2*} = \frac{1}{2}\sigma_l^2 + \frac{1}{2}\sigma_u^2$. This can serve as a much simpler heuristic for making a choice of convex subset of the reliability region to constrain the planning problem.

With $\bar{\sigma}_t^2$ selected, we can apply the definitions of $\bar{m}_t$ and $\bar{b}_t$ above and approximate the set of feasible n with the following linear constraint:

$$\mu(n) - \frac{\alpha}{2\bar{\sigma}_t}\sigma^2(n) \geq \frac{\alpha}{2}\bar{\sigma}_t + L_t \tag{2.18}$$

As seen in Figure 2.1, this approximation is most accurate when $\sigma^2(n)$ is close to the choice of $\bar{\sigma}_t^2$, and less accurate (more conservative than the original constraint) farther away. In practical terms, this means that a system design using only the least-variable or most-variable class of thermal generators is more likely to be overbuilt relative to one using a mix of generation technologies. If only a single class of thermal generators is considered in the optimization, the intersection of the reliability boundary and potential build space defines a unique choice of $\bar{\sigma}_t^2$, and the approximated linear constraint is mathematically equivalent to the non-convex reliability constraint, within the available build space.

This approach is attractive given its ability to consider explicit probabilistic reliability criteria in a concise mathematical representation. It is, however, subject to serious caveats. Most fundamentally, the method assumes that the relevant discrete Binomial distributions $B(n,p)$ can be satisfactorily approximated as continuous Normal distributions $N(\mu,\sigma^2)$. This approximation is most accurate for near-symmetric Binomial distributions ($p \approx 0.5$), which would imply a much lower level of reliability than is typical of modern generating units, where typically $p \geq 0.9$. This higher reliability significantly skews the Binomial distribution and degrades the Normal approximation. A rule of thumb is that any continuous outcome within three standard deviations of the distribution mean should still be in the range of valid binomial outcomes, which corresponds to the criteria $n > 9\frac{1-p}{p}$ and $n > 9\frac{p}{1-p}$. For a class of units with $p = 0.9$, this implies a need to consider at least 81 built units (in that class alone) before the approximation can be considered sufficiently accurate. This method is therefore not appropriate for representing smaller systems or those having classes of generators with only a few units.

Furthermore, much like traditional convolution-based RA assessment methods [4], the computational tractability of this approach depends on a single-area and time-independent representation of system operations. More detailed intertemporal and interregional operating characteristics, such as storage and transmission constraints, cannot be represented.

Finally, the constraint derivations provided here are all specific to calculating single-

period LOLP: considering alternative risk metrics would require a fundamentally different mathematical representation, where possible at all.

2.5 Empirical Tests

2.5.1 Test Systems

Each of the alternative thermal outage representations was applied to solve power system planning problems for different sizes of system (500 MW, 5 GW, and 50 GW peak demand). In each case, the model could select from two classes of generators: a generic 'baseload' class, characterized by a larger unit size, higher per-MW capital cost, and lower operating costs, and a generic 'peaking' class, with smaller unit size and per-MW capital cost, and higher operating costs.

To focus on the impact of the thermal generator representations, a very simple four-timeslice operations model was considered, with no renewable generation or storage expansion options. The peak demand period LOLP was constrained to not exceed 1%. For the Endogenous Monte Carlo approach, 50 different solutions were generated for each system size in order to capture variability in outcomes resulting from different random outage draws.

2.5.2 Normal Approximation Validity

Figure 2.2 shows the resulting discrete available capacity distribution for each system developed under the Normal Approximation method, as well as the continuous Normal approximations corresponding to those buildouts. As expected, the quality of the approximation increases with system size as more discrete units in each class of generator are built.

By linearity of expectation, the expected values of the discrete and continuous distributions will always match (within rounding errors associated with quantizing continuous buildouts). However, the high reliability of individual units introduces a skew to the discrete (convolved Binomial) distributions, such that their mode exceeds their expected value. The Normal approximation distributions are symmetrical and so are unable to reproduce this skew.

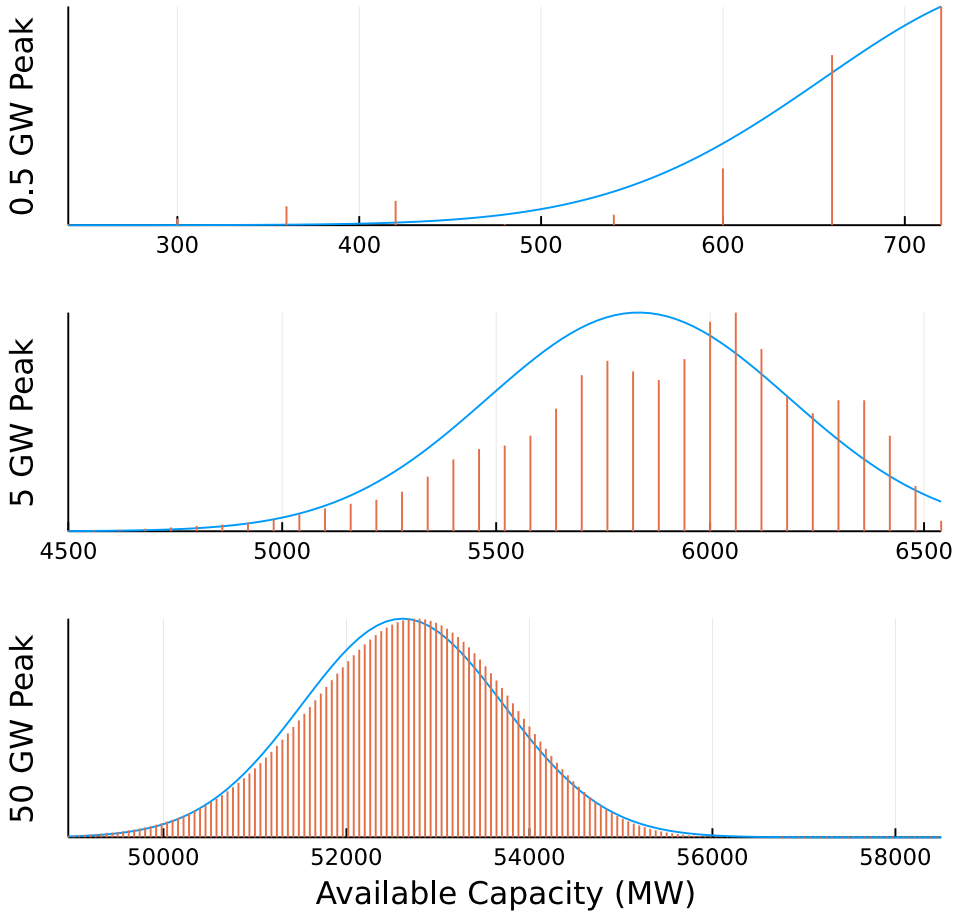

Figure 2.2: Discrete distributions of total available generator capacity given build decisions (rounded to the nearest full unit) for three system sizes, with the corresponding Normal approximations that informed those decisions.

2.5.3 Endogenous Risk Estimates

Figure 2.3 compares the accuracy of each solution's internal peak-demand LOLP estimate under the endogenous Monte Carlo approach (as a function of sample size) with the Normal Approximation method. As expected, Monte Carlo estimates are inaccurate and biased downwards at smaller sample sizes, but improve as more outage conditions are considered. They require many hundreds of samples in order to consistently outperform the Normal Approximations.

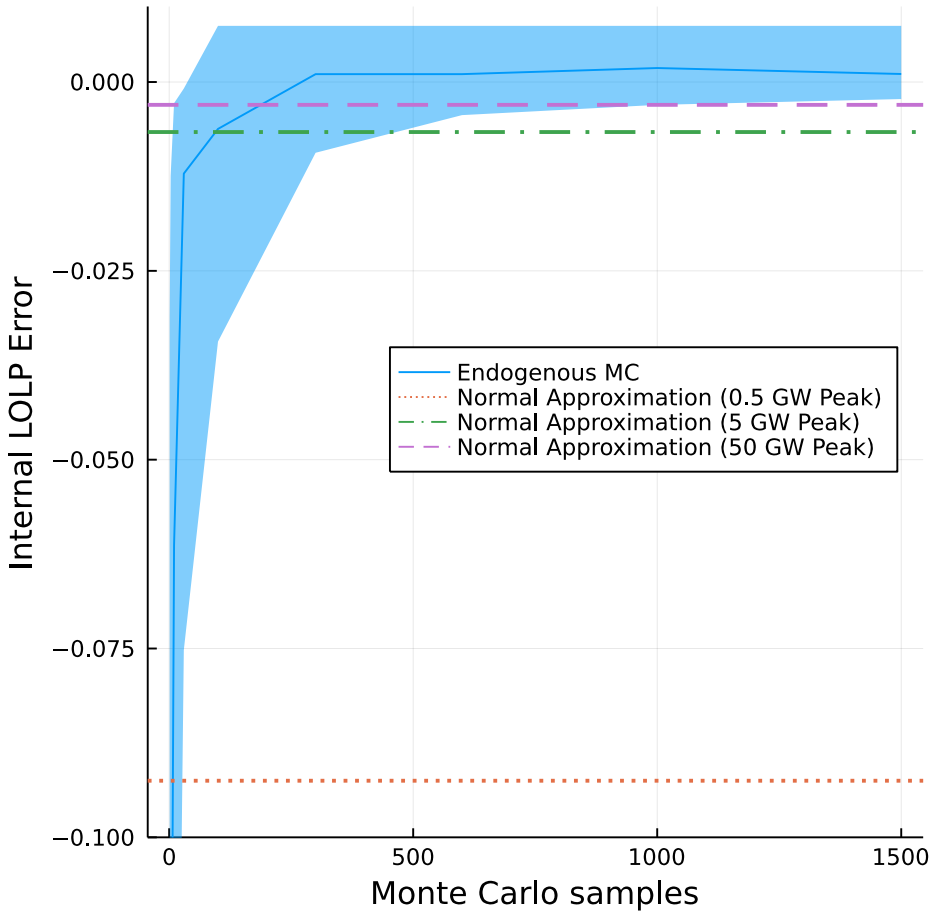

Figure 2.3: Error in each model's LOLP estimate for the peak demand period. The solid Monte Carlo line represents the median LOLP estimate error, while the band around it represents the 5th to 95th percentile range of outcomes. The Reserve Margin Iteration approach does not generate an internal LOLP estimate and so is not plotted.

Unsurprisingly, the Normal Approximation delivers better results for larger systems where the approximation is more accurate. That method's risk estimate for the 0.5 GW peak system is particularly poor. In addition to the poor quality of the continuous approximation, at this size of system relative to the size of individual generating units, the solution quality is highly susceptible to rounding error. In this instance the model chose to build partial units in both unit classes, which were 'rounded off' when formulating the discrete capacity distribution for ground-truth LOLP calculations. The strong rounding impact at this system

size is also visible in Figure 2.2, where the Normal Approximation's expected available capacity ends up larger than the system's post-rounding total installed capacity.

2.5.4 Computational Performance

Unsurprisingly, the smaller, convex models (Normal Approximation and Reserve Margin Iteration) solve many orders of magnitude faster (on the order of milliseconds) than the larger, non-convex Endogenous Monte Carlo models (which solve in seconds or minutes, depending on the number of samples considered). Furthermore, the latter approach adds additional binary build decision variables as the system size increases, further limiting computational scalability, while the number of variables and constraints in the convex models stay constant.

The Reserve Margin results here only consider the time needed to perform the final optimization, and not the preliminary solves used to find the PRM value corresponding to the LOLP target. In practice, it is unlikely a system planner would re-run their expansion model iteratively: instead, a more efficient resource adequacy model may be used to develop a PRM target for the system, or a rule-of-thumb PRM may be used and only adjusted if subsequent analysis reveals adequacy issues with the build plan.

2.5.5 Investment Outcomes

Figure 2.5 shows the range of build decisions (investments in baseload and peaking generator capacities) across the different adequacy representations. Endogenous Monte Carlo build decisions exhibit some variability and bias towards underinvestment when working with limited samples, although this spread diminishes as more samples are considered.

An important corollary of explicit probabilistic adequacy considerations is that the Endogenous Monte Carlo and Normal Approximation methods can automatically 'choose' an appropriate PRM, rather than having this prescribed as a parameter. In the 500 MW system, the models build twice as much capacity as peak demand, recognising the small number of generators in the system and a corresponding lack of redundancy during outages. In the 50 GW system (with more discrete units), the PRM falls to around 17%, which is

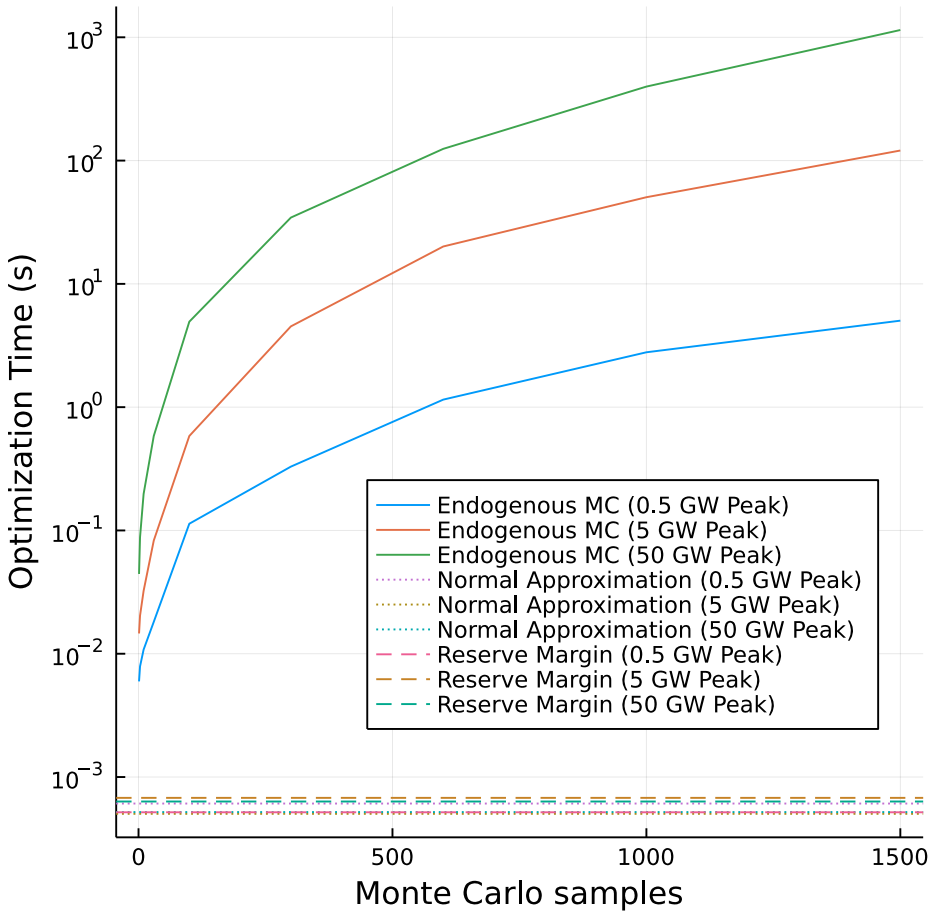

Figure 2.4: Average optimization solve time for each adequacy representation and system size.

comparable to target values for real grids.

Interestingly, the observed results suggest that the Normal Approximation and Reserve Margin methods may exhibit a systematic bias towards building more baseload units and fewer peaking units. This may be related to their deterministic operational representation which, unlike the Endogenous Monte Carlo method, does not consider the operating cost implications of random generator outages, but instead assumes uniformly-derated capacity availability.

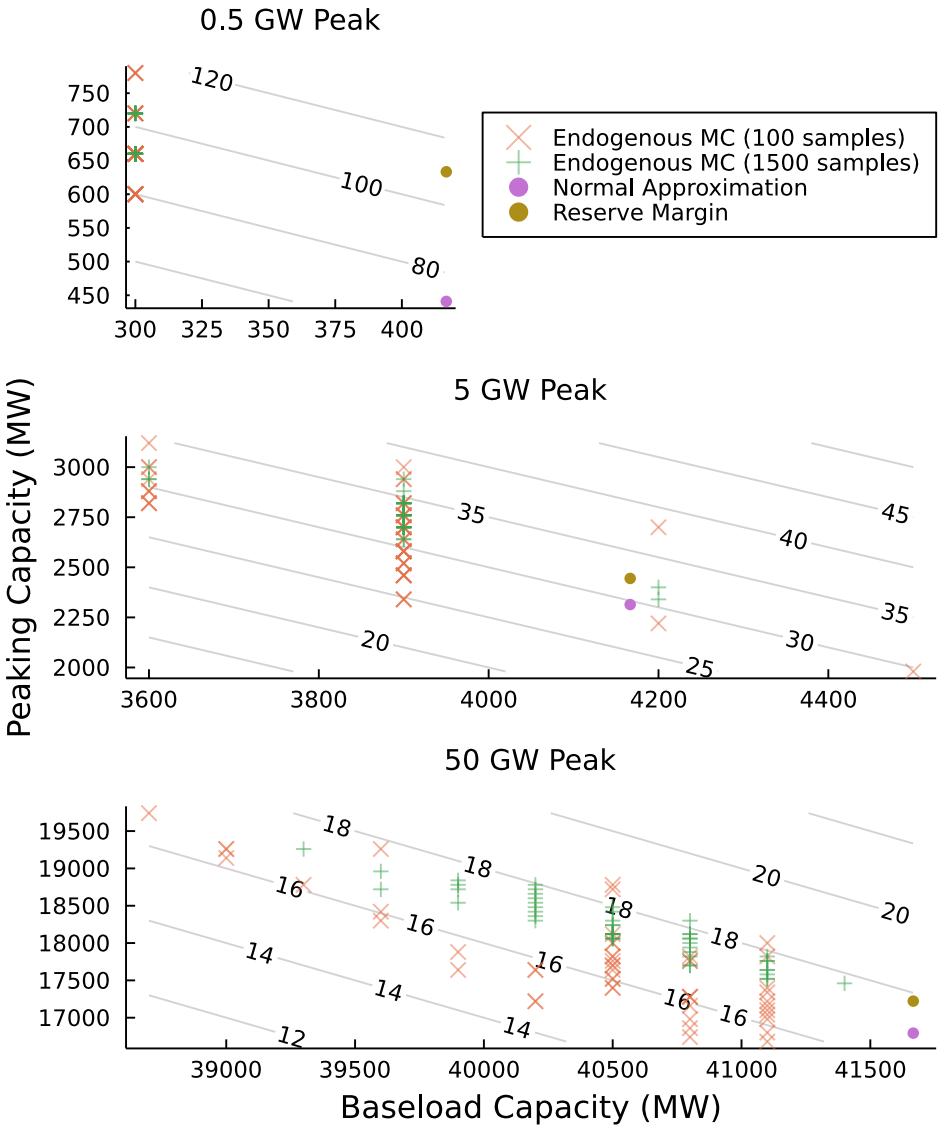

Figure 2.5: Investment outcomes under different adequacy representations and sample sizes. Grey contour lines denote the built system's planning reserve margin. The Endogenous Monte Carlo methods return different results as a function of different sets of random outage draws.

2.6 Conclusion

Accurately representing thermal generator outage impacts in CEMs is challenging. The three thermal outage representations and corresponding RA constraints considered here each

provide benefits over a conventional reserve margin approach, but no one option dominates the other.

The Endogenous Monte Carlo approach is able to simultaneously consider operational details (such as interregional power transfers and intertemporal storage constraints) and stochastic outages, but is prone to bias with small sample sizes, and has potentially prohibitive computational costs and optimization outcomes that may be sensitive to the random outage inputs. The method works best for small systems where a large sample size remains computationally feasible.

The Normal Approximation method yields a much more tractable optimization problem and consistent decisions, but cannot consider potentially relevant details of system operations, and relies on multiple layers of approximations that may not always be valid. As such it is better suited to applications in systems with a large number of similar generators, few energy-limited resources, and few binding transmission constraints.

In cases where transmission and storage cannot be neglected, but computational limits prevent the application of Endogenous Monte Carlo sampling, iterative tuning of the traditional planning reserve margin remains a simple, reliable, option. If the preselected reserve margin is chosen correctly, it can achieve the desired reliability outcome – although the model itself has no endogenous understanding of this objective, so multiple solve iterations may be required. Reserve Margin Iteration is the only method studied here that can guarantee the adequacy of the final result (since a full adequacy assessment is included in the design process).

This work has focused on improving the representation of thermal unit outages in system planning optimizations. The methods considered here are compatible with modeling variable and energy-limited resources as well, although this requires considering adequacy contributions in many more time periods than just the peak load period. Chapter 3 will investigate how to efficiently incorporate these factors in an iterative optimization framework.

Chapter 3

REPRESENTING ADEQUACY CONTRIBUTIONS OF VARIABLE AND ENERGY-LIMITED RESOURCES

Unlike thermal generators, VRE resources such as wind and solar photovoltaic generation are built from smaller mechanically-independent components (e.g. individual wind turbines) and may be much less likely to randomly fail (e.g. solid-state electronics). As a result, levels of VRE investment can be approximated as continuous, and resources are not prone to discrete availability state transitions. Instead, availability uncertainty across nearby resources is dominated by their direct dependence on shared weather conditions, which determine their level of generation within a continuous and highly correlated range of values.

This continuous, correlated nature of variable resource availability is much more conducive to representation in a convex optimization framework. While the chaotic atmospheric processes that govern resource availability are much more difficult to model than discrete random outages in thermal units, scenario-generating simulations can be provided as static parameters to an optimization model, and so need not complicate the internal problem representation.

Given the correlated nature of the availability of specific units, the aggregate generation from a technology class in a given region at a given time can be readily assessed as a linear function of total capacity investment, eliminating the need to consider individual units and allowing the formulation of smaller convex expansion problems. However, direct consideration of system adequacy in the presence of variable resources requires considering the system's balancing ability in many more operating periods than is necessary for uniformly-available thermal generating resources, and temporal coupling between periods becomes important to understand the availability of energy-limited resources such as storage.

As variable renewable resources started being incorporated into system planning, methods to determine how much these resources should be allowed to "count" in a capacity-based

adequacy paradigm became necessary. Clearly, a wind or solar facility cannot be expected to dependably generate at its nameplate capacity during peak load periods - but it is also unlikely to not be generating at all. The concept of capacity accrediation was therefore developed to assess the adequacy contribution ("capacity credit") of different resources and allow these new technologies to be incorporated into traditional PRM-based planning processes. Unfortunately, the capacity credit of a resource is highly dependent on the context of the system in which it's being studied.

Most resources exhibit some degree of diminishing marginal capacity credit. For example, on a system with afternoon demand peaks, solar PV may provide a generation profile that's well-aligned with the system's capacity needs. However, as more solar is added to the system, a surplus of solar generation during the day may shift the periods of greatest relative risk to evening hours after sunset, at which point additional solar PV investments would do very little to improve system adequacy. Cross-resource interaction effects are also important to consider: for example, adding additional storage to the system may allow surplus solar generation to be shifted to the new evening peak, increasing the apparent capacity credit of solar once again.

Given these interactive dynamics, resource capacity credits need to be constantly recalculated as the potential resource mix changes. Unfortunately, the probabilistic capacity accreditation methods that most accurately reflect a resource's incremental contribution to system adequacy can be slow and cumbersome to compute, as they require iteratively assessing the adequacy of the system to map the benefits provided by each candidate resource into the equivalent contribution that would be provided by a non-variable resource.

The need to continuously recompute capacity credits coupled with the computational burden of doing so has inspired a long history of efforts into developing and assessing the accuracy of capacity credit approximation methods. Milligan and Parsons [16] provide one early example, with subsequent examples spanning multiple decades and considering capacity accreditation as a more general objective in of itself [17, 18], or more explictly as a means to an end in the planning process [19]. With the emerging importance of storage in system adequacy, newer approximation methods became necessary to better capture the adequacy contribution of those technologies [20, 21] and motivated work to understand how

that contribution may change under alternative assumptions of how storage is dispatched [22].

Capacity credits are ultimately linear approximations of an underlying nonlinear function which maps investments in individual resources to a system-level measure of adequacy. To help address the fact that any single capacity credit used in a planning exercise becomes inaccurate as soon as the system changes, recent work has also sought to precompute system adequacy under many alternative resource portfolios and using those to estimate dynamic capacity credits endogenously in planning models [23, 24].

There has been comparatively less work into more radical alternatives to the PRM-based adequacy paradigm, in spite of growing acknowledgement that 'every hour matters' in adequacy assessment, and de-emphasis of PRM-based risk assessments in real systems in favor of more rigorous probabilistic assessments. There are, however, a few examples available.

To better capture the contribution of storage and demand response, Hawaiian Electric has starting developing resource plans that satisfy system operations with an "energy reserve margin" applied [25], rather than planning against a capacity margin. Mertens et al. [26] proposed more explicitly representing critical peak periods in a planning model, decoupling operational representations used for estimating system cost from those used for testing system adequacy. The MIT Future of Energy Storage study [27] iterated between planning and deterministic operations models to identify key risk periods to include in the planning problem, rather than simply enforcing a planning reserve margin and raising it when system adequacy was found to be deficient.

Section 3.1 provides a more detailed discussion of two iterative adequacy constraint frameworks based on capacity credits and PRMs. Section 3.2 then describes a strategy for iteratively adjusting risk periods, not reserve margins or capacity credits, to capture the system's adequacy needs (including potential dependencies on seasonal energy shifting) and choose investments that satisfy those needs.

Capturing the capabilities and constraints of longer-timescale energy shifting is a common challenge. To that end, Section 3.3 develops a "sparse" multi-year chronological representation of storage state-of-charge evolution. Section 3.4 then discusses how storage sizing

decisions are taken under a PRM adequacy framework and how an iterative risk period selection approach can simplify the sizing process and lower system costs.

Section 3.5 applies all these methods to the problem of planning a single-region system comprised solely of wind, solar, run-of-river hydro generation, and energy storage. The process is then extended to a multi-region, national-scale dataset to study how risk period iteration requirements scale with temporally-diffuse shortfall conditions arising from load and resource diversity. Finally, Section 3.6 provides concluding thoughts.

3.1 Capacity Credit Approaches to Adequacy

To test the performance of the risk period iteration approach described later in this section, we will compare it to two alternative applications of the capacity-based planning reserve margin adequacy paradigm. Since we will use total system cost as a measure of solution quality, these solutions also iterate in an attempt to avoid an overbuilt (or underbuilt) system, rather than performing a single pass and using a deliberately-large planning reserve margin to (hopefully) achieve an adequate system, as is often done in current practice.

For the purposes of this work, we define capacity credit as a variant of effective load carrying capability (ELCC) [9]. For a given normalized hourly load profile (with magnitude varying between 0 and 1), a resource portfolio is tested to determine the maximum scaling factor that can be applied to the profile before the system is unable to serve all energy demand. The adequacy assessment applies periodic boundary conditions such that the state-of-charge of each storage device at the end of the planning horizon (one year) matches the initial state of charge for each device. The fractional capacity credit of a given resource is then defined as one tenth of the incremental increase in the load scaling factor that can be accommodated given the addition of 10 MW of new capacity of the resource in question.

3.1.1 Planning Reserve Margin Augmentation

The first planning reserve margin / capacity credit implementation we consider is very similar to the approach outlined in Section 2.2, where adequacy is achieved by tuning the system's planning reserve margin. We begin by establishing a base system: in brownfield

expansion situations, this base system would simply be the existing generating portfolio, and the marginal capacity credit of resources relative to this asset mix could be calculated.

In a greenfield expansion, as will be considered here, variable generating resources are initially assigned a capacity credit equal to their annual capacity factor, while storage devices are assigned a capacity credit of 100%. A base system is then determined by running the planning optimization with the constraint that the sum of the built resources' capacity contributions meets or exceeds the peak system load (no additional planning reserve margin is applied).

This base system will typically not be resource adequate: in that case, the base system's ELCC is calculated, as well as the marginal capacity credit for each candidate resource in the context of the base system. The base system portfolio is then fixed and additional resource investments are selected by the planning model based on the newly-calculated capacity credits.

In this method, if the resulting system remains inadequate, the system's planning reserve margin is incremented by two percentage points and the additional resource investments beyond the base system are re-selected. The PRM is repeatedly increased (and capacity credits are left unchanged) until a resource adequate portfolio has been chosen. This process is visualized in Figure 3.1.

Clearly, this approach is somewhat naive when planning systems with signficant variable and energy-limited resources, as it increases the magnitude of investments made without accounting for how the capacity credit of those resources changes as more are provisioned. However, it is conceptually simple, involves only a single set of capacity credit calculations, and reflects an intutive understanding of how planning reserve margins are applied in practice (increasing the system's planning reserve margin in an attempt to increase the system's level of adequacy).

3.1.2 Capacity Credit Iteration

A more sophisticated application of the planning reserve margin framework would consider how the capacity credit of different resources evolves as a function of system composition.

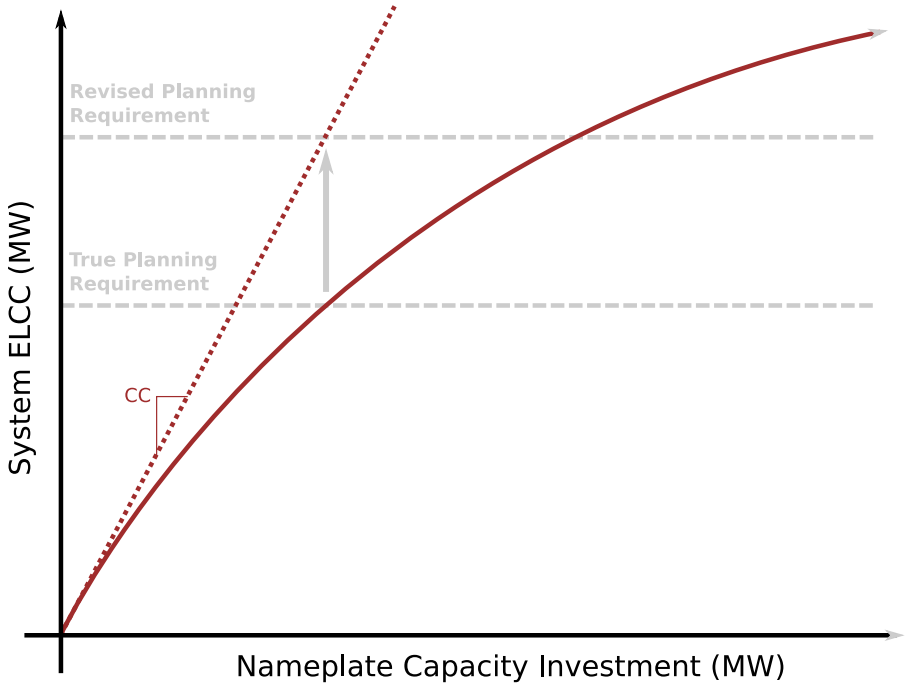

Figure 3.1: Visual depiction of the planning reserve margin augmentation approach to enforcing system adequacy. The PRM is incrementally increased until the system is driven to procure sufficient resources to serve load in all time periods.

As visualized in Figure 3.2, once the base system has been established (as above), and the initial supplemental investments have been selected, the marginal capacity credit of new resources can be recalculated in the context of the base system plus the supplemental investments. Rather than re-selecting different supplemental investments, the first tranche can be locked in (as the base system already is), and a second tranche of supplemental investments selected based on the updated capacity credits.

This process of additional investment and capacity credit recalculation can be repeated as many times as necessary until the system is able to serve the desired load level. Since the final portfolio selected is rooted in a more nuanced understanding of the evolving marginal benefit of each resource, the method can be expected to achieve system adeqacy at a lower total cost than the PRM-augmentation approach above. However, it requires repeated recalculation of resource capacity credits, which requires more computational effort than

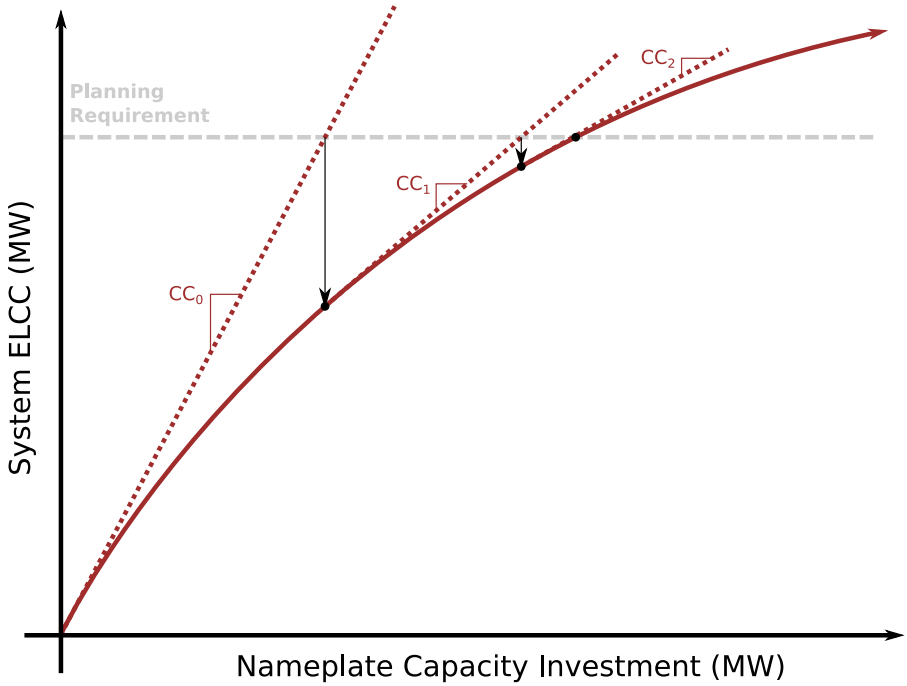

Figure 3.2: Visual depiction of the capacity credit iteration approach to enforcing system adequacy. As additional tranches of resources are added to the system, the marginal capacity credit of each resource is recalculated to inform the next tranche of procurement. The iteration continues until sufficient resources have been added to meet demand in every time period.

simply resolving the planning problem with an increased PRM.

3.2 Risk Period Iteration

Even the most sophisticated capacity credit approaches still depend on exogenously calculated linearizations of each resource class' marginal contribution to system adequacy, and disregard the nature of specific shortfalls experienced by the system, which could better inform least-cost mitigation strategies.

Ideally, a planning model would include a representation of system operations that spanned every hour of the year, such that shortfall periods could be internally identified and eliminated via additional resource investments. Unfortunately, including large numbers of

operating periods in planning problems quickly becomes computationally burdensome and is often impractical for all but the smallest of systems.

Fortunately, in practice only the most stressful periods of the planning horizon drive the need for resource adequacy investments. If those periods were known in advance they could be represented explicitly in the planning problem, yielding a resource-adequate system design from a much more compact optimization problem. In fact, this is the same logic behind the traditional planning reserve margin constraint used in planning models: when generating resources have uniform availability across the planning horizon, the period of greatest system stress is simply the period with peak load.

In systems dominated by variable and energy-limited resources, the periods of greatest system stress become a function of which resources are chosen to build: a system served primarily by solar PV will be at greatest risk overnight, while a system with large amount of wind generation will be at higher risk during still periods. Storage complicates these considerations further, as the availability of storage resources to generate at a given point in time depends on the ability of storage to charge in preceding periods. The chronological ordering of periods therefore becomes important to consider as well.

Instead of trying to predict which periods represent the greatest system risk, this chapter considers an iterative strategy to dynamically select risk periods based on an exogenous, all-hours resource adequacy evaluation of the system portfolio selected by the planning model. This ensures the final system selected by the model is resource adequate (since iteration will continue until such a condition is satisfied) while using the result of each intermediate adequacy assessment to inform the periods explicitly represented in the next re-solved planning optimization.

The specific period selection strategy used in this work is as follows:

1. Perform a full chronological resource adequacy assessment of the candidate system design to determine if it is resource adequate. If it is, use this as the final portfolio. If it isn't, determine the temporal distribution of unserved energy.

2. Add an explicit representation of the day with the most unserved energy to the planning model. If this day has already been included, add the day with the most unserved

energy that has not yet been included. If all days with unserved energy have been included, do not add a day in this step.

3. Add an explicit representation of the day *before* the day with the most unserved energy to the planning model. If this day has already been included, add the closest day prior to the max unserved energy day that has not yet been added

4. Re-solve the planning problem with these new risk days included, and return to step 1.

Note that in a system with purely diurnal storage utilization, step 3 would not be required: representing the highest risk days would be sufficient to ensure that the planning optimization yielded a resource-adequate solution. However, if energy shifting can occur over long time periods, it may occur that a stress day is preceded by a day which is more demanding than the "representative" day the planning model sees. Serving load on this day may require discharging additional storage, reducing the available energy in the subsequent stress period below what is anticipated by the planning model. Explicit representations of (potentially non-shortfall) periods preceding load dropping periods therefore become important to consider as well.

There are many alternative strategies available for choosing specific periods to augment the planning model's operations representation model. Much work remains to be done in understanding the strengths and weaknesses of different approaches and their impact on overall solution solve time and quality. In particular, the approach suggested here is a basic heuristic, and alternative strategies with stronger theoretical grounding could be particularly valuable.

3.3 Temporally-Coupled Representative Periods

A common means of representing the constraints and capabilities of energy storage in planning models is to include operational representations for a limited set of temporally-contiguous timesteps (e.g. sets of 24-hour periods grouped into "representative days"). In the presence of intertemporal constraints (such as storage state of charge or generator ramp

rates) periodic boundary conditions are typically enforced such that the system ends the period in the same state it began in. If multiple periods are used to represent system operations (for example, one representative day for each season of the year), each operating period's dispatch is considered fully independent of every other.

While using representative days with periodic boundary conditions can be an effective means of representing energy shifting capabilities in systems where storage is used exclusively diurnally, this representation cannot directly capture the potential for candidate resources to move energy across periods of time that exceed the length of the individual representative periods. This becomes problematic when planning systems with very high levels of variable renewable resources, which often depend on seasonal energy shifting to balance supply and demand across the entire year. To address this challenge, in this section we develop a "sparse" chronological storage representation that allows tracking and constraining the evolution of a storage device's charge state over an arbitrarily-large number of time periods, while only explicitly representing system operating conditions and dispatch decisions for a finite set of time periods.

3.3.1 Modeling consecutive identical dispatch days

We start with one day (24 hours, or more generally, T timesteps) of dispatch decisions to be repeated N times in a row, providing a reduced-form representation of N days of operations. The primary purpose of this is to reduce the number of dispatch decision variables from TN to just T: $p_t \; \forall t \in T$.

Typically, p_t would be constrained such that $0 \leq e_0 + \sum_{i=1..t} p_i \leq E \; \forall t \in T$, respecting the storage device's min and max state of charge constraints. However, since the dispatch variables are applied in multiple different days, each with a potentially different e_0, and e_0 is a function of dispatch decisions in previous days, it is more convenient to consider the device state of charge relative to its starting point, rather than in terms of the absolute state of charge. In particular, we can define $\Delta e = \sum_{t \in T} p_t$ as the net change in state-of-charge at the end of the day, and $\lfloor e \rfloor$ and $\lceil e \rceil$ as the net change in state-of-charge during the lowest and highest state-of-charge periods during the day. The latter two can be enforced as:

$$\lfloor e \rfloor \leq \sum_{i=1..t} p_i \ \ \forall t \in T \tag{3.1}$$

$$\lceil e \rceil \geq \sum_{i=1..t} p_i \ \ \forall t \in T \tag{3.2}$$

We can therefore formulate the following set of constraints for each of the N consecutive days, with e_0 now representing stored energy at the beginning of the entire sequence of days:

$$0 \leq e_0 + (n-1)\Delta e + \lfloor e \rfloor \ \ \forall n \in 1..N \tag{3.3}$$

$$e_0 + (n-1)\Delta e + \lceil e \rceil \leq E \ \ \forall n \in 1..N \tag{3.4}$$

When $\Delta e \geq 0$, day N will see the highest absolute state of charge, and day 1 will see the lowest. By the same logic, when $\Delta e \leq 0$, day 1 will see the highest absolute state of charge, and day N will see the lowest. We can therefore reduce the constraint set from $2T + 2N$ elements to $2T + 4$:

$$0 \leq e_0 + \lfloor e \rfloor \tag{3.5}$$

$$0 \leq e_0 + (N-1)\Delta e + \lfloor e \rfloor \tag{3.6}$$

$$e_0 + \lceil e \rceil \leq E \tag{3.7}$$

$$e_0 + (N-1)\Delta e + \lceil e \rceil \leq E \tag{3.8}$$

In the special case of $N = 1$, the constraint set further reduces to the traditional $2T$ elements.

This allows us to compactly represent state-of-charge evolution over an arbitrarily-long sequence of assumed-identical dispatch days. Next, we consider how different such sequences can be combined together to produce a sparse chronology over heterogeneous dispatch days.

3.3.2 General chronology via heterogenous day sequences

We now consider sequencing different collections of distinct days together to represent a more general period of time. We adjust our notation to represent a particular hourly dispatch decision in one of D different archetypical days as p_{td}, and the the min, max, and net state-of-charge change across an archetypical day as $\lfloor e \rfloor_d, \lceil e \rceil_d, \Delta e_d$. The overall operating period considered is divided into P partitions, where each "true" day in one partition is represented by the same archetypical day d_p. One archetypical day may be associated with multiple partitions, such that $P \geq D$. Each partition can have a unique sequence length N_p and starting state-of-charge e_{0p}.

The starting energy for a partition is determined recursively from the partition that precedes it:

$$e_{0,p+1} = e_{0,p} + N_p \Delta e_{d_p} \tag{3.9}$$

The recursion can be terminated by either exogenously specifying the device's state of charge at the beginning of the overall time horizon, or by applying periodic boundary conditions such that the starting state of charge is equal to the ending state of charge in the final chronological sequence. We adopt the latter approach in this work.

Finally, since each partition has a different starting state-of-charge, the linking constraints must be imposed on each partition individually:

$$0 \leq e_{0p} + \lfloor e \rfloor_{d_p} \forall\ p \in P \tag{3.10}$$

$$0 \leq e_{0p} + (N_p - 1)\Delta e_{d_p} + \lfloor e \rfloor_{d_p} \forall\ p \in P \tag{3.11}$$

$$e_{0p} + \lceil e \rceil_{d_p} \leq E\ \forall\ p \in P \tag{3.12}$$

$$e_{0p} + (N - 1)\Delta e_{d_p} + \lceil e \rceil_{d_p} \leq E\ \forall\ p \in P \tag{3.13}$$

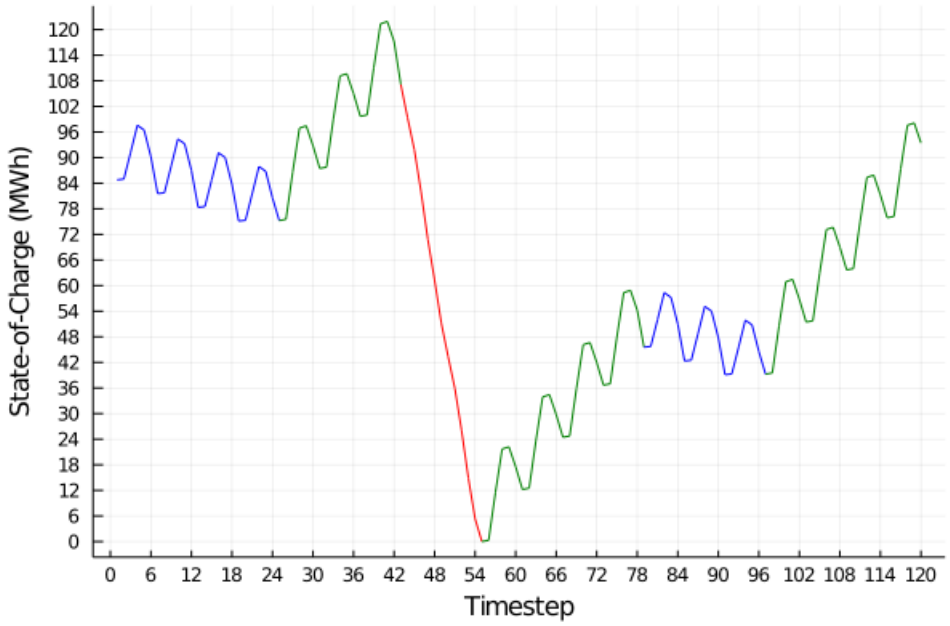

Figure 3.3: Sequential evolution of a storage device's state-of-charge over time, as represented through three representative "days" (blue, green, or red) grouped into six temporal partitions. 20 sequential "days" (each comprised of six timesteps) can be represented with only three days (18 timesteps total) of dispatch decisions, plus a set of linking variables / constraints for each partition.

Figure 3.3 provides a graphical representation of state-of-charge evolution under these constraints. In summary, we are able to represent the full-horizon state-of-charge evolution of a storage device with:

- TD dispatch variables (p_{td})
- P boundary condition energy variables (e_{0p})
- D state-of-charge evolution variables (Δe_d)
- 2D state-of-charge bounding variables ($\lceil e \rceil_d, \lfloor e \rfloor_d$)
- P boundary condition equality constrants (enforcing the definition of e_{0p})

- D state of charge evolution definitional constraints (enforcing the definition of Δe_d)
- 2TD relative minima and maxima inequality constraints (enforcing the definitions of $\lceil e \rceil_d$ and $\lfloor e \rfloor_d$)
- 4P state of charge constraints

T is the number of dispatch periods in each archetypical day (usually 24), D is the number of archetypical days considered, and P is the number of partitions considered, with $P \geq D$. This gives a total of TD + P + 3D variables and 2TD + 5P + D constraints. If a fixed intial state of charge is used for the device, making the definitional recursion for e_{p0} finite, the e_{p0} equality constraints can be directly substituted into other constraints, reducing the problem to TD + 3D variables and 2TD + 4P + D constraints.

With the typical value of T = 24, this formulation represents 27D + P variables and 49D + 5P constraints per storage device, indicating that new partitions that re-use existing archetypical days can be added with relatively low cost to model parsimony.

3.4 Storage Sizing Considerations

Storage resource investments are made in terms of both capacity and energy, which presents challenges for assigning capacity credits in a purely capacity-based reserve margin adequacy framework. Once option would be to assign one capacity credit to additional investments in storage power, and a second for storage energy, but as the two resource parameters are highly interactive these values would become inaccurate under additional investments even more quickly than other capacity credits, and imply potentially nonsensical phenomena, such as the ability to serve more load by investing in additional energy capacity when the technology's adequacy contribution has become power-constrained.

Instead, a common workaround is to define multiple resource classes for a single storage technology, representing different power-to-energy ratios separately and assigning each its own capacity credit (for example, the capacity credit of a two-hour storage device might be 25%, while a four-hour device is 50% and an eight-hour device is 100%). This discretization of power/energy durations increases the size of the planning problem (since

each duration represented requires its own set of planning and operations decision variables and constraints) and restricts the range of possible power/energy investments, potentially increasing system costs.

If the capacity credit / planning reserve margin framework is no longer necessary (as is the case when risk period iteration is used), this workaround can be eliminated and storage resources can be directly sized in terms of both power and energy, reducing the problem size and potentially improving the solution quality. To investigate the benefit of this approach in this work we consider risk period iteration using both quantized energy/power ratios and direct storage sizing, and compare the impact on solution quality and solve time.

3.5 Empirical Tests

3.5.1 Adequacy Formulation Comparisons

To explore the practical implications of these alternate resource adequacy constraints, we apply them here to an expansion problem based on operational data from the IEEE RTS-GMLC test system [28], with annualized technology investment costs derived from the NREL Annual Technology Baseline [29]. The capacity expansion problem is tasked with selecting a least-cost portfolio of wind, solar PV, run-of-river hydro, and lithium ion battery storage capable of fully serving a given load profile. For simplicity we assume all problem parameters to be fully deterministic (eliminating the need for probabilistic adequacy assessment).

Three different styles of resource adequacy constraints are considered: the planning reserve margin augmentation approach described in Section 3.1.1, the capacity credit iteration method described in Section 3.1.2, and the risk period iteration approach described in Section 3.2. A fourth method that uses risk period iteration and directly specifies the battery storage power and energy characteristics, as discussed in Section 3.4, is also tested. The sparse chronological linking constraints described in Section 3.3 are applied in all cases.

Each method is assessed using seven choices of representative periods:

1. One representative day for each season (4 total)

2. Two representative days (weekday + weekend) for each season (8 total)

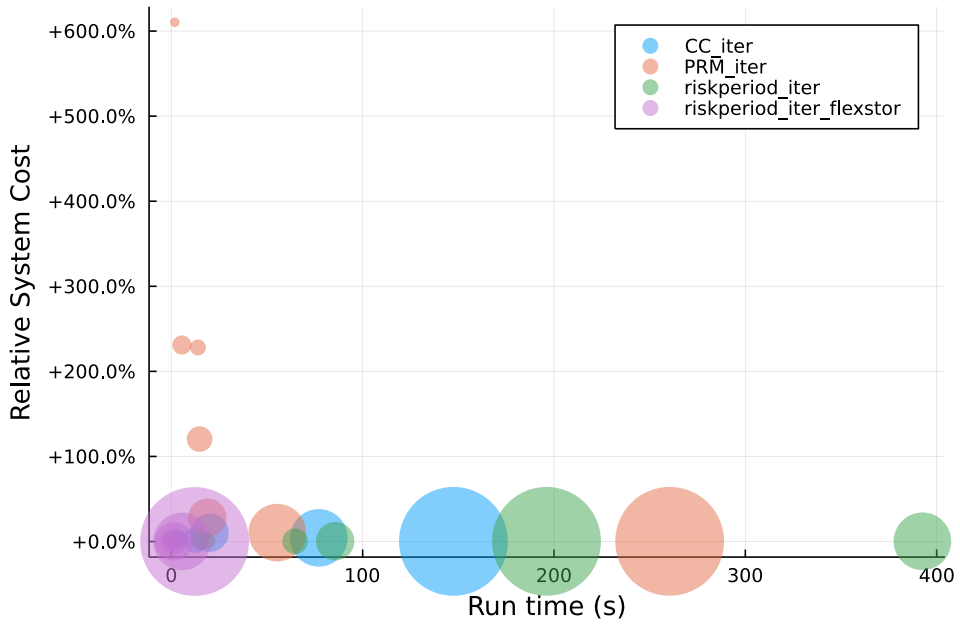

Figure 3.4: Solution quality vs solve time for each of the four problem formulations considered. Marker size is proportional to the number of initial representative periods used in that solution.

3. One representative day for each month (12 total)

4. Two representative days (weekday + weekend) for each month (24 total)

5. One representative day for each week (52 total)

6. Two representative days for each week (104 total)

7. Explicit representation of every day of the year (365 total)

In the risk period iteration cases, actual data for the selected risk days replaces the representative data initially assigned to the days in question.

Figure 3.4 compares the performance of the four different methods in terms of both time to develop a final solution and the quality of that solution, represented as system cost relative to the lowest-cost solution identified. The marker sizes are proportional to

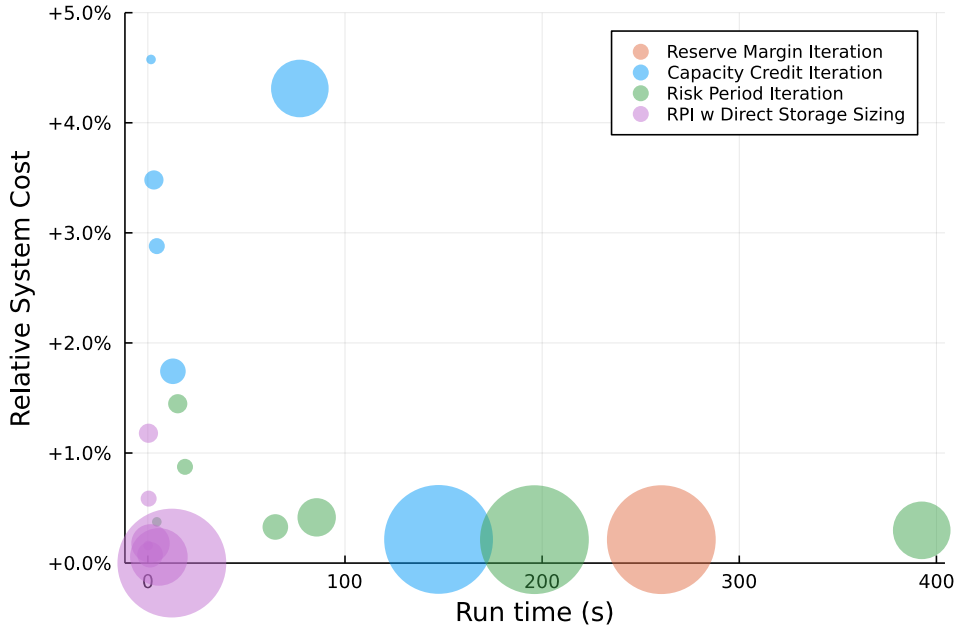

Figure 3.5: Solution quality vs solve time for each of the four problem formulations considered, limited to outcomes within 5% of the lowest-cost solution. Marker size is proportional to the number of initial representative periods used in that solution.

the number of representative days used in each case. As expected, the PRM augmentation method (represented in orange) solves relatively quickly, but the quality of the identified solution is often very poor, given the large quantity of resources that must be added to serve all demand and the method's lack of accurate insight into the marginal value of those resources.

If we focus only on results within 5% of the lowest-cost solution (as shown in Figure 3.5), we eliminate all but one of the PRM augmentation results, as well as one of the capacity credit iteration results. The relationship between the other three methods tested becomes more clear: the capacity credit iteration method, while much better performing than the PRM augmentation approach, remains dependent on coarser approximations of the true adequacy contribution of candidate resources, and so is unable to deploy them efficiently, resulting in higher system costs.

The risk period iteration approaches yield the highest quality and most consistent so-

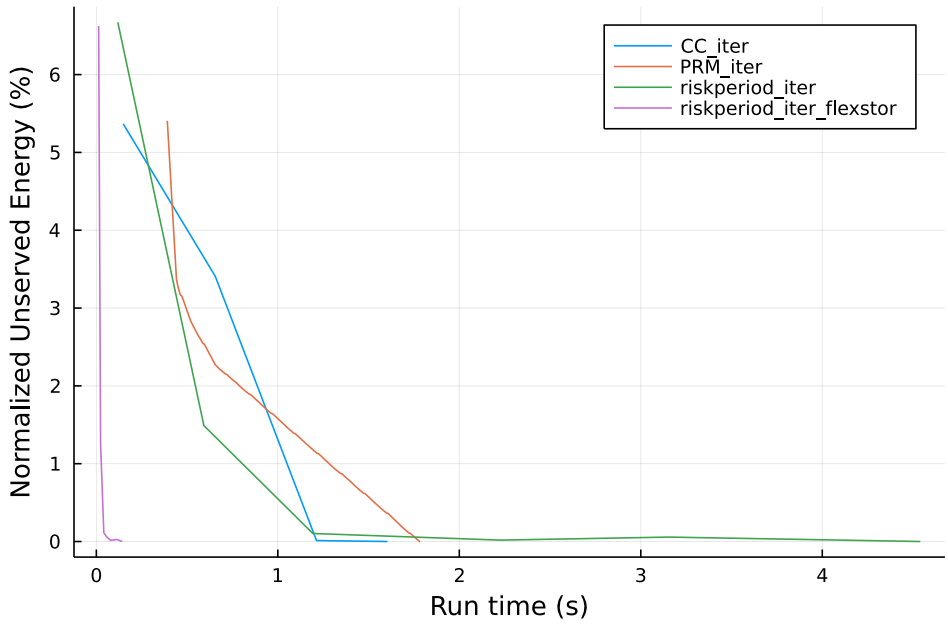

Figure 3.6: Adequacy improvements as a function of iteration time for each of the four problem formulations considered, using one representative day per season.

lutions across different choices of representative periods, although this comes with higher computational costs, as they require continuously increasing the size of the optimization problem to be solved. Fortunately, direct selection of the storage size parameters eliminates the need for representing multiple alternative classes of storage technology, reducing the necessary solve time below even the PRM-based methods while simultaneously improving its solution quality, allowing this approach to dominate the alternatives.

Finally, we can consider the breakdown of where time is spent in developing each solution. Figure 3.7 depicts the process iterating between solving the capacity expansion optimization problem (blue) and resource adequacy assessment and capacity credit calculations (orange). All results are based on using four seasonal representative days.

The PRM augmentation approach takes many iterations to find an appropriate PRM value, but as the optimization re-solves are trivial (simply acquire more of the resource that provides the lowest-cost capacity contribution) and no new capacity credit calculations are needed, each iteration can be performed very quickly.

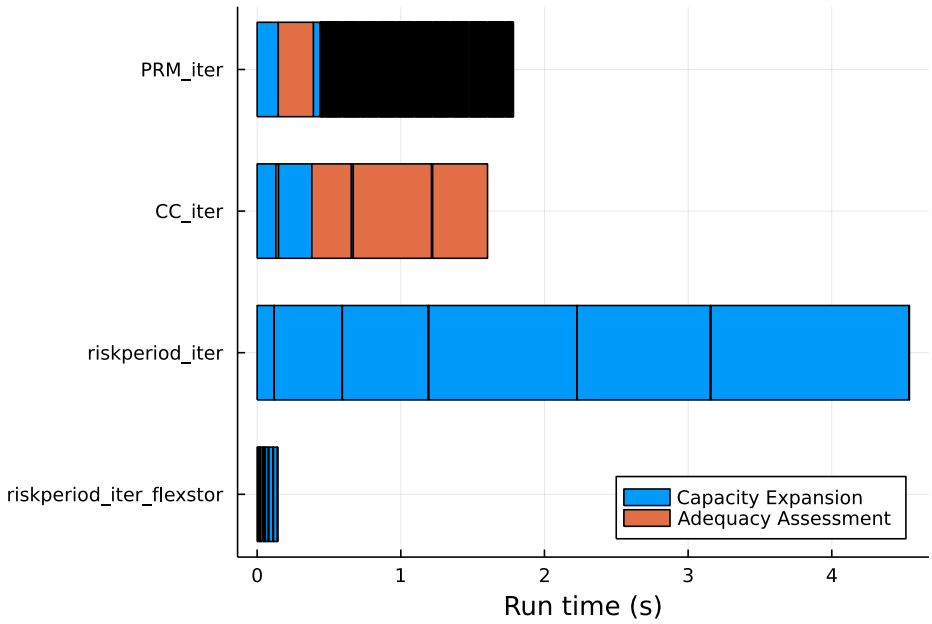

Figure 3.7: Iteration progress though time for each of the four problem formulations considered, using one representative day per season.

Conversely, the capacity credit iteration approach requires recalculating capacity credits for each resource class at every step, which dominates the solve time as the ELCC-based capacity credit calculation applied here involves repeatedly solving system adequacy assessments, while the optimization re-solves remain trivial. Alternative capacity credit heuristics could greatly reduce the overall solution time required, but any inaccuracies in those values would also reduce the already-suboptimal quality of solutions from this method.

The risk period iteration method does not require calculating capacity credits, and so is dominated by optimization problem solve time. When multiple storage classes are required to represent alternate power/energy ratios, this method is the slowest of the four considered, but when the problem is simplified by direct storage sizing, the problem solves the fastest by a large margin. It should also be noted that the process of modifying the optimization problem currently discards previous solution data for the sparse chronology linking constraints, so these solve times (particularly for the fixed power/energy ratio storage case) could be further improved in the future.

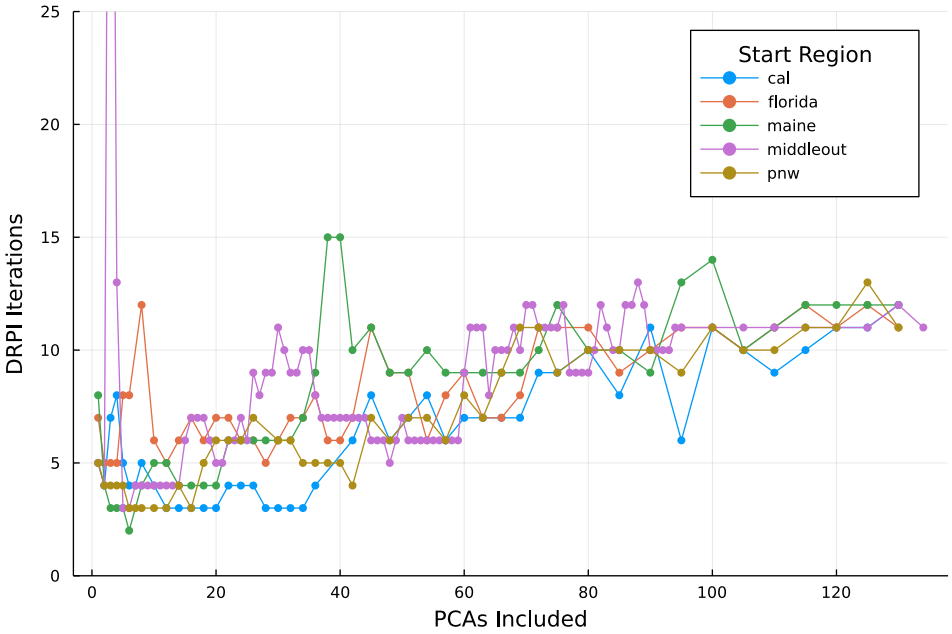

Figure 3.8: Number of dynamic risk period identification iterations required to achieve a resource adequate system, as a function of the number of geographic regions (ReEDS "PCAs") represented in the optimization. The systems tested expand outwards from five initial regions in different parts of the contiguous United States.

3.5.2 DRPI Scaling under Diffuse Risk

While dynamic risk period identification presents promising advantages over traditional capacity-based planning frameworks, it may lack efficiency when considering temporally-diffuse risk distributions. In a system where the entire grid experiences shortfall risk at the same time, endogenizing that risk in the optimization problem only requires adding one new risk period, providing a substantial computational advantage. However, in the pathological case that risk is broadly distributed across many time periods and driven by diverse underlying conditions, a very large number of risk periods may need to be identified and included in order to expose all of the system's failure modes to the optimzation (such that they can be appropriately mitigated).

In particular, multi-region, geographically-expansive power systems may suffer from such decorrelated risks. As the modeled grid extends to encompass multiple climates and time

zones, the likelihood that different parts of the system experience shortfall risk at different time of day and year increases substantially. With dynamic risk period identification, discovering and includings all of these time periods involves repeatedly solving an increasingly large optimization problem, which could become computationally infeasible.

To investigate the practical implications of such temporally-diffuse risk, interregional transmission constraints and seven years of regional wind, solar, and load profiles were extracted from NREL's ReEDS model [8] and used to parametrize a planning problem formulated with the DRPI approach. ReEDS models 134 regions spanning the entirety of the contiguous United States, and so its multi-year dataset captures diverse adequacy risks across summer- and winter-peaking regions, weather systems, climate trends, and time zones.

To understand how DRPI iteration requirements might evolve as a function of increasing risk diversity, five model regions across the US were selected for solving as single-region systems. Neighbouring regions to each were incrementally added to form increasingly-large problems and study how the number of iterations needed to design a resource adequate system may increase with geographic diversity.

Figure 3.8 shows that increasing the number of regions considered generally required more iterations to solve, as expected. However, this increase is relatively insensitive to system size, with the iteration count only growing by a factor of 2-3 while the geographic scope and region count rose by multiple orders of magnitude. This suggests that, even though risk drivers do become more diverse at larger geographic scales, important correlations remain such that adding a risk period based on shortfalls in one region still provides much of the information necessary to help mitigate similar risks elsewhere in the system.

3.6 Conclusion

This work has compared solution quality and solve time for four alternative problem formulations for selecting a resource-adequate, least-cost generation portfolio comprised exclusively of variable renewable resources and storage. Dynamic risk period identification consistently identified lower-cost systems than the two alternatives using planning reserve margins and capacity credits to enforce system adequacy. This technique also enabled siz-

ing storage power and energy parameters directly, which further reduced system costs while also dropping solution time below that of the capacity credit-based techniques.

Further tests on a large national-scale dataset suggested that the number of iterations required to achieve a resource adequate system scales well with increasing system size, even as greater geographic diversity decorrelates the periods in which different regions of the system experience shortfall risk.

While these results are promising, they also suggest a number of potential areas for future research. This work selected the initial representative periods used by the model somewhat arbitrarily: a more data-driven approach to period selection may better capture more diverse operating conditions in fewer periods, potentially reducing problem size and/or the number of risk period iterations required. In systems with non-zero marginal cost resources, this would also improve the model's internal estimate of system operating costs. The computational efficiencies associated with grouping temporally-contiguous periods in the sparse storage chronology representation would also need to be accounted for in any such process.

Finally, the procedure by which specific days were identified as risk periods and included in the planning model was also somewhat arbitrary, and could likely be improved. Other information beyond the timing and magnitude of shortfall events could be incorporated in the decision process, such as the timing and magnitude of system surplus, and the nature of constraints on energy storage during shortfall events (i.e., whether the storage is energy-limited or capacity-limited).

Chapter 4

REPRESENTING PROBABILISTIC RISK IN DETERMINISTIC OPTIMIZATION VIA RISK CURVES

CEMs have traditionally only considered deterministic adequacy heuristics such as planning reserve margins (PRMs), with modern resource adequacy checks on the results done (if at all) using external Monte Carlo simulation tools [30] to calculate probabilistic risk metrics such as expected unserved energy (EUE), loss-of-load probability (LOLP), and loss-of-load expectation (LOLE) [7, 6]. While Section 2.3 and other academic research has considered stochastic optimization approaches to endogenize such probabilistic criteria within the planning optimization [31, 32, 33], these formulations can be computationally impractical to solve for large-scale systems.

As discussed in Chapters 2 and 3, a more pragmatic alternative may be to iterate between capacity expansion and resource adequacy steps, either adjusting deterministic criteria based on adequacy outcomes and re-optimizing the system design [34], or building capacity in smaller tranches until adequacy criteria are met [35], eschewing an optimized least-cost solution.

Section 4.1 of this chapter elaborates further on the traditional PRM approach and describes a more modern energy-based alternative. Section 4.2 introduces a new adequacy constraint formulation to approximately enforce probabilistic criteria within a deterministic optimization framework. Section 4.4 describes the design and outcomes of an empirical case study testing this formulation, and Section 4.5 provides concluding remarks, including opportunities for future work.

4.1 Energy Reserve Margins

In traditional capacity expansion models, resource adequacy requirements are enforced though planning reserve margin (PRM) constraints, where the sum of capacity contributions from all resources in a region are required to exceed expected peak load by some

ratio (the exogenously-specified PRM). While conceptually-simple when estimating system adequacy in the context of uniformly available thermal generation and no transmission constraints, this approach requires significant additional complexity to satisfactorily represent transmission congestion and emergent interaction effects between variable renewable and energy-limited resources.

More recently, energy reserve margin (ERM) constraints have developed as an alternative CEM adequacy framework to more directly capture the contributions of wind, solar, storage, and transmission resources. In this approach, a parallel, non-economic, chronological storage and transmission dispatch is considered directly within the model, requiring that those resources could be operated to distribute a minimum level of energy-backed[1] capacity surplus to every region and time period in the planning horizon. Resource contributions in every timestep are considered directly in the model, eliminating the need for capacity accreditation. Variations of ERM constraints have been applied in a number of different planning exercises studying high-renewable futures [25, 36].

While ERMs provide a more elegant heuristic than PRMs for system adequacy when planning systems with high levels of variable and/or energy-limited resources, the size of the margin must still be specified exogenously. Since the ERM required to produce a desired probabilistic adequacy outcome will vary as a function of system size and portfolio composition, this value is typically selected based on a parameter sweep or bisection which requires repeatedly re-solving the optimization problem [37] to find an ERM level that will meet stated probabilistic adequacy criteria without increasing system costs more than necessary.

4.2 Endogenous EUE Estimation

In this chapter, we develop an alternative resource adequacy constraint formulation for capacity expansion models that captures the same spatiotemporal energy shifting dynamics as the ERM approach, but eliminates the need to specify an exogenous reserve margin

[1]This is an important distinction relative to normal operating reserve requirements, in which energy-limited resources are not actually called upon and so can contribute reserve capacity in more time periods than they would have energy to deliver.

parameter. This is accomplished by incorporating an endogenous estimate of shortfall risk as a function of energy surplus available in each hour and time period, based on a prior probabilistic risk assessment of the study system. This endogenous estimate can be used to directly formulate a constraint on the probabilistic risk faced by the system.

4.2.1 Estimating Incremental Capacity Impacts

The first step in the process is to perform a probabilistic adequacy assessment based on some initial assumption about the final system design. Of course, since the capacity expansion has not yet been run, there can be no expectation that this assumption is accurate, only that it be sufficient to approximately identify distributions of risk across regions and periods. When expanding an existing system, this approximated representation may simply be the existing system with expected load growth or generator retirements applied, but with no new capacity investments in place. In "greenfield" expansion scenarios with no existing resource portfolio, this approximation could be a manually-specified system design, or the result of bootstrapping an initial capacity expansion under an alternate adequacy framework (including possibly no adequacy framework at all).

This assessment will yield probabilistic shortfall and surplus samples for each region and period, which can be used to calculate spatiotemporally-resolved loss-of-load probabilities, and estimate how those probabilities would have been different had capacity been added or removed from that region and period. For example, if the probabilistic simulation involved 100 Monte Carlo replications, and in one region and time period two of those replications experienced shortfall while the remaining 98 experienced surplus, that region-period would have an LOLP of 2%. If the shortfall magnitudes observed were 100 MW and 150 MW, then we can estimate that making 50 MW of surplus available in that region-period would not change LOLP, but that making between 100 and 150 MW available would reduce LOLP to 1%, eliminating the smallest shortfall sample but not the second-smallest.

The same logic applies to surplus samples. Continuing from the previous example, if the smallest-magnitude sample in the 98-element surplus set is 140 MW, we can estimate that for the first 140 MW of surplus removed (either by reducing capacity investments in

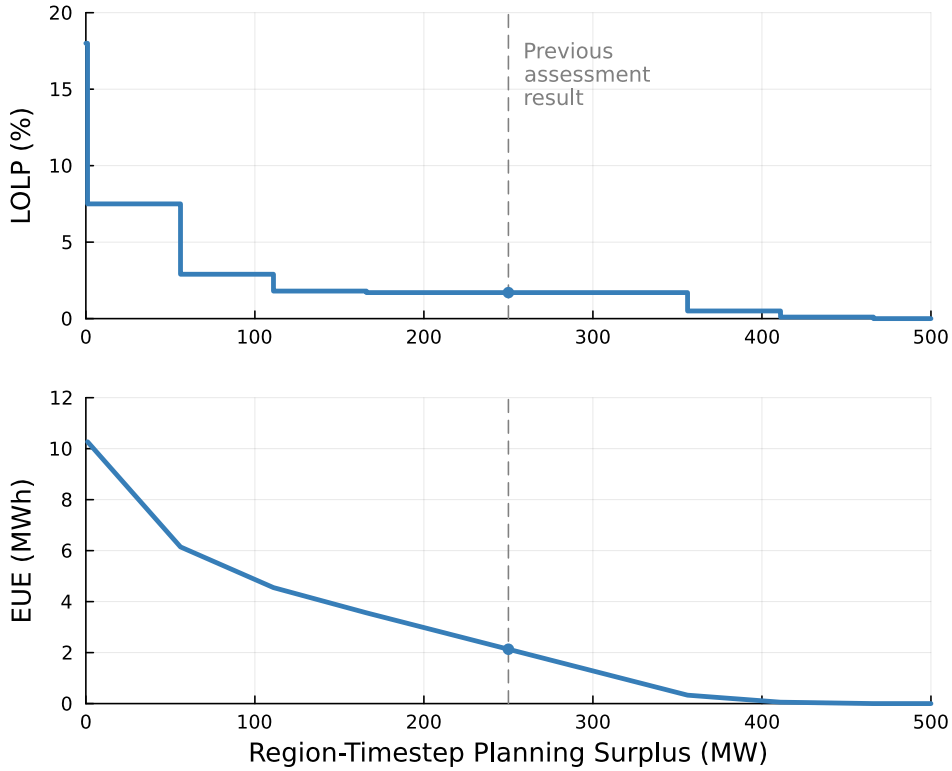

Figure 4.1: Region- and period-specific LOLP and EUE estimates as functions of energy-backed expected surplus.

that region-period, or shifting the surplus energy to another region-period via transmission and storage), the LOLP will remain at 2%. If the next-smallest surplus is 200MW, we can estimate that removing between 140 and 200 MW would increase region-period LOLP to 3%.

Once the approximate adequacy impact of incremental capacity additions or removals has been established for each region and timestep, relative to some initial system assumption, we can formulate an estimator function for region-period LOLP as a function of absolute expected energy surplus in that region-period. To do this, we simply calculate the region-period's expected energy surplus in the initial design, and apply this value as an offset to the relative capacity changes discussed above. To continue the first example above, if the initial system had a surplus of 250 MW in a region-period, corresponding to a 2% LOLP

with 100 MW as the smallest observed shortfall and 140 MW as the smallest observed surplus out of 100 samples, we could estimate that increasing the surplus to 350 MW would decrease LOLP to 1%, while decreasing surplus to 110 MW would increase LOLP to 3%. An empirical example of an LOLP estimator that roughly aligns with these values is shown in the top panel of Figure 4.1.

4.2.2 Internal EUE Approximation

The LOLP estimator derived above is non-convex and therefore not ideal for direct embedding in a mathematical optimization framework. However, as LOLP is the negative derivative of EUE with respect to incremental capacity in a region-period [38], this stepwise, monotonically-decreasing function can be integrated to obtain an estimator function for EUE that is both convex and piecewise-linear, making it highly amenable to inclusion in a linear optimization framework. In this case the constant of integration should be chosen such that the LOLP and EUE functions reach zero at the same surplus level, which is equivalent to enforcing that the EUE estimator reproduces the observed EUE level of the initial system assumption. An empirical example of such an EUE estimator is shown in the bottom panel of Figure 4.1.

We can now endogenize EUE in a linear optimization framework by defining it as a decision variable U_{rt} dependent on energy surplus X_{rt} for each region and time period. Since lower EUE values are more desirable in our expansion problem and the estimator is a convex function, it is sufficient to lower bound U_{rt} with every line making up the piecewise-linear estimator function:

$$U_{rt} \geq U^{\max}_{rtm} - U'_{rtm} X_{rt} \;\; \forall \;\; r \in R, t \in T, m \in M_{rt} \tag{4.1}$$

where U'_{rtm} and $U^{\max}_{rtm}$ are the (negative) slope and y-intercept, respectively, of the individual line segments m in the set M_{rt} of line segments. A complete description of the set notation used here is provided in the appendix.

With U_{rt} defined in this way, and by the linearity of expectation, we can directly enforce upper bounds on arbitrary aggregations of estimated region-timestep EUEs. One basic

example would be to impose regional EUE limits $U_r^{\max}$ as:

$$\sum_t U_{rt} \leq U_r^{\max} \ \forall \ r \in R \tag{4.2}$$

4.2.3 Iterative Re-solves

Since the LOLP and EUE functions derived above only provide local approximations of system adequacy impacts, their accuracies may be improved by recalculating the functions based on an enhanced initial system approximation. In particular, there may be value in re-solving the optimization multiple times, each time using the solution from the previous step as the basis for regenerating the adequacy estimator. For example, if the initial result is overly adequate, reparametrizing the model with the results of that assessment will provide the optimization with signals about potential unnecessary surplus and provide a principled means of targeting a leaner system design.

However, it should be noted that in the presence of discrete thermal generator investments subject to discrete forced outages, neither the cost optimization nor the ensuring adequacy assessment are continuous functions, and as such there can be no expectation that repeated re-optimizations will lead the solution to improve monotonically nor converge to some stable fixed point. It is therefore challenging to define a clear stopping criteria for such an iterative process.

4.3 Capacity Expansion Problem Formulation

To test the endogenous EUE approach outlined above relative to an ERM-based adequacy framework, we implement a simple capacity expansion model as follows. $g \in G$, $s \in S$, $r \in R$, $i \in I$, and $t \in T$ represent elements and sets of generator classes (technologies x regions), storage classes (technologies x regions), regions, transmission interfaces between regions, and operating time periods, respectively. We seek to minimize the sum of annualized capital costs and hourly operating costs incurred by generating capacity investment decisions N_g, storage capacity and energy investment decisions P_s and E_s, and generator operating decisions p_{gt}, as follows:

$$\min \sum_{g} C_g^c N_g + \sum_{s} (C_s^c P_s + C_s^e E_s) + \sum_{g,t} C_g^o p_{gt} \tag{4.3}$$

Here, C_g^c, C_s^c, and C_s^e, are capital cost parameters corresponding to per-unit generator costs, per-MW storage capacity costs, and per-MWh storage energy costs. C_g^o is a generator's per-MWh operating cost. Generator investment and dispatch constraints are defined as:

$$0 \leq N_g \leq N_g^{\max} \ \forall \ g \in G \tag{4.4}$$

$$0 \leq p_{gt} \leq N_g P_{gt} \ \forall \ g \in G, t \in T \tag{4.5}$$

where $N_g^{\max}$ is an upper bound on the number of generating units of class g that can be constructed, and parameter P_{gt} is the average capacity availability of a single unit of class g at time t. For thermal generating units, we additionally impose that $N_g \in \mathbb{Z}$, while non-thermal units (e.g., variable renewables) can be sized in continuous increments.

Storage resource investments and operations are constrained as:

$$0 \leq P_s \leq P_s^{\max} \ \forall \ s \in S \tag{4.6}$$

$$0 \leq E_s \leq E_s^{\max} \ \forall \ s \in S \tag{4.7}$$

$$-P_s \leq p_{st} \leq P_s \ \forall \ s \in S, t \in T \tag{4.8}$$

$$0 \leq e_{st} \leq E_s \ \forall \ s \in S, t \in T \tag{4.9}$$

$$e_{st+1} = e_{st} - p_{st} \ \forall \ s \in S, t \in T \tag{4.10}$$

where P_s and E_s represent storage power and energy capacity investments, parameters $P_s^{\max}$ and $E_s^{\max}$ are upper bounds on those values, p_{st} represents storage dispatch (positive

for discharging and negative for charging), and e_{st} tracks a storage class' aggregate state of charge. In this work we set $e_{s0} := 0 \ \forall \ s \in S$, although periodic boundary conditions across the full operating horizon could be applied instead.

The model uses a zonal transmission network representation, limiting power flows between regions but not within them. Interregional power flows f_{it} are constrained as:

$$-F_i \leq f_{it} \leq F_i \ \forall \ i \in I, t \in T \tag{4.11}$$

where parameter F_i denotes the interface flow limit. The power balance constraint for each region is then given as follows:

$$\sum_{g \in \text{gens}(r)} p_{gt} + \sum_{s \in \text{stors}(r)} p_{st} - \sum_{i \in \text{from}(r)} f_{it} + \sum_{i \in \text{to}(r)} f_{it} = L_{rt} \quad \forall \ r \in R, t \in T \tag{4.12}$$

where L_{rt} is the electrical load in region r and period t, gens(r) and stors(r) are the sets of generator and storage classes associated with region r, and from(r) and to(r) are the sets of interfaces originating and terminating in region r. To capture storage and transmission impacts on system adequacy, we define a parallel set of "reliability dispatch" decision variables $\tilde{p}_{st}$, $\tilde{e}_{st}$, and $\tilde{f}_{it}$, with corresponding constraints matching (4.6)–(4.11). We can then calculate an expected energy-backed capacity surplus X_{rt} for each region and time period as:

$$X_{rt} = \sum_{g \in \text{gens}(r)} P_{gt} + \sum_{s \in \text{stors}(r)} \tilde{p}_{st} - \sum_{i \in \text{from}(r)} \tilde{f}_{it} + \sum_{i \in \text{to}(r)} \tilde{f}_{it} - L_{rt} \quad \forall \ r \in R, t \in T \tag{4.13}$$

We can now exogenously define an energy reserve margin $X_{rt}^{\min}$ for each region and timestep as

$$X_{rt} \geq X_{rt}^{\min} \ \forall \ r \in R, t \in T \tag{4.14}$$

and tune this parameter to enforce system adequacy. Alternatively, with X_{rt} defined, we can define an EUE estimate variable U_{rt} and constrain it with the relevant piecewise estimator function as:

$$U_{rt} \geq U_{rtm}^{\max} - U'_{rtm} X_{rt} \ \forall \ r \in R, t \in T, m \in M_{rt} \tag{4.15}$$

where U'_{rtm} and $U_{rtm}^{\max}$ are the (negative) slope and y-intercept, respectively, of the individual line segments m in the set M_{rt} making up the piecewise-linear convex estimator function $U_{rt}(X_{rt})$. In theory, m could be as large as the number of Monte Carlo samples included in the preceding adequacy assessment step, but since many samples often result in the same adequacy outcome, in practice it is far smaller.

Finally, we can endogenously enforce regional EUE limits $U_r^{\max}$ as:

$$\sum_t U_{rt} \leq U_r^{\max} \ \forall \ r \in R \tag{4.16}$$

These limits can be specified directly, or calculated from an NEUE target and total demand in region r.

4.4 Empirical Case Study

We parameterize the capacity expansion model described above with data derived from the RTS-GMLC test system [28]. Specifically, we define three regions with load time series corresponding to the three RTS regions, and allow power transfers between these regions based on the sum of thermal flow limits on interregional lines.

For simplicity, we aggregate plant-level wind and solar generation profiles by region and normalize these to a single regional availability time series for each variable resource. We also allow the model to build discrete combined-cycle and combustion turbine gas units with parameters for production costs, unit sizes, and reliability statistics taken from the

RTS dataset. Annualized investment costs for all of these resources, as well as lithium-ion battery storage, are taken from NREL's 2023 Annual Technology Baseline [29], considering both capital as well as fixed operations and maintenance costs.

We solve the capacity expansion model under three different kinds of resource adequacy constraints (or lack thereof). In each case, sample-level, spatiotemporally-resolved probabilistic adequacy results are obtained using NREL's PRAS tool [39] to perform 1000 chronological Monte Carlo simulations considering stochastic thermal generator outages and the same zonal transmission representation as the CEM.

In the first case, "Expectation Only", the system only needs to meet hourly system demand based on expected unit availabilities (hourly time series for wind and solar resources, and unforced capacity for gas generators). In the second case, "Lowest Adequate ERM", we identify the lowest energy reserve margin requirement that yields a system meeting a 1 part-per-million (ppm) normalized EUE (NEUE) adequacy criterion in each region. In the third, "Endogenous EUE", we use the adequacy assessment results from the Expectation Only buildout to parameterize the internal EUE estimator function, and endogenously enforce a 1 ppm NEUE constraint for each region. We report both the initial result obtained, as well as the best result after performing iterative re-solves. System cost and adequacy outcomes for these four different capacity expansion results are summarized in Table 4.1, with technology-level investment breakdowns visualized in Figure 4.2.

4.4.1 Expectation Only (no explicit adequacy constraints)

As expected, running the optimization such that load only needs to be served on an expected value basis, equivalent to enforcing a 0% hourly energy reserve margin, results in less resource investment overall and so yields the lowest system levelized cost of electricity (LCOE), at 38.5 $/MWh. However, the probabilistic adequacy risk associated with this design is orders of magnitude larger than the stated 1 ppm NEUE target. LOLE is also many times larger than commonly-accepted risk thresholds, which are often on the order of 1-3 event-hours per year [7].

This outcome reiterates the danger of only planning a system against average resource

Table 4.1: Test System Costs and Adequacy Outcomes

	Expectation Only	Lowest Adequate ERM	Endog. EUE (Initial)	Endog. EUE (Final)
Optimizations	1	12 (+1)	1 (+1)	3 (+1)
LCOE ($/MWh)	38.5	40.4	40.4	40.2
NEUE (ppm)	76 ± 2	0.21 ± 0.08	0.22 ± 0.07	0.4 ± 0.1
LOLE (event-h/y)	16.9 ± 0.4	0.08 ± 0.02	0.07 ± 0.02	0.12 ± 0.02

availabilities, even if hourly time series and chronological operations are considered. Deterministic capacity expansion exercises should always include probabilistic assessments on outcomes, even if only in post-processing, to more fully account for stochastic phenomena that are not considered in the optimization process. In this case, unmodeled stochasticity in gas plant forced outages resulted in unacceptable adequacy outcomes, despite the system appearing to never experience shortfall when considering average availability factors applied to those units.

4.4.2 Lowest Adequate ERM

Once the Expectation Only case indicated that a 0% ERM was insufficient to meet the stated probabilistic adequacy criteria, the hourly ERM requirement was steadily increased in 1% increments until the optimization produced a solution with acceptable shortfall risk. For this test system, a 12% ERM was necessary to achieve this outcome. This yielded a solution with increased capacity investments and thus a larger system LCOE (40.4 $/MWh) than the Expectation Only case.

The ERM search required for this approach involves solving the expansion problem under many different reserve margin levels, in order to find the choice that maps to desired

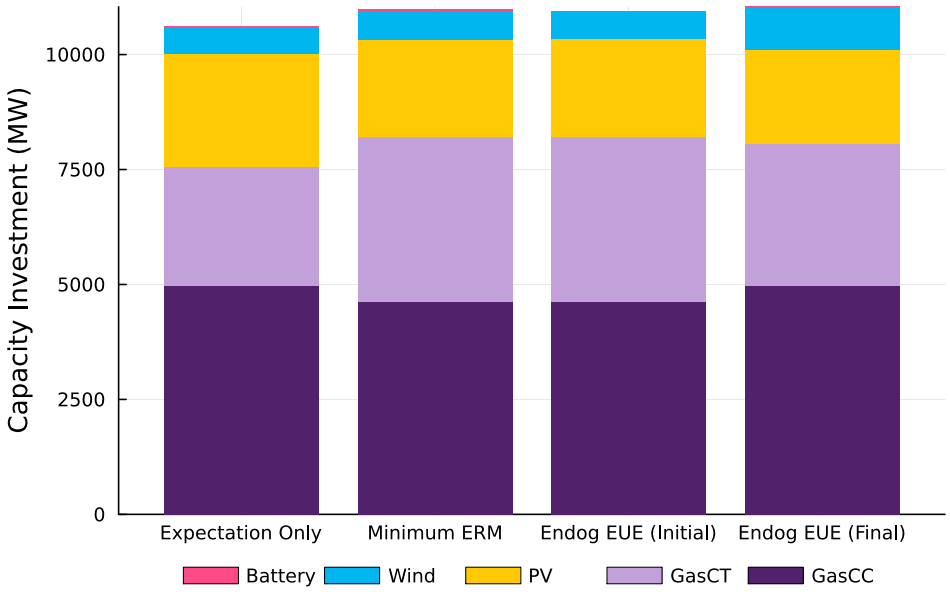

Figure 4.2: Technology investment decisions for each adequacy constraint method.

probabilistic outcomes. While these solves could be parallelized by speculatively solving the model under many ERM levels at once and choosing the lowest that meets adequacy criteria, this requires substantial computational resources. In this case, the optimizations were performed sequentially on a single machine, and so required 12 serial optimization solves to reach the final result (not including the original Expectation Only / 0% ERM solve). Given the discrete thermal build decisions involved, there is no guarantee that system adequacy will improve monotonically with increasing ERM, complicating the potential use of bisection strategies to reduce the number of optimizations required.

4.4.3 Endogenous EUE

As this comparison was a greenfield capacity expansion exercise, the endogenous EUE method's internal estimator functions were bootstrapped using the adequacy assessment results from the (inadequate) Expectation Only case. With these parameters in place, the very first endogenous EUE solve was able to produce a system design that was near-identical

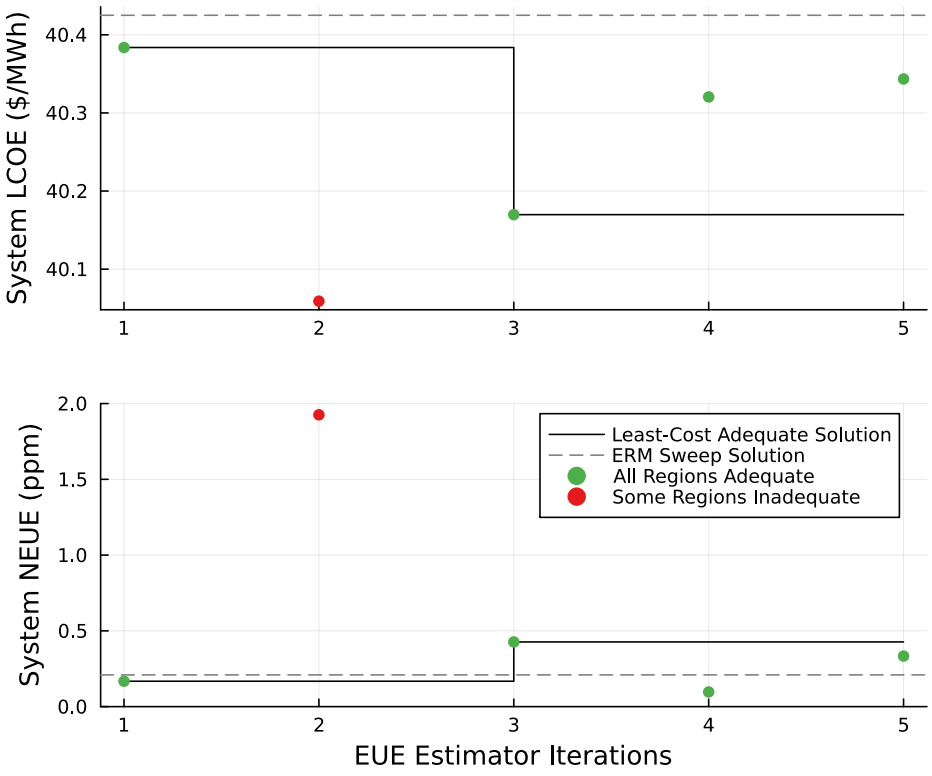

Figure 4.3: System cost and reliability across iterative re-solves, under evolving parameterizations of the endogenous EUE estimator.

to the 12% ERM result.

Since this solution yielded an NEUE value (0.2 ppm) that was still comfortably within the adequacy criteria, ten iterative resolves were performed in an attempt to find a lower cost solution that still delivered acceptable adequacy outcomes. System cost and NEUE outcomes for the first five of these iterations are shown in Figure 4.3. The third iteration was able to reduce the system cost by 0.5% while only increasing NEUE to 0.4 ppm, and no subsequent resolve improved this further. This suggests that, at least for this specific system and bootstrap condition, the initial estimator parameters were not particularly inaccurate relative to subsequently-available values, limiting the potential of iterative resolves to improve solution quality.

4.5 Conclusion

This chapter has demonstrated that ignoring stochastic generator availability in capacity expansion modeling introduces the potential for orders of magnitude more shortfall risk than desired, even if the system appears adequate on an hourly expected value basis. This underscores the importance of both enforcing supplementary resource adequacy constraints inside capacity expansion models and performing probabilistic resource adequacy checks to verify the reliability of deterministic capacity expansion model solutions.

To better address the need for adequacy-aware capacity planning, this work has developed a novel approach to endogenously estimate probabilistic resource adequacy outcomes within a computationally-tractable deterministic optimization framework. These endogenous EUE constraints achieve similar cost and reliability outcomes to existing approaches based on reserve margin parameter sweeps, but only require a single optimization solve. Iterative refinements to the linear EUE estimator function have potential to further improve solution quality, although empirical results from one specific test system suggest the incremental benefit may be limited.

There are a number of possible enhancements to this method that merit future consideration. While this work developed piecewise-linear EUE estimator functions that encode exact shortfall/surplus outcomes from a prior resource adequacy assessment, approximation functions based on fewer line segments may be able to reduce problem size without substantially impacting the accuracy of estimated adequacy outcomes. This method could also be augmented with techniques developed in [33] to procure surplus in terms of standard deviations above expected availability, rather than absolute energy margins, in order to account for changes in surplus variance associated with alternate thermal generator investments. It may also be possible to incorporate endogenous risk estimator functions for other (non-EUE) metrics, although such functions would necessarily be less-accurate approximations due to the non-convex nature of most alternate metrics of interest (loss-of-load events, loss-of-load hours, loss-of-load days, etc).

This method likely also presents significant complementarities with iterative stress period identification methods for problem size reduction [40], where shortfall and surplus data

from the previous iteration's adequacy assessment could be used to parameterize the next optimization's EUE estimator. Chapter 5 will consider this further. Future work should also assess this method's performance on larger systems with more diverse resource characteristics and long-range transmission congestion, which often complicates regional reserve margin tuning - this method's more direct approach may be disproportionately beneficial in such situations.

Finally, this work has focused solely on using the endogenous risk estimate to enforce an engineering design constraint. Taken with the system's value of lost load, the EUE estimate could alternatively be incorporated into the problem's objective function to more directly balance the societal cost of additional investments against the benefits of reduced shortfall risk.

Chapter 5

A UNIFIED ITERATIVE FRAMEWORK FOR MODERN POWER SYSTEM PLANNING

Chapter 2 of this dissertation considered multiple techniques for embedding probabilistic considerations in capacity expansion optimization problems, and concluded that iterating with a dedicated resource adequacy model represented a pragmatic and computationally efficient solution. Chapters 3 and 4 then built on this conclusion to develop techniques to better represent variability and spatial coupling (Section 3.2), temporal coupling (Section 3.3), and uncertainty (Section 4.2) in resource availability. In this chapter, we integrate all of these approaches into a unified mathematical framework and discuss computational strategies for solving this class of problem more efficiently.

Section 5.1 of this chapter reviews why conventional capacity-oriented resource adequacy heuristics struggle in planning modern power systems, and presents known solutions (endogenous risk curves, adaptive stress period planning, and sparse storage chronology) to those problems. Section 5.2 introduces three novel techniques to integrate these approaches to enforce probabilistic resource adequacy constraints in a computationally-efficient, deterministic portfolio optimization framework. Section 5.4 describes three new test systems, when are then used in Section 5.5 to investigate how expansion outcomes are impacted by the multiple problem parameters introduced with the enhanced risk curve methods, considering tradeoffs between problem runtime and solution quality. Section 5.6 discusses opportunities for future work and provides concluding remarks.

5.1 Motivation and Background

5.1.1 Evolving Risk Drivers

It has long been understood that PRMs are insufficient to capture probabilistic shortfall risk, even in historical power systems where adequacy risk was driven entirely by thermal

generator outages [41]. While probabilistic studies could help quantify uncertainty in supply availability, embedding these considerations directly into portfolio selection processes was not computationally feasible. Instead, probabilistic outcomes were used to parametrize heuristics (i.e. PRM magnitudes and capacity accreditation) that could be more readily applied inside a linear planning framework.

The energy transition has introduced a number of new considerations that complicate traditional capacity-focused adequacy assessment. Quantifying uncertainty in thermal generator availability during peak demand periods is no long sufficient as increased integration of variable renewable generation and grid-coupled energy storage, as well as the increasing frequency of extreme weather events, have introduced new dynamics that must be considered in a credible resource adequacy assessment [42].

Supply variability across time

Peak load conditions are no longer the sole driver of system risk. Both variable renewable and thermal technologies have time-varying outputs that need to be considered. How these availabilities correlate (or not) with demand levels has critical impacts on system adequacy.

Spatial coupling

Cost effective solutions increasingly involve power systems larger than weather systems to diversify against localized variations in supply and demand under both normal conditions (across time zones and climate regions) and extreme conditions (including cold snaps, heat waves, wind and solar droughts, and natural disasters). The benefits, constraints, and risks associated with wide-area resource sharing using long-distance transmission need to be explicitly considered.

Temporal coupling

As new energy storage technologies become increasingly economically competitive, they play a more prominent role in enabling cost-effective generation portfolios with larger shares of variable resources. Energy shifting can occur not only within single days but across weeks

or even seasons, meaning that near-term operational decisions about whether to charge or discharge a storage device can have adequacy implications days, weeks, or months into the future.

5.1.2 New methods for representing new risks

While traditional PRM approaches struggle to accurately represent traditional uncertainty considerations, they are even less well-suited to capture the new dynamics above. One solution to this has been to apply an increasingly complicated set of elaborations to traditional techniques, often based on joint capacity accreditation of many speculative resource portfolios. While they provide improvements, these methods are computationally intensive and still struggle with the fundamental limitations of a capacity-oriented adequacy paradigm [43]. Recognizing this fact, a range of more radical changes have been proposed to individual aspects of the challenge. However, short of large-scale stochastic optimization facilitated by decomposition techniques and supercomputing, we are unaware of any approach to date that provides a unified framework for addressing these new interacting adequacy dynamics.

Endogenous risk curves

Chapter 4 and [44] previously introduced the concept of risk curve iteration as a means to endogenously approximate supply availability uncertainty without resorting to stochastic optimization. With this approach, the average capacity surplus in each region and timestep is considered independently, and mapped to expected unserved energy (EUE, [7]) for that same region and timestep, based on a piecewise-linear estimator function ("risk curve"). This risk curve is empirically derived from a probabilistic adequacy assessment of the last candidate solution. These deterministic estimates of EUE outcomes, rather than stochastically-sampled generator availability inputs, can then be aggregated together to form the basis for enforcing approximate probabilistic adequacy constraints.

We briefly review how such curves are generated. First, a previously-available solution candidate is assessed in a dedicated probabilistic resource adequacy simulation. In the first iteration, the existing solution may be obtained from the starting resource portfolio before

any new investments are made, or it may be the result of an initial capacity expansion performed without resource adequacy constraints. A baseline average shortfall or surplus level is recorded for each region and timestep. Then, sample-level realizations of surplus and shortfall results are used to perturb this point estimate into an estimate of loss of load probability (LOLP) as a function of available capacity surplus (see Figure 5.1, panels a and b).

For example, if in a given place and time the previous portfolio provided 100 MW of surplus capacity on average, and 10% of probabilistic samples experienced shortfalls, all of which had a magnitude of 50 MW or less, the original system would report an LOLP of 10%. We assume a different portfolio with 150 MW average surplus in the same place and time would drop LOLP to zero, since 50 MW more is, on average, enough to eliminate every previously-observed shortfall event. This is, of course, an approximation: in reality, sometimes not all of the 50 MW of added "average" supply will be available during the event conditions, while at other times more than 50MW may be.

Since EUE is defined as the product of likelihood and magnitude of shortfall events, we can integrate the stepwise LOLP function by negative surplus to get a function estimating EUE as a function of surplus. The integration constant is set such that LOLP and EUE reach zero at the same surplus level. (see Figure 5.1, panel c). Since the LOLP function was stepwise and monotonically non-increasing with respect to surplus, the negatively-integrated EUE function is piecewise-linear, monotonically decreasing, and convex with respect to surplus. This allows such functions to serve as lower bounds on EUE decision variables in a linear program. With EUE defined in this way for every region and timestep, aggregate EUE metrics are easily upper bounded in a cost-minimization problem by summing individual EUE variables across space and/or time.

While Chapter 4 demonstrated that the risk curve approach is able to produce cost-effective systems that meet probabilistic adequacy criteria without the need for stochastic optimization, capacity accreditation, or exogenously tuned reserve margins, it did so by representing power system operations with full chronology in order to properly capture time-varying relationships between electricity supply and demand. While this "brute-force" temporal representation may be tenable when considering a single weather year, a

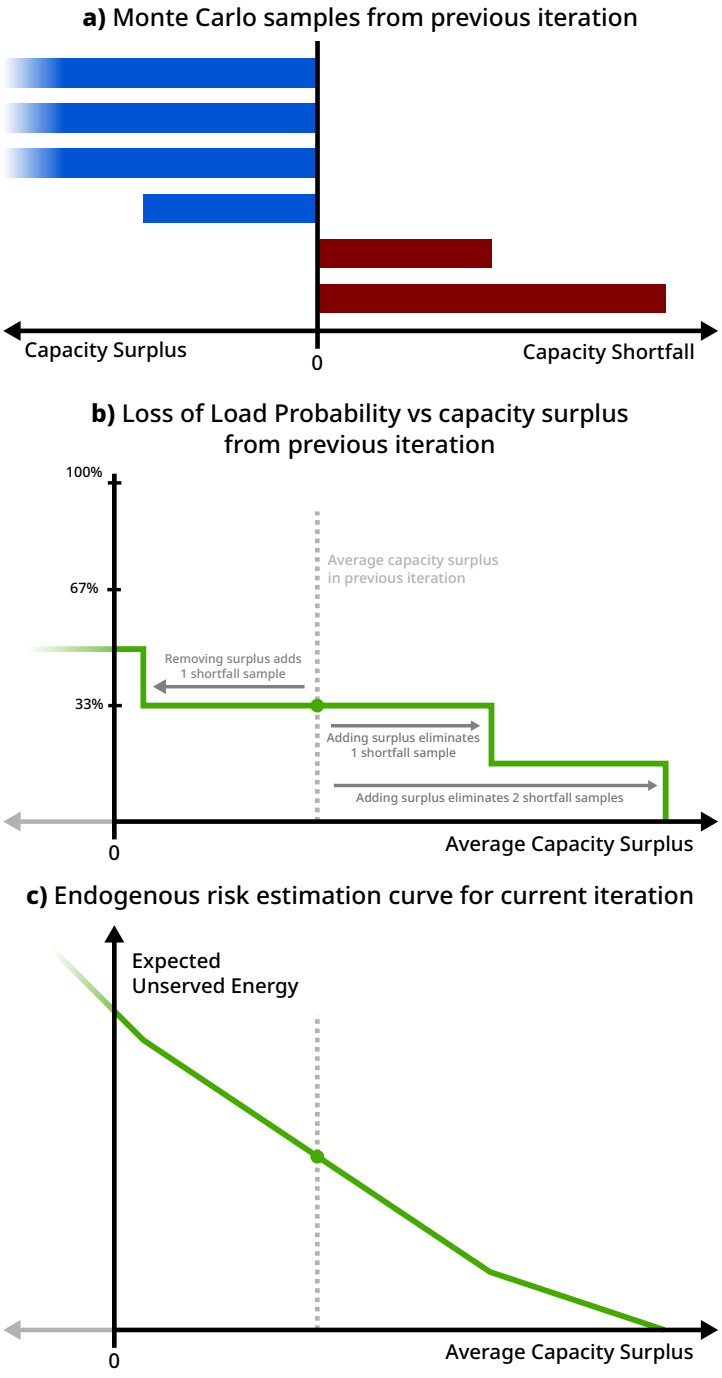

Figure 5.1: A risk curve for a specific timestep and region is generated by mapping from raw Monte Carlo sample data (top) to LOLP as a stepwise function of surplus (middle) to EUE as a piecewise-linear, convex function of surplus (bottom).

geographically-concise system, and only a few technology investment options, scaling to larger expansion problems quickly becomes challenging.

Time series aggregation techniques are a well-studied means to substantially reduce the temporal representation of the problem, and thus overall problem size, without significantly distorting the ultimate investment decisions. While a large body of literature exists on techniques for selecting "representative" time periods to estimate system operating costs [45], the resource adequacy challenges that drive investment needs happen by definition during outlier conditions that can look very different from representative periods.

Adaptive stress period planning

Adaptive stress period planning (ASPP), as previously discussed in Section 3.2, pairs representative periods, which represent "normal" operating conditions, with emergent stress periods, which represent "worst-case" outlier conditions that the system must be able to ride through in order to maintain resource adequacy [43]. Since resource adequacy risks tend to be driven by a small set of critical operating conditions, augmenting a limited set of representative periods with just a few key stress periods can be all that's needed to drive a planning model towards a resource-adequate solution.

Figure 5.2 provides an example of the iterative process by which such periods are identified, starting with five periods: one representative period for each of the four seasons, and a fifth initial stress period that corresponds to the highest-energy-demand day in the operations dataset. Over three subsequent iterations, new stress periods are added to the problem to explicitly represent risky system conditions. The initial solution relies almost exclusively on variable renewables and storage, and so encounters significant shortfall risk on a winter day with low wind and solar output, even though demand is lower than any of the representative days. In the second iteration, this day is included in the optimization problem, leading to a solution with much more thermal generating capacity. In subsequent iterations, specific days with low wind output and high demand are identified as challenging conditions that need to be planned against in order to achieve a resource-adequate system.

While ASPP provides an efficient alternative to full-chronology operations representa-

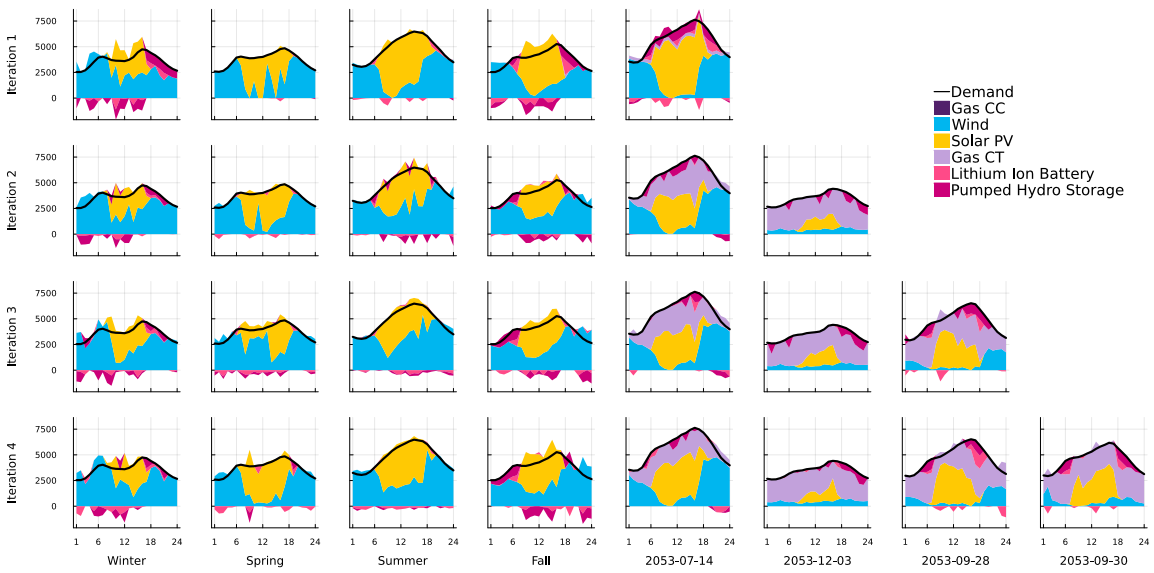

Figure 5.2: Example of dispatch stacks for the operating conditions seen by a CEM over the course of an adaptive stress period planning process. The first iteration uses a predetermined set of representative periods and a single stress period, and new stress periods are incrementally added to the CEM based on times with the most system risk in the previously-identified solution. Negative dispatch corresponds to storage charging.

tions in capacity expansion models, taken alone it still fails to encode probabilistic risks. While being able to balance supply and demand *on average* during stress periods is certainly better for system adequacy than not, it provides no guarantee that the risk of shortfall satisfies a given probabilistic criteria. As demonstrated in [44], even full-chronology capacity expansion models that ignore supply uncertainty can yield portfolios with an order of magnitude more risk than deemed acceptable.

Sparse storage chronology

Another traditional challenge with using representative periods is the limited ability to represent storage's potential to shift energy outside of the representative periods themselves. The techniques developed in Section 3.3 and [46] allow representative periods to be chronologically-coupled without requiring an explicit representation of all time periods to enforce state-of-charge bounds. These "sparse chronology" methods provide a more accu-

rate representation of the capabilities and limitations of longer-duration storage and allow for a fuller valuation of such technologies' potential contributions to resource adequacy.

5.2 Enhancing Probabilistic Risk Curves

The endogenous risk curve formulation, while powerful, has the potential to introduce a large number of new constraints to the optimization problem. Every region and time period in the full operations dataset has the potential to introduce a piecewise linear curve with the same number of line segments as the number of Monte Carlo samples in the underlying adequacy assessment. In the worst case, the probabilistic assessment produces a unique shortfall/surplus value in each of its n samples, resulting in an LOLP function with n steps and an EUE function with n different segments. In this section, we present multiple techniques for controlling this potential growth in problem size and improving the solve time of the resulting optimization problem.

5.2.1 Risk Curve Aggregation

Since most time periods are removed from the optimization problem during time series aggregation, most risk curves are as well, leading to a substantial underestimation of system risk if only the remaining period's curves are considered in endogenous aggregate EUE constraints. To compensate for this, we define a many-to-one mapping of represented periods to representative periods, and then add together the LOLP functions from the multiple represented timesteps into a single LOLE function for the representative timestep (see Figure 5.3). The multi-period LOLE estimate from this function can be used to derive a multi-period EUE function in exactly the same way that the original LOLP functions were transformed into single-period EUE functions. These multi-period EUE functions have the same helpful properties of piecewise-linearity and convexity.

The number of line segments in a multi-period EUE curve is upper-bounded by the sum of the number of line segments in the constituent single-period curves. However, there is also much more potential for samples with identical surplus across a set of similar represented periods, causing previously-distinct line segments from different periods to collapse into a single line segment in the multi-period curve. As the number of aggregated periods increases,

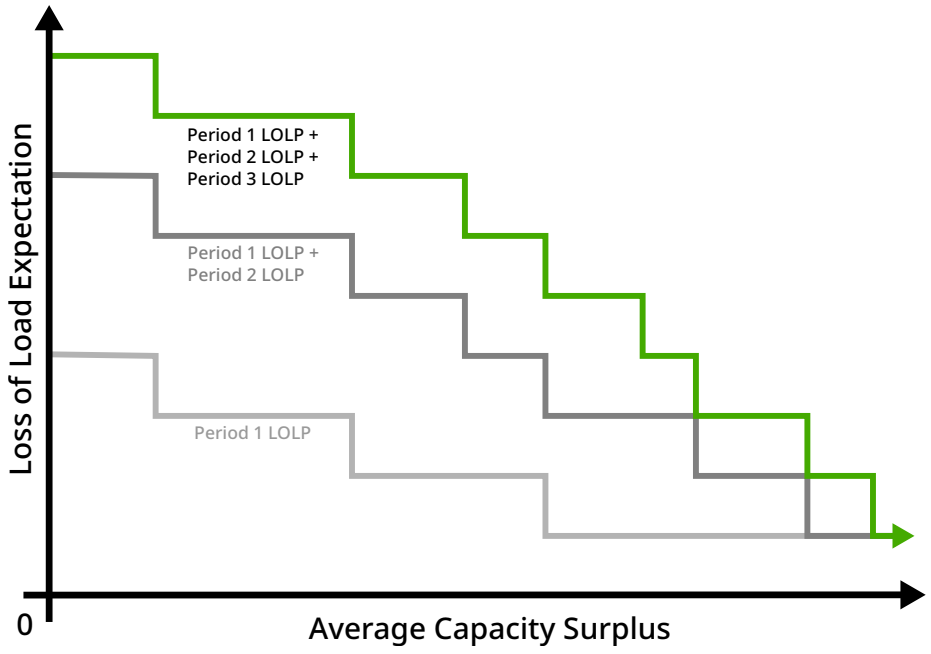

Figure 5.3: Example of combining LOLP curves from separate timesteps in an adequacy model into an aggregate LOLE curve for a single representative timestep in a CEM.

the likelihood of such collisions does as well, potentially driving the number of optimization constraints required for a multi-period curve to be proportionately smaller than the number of segments in each of the constituent single-period curves.

5.2.2 Risk Curve Reduction

There is also an opportunity to further reduce the number of constraints required to encode risk curves by removing line segments that only introduce minor deviations in the curve from the original. Some constraints are only active for a very small range of capacity surpluses, and removing them only impacts the EUE estimate slightly, in that small surplus range. For example, Figure 5.4 shows an example where two of the six constraints that make up the risk curve can be eliminated while introducing very little error to the estimator function.

By the convexity of the risk curves, we know that removing constraints always results in an underestimation of EUE relative to the original function. Furthermore, the maximum magnitude of this overestimation is simply the difference between the original and new curves at the point where the line segments on either side of the removed line segment intersect. We can easily calculate the worst-case error introduced by removing any single line, choose the line with the smallest value, and then repeat. Since by definition the worst case error for each selection always increases (since we're picking the smallest errors first), the worst case error of the overall function is simply the worst case error from the last line removal.

Since we typically only care about aggregate EUE metrics, we can sum the worst-case errors across all of the functions being aggregated into the metric to get the worst-case error in the overall metric estimation. If we define an upper bound on such error, we can remove line segments (from whichever function is the best candidate / introduces the least error) until the aggregate worst-case error would exceed the threshold. More error tolerance therefore allows us to remove more line segments / constraints.

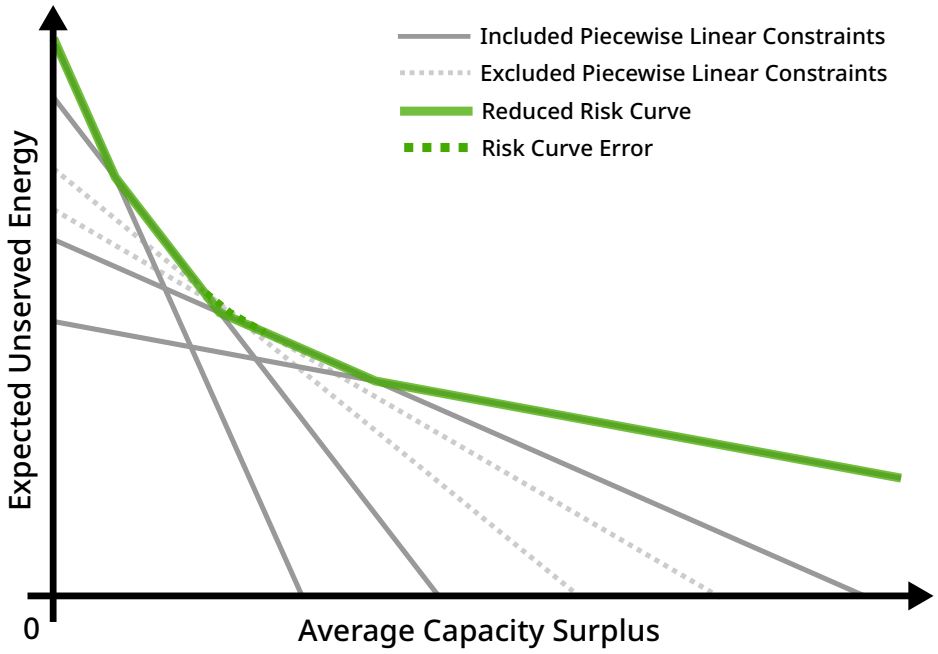

Figure 5.4: Example of eliminating two out of six line segments from a piecewise-linear risk curve, reducing CEM problem size with minimal impact to EUE estimator accuracy.

5.2.3 Risk Curves with Adaptive Stress Period Planning

While aggregating stress curves together for representative periods provides the optimization problem with information on the shortfall risk that exists in unmodeled periods, it does not inherently provide information on the context in which specific risk occurs, nor what particular options may exist to cost-effectively mitigate that risk. As a result, targeted solutions that address adequacy needs during very specific situations may be overlooked in favor of more generic (and expensive) options. Conversely, solutions may mistakenly rely on resources that appear to provide adequacy benefits during typical operating conditions, but are not actually available during more challenging conditions.

As discussed previously, ASPP solves this exact problem. While alone the technique lacks information on uncertainty of supply availability, it is extremely complementary with the use of risk curves, since both methods work iteratively by evaluating solutions from a capacity expansion model in a dedicated resource adequacy simulation. After a candidate resource portfolio has been identified, probabilistic, full-chronology adequacy assessment results can be applied for two purposes simultaneously. First, time periods with the greatest overall shortfall risk are identified and, if not yet be represented in the optimization, used to augment the problem. Second, sample-level results are used to generate risk curves for both representative and stress periods in the augmented problem, encoding the missing probabilistic data in the problem. Each timestep in the full-chronology data is mapped to either a representative period or a stress period, with no distinction between the two classes in the underlying formulation, other than that risk curves in representative periods will be derived from samples from many different timesteps, while risk curves in stress periods may only correspond to data from one timestep, that of the stress period itself.

With ASPP capturing the key average supply and demand conditions that drive adequacy needs, and risk curves approximating the levels of investment needed to meet probabilistic metrics during those specific periods, the investment optimization problem has all the information necessary to identify a cost-effective portfolio design that meets probabilistic adequacy criteria.

Furthermore, these two approaches complement each other in a way that other iterative

techniques may not. For example, as discussed in Sections 2.2 and 4.1, another mechanism available to calibrate a deterministic planning model to probabilistic adequacy criteria is to enforce and iteratively tune a reserve margin, increasing it when a resource adequacy assessment deems the system unreliable and decreasing it when the system is overbuilt [34]. This technique interacts badly with ASPP, since adding a new stress period to the planning model invalidates any previous calibration of the reserve margin; a large reserve margin is needed when stress conditions aren't being explicitly planned for, but enforcing that reserve level during stress conditions too will lead to overprocurement. With risk curves, adequacy estimates already encompass risk from all time periods considered in the adequacy model, and are derived independently for each period included in the optimization, meaning that adding a new stress period to the planning problem simply breaks out that period's risk curve data into its own single-period curve, without invalidating the risk curves of other time periods.

5.2.4 Numerical Considerations

While the risk curve approach enables formulating a problem that considers probabilistic adequacy considerations far more compactly than stochastic optimization, it also introduces a new opportunity for numerical issues in the optimization process, which requires careful treatment. Specifically, the slopes of the line segments in the the piecewise-linear EUE functions, and therefore entries in the optimization problem's constraint matrix, can take on values spanning many orders of magnitude, which has the potential to cause numerical instability issues and degrade the performance of the optimization solver [47].

Since these slopes correspond to negative LOLP or LOLE values, the lower bound on their magnitude drops as more samples are used in the probablistic adequacy assessment step. For example, the smallest non-zero LOLP (and thus LOLE) resolvable by a 100-sample simulation is 10^{-2}, while a 1000-sample simulation could report an LOLP of 10^{-3}. Since the domain of the EUE function spans all non-negative surplus levels, there will always be a line segment in the function with the minimum-magnitude slope (LOLP or LOLE), unless the worst-case shortfall observation happens to occur in more than one sample.

A slope's upper bound is governed by a different factor – the number of timesteps aggregated together into a single multi-period curve. Surplus in any one timestep can correspond to at most an LOLP of 1.0, but when n periods are aggregated together the upper bound grows proportionately to an LOLE of n event-periods. Line slopes / LOLEs in this order of magnitude are very likely at low surplus levels, unless variance in total availability of supply happens to be very small.

Given these bounds, using a small number of representative periods (4-10), a moderate number of probabilistic samples (100-1000), and a reasonable number of alternate weather years (5-10) can easily introduce a range of slopes spanning 4-6 orders of magnitude (e.g., 10^{-3} to 10^{3}). Taken alone, this is just within the range of values modern solvers using double-precision floating point arithmetic are well equipped to handle. However, these EUE function slopes are typically multiplied by other wide-ranging values to form the elements of the final constraint matrix, further exacerbating the issue. For example, wind and solar build decisions may be multiplied by hour availabilities ranging from 10^{-6} to 10^{0}, while in the same constraint integer thermal build decisions are multiplied by nameplate capacities on the order of 10^{2}. Multiplying these values with the curve slopes would yield values ranging from 10^{-9} to 10^{5}, spanning 14 orders of magnitude.

Techniques such as rounding small non-zero variable generation availabilities to zero and selecting the units of decision variables strategically can help reduce the range of coefficients considerably. For the test systems described in the next section, such strategies were able to reduce the range of coefficients in the final constraint matrix from ∼15 orders of magnitude to ∼5.

5.3 Capacity Expansion Problem Formulation

The concepts described above can be integrated into a unified mathematical framework as follows. Z represents the set of all zones (regions) in the system, and I represents the set of all transmission interfaces between zones. I_z is the subset of interfaces with forward direction into $z \in Z$, while I_{-z} is the subset of interfaces with forward direction out of z. H, V, and S represent the sets of all thermal generation classes, variable generation classes, and storage classes respectively, where a "class" defines a specific technology in a specific

zone. H_z, V_z, and S_z represent the subsets of classes associated with zone z. Instances of thermal generation, variable generation, or storage classes must be built at specific sites q from one of the sets Q_h, Q_v, or Q_s, respectively.

P represents the set of operating periods, where each operating period $p \in P$ represents a set T_p of temporally contiguous timesteps t. R is the set of repetitions, where a repetition $r \in R$ is a chronological sequence of the same period p repeated N_r times. The repeated period associated with r is denoted as p_r, and periods have weights Θ_p denoting how many times they are repeated across all repetitions. Finally, J_{zt} is the set of risk curve line segments j for a given zone z and timestep t.

Expansion decision variables are $N_q \in \mathbb{Z}$ for discrete thermal generation units, C_q for variable generation and storage capacity, E_q for storage energy, and C_i for interregional transmission. Nameplate thermal site capacity C_q is calculated by scaling unit count N_q by unit size Φ_h. Dispatch decision variables are g_{ht}, g_{vt}, c_{qt}, d_{qt}, and f_{it} for thermal and variable generation, storage charging and discharging, and transmission flow, respectively.

We seek to minimize overall annualized system cost as follows:

$$\begin{aligned}\min \sum_{h\in H, q\in Q_h} K_h^c \Phi_h N_q + \sum_{v\in V, q\in Q_v} K_v^c C_q + \\ \sum_{s\in S, q\in Q_s} \left(K_s^c C_q + K_s^e E_q\right) + \sum_{i\in I} K_i^c C_i + \\ \sum_{p\in P, t\in T_p} \Theta_p \Bigg(\sum_{h\in H} k_h g_{ht} + \sum_{v\in V} k_v g_{vt} + \\ \sum_{s\in S, q\in Q_s} k_s \left(c_{qt} + d_{qt}\right) \Bigg) \end{aligned} \tag{5.1}$$

Here, K_h^c, K_v^c, K_s^c, K_s^e, and K_i^c are cost parameters corresponding to thermal and variable generation capacity, storage capacity and energy, and transmission capacity. They incorporate both annualized capital costs and fixed annual operating costs. k_h, k_v, and k_s are variable operating costs for thermal generation, variable generation, and storage, incorporating the cost of fuel and/or variable operations and maintenance.

Build decisions are constrained by upper investment bounds $\overline{N}_q$, $\overline{C}_q$, $\overline{E}_q$, and $\overline{C}_i$:

$$
\begin{aligned}
0 \leq N_q \leq \overline{N}_q \ \ \forall \ \ h \in H, \ q \in Q_h \\
0 \leq C_q \leq \overline{C}_q \ \ \forall \ \ v \in V, \ q \in Q_v \\
0 \leq C_q \leq \overline{C}_q \ \ \forall \ \ s \in S, \ q \in Q_s \\
0 \leq E_q \leq \overline{E}_q \ \ \forall \ \ s \in S, \ q \in Q_s \\
0 \leq C_i \leq \overline{C}_i \ \ \forall \ \ i \in I
\end{aligned}
\tag{5.2}
$$

Generation, storage dispatch, and transmission flows are constrained by build decisions across all sites. For all $p \in P, \ t \in T_p$:

$$
\begin{aligned}
0 \leq g_{ht} \leq \Phi_h \sum_{q \in Q_h} \alpha_{qt} \left(N_q^0 + N_q \right) \ \ \forall \ \ h \in H \\
0 \leq g_{vt} \leq \sum_{q \in Q_v} \alpha_{qt} \left(C_q^0 + C_q \right) \ \ \forall \ \ v \in V \\
0 \leq c_{qt} \leq C_q^0 + C_q \ \ \forall \ \ s \in S, \ q \in Q_s \\
0 \leq d_{qt} \leq C_q^0 + C_q \ \ \forall \ \ s \in S, \ q \in Q_s \\
-\left(C_i^0 + C_i \right) \leq f_{it} \leq C_i^0 + C_i \ \ \forall \ \ i \in I
\end{aligned}
\tag{5.3}
$$

N_q^0, C_q^0, and C_i^0 represent existing infrastructure at a site or interface, while α_{qt} is the average resource availability at a specific site and timestep.

Next, to develop the sparse chronological representation of storage, we define decision variable Δe_{qp} as the change in state of charge for a storage site q over a single operating period p, and decision variables $\underline{e}_{qp}$ and $\overline{e}_{qp}$ as the minimum and maximum states of charge in the period, relative to the period's starting energy. Taking parameter ε_s as the round-trip efficiency of storage class s, we constrain the variables to these definitions with, for all $s \in S$, $q \in Q_s$, $p \in P$:

$$
\begin{aligned}
\Delta e_{qp} = \sum_{t \in T_p} \left(\varepsilon_s^{\frac{1}{2}} c_{qt} - \varepsilon_s^{-\frac{1}{2}} d_{qt} \right) \\
\underline{e}_{qp} \leq \sum_{t'=1}^{t} \left(\varepsilon_s^{\frac{1}{2}} c_{qt'} - \varepsilon_s^{-\frac{1}{2}} d_{qt'} \right) \leq \overline{e}_{qp} \ \ \forall \ \ t \in T_p
\end{aligned}
\tag{5.4}
$$

Taking decision variable e^0_{qr} as the absolute state of charge of a storage site q at the beginning of repetition r, and parameter Γ_r as the length (in periods) of that repetition, we constrain the state of charge of each storage site to always be within physical energy limits for each $s \in S$, $q \in Q_s$, $r \in R$:

$$\begin{aligned} e^0_{qr} &= e^0_{qr-1} + \Gamma_{r-1}\Delta e_{qp_{r-1}} \\ 0 &\leq e^0_{qr} + \underline{e}_{qp_r} \\ 0 &\leq e^0_{qr} + (\Gamma_r - 1)\Delta e_{qp_r} + \underline{e}_{qp_r} \\ e^0_{qr} + \overline{e}_{qp_r} &\leq E^0_q + E_q \\ e^0_{qr} + (\Gamma_r - 1)\Delta e_{qp_r} + \overline{e}_{qp_r} &\leq E^0_q + E_q \end{aligned} \tag{5.5}$$

In this work we take $e^0_{q0} = 0$, although periodic boundary conditions are also possible. A more detailed explanation of these constraints is available in Section 3.3 or [46]. With electrical load given as L_{zt}, we enforce average power balance (for the purposes of estimating operating costs) for $z \in Z$, $p \in P$, $t \in T_p$:

$$\sum_{h \in H_z} g_{ht} + \sum_{v \in V_z} g_{vt} + \sum_{s \in S_z, q \in Q_s} (d_{qt} - c_{qt}) + \sum_{i \in I_z} f_{it} - \sum_{i \in I_{-z}} f_{it} = L_{zt} \tag{5.6}$$

Finally, we enforce the probabilistic adequacy constraints. To capture storage and transmission impacts on adequacy we define a parallel set of "reliability dispatch" decision variables $\tilde{c}_{qt}$, $\tilde{d}_{qt}$, $\Delta\tilde{e}_{qp}$, $\tilde{\overline{e}}_{qp}$, $\tilde{\underline{e}}_{qp}$, and $\tilde{f}_{it}$, with a parallel constraint set matching the relevant economic dispatch constraints from (5.3)–(5.5). We then mirror (5.6) to calculate an expected energy-backed capacity surplus x_{zt} and apply it to estimate EUE u_{zt} for all $z \in Z$, $p \in P$, $t \in T_p$:

$$x_{zt} = \sum_{h \in H_z, q \in Q_h} \alpha_{qt} \Phi_h \left(N_q^0 + N_q \right) \; + \sum_{v \in V_z, q \in Q_v} \alpha_{qt} \left(C_q^0 + C_q \right) \; + \sum_{s \in S_z, q \in Q_s} \left(\tilde{d}_{qt} - \tilde{c}_{qt} \right) \; + \sum_{i \in I_z} \tilde{f}_{it} - \sum_{i \in I_{-z}} \tilde{f}_{it} - L_{zt} \quad (5.7)$$

$$u_{zt} \geq m_{ztj} x_{zt} + b_{ztj} \;\; \forall \;\; j \in J_{zt} \quad (5.8)$$

where m_{ztj} and b_{ztj} are the slope and y-intercept of the line segments in the endogenous risk curve for a given location and timestep. More details on this risk curve formulation are available in Section 4.2 or [44]. We can now enforce our probabilistic adequacy criteria, an EUE limit $\overline{u}_z$ for each zone $z \in Z$:

$$\sum_{p \in P, t \in T_p} u_{zt} \leq \overline{u}_z \quad (5.9)$$

5.4 Test System Development

To verify the enhanced scalability of these unified adequacy enforcement techniques, we develop three test systems representing diverse climate regions of the United States: the Great Plains, the Southeast, and the Southwest. Each system is divided into four regions based on the same balancing areas as the ReEDS capacity expansion model [8] and uses the same seven years of hourly load, wind, and solar input data (2007-2013) provided by that model.

The Great Plains system is patterned on historical weather and load data from Kansas and Nebraska (ReEDS regions p39, p40, p52, and p53). This system is characterized by a high quality wind resource and lower seasonal load variability than the other regions. The Southeast system adapts data from Georgia, South Carolina, and parts of North Carolina (ReEDS regions p94, p95, p96, and p97), and has the highest winter (November to March) peak load of the three regions as well as the lowest average wind and solar availability.

Table 5.1: Load and resource characteristics across the three test systems.

		Great Plains	Southeast	Southwest
Load (MW)	Average	4000	4000	4000
	Summer Peak	6736	7247	7739
	Winter Peak	5867	6141	5314
Wind	Locations	55	115	70
	Average Capacity Factor	55%	40%	45%
Solar PV	Locations	13	32	19
	Average Capacity Factor	30%	28%	36%

The Southwest system is derived using data from Southern California, Nevada, and Arizona (ReEDS regions p10, p13, p27, and p28), and has the best solar resource, the highest summer (April to October) peak load, and the most modest winter load of the three systems. More details on the characteristics of each system are available in Table 5.1.

Regional load in each system is scaled proportionately such that the average system-wide demand across each seven year dataset is 4000 MW, while preserving interregional differences in demand (such that some regions within a system have much more demand than others). The zonal topology of each system is shown in Figure 5.5, with the magnitude of annual energy demand for each region proportional to the size of the zone's node in the network.

To identify candidate sites for wind and solar expansion, we leverage the same resource data grid cell clustering performed by the reV model [48] as an upstream step to ReEDS. In ReEDS, these cell clusters would be further aggregated into resource supply curves, but here we instead directly provide the discrete cell clusters as build options to the capacity expansion model. At the computational cost of additional investment decision variables, we gain enhanced spatial resolution which provides the optimization with more granular information on resource availability in specific locations, allowing it to choose specific sites based on resource diversity or production potential during key times of system need.

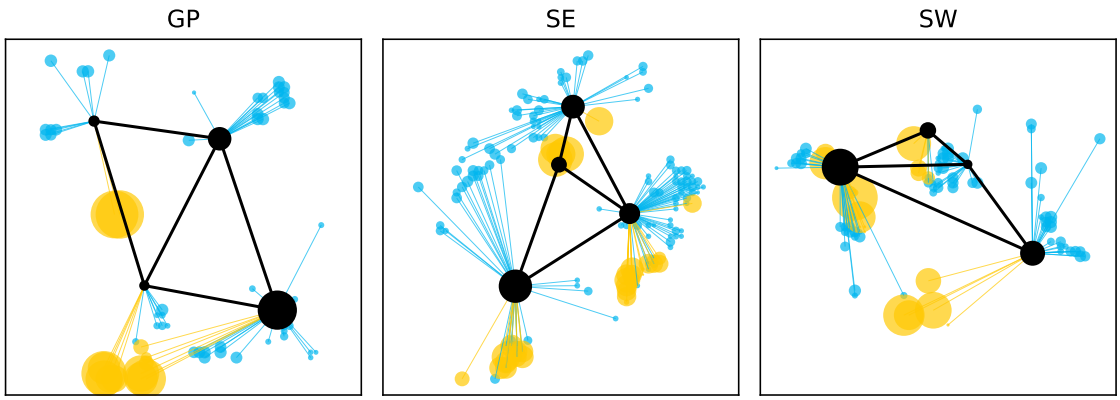

Figure 5.5: Zonal network topology for each of the three test systems. Black nodes represent the four load zones in each system, with the size of each node corresponding to that region's annual energy demand. Black edges are tranmission interfaces between zones. The blue and yellow nodes correspond to potential wind and solar build sites, with their sizes proportional to capacity available to build at that location.

Ideally, every potential wind and solar build location would be considered as a candidate in the capacity expansion process. Since this full candidate set yields far more potential renewable generation than the system is able to use (with potential generation exceeding electrical energy demand by orders of magnitude), we filter the candidate list to include only enough resource sites to meet 1.5x the system's overall energy demand with a single technology (wind or solar PV). Since the average demand of each system is 4000 MW, this means we consider just enough candidate sites that, if they were all fully built out, they would generate at an average level of 6000 MW. This reduced candidate set is still far beyond what the system needs – ignoring storage losses, in theory even a 100% renewable energy system would only need to procure 1/3 of the energy generation potential available. However, this reduced set eliminates the vast majority of the build locations and so drastically reduces the number of investment variables in the problem. The final counts of candidate resource locations for each system are shown in Table 5.1 and their geographic distribution is shown in Figure 5.5.

Resource sites are considered for inclusion in the reduced candidate set in decreasing order of average capacity factor. To maintain spatial diversity, no one region is allowed to

contribute more than 1/3 of the potential energy generation (that is, for a given technology, once enough sites from a single region have been added to the candidate set such that the potential average generation from all those sites exceeds 2000 MW, no further sites from that region are considered).

Four generation technologies and one storage technology are available in the portfolio selection problem: wind, solar PV, natural gas combined cycle generators, natural gas combustion turbines, and lithium ion batteries. Capital costs as well as fixed and variable operating costs for each of these technologies are taken from the NREL Annual Technology Baseline (ATB) [29]. Natural gas unit sizes, heat rates, mean times to failure, and mean times to repair are taken from the RTS-GMLC test system [28]. Natural gas fuel prices are taken from ReEDS. Since we use the same region definitions as ReEDS, we use the same interregional transmission expansion costs as well [8].

Finally, to demonstrate the use of the sparse storage chronology representation and incentivize economic solutions that may be more challenging to plan from a resource adequacy perspective, we allow up to 1GW/20GWh of pumped storage hydro to be built in the two systems where such a technology would be geographically plausible (Southeast and Southwest).

5.5 Empirical Tests

Endogenous risk curve aggregation and reduction introduce a number of new design parameters for the optimization problem, which we study here using the three test systems described above. In particular, risk curve error tolerance, Monte Carlo sample count, and the number of representative periods used all have the potential to influence the problem size, solve time, and solution quality of the capacity expansion optimization problem.

We formulate and solve iterative capacity expansion optimization problems applying all of the techniques discussed in the previous sections, using the linear program formulation provided in Section 5.3 for capacity expansion optimization and NREL's Probabilistic Resource Adequacy Suite [39] to assess the adequacy of candidate systems. We run portfolio selections for a range of EUE error tolerance and probabilistic sample count parameters in order to understand the practical impact of these parameters on the size of the optimization

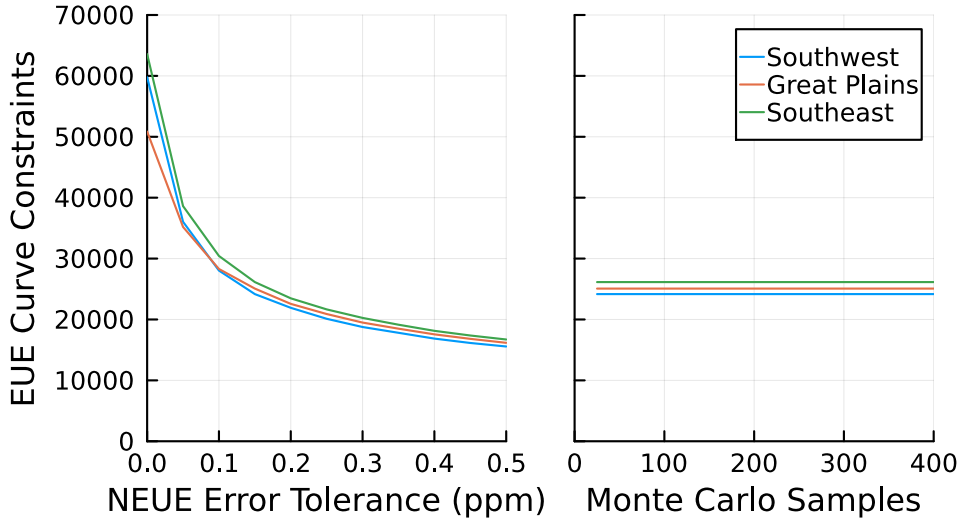

Figure 5.6: Number of constraints required to implement endogenous risk curves for four seasonal representative periods, as a function of relaxed NEUE error tolerance (left, 400 samples) and the adequacy assessment's Monte Carlo sample size (right, 0.15 ppm error tolerance).

problem and quality of the solution.

As expected, relaxing the error tolerance of the EUE estimator curves allows for more line segments to be removed from the problem and significantly reduces the number of constraints required to represent the curves. This is particularly impactful when the EUE curve constraints account for the majority of the constraints in the problem, as is the case here. As shown in Figure 5.6, many of the constraint elimination benefits can be achieved with a relatively small relaxation: across all three test systems, a 15% error tolerance (0.15ppm relative to a 1ppm NEUE target) consistently enables eliminating half of the EUE curve constraints.

While adding more Monte Carlo samples to the probabilistic adequacy assessment increases the theoretical upper bound on the number of line segments in the risk curves, in practice we observe almost no impact at all, indicating that the number of unique shortfall and surplus levels observed saturates very quickly in the systems studied. Figure 5.6 illustrates this for the case of 15% error tolerance, but the pattern persists across other

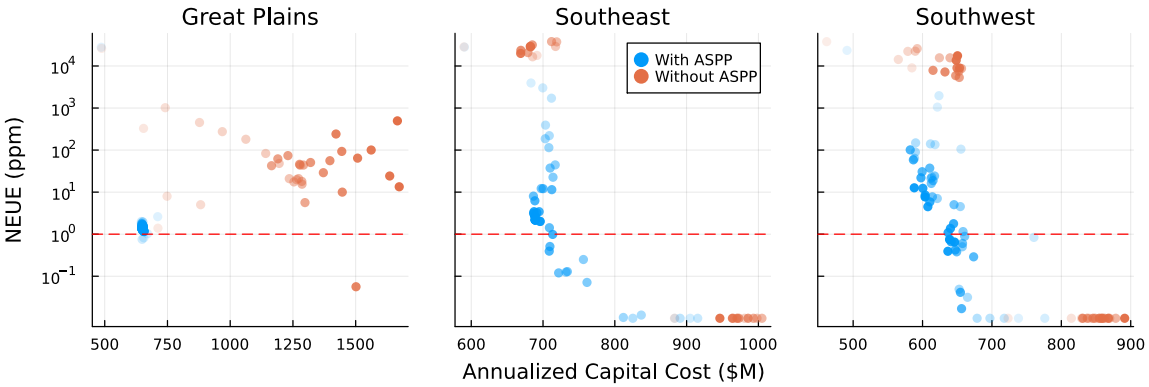

Figure 5.7: Capital costs and adequacy of all intermediate solutions for three test systems, iterating with either endogenous risk curves alone (orange), or endogenous risk curves combined with ASPP (blue). Points are shaded darker as iterations progress.

tolerance levels as well (including no curve reduction at all).

We also consider the extent to which representative periods with aggregated risk curves alone are sufficient to produce cost-efficient, resource adequate outcomes. As seen in Figure 5.7, in the Southeast and Southwest systems applying aggregated risk curves without adaptive stress periods is able to identify resource adequacy portfolios, but at a much greater cost than when using stress periods. In these situations, the iteration oscillates between underbuilt and overbuilt solutions and fails to converge to the risk target. The Great Plains systems exhibit a different failure mode, gradually improving reliability as iterations progress, but doing so extremely inefficiently. In all of these cases, combining adaptive stress periods with endogenous risk curves is sufficient to drive the iteration process to converge to lower-cost solutions at or near the target reliability level.

Finally, we consider the impact of the number of representative periods on the model's ability to identify cost-effective, resource adequate solutions. We see that a small number of iterations are sufficient to identify resource adequate solutions and that continued iterations are able to significantly reduce costs while maintaining adequacy, regardless of the number of initial representative periods. From a resource adequacy and capital cost perspective, the combination of endogenous risk curves and adaptive stress periods eliminates the need for larger numbers of representative periods. As demonstrated in Figure 5.8, using a small

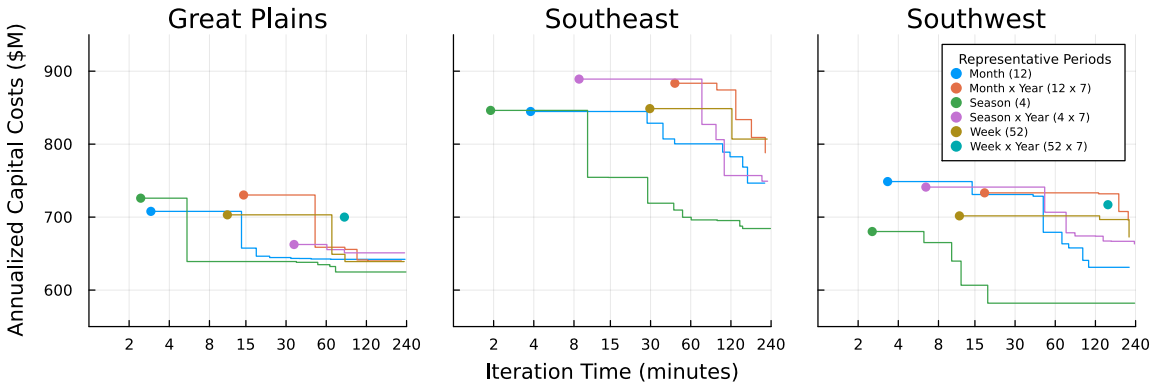

Figure 5.8: Best-so-far capital costs for solutions as a function of iteration time, across three test systems and six choices of initial representative periods. Only solutions that meet prescribed adequacy criteria are considered.

set of representative periods to keep the problem size small can yield faster initial solutions and rapid cost improvements before a larger set of representative periods has yielded even its first higher-cost resource adequate solution.

5.6 Conclusion

This work has developed a unified, computationally-efficient framework for planning power systems to probabilistic adequacy criteria by integrating and enhancing endogenous risk curve, adaptive stress period planning, and sparse storage chronology techniques. These techniques were demonstrated on three geographically-diverse test systems where they were able to rapidly identify initial system designs that met resource adequacy criteria and quickly reduce the cost through continued iteration.

In spite of the enhancements described in this work, the endogenous risk curves remain a first-order estimate of system risk as a function of average surplus capacity. More sophisticated approximations that consider potential variability around that average, such as those developed in [33], would improve the robustness of the risk curves under varying system designs and provide the expansion problem with novel insights into the reliability benefits of building many smaller thermal units as compared to a few larger ones.

Finally, this work has focused on improving capacity expansion models' ability to ef-

ficiently balance probabilistic resource adequacy targets and infrastructure capital costs. However, operating costs are another important factor in identifying cost-effective power system designs. While there is a large body of literature available on time series aggregation techniques for improving operating cost estimates in capacity expansion models [45], that work typically focuses only on clustering time periods with similar weather and load characteristics and choosing the best representatives for each cluster. Sparse chronology methods for long duration storage modeling, such those used here, depend on contiguous repeated blocks of the same representative day to maximize computational efficiency, adding a new discrete dissimilarity measure to consider when assessing the fitness of a clustering assignment. More work is needed to understand the computational tradeoffs between the number of representative periods in a problem formulation and the number of transitions between such periods in a chronological sequence. Insights from that work should then be applied to identify effective clustering strategies for balancing discrete temporal adjacency against more traditional similarity measures based on Euclidean distances.

Chapter 6
CONCLUSIONS AND FUTURE WORK

This dissertation constitutes a sequence of research efforts advancing the state-of-the-art in power system capacity expansion modeling subject to probabilistic resource adequacy criteria. This includes enhancing the ability of planning models to endogenously understand uncertainty arising from probabilistic thermal generator outage risks (Chapter 2), dynamically identifying and considering periods of elevated risk arising from variable and energy-limited resources (Chapter 3), developing endogenous risk curves to produce probabilistic shortfall estimates inside a deterministic optimization (Chapter 4), and integrating these various techniques into a single mathematical framework that can be solved efficiently (Chapter 5). While this work represents a useful contribution in its own right and provides a number of practical methods that are ready for adoption by power system planners, it also raises many more research questions.

For example, there are multiple aspects of the adaptive stress period identification process described in Chapter 3 and extended in Chapter 5 that could be studied further and likely improved upon. The risk period iteration process requires a policy for identifying which periods to add to the planning problem in each iteration: the particular approach used in this work is a basic heuristic based on total EUE within a candidate stress period. It is reasonable to assume that more sophisticated methods using more information about the nature of observed shortfalls, and contributions of energy-limited resources during those shortfalls, could more efficiently identify the key risk periods faced by the system. In particular, for systems with significant potential for shifting energy across time (i.e., high levels of long-duration energy storage), it is entire possible that underlying sources of system stress are temporally distant from the time periods in which that stress is manifested as unserved energy. More sophisticated approaches that trace this causal link could enable more strategic stress period selection, reducing the number of iterations "wasted" by adding

risk periods that don't materially change the intermediate solution and keeping the overall problem size smaller.

Similarly, the temporal partitioning schemes used in Chapters 3 and 5 to select representative days from the system's operating horizon are extremely basic, considering only calendar-based characteristics of periods (grouping by weeks, months, seasons, and/or years). More sophisticated clustering that considers resource and load characteristics on a day-by-day basis, coupled with an incentive to assign temporally-adjacent days to the same partition, could more accurately capture the system's true ability to shift energy across time from the outset, resulting in both fewer total iterations required to achieve resource adequacy, and a smaller problem size overall. Furthermore, individual stress periods are always assigned to their own clusters, even though there may be many time periods with similarly-challenging conditions that are better represented by a stress period than a representative period. Dynamically reclustering periods as new stress periods are identified could help improve a CEM's internal representation of operating costs and shortfall risks.

There are also potential enhancements available to the endogenous risk curves developed in Chapter 4 and extended in Chapter 5. Those risk curves estimate EUE based solely on average available capacity in a location and timestep. However, while adding new resources with uncertain availability increases that average availability, it also increases the variance around that average. The current approach estimates EUE by assuming the existing available capacity distribution is merely shifted as capacity is added or removed, and ignores any changes to the shape of the overall distribution. These variance changes can greatly influence the assessed reliability of the system, for example when replacing multiple small generators with a single large generator, even if it provides the same average available capacity. To capture these higher-order considerations, the adequacy estimator curve could be extended to a higher dimensional surface that considers the direct adequacy impact of adding or removing different distinct classes of resources. Since the higher-dimensional function would be able to distinguish between different portfolios with the same average capacity, results from previous probabilistic assessments could be accumulated as cutting planes approximating a universal EUE estimator function, rather than discarding previous results each iteration and re-computing the EUE curves based on the last iteration's results

alone.

Finally, while the methods developed in this work present compelling theoretical advantages over the state-of-the-art, there remains much work to be done in proving out their value in real-world case studies. Additional practical analysis that demonstrates how these methods provide consistent computational advantages and/or higher quality solutions relative to existing industry standard approaches will be critical to encouraging adoption of these methods among practitioners. While such comparisons could be done in a purely academic environment, working more closely with potential software tool developers and end users can help accelerate the diffusion of these concepts and ultimately accelerate moving industry planning practices closer to where they need to be to confidently and responsibly navigate the ongoing energy transition.

BIBLIOGRAPHY

[1] W. Cole, B. Frew, T. Mai, Y. Sun, J. Bistline, G. Blanford, D. Young, C. Marcy, C. Namovicz, R. Edelman, B. Meroney, R. Sims, J. Stenhouse, and P. Donohoo-Vallett. Variable renewable energy in long-term planning models: A multi-model perspective. Technical Report NREL/TP-6A20-70528, National Renewable Energy Laboratory (NREL), Golden, CO, November 2017.

[2] P. Maloney, P. Chitkara, J. McCalley, B.F. Hobbs, C.T.M. Clack, M.A. Ortega-Vazquez, A. Tuohy, A. Gaikwad, and J. Roark. Research to develop the next generation of electric power capacity expansion tools: What would address the needs of planners? *International Journal of Electrical Power & Energy Systems*, 121:106089, 2020.

[3] E. Lannoye, I. Danti Lopez, and G. de Mijolla. Resource adequacy philosophy: A guide to resource adequacy concepts and approaches. Technical Report 3002024368, Electric Power Research Institute (EPRI), 2022.

[4] R. Billinton. *Power System Reliability Evaluation.* Gordon and Breach, New York, NY, 1970.

[5] R. Billinton and W. Li. *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods.* Plenum Press, New York, NY, 1994.

[6] Gord Stephen, Simon H. Tindemans, John Fazio, Chris Dent, Armando Figueroa Acevedo, Bagen Bagen, Alex Crawford, Andreas Klaube, Douglas Logan, and Daniel Burke. Clarifying the interpretation and use of the LOLE resource adequacy metric. In *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 1–4, June 2022.

[7] G. de Mijolla. Resource adequacy for a decarbonized future: A summary of existing and proposed resource adequacy metrics. Technical Report 3002023230, Electric Power Research Institute, 2022.

[8] K. Eurek, W. Cole, D. Bielen, N. Blair, S. Cohen, B. Frew, J. Ho, V. Krishnan, T. Mai, B. Sigrin, and D. Steinberg. Regional Energy Deployment System (ReEDS) model documentation: Version 2016. Technical Report NREL/TP-6A20-67067, National Renewable Energy Laboratory (NREL), Golden, CO, November 2016.

[9] S. Zachary and C. J. Dent. Probability theory of capacity value of additional generation. *Proc. IMechE Part O: J. Risk and Reliability*, 226:33–43, July 2011.

[10] NERC Reliability Assessment Subcommittee. 2019 long-term reliability assessment. Technical report, North American Electric Reliability Corporation, 2019.

[11] B. Frew, G. Stephen, D. Sigler, J. Lau, W. B. Jones, and A. Bloom. Evaluating resource adequacy impacts on energy market prices across wind and solar penetration levels. *The Electricity Journal*, 32:106629, October 2019.

[12] NERC Probabilistic Assessment Working Group. Probabilistic adequacy and measures. Technical reference report, North American Electric Reliability Corporation, July 2018.

[13] A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.

[14] B. S. Palmintier and M. D. Webster. Heterogeneous unit clustering for efficient operational flexibility modeling. *IEEE Transactions on Power Systems*, 29:1089–1098, May 2014.

[15] E. Kreyszig. *Advanced Engineering Mathematics, 9th Edition.* John Wiley & Sons, Hoboken, NJ, 2006.

[16] M. Milligan and B. Parsons. A comparison and case study of capacity credit algorithms for intermittent generators. Technical Report NREL/CP-440-22591, National Renewable Energy Laboratory, Golden, CO, March 1997.

[17] S. H. Madaeni, R. Sioshansi, and P. Denholm. Comparison of capacity value methods for photovoltaics in the western united states. Technical Report NREL/TP-6A20-54704, National Renewable Energy Laboratory, Golden, CO, July 2012.

[18] J. Jorgensen, S. Awara, G. Stephen, and T. Mai. Comparing capacity credit calculations for wind: A case study in texas. Technical Report NREL/TP-5C00-80486, National Renewable Energy Laboratory, Golden, CO, September 2021.

[19] G. Stephen, E. Hale, and B. Cowiestoll. Managing solar photovoltaic integration in the western united states: Resource adequacy considerations. Technical Report NREL/TP-6A20-72472, National Renewable Energy Laboratory, Golden, CO, January 2021.

[20] E. Hale, B. Stoll, and T. Mai. Capturing the impact of storage and other flexible technologies on electric system planning. Technical Report NREL/TP-6A20-65726, National Renewable Energy Laboratory, Golden, CO, May 2016.

[21] A. Mills and P. Rodriguez. A simple and fast algorithm for estimating the capacity credit of solar and storage. *Energy*, 210:118587, November 2020.

[22] G. Stephen, T. Joswig-Jones, S. Awara, and D. Kirschen. Impact of storage dispatch assumptions on resource adequacy and capacity credit. In *17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Manchester, UK, June 2022.

[23] J. Nelson and B. Heath. ELCC surface in resource planning: Dynamic capacity contribution for energy-limited resources. In *IEEE Resource Adequacy Working Group Annual Meeting*, July 2021.

[24] Northwest Power and Conservation Council. Associated system capacity contribution. `https://www.nwcouncil.org/2021powerplan_associated-system-capacity-contribution/`, 2021. Accessed: 2022-08-22.

[25] Hawaiian Electric Company. Energy reserve margin criteria analysis. `https://www.hawaiianelectric.com/documents/clean_energy_hawaii/integrated_grid_planning/stakeholder_engagement/technical_advisory_panel/20211101_tap_meeting_presentation_materials.pdf`, November 2021. Accessed: 2022-08-22.

[26] T. Mertens, K. Bruninx, J. Duerinck, and E. Delarue. Adequacy aware long-term energy-system optimization models considering stochastic peak demand. *Advances in Applied Energy*, 4:100072, November 2021.

[27] MIT Energy Initiative. The future of energy storage. Technical report, Massachusetts Institute of Technology, June 2022.

[28] C. Barrows, A. Bloom, A. Ehlen, J. Ikäheimo, J. Jorgenson, D. Krishnamurthy, et al. The IEEE Reliability Test System: A proposed 2019 update. *IEEE Transactions on Power Systems*, 35:119–127, January 2020.

[29] National Renewable Energy Laboratory. Annual technology baseline. `https://atb.nrel.gov/electricity/2023/data`, July 2023.

[30] G. E. Haringa, G. A. Jordan, and L. L. Garver. Application of Monte Carlo simulation to multi-area reliability evaluations. *IEEE Computer Applications in Power*, 4:21–25, January 1991.

[31] Xavier Blanchot, François Clautiaux, Aurélien Froger, and Manuel Ruiz. Modeling and solving a stochastic generation and transmission expansion planning problem with a "loss of load expectation" reliability criterion. *hal-03957750*, 2023.

[32] Seyyed A. Rashidaee, Turaj Amraee, and Mahmud Fotuhi-Firuzabad. A linear model for dynamic generation expansion planning considering loss of load probability. *IEEE Transactions on Power Systems*, 33(6):6924–6934, Nov 2018.

[33] G. Stephen and D. Kirschen. Enhanced representations of thermal generator outage risk in capacity expansion models. In *17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Manchester, UK, June 2022.

[34] B. Frew, G. Stephen, D. Sigler, J. Lau, W. B. Jones, and A. Bloom. Evaluating resource adequacy impacts on energy market prices across wind and solar penetration levels. *The Electricity Journal*, 32:106629, October 2019.

[35] Joe Oteng-Adjei, Abdul-Majid Issah Malori, and Emmanuel Kwaku Anto. Generation system expansion planning using loss of load expectation criterion. In *2020 IEEE PES/IAS PowerAfrica*, pages 1–5, Aug 2020.

[36] James H. Williams, Ryan A. Jones, Ben Haley, Gabe Kwok, Jeremy Hargreaves, Jamil Farbes, and Margaret S. Torn. Carbon-neutral pathways for the United States. *AGU Advances*, 2(1):e2020AV000284, 2021.

[37] Energy+Environmental Economics. Hawaiian electric resource adequacy workplan update. `https://www.hawaiianelectric.com/a/12540`, July 2023. Accessed: 2024-01-12.

[38] Stan Zachary, Amy Wilson, and Chris Dent. The integration of variable generation and storage into electricity capacity markets. *The Energy Journal*, 43(4), 2022.

[39] G. Stephen. Probabilistic Resource Adequacy Suite (PRAS) v0.6 model documentation. Technical Report NREL/TP-5C00-79698, National Renewable Energy Laboratory (NREL), May 2021.

[40] Björn Bahl, Alexander Kümpel, Hagen Seele, Matthias Lampe, and André Bardow. Time-series aggregation for synthesis problems by bounding error in the objective function. *Energy*, 135:900–912, 2017.

[41] Giuseppe Calabrese. Generating reserve capacity determined by the probability method. *Transactions of the American Institute of Electrical Engineers*, 66(1):1439–1450, 1947.

[42] Derek Stenclik, Aaron Bloom, Wesley Cole, Gord Stephen, Armand Figueroa Acevedo, Rob Gramlich, Chris Dent, Nick Schlag, and Michael Milligan. Quantifying risk in an uncertain future: The evolution of resource adequacy. *IEEE Power and Energy Magazine*, 19(6):29–36, 2021.

[43] Jess Kuna, Gord Stephen, and Trieu Mai. Beyond capacity credits: Adaptive stress period planning for evolving power systems. Technical Report NREL/TP-6A20-73067, National Renewable Energy Laboratory (NREL), 2024.

[44] Gord Stephen and Daniel Kirschen. Endogenizing probabilistic resource adequacy risks in deterministic capacity expansion models. In *2024 18th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 1–6, June 2024.

[45] Holger Teichgraeber and Adam R. Brandt. Time-series aggregation for the optimization of energy systems: Goals, challenges, approaches, and opportunities. *Renewable and Sustainable Energy Reviews*, 157:111984, 2022.

[46] Yunzhi Chen, Brian Sergi, Jonathan Ho, Gord Stephen, Wesley Cole, and Kody M. Powell. Sparse chronology strategy for integrating seasonal energy storage in capacity expansion models. 2024.

[47] Ed Klotz. *Identification, Assessment, and Correction of Ill-Conditioning and Numerical Instability in Linear and Integer Programs*, chapter Chapter 3, pages 54–108. Institute for Operations Research and the Management Sciences, 2014.

[48] Galen Maclaurin, Nicholas Grue, Anthony Lopez, Donna Heimiller, Michael Rossol, Grant Buster, and Travis Williams. The renewable energy potential (rev) model: a geospatial platform for technical potential and supply curve modeling. Technical Report NREL/TP-6A20-73067, National Renewable Energy Laboratory (NREL), 2021.