Comparison of Different Approaches for Solar PV and Storage Sizing

Fiodar Kazhamiaka, Yashar Ghiassi-Farrokhfal, Srinivasan Keshav, and Catherine Rosenberg

***Abstract*—We study the problem of optimally and simultaneously sizing solar photovoltaic (PV) and storage capacity in order to partly or completely offset grid usage. While prior work offers some insights, researchers typically consider only a single sizing approach. In contrast, we use a firm theoretical foundation to compare and contrast sizing approaches based on robust simulation, robust optimization, and stochastic network calculus. We evaluate the robustness and computational complexity of these approaches in a realistic setting to provide practical, robust advice on system sizing.**

I. Introduction

In the last few years, the prices of solar panels and storage have dropped dramatically, putting them in reach of many consumers. Companies such as Trina, Yingli, and Canadian Solar offer solar panels at a cost of less than USD 0.5/Watt, and companies such as Tesla, Sonnen, and Moixa provide off-the-shelf (albeit expensive) storage solutions.

Consider an entity that wants to purchase and install solar PV panels and storage in order to partly or completely offset grid usage[1]. How much of each should they buy? If the budget is not a constraint, then both can be generously sized, with ample slack capacity. However, given the high cost of storage, budget is often a binding constraint. Thus, we would like to provide practical guidance on the smallest possible *sizing*[2] to adequately meet the anticipated load. This is the subject of our work.

We expect many entities to face such a sizing problem in the future. These include individuals, small companies, and building operators faced with the rising cost of grid-provided electricity.

While prior work on this topic offers some insights, researchers typically consider only a single sizing approach [1], [2], [3], [4], [5], [6]. Moreover, the approaches advocated by some past researchers results in sizing decisions that may not be robust to perturbations in the inputs. In our work, we attempt to provide *practical*, *robust* advice on system sizing. To do so, we compare and contrast multiple sizing approaches, extending well-known approaches as necessary to reduce them to practice.

The approaches we study use historical solar generation and electricity consumption (load) time series as input to compute the sizing. Given that this data is difficult to obtain for a horizon long enough to adequately capture the non-stationarity of the underlying stochastic processes, we assume that past and future samples are drawn from the same distribution. Nevertheless, any practical data-driven approach must take steps to prevent overfitting to historical data.

In our work, we tackle the problem of overfitting by extending three approaches to computing a *robust* sizing: simulation, mathematical programming, and stochastic network calculus. With simulation and mathematical programming approaches, we compute a robust sizing by using upper probability bounds on the sizings that meet the performance requirements on historical data; with stochastic network calculus, roughly speaking, we reduce the available data to a set of representative features that are then used to compute probability bounds on the performance targets achieved by any given sizing (this statement is made more precise later in the paper).

We make three key contributions:

- We provide a firm theoretical foundation for robust and practical sizing of both solar PV generation and storage based on three approaches: simulation, optimization, and stochastic network calculus.
- We make contributions to the state-of-the-art in stochastic network calculus.
- We evaluate the robustness and computational complexity of these approaches in a realistic setting.

We have publicly released the program modules for computing robust PV-storage system sizing via simulation and stochastic network calculus [7].

II. Related Work

Prior work on sizing approaches for energy storage in the presence of renewable energy sources can be grouped into three main classes: mathematical programming, simulation, and analytical methods. We sketch these approaches here, with a survey of representative work, deferring details of each approach to Section V.

A. Mathematical Programming

There exist many methods for solving sizing optimization problems. In this paper, we focus on mathematical programming, which is a scenario-based approach. It requires modelling the system as a set of parameters and variables that

F. Kazhamiaka and S. Keshav are with the Cheriton School and Computer Science, University of Waterloo.
Contact E-mail: fkazhami@uwaterloo.ca

Y. Ghiassi-Farrokhfal is with the Rotterdam School of Management at Erasmus University.

C. Rosenberg is with the Electrical and Computer Engineering department at the University of Waterloo.

[1]The former case corresponds to that of an entity that remains grid-connected but wants to reduce its overall cost for electricity and the latter corresponds to an off-grid scenario. We treat them both identically in our work.

[2]By sizing, we refer to the power/energy size of the storage in kW/kWh and the size of solar generation in kWp.

are constrained to represent the capabilities of the physical system being modelled and an objective function representing the system target. Importantly, it typically does not model the operating policy; instead, the optimal operation is an output of the optimization program, and is dependent on the inputs. An algorithm, or *solver*, is used to search the space of feasible solutions to find the one which maximizes (or minimizes) the objective function for the given parameters. For example, in Reference [8], the problem of sizing a battery to meet the energy demands of a microgrid is formulated as a mixed-integer linear program. In Reference [9], the problem of sizing batteries and solar panels under a fixed budget to maximize the revenue of a solar farm is formulated as a non-linear optimization problem, which is linearized to reduce the solving time.

Another notable optimization approach is to formulate a *robust optimization* problem [10], in which the objective function is optimized even when the inputs are perturbed. We do not cover robust optimization in this paper; rather, we present a simpler approach to dealing with uncertainties in the input parameters.

B. Simulation

Simulations are *scenario-based* sizing approaches that provide optimal system sizing for a given trajectory (i.e, a time series) for load and PV generation. They are versatile: a simulation program can evaluate different combinations of PV panel and battery sizes, calculating metrics such as loss of load probability (LOLP) [4], expected unserved energy (EUE), and operating cost [11]. The simulated system can be operated using virtually any operating strategy, such as those proposed in [3], [6], [2], [4], and can implement complex battery models [11].

C. Analytical Methods

Inspired by the analogy between energy buffering by batteries and data buffering in computer networks, a variety of analytical methods have been proposed for storage capacity sizing in the literature. For example, in Reference [12] the system is modelled as a cyclic non-homogenous Markov chain, and the authors propose a steady-state analysis to determine whether a given system size is sufficient to meet a target LOLP. In Reference [13], the authors use a probabilistic tail bound on the aggregate of many regulated energy demand loads to jointly size the battery capacity and transformers for a certain LOLP in a residential setting.

Among existing analytical approaches, stochastic network calculus (SNC) [14] has shown great robustness and accuracy. This approach has been used in several applications: battery sizing to reduce reliance on diesel generators in rural areas with unreliable grid connections [15], energy demand management in a fleet of electric car charging stations [16], gaining energy flexibility through heating/cooling systems in data centres [17], supply-demand matching for prosumers [18], [19], [20], and profit maximization for renewables in electricity markets [21].

Applying stochastic network calculus to energy systems has some subtleties, due to the unique statistical properties of the underlying energy processes and the storage model in use. This has led to a series of incremental improvements in this field of research. The idea of using stochastic network calculus for energy systems was proposed in [19], where the authors assume ideal storage devices and use affine functions to separately model the long-term behavior of each of energy demand and energy supply. In Reference [18], the authors improve this approach by assuming a more realistic storage model and more complicated uni-variate envelopes for energy demand and supply. It is shown in [21] that uni-variate envelopes cannot properly capture the statistical properties of solar power due to its substantial seasonality; hence, introducing bi-variate envelopes to separately model the long term behavior of energy demand and supply. In this paper, we advance the state-of-the-art as discussed in Section VII-B.

III. Goal

At a high level, the goal of our work is to provide robust, practical advice on how to size both solar panels and storage to partly or completely offset grid usage. This section discusses the inputs and objective of this sizing problem.

A. Inputs

It is reasonable to assume that an entity making a sizing decision would have access to a representative set of load traces, especially with the widespread deployment of smart meters that typically measure hourly load[3]. It is also possible to obtain hourly solar radiation traces in the geographical location of the entity, for most parts of the world [22], and calculate the corresponding power generated from PV panels with reasonable accuracy [23] .

In keeping with prior work, we make the assumption that these historical traces are generally representative of loads and generation. Nevertheless, the future will never exactly mimic the past; if it did, we would be able to make decisions with perfect information. Thus, the sizing decision must be robust to perturbations in the inputs, i.e., to 'small' changes in the solar irradiation or loads (we make this precise in Section IV).

In addition to generation and load traces, we need two other inputs. First, we need to know how a decision is made to either inject power into or withdraw power from the storage system. This *operating policy* can be quite complex, and is the subject of much research [3], [12], [24], [9]. Nevertheless, simple rules such as 'store excess solar energy' and 'discharge the store when solar generation is less than the load' are often adequate for most situations. We assume that, for the case of simulation and stochastic network calculus approaches, such an operating policy is provided to the sizing decision-maker. Second, it is necessary to model the behaviour of a storage system in response to power injection and discharge. We use a recently-proposed storage model in our work [25].

To summarize, we assume that the sizing decision-maker has access to the following inputs:

[3]Finer-grained traces would, of course, be good to have, but unlikely to be available in practice.

- A representative set of *solar traces* $S = \{S^i\}$ (for now, think of them as one trace per-year, but we discuss this point in more detail in Section IV-C).
- A representative set of load traces $D = \{D^j\}$ that constitute a set of *load scenarios*. Each load trace needs to be of the same time duration as the solar traces.
- An *operating policy*: for the simulation and stochastic network calculus approaches, the set of rules that determine when the store is charged or discharged.
- A *storage model*, along with all associated model parameters: given the current state of charge, and the applied power, this is a set of equations that computes the new state of charge.

B. Sizing Objective

Given the inputs in Section III-A, our objective is to compute the "best" sizing for solar PV panels and the storage capacity. What constitutes the best choice will depend on the situation at hand. Several quality metrics are plausible[4]:

- **Minimize LOLP:** This is the probability that the system is unable to meet the load from *solar* generation. This probability can be numerically estimated as the ratio of the time period during which the load is unmet from solar generation to the total time period under consideration.
- **Minimize expected unserved energy (EUE):** This is the total amount of load (energy) that cannot be delivered from solar generation during the period under consideration. If this load is not met from the grid, there will be user discomfort.
- **Minimize financial cost:** This is the dollar cost of purchasing the solar panel and storage system, as well as the cost of purchasing, as necessary, electricity from the grid, at its currently prevailing price. It can be viewed either as a one-time capital cost added to a periodical operational expense, due to potential purchases from the grid and the eventual degradation of the equipment from wear and tear. Note that if we can associate a cost to meeting unmet load from the grid or a diesel generator, then the cost-minimization objective incorporates the objective of minimizing unmet load.
- **Maximize robustness:** This is the degree of sensitivity of the sizing to perturbations in the input. Intuitively speaking, we wish to pick an approach such that small perturbations in the inputs result in only a small perturbation in the sizing [26]. We discuss this point in greater detail in Section IV.
- **Minimize computation time:** We expect that the sizing decision will be made on behalf of a system purchaser by a sizing decision maker. The computation cost of each such decision, therefore, should not be onerous.

In many cases, there will be a trade-off between cost on the one hand, and LOLP/EUE and robustness on the other. Moreover, robustness and computation cost go hand-in-hand, since to get robust results we (generally) have to process more data and hence perform more computation. In this work, for concreteness, we focus on minimizing the cost of solar PV and storage, subject to meeting a certain LOLP or EUE constraint.

Traditionally, the LOLP/EUE target is specified together with a length of time over which this criteria should be met [27]. For example, a common loss-of-load target for reliable grid-scale electrical systems is one day over a period of 10 years, corresponding to an LOLP target of 0.000274. Such a high level of reliability makes sense when frequent or prolonged loss-of-load events correspond to millions of dollars in losses to the economy supported by the electrical system. Achieving this level of reliability equates to sizing for the worst-case behaviour with virtually 100% confidence, requiring expensive systems that are oversized for the average behaviour but are nevertheless cheaper than the cost of loss-of-load events.

In contrast, smaller microgrids such as a house can tolerate higher LOLP due to smaller penalties associated with loss-of-load events, the ability to easily shut off electrical appliances at times of high load with negligible costs, and the availability of the grid and perhaps a diesel generator to offset some of these events. For the same reasons, sizing the system for the worst case would be a sub-optimal financial decision. Hence, a home-owner could desire a system that achieves target LOLP of 5% over all 90-day periods, with a confidence of 95%. Mathematically, this corresponds to a system which gives $\mathbb{P}(LOLP \leq 0.05) \geq 0.95$ over any 90-day period. We will refer to the time interval, the LOLP/EUE target, and the confidence as the quality of service (QoS) target.

Using other optimization objectives is also possible, and discussed at greater length in Section VII.

IV. The Impact of Non-Stationarity

A key insight in our work is that the traces which serve as input to any sizing approach may neither be stationary nor representative of the future. We discuss this next.

A. Traces, Trajectories, and Stochastic Processes

A solar or load trace with T entries of the form ($time$, $value$) is a trajectory instantiated from a stochastic process, which is defined as a set of random variables indexed by time. That is, $S^i(t)$, the t^{th} element of the i^{th} solar trace (resp. $D^j(t)$, the t^{th} element of the j^{th} load trace) is a value assumed by the random variable $S(t)$ (resp. $D(t)$) from a corresponding distribution. Hence, we can fully characterize the *historical* solar (resp. load) stochastic process by defining joint distribution of a set of T random variables, one for each time step. Assuming independence of each time step, we can decouple these distributions, allowing us to use the set S (resp. D) of solar generation (resp. load) traces to estimate parameters for each of the T distributions. For example, the numerical mean of the t^{th} time step of the set of traces can be viewed as an estimate of the mean of the t^{th} distribution and the sample variance of this set is an estimate of its variance. Thus, with sufficient data, we can use standard statistical techniques to find the best-fitting distributions that characterize a set of traces.

[4]For each application, one or multiple of these items can serve as objectives and one or multiple others as constraints.

Given this characterization of historical stochastic processes, what can we say about the future? Suppose that the generation and load stochastic processes were time-invariant. Then, once the historical processes are characterized, the future is also 'known' in that we can generate potential future trajectories by generating a random value per time step from the corresponding distribution. We can then choose a sizing that meets our sizing objectives not just for historical trajectories, but also for potential future trajectories.

However, this naive approach has three problems. First, even assuming independence of time steps, it is onerous to define T separate distributions, since T can be very large, on the order of 10,000 – 100,000 values. Second, there is no guarantee that a stochastic process parametrized based on historical traces will adequately represent the future. Third, we do not have any guidelines on how much data is 'enough.' To solve these problems, we need to take a closer look at the generation and load stochastic processes.

B. Causes of Non-Stationarity

A key observation is that both the solar and load stochastic processes are *non-stationary*[5] due to three effects:

1) **Diurnality.** For example, the distribution of the r.v. $S(t)$ corresponding to a time slot t at night will differ from the distribution of an r.v. corresponding to a time slot at mid-day.
2) **Seasonality.** For example, the distribution of the r.v. $S(t)$ corresponding to a time slot t at mid-day in winter will differ from the distribution of an r.v. corresponding to a time slot at mid-day in summer.
3) **Long-term trends.** For example, the distribution of the r.v.'s $D(t)$ and $S(t)$ corresponding to a time slot t at the start of a trace may differ from their distribution for a time slot later in the trace.
4) **Autoregressivity.** For example, the distributions of the r.v.'s $S(t)$ and $D(t)$ are dependent on the values taken by the respective r.v.'s $S(t-1)$ and $D(t-1)$.

Non-stationarity should be taken into account upon characterizing historical generation and load stochastic processes.

C. Stochastic Process Parametrization

Recall that the parameters of the stochastic process, i.e., corresponding to each of the T distributions constituting the process, are derived from solar and load traces. Given that the process has both diurnal and seasonal non-stationarity effects, the solar and load traces must be both *detailed* enough and *long* enough to capture all three effects. More precisely, :

- The traces should have sufficient temporal resolution to capture diurnal changes. That is, the time step should be sufficiently small to have an adequate number of values for each part of the day.
- The traces should be long enough to capture seasonality, i.e., at least one year in duration, if not longer.
- The traces should be long enough to capture any long-term trends in load (we assume that solar generation is stationary at the time scale of a year).
- There should be enough traces in the set of traces so that there are sufficient samples to adequately estimate the parameters of each distribution.

Ideally, we would have access to per-minute load and generation traces spanning several decades. Then, setting $T = 60 \times 24 \times 365 = 525,600$, i.e., per minute of the year, we would obtain multiple sample values for each time step, allowing us to estimate, with adequate confidence, the parameters of each of the T solar generation distributions, and potentially long-term trends in the load (for example by fitting a linear regression to the residual after accounting for diurnal and seasonal effects).

In practice, it is unlikely that such data traces are available. Load and solar generation are often measured at a time scale of 30 minutes or longer, and it is rare to have more than a year or two of data. Therefore, we resort to the following pragmatic approach that, in our experience, works reasonably well: create a set of sub-samples from data traces that spans multiple seasons. More precisely, given a dataset in the form of a time series that is at least one year long and starts and ends on the same day, we join the ends of the data to create a circular time series. We then sample blocks of X days, where X is the time interval in the QoS target, from the time series by choosing random start times. We call each of the n samples a 'scenario', and treat their ensemble as an estimate of expected future scenarios.

We now discuss three robust sizing approaches that base their sizing decisions on these 'traces of the future.' We defer a discussion on how to evaluate the robustness of these approaches to Section VI.

V. Robust Sizing Approaches

The sizing decision, that is, choosing the size of the store B (in kW/kWh) and of the solar panels C (in kW) to meet one or more of the objectives discussed in Section III-B, can be made using many different approaches. In this section, we present three representative approaches.

We make the following assumptions:

- For simplicity, and for reasons of space, we assume that the goal is to find the minimum-cost storage and solar PV sizes that meet a certain LOLP or EUE QoS criterion over a given time duration of X days.
- We assume that we have available a solar generation trace S and load trace D corresponding to the same time interval and with a length of at least one year. As already discussed, from these traces we can obtain n scenario samples of X days of solar generation and the corresponding load.
- We only size the storage system for energy, not for power, since sizing for power is typically trivial (the power rating of the storage system must exceed the sum of power draws of the set of simultaneously active load components).
- We assume the storage system energy capacity can only take one of b different values and that the solar panel size

[5]Roughly speaking, this means that statistics computed from two different random sub-samples of the traces can differ.

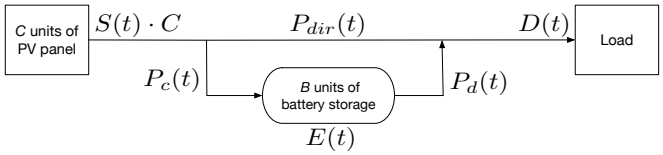


Figure 1: System diagram

can only take on one of c different values. Hence, sizing a system is essentially conducting a grid search through bc pairs of solar PV panel sizes and storage capacity sizes and determining the optimal sizing, i.e., the sizing with the minimum cost which has some guarantee of satisfying the target LOLP/EUE.

- We assume that if a certain combination of storage and PV values results in a certain LOLP/EUE, then larger values of either storage or PV will always result in lower values of LOLP/EUE. This allows us to use a greedy grid search heuristic.

We denote the number of time steps in the load and solar generation traces by T. π_B is the price for one unit of battery (i.e., 1 cell), and π_C is the price for one unit of PV panel. We normalize the solar generation S^i trace, so that it represents the generation from a single PV panel unit. Finally, the LOLP target is denoted ϵ, and the EUE target, expressed as a fraction of the total load, is denoted θ.

A. Optimization

In this approach, we formulate an optimization program for solar panel and battery sizing with the objective of minimizing the capital cost of the system, subject to physical system and LOLP/EUE constraints. We do not specify the operating policy, leaving this decision to the optimization solver. This allows us to compute the best possible sizing in the case of optimal operation. In this sense, although the sizing decision made by the optimization program is a potentially-unattainable lower bound, it measures the level of sub-optimality in the operating policy used in the two other approaches.

Our approach has two phases. In the first phase, for each scenario and for each of the b potential battery sizes, we compute the optimal solar panel size C, assuming optimal operation. This gives us n *sizing curves* defined by the interpolation of b sizings that were computed for each scenario (see Figure 2). In the second phase, we use a technique based on the Sample Univariate Chebyshev bound [28] to compute a *robust* sizing that is insensitive to the details of individual traces. We discuss each phase in turn.

1) Phase 1: For phase 1, define P_c to be the charging power, P_d to be the discharging power, P_{dir} to be the power that flows directly from PV panels to load, and E to be the energy content. The size of the battery is B and the generation capacity of the panels is C. Figure 1 shows a labelled system diagram.

The battery model used here is Model 1* from [25] with the following parameters: η_c (resp. η_d) the charging (resp. discharging) efficiency, α_c (resp. α_d) the charging (resp. discharging) rate limit, u_1, v_1, u_2, v_2 used to characterize the power-dependent lower and upper limits on the energy content (see constraint (6)). The energy content at the end of time slot t is denoted $E(t)$, and the initial energy content is U. Recall that the data trace for solar generation S is for one unit of PV panel and that D denote the household electricity load. The duration of a time-step is T_u and the number of time-steps in a data trace is T.

We will first present the full formulation with an LOLP constraint, and then show how it can be modified to work with a EUE constraint. Given a scenario ($S(t)$), ($D(t)$), and storage parameters B, η_c, η_d, α_c, α_d, u_1, v_1, u_2, v_2, U, and trace parameters T_u and T, the problem can be formulated as:

$$\min_{\substack{C,P_c,P_d,\\ P_{dir},I,\gamma,E}} C \tag{1}$$

subject to

$$P_c(t) + P_{dir}(t) \leq S(t)C \quad \forall t \tag{2}$$
$$P_{dir}(t) + P_d(t) = D(t) - \delta(t) \tag{3}$$
$$E(0) = U \tag{4}$$
$$E(t) = E(t-1) + P_c(t)\eta_c T_u - P_d(t)\eta_d T_u \quad \forall t \tag{5}$$
$$u_1 P_d(t) + v_1 B \leq E(t) \leq u_2 P_c(t) + v_2 B \quad \forall t \tag{6}$$
$$0 \leq P_c(t) \leq B\alpha_c \quad \forall t \tag{7}$$
$$0 \leq P_d(t) \leq B\alpha_d \quad \forall t \tag{8}$$
$$I(t) \in \{0,1\} \quad \forall t \tag{9}$$
$$B, C, P_{dir}(t), \delta(t), E(t) \geq 0 \quad \forall t \tag{10}$$
$$1/T \sum_{t=1}^{T} I(t) \leq \epsilon \tag{11}$$
$$I(t) \leq \delta(t) Z \quad \forall t \tag{12}$$
$$\delta(t) \leq I(t) D(t) \quad \forall t \tag{13}$$
$$P_c(t) P_d(t) = 0 \ \forall t \tag{14}$$

where constraint 2 states that the sum of what goes into the battery and directly towards the load is bounded by the solar generation, $\delta(t)$ is the load that is not met from solar generation at time t (it is always $\leq D(t)$), constraints 5–8) represent the battery model, and $I(t)$ is an binary variable used to indicate if the load is met or not in time-step t ($I(t) = 1$ means the load is not met). Constraint 12 ensures that $I(t)$ is zero if $\delta(t) = 0$ (Z is a large positive constant), constraint 13 ensures that $I(t)$ is one if $\delta(t) > 0$ and constraint (11) is the LOLP constraint. Constraint 14 forbids simultaneous charging and discharging and it was shown in [9] that it can be ignored which makes the problem an Integer Linear Program (ILP). Note that in this problem B and C are real numbers, i.e., they are not limited to the pre-defined values used for the other two approaches.

To express an EUE constraint, this problem formulation can be modified as follows. Replace Constraints 11–13 with:

$$\sum_{t=1}^{T} \delta(t) \leq \theta \sum_{t=1}^{T} D(t) \tag{15}$$

Note that the formulation with the EUE constraint is a Linear Program (LP), which can be solved much more efficiently than an ILP.

We use this mathematical program to compute the smallest[6] C for each of the b battery sizes so that the system meets the QoS target; these points define a curve in a (B, C) space. We denote each curve as K_i corresponding to the ith scenario, for a total of n curves.

2) Phase 2: In phase 2, we use the n sizing curves obtained in phase 1 to compute a probability bound on the system size with a given measure of confidence.

First, for each of the b values of B', we construct the set $L_{B'}$ consisting of points in the (B, C) space along the intersection of the line at $B = B'$ and each curve K_i.

$$L_{B'} = \{C'' : C'' = K_i(B')\} \tag{16}$$

Each set of points can be viewed as samples from a distribution defined by the sizing curves. Denote the size of the set $|L_{B'}| = N_{B'}$. Not all sizing curves have a value defined at B', so $N_{B'} \leq n$. We can compute a sample Chebyshev bound, as formulated in [29], on the C values as follows:

$$\mathbb{P}\{|C - \mu_{C,N_{B'}}| \geq \lambda \sigma_{C,N_{B'}}\} \leq \min\left(1, \frac{1}{N_{B'}+1}\left\lfloor \frac{(N_{B'}+1)(N_{B'}^2 - 1 + N_{B'}\lambda^2)}{N_{B'}^2 \lambda^2} \right\rfloor\right) \tag{17}$$

The inequality above is a bound on the probability that the difference between some future value of C for the corresponding B' from the estimated mean $\mu_{C,N_{B'}}$ exceeds a factor λ of the estimated standard deviation $\sigma_{C,N_{B'}}$. To use this inequality to compute a sizing, we first find the smallest λ that satisfies our confidence measure γ:

$$\min_{\lambda}\left(\frac{1}{N_{B'}+1}\left\lfloor \frac{(N_{B'}+1)(N_{B'}^2 - 1 + N_{B'}\lambda^2)}{N_{B'}^2 \lambda^2} \right\rfloor\right) \leq 1 - \gamma \tag{18}$$

Next, we rearrange the inequality in the LHS of Eq. 17 to obtain a robust value $C^*_{B'}$ using the λ that satisfies Eq. 18:

$$C^*_{B'} = \mu_{C,N_{B'}} + \lambda \sigma_{C,N_{B'}} \tag{19}$$

The resulting set of points $(B', C^*_{B'})$ can be interpolated to define a curve which we call the *Chebyshev curve on* C, since each point on the curve is a Chebyshev bound on C values. Similarly, we can construct a *Chebyshev curve on* B by computing Chebyshev bounds on the following sets for each of the c values of C':

$$L_{C'}\{B'' : Z_i(B'') = C'\} \tag{20}$$

The upper envelope of these Chebyshev curves represents system sizings which are *robust* with respect to both B and C with confidence measure γ. We use the least-cost system along the upper envelope as the robust sizing.

If we are confident that the estimated mean and standard deviation have converged to the population mean after n samples, we can obtain a tighter bound characterized by a value of λ that satisfies the following:

$$(1 - \lambda)^{-1} = 1 - \gamma \tag{21}$$

Using Eq. 21 in place of Eq. 18 corresponds to the classical *analytical* Chebyshev bound which assumes that the population mean and standard deviation are known.

3) Computation cost: The two Chebyshev curves can be computed with $O(bn + cn)$ computations, hence the computation time of this approach is dominated by the computation of the sizing curves via the optimization program. The inputs to the optimization program are the solar and load traces, each of size $O(T)$, for a total size of $O(T)$. Asymptotically, this is also the number of variables in the program. Denoting by $Q = O(T)$ the number of variables and $L = O(T)$ as the number of bits of input to the algorithm, even for an LP, which is far more computationally efficient than an ILP, the best-known approach, the interior-point method, requires a runtime of $O(Q^{3.5}L^2 \cdot logL \cdot loglogL) \approx O(T^{5.5} \cdot logT)$ [30]. Since we need nb such runs, and our problem is integer, the total complexity is lower bounded by $O(nbT^{5.5} \cdot logT)$.

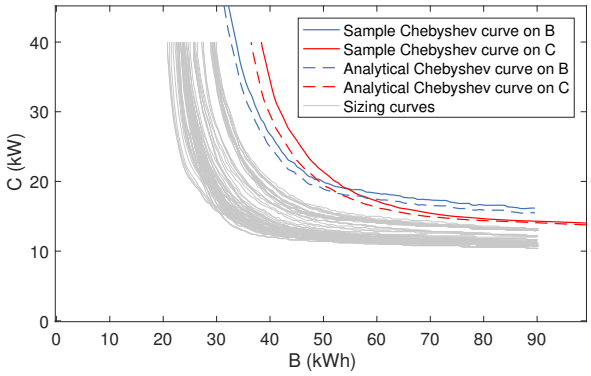


Figure 2: Fifty sizing curves as well as their sample and analytical Chebyshev curves for $\gamma = 0.95$.

B. Simulation

In this approach, we run system simulations to construct the sizing curves for each of the n scenarios. Specifically, for each scenario, for each potential sizing choice, and for each time step $t \in [1, T]$, we determine the availability of solar power $S(t)$ and the load $D(t)$. Depending on these values, the storage and PV sizes under test, and the given operating policy, we use the storage model to either charge or discharge the store, updating its SoC as in Eq. 5. If we find that there is a need to discharge the store, but its SoC is zero, then we mark that the load is unmet from solar generation for this time step. At the end of each simulation, we empirically compute the LOLP ϵ or EUE θ for this sizing.

We use a search algorithm that, given a scenario, efficiently searches the (B, C) space to compute the sizing curve:

Step 1: For $C = C^{max}$, start at $B = B^{min}$ and increment B to find the smallest value of B such that the system satisfies the target performance requirement, and store (B, C).

[6]The solution gives us the optimal C as a real number, which we round up to the nearest potential C value among the c possibilities.

Step 2: Decrement C, then start from the most recent value of B and increment it until the system satisfies the performance requirement, and store (B, C).
Step 3: Repeat previous step through $C = C^{min}$. Stored (B, C) pairs give us the sizing curve.

This algorithm first finds the edge of the curve at $C = C^{max}$, then traces the rest of the curve using at most $b + c$ simulations. We then use phase 2 as described in Section V-A2 to process these curves and compute a robust sizing from the upper envelope of the two Chebyshev curves.

Note that the computation cost of this approach is $O(nT(b+c))$, since each time step takes $O(1)$ computation time, there are T steps per simulation, and $(b + c)n$ simulations.

C. Stochastic Network Calculus (SNC)

Unlike the mathematical programming and simulation approaches, which can compute either LOLP or EUE metrics with minor changes, stochastic network calculus has significant differences in the mathematical formulation used to compute the sizing for each metric. Given a sizing, a QoS target and a set of scenarios, the SNC approach computes a loss bound on each scenario. We consider a sizing to be valid if the percentage of scenarios that meet the LOLP or EUE target meets the confidence target. The algorithm used to efficiently search the (B, C) space for robust sizings is the same as the one used to compute sizing curves as described in Section V-B.

The formulation is complex, so for each metric we follow the full formulation with a summary of the algorithm for computing the loss bound on each scenario.

1) SNC for LOLP: In this approach, we characterize the net power arrival to the battery using lower and upper bounds computed on the ensemble of input traces. Then, we use stochastic network calculus to compute the LOLP for each choice of storage and solar panel size (B, C). The output is a statistical characterization of the LOLP ϵ as a function of the selected choices of (B,C). We then use the greedy heuristic discussed in Section V-B to compute the least-cost sizing that meets the LOLP criterion. Since SNC sizing is known to be robust to small perturbations in the input traces, we view this least-cost sizing as being robust to the input traces.

Denote by $P_c(t)$ and $P_d(t)$, respectively, the charging and discharging power from and to the battery, given by an operating policy corresponding to:

$$P_c(t) = \min\left([S(t)C - D(t)]_+, \alpha_c B\right) \tag{22}$$
$$P_d(t) = \min\left([D(t) - S(t)C]_+, \alpha_d B\right) \tag{23}$$

That is, we assume that the operating policy is as follows: the battery is charged whenever the generation $S(t)$ exceeds the load, and discharged otherwise, with a bound $B\alpha_c$ on the charge power and a bound $B\alpha_d$ on the discharge power (matching Eq. 7-8). Different operating strategies will require these equations to be modified appropriately.

Define the net power inflow to the battery at any time as the overall net equivalent power injected to the battery, which is

$$P_{net}(t) = \eta_c P_c(t) - \eta_d P_d(t) \tag{24}$$

Note that at any time instant t, $P_c(t) \cdot P_d(t) = 0$, and $P_{net}(t)$ can be expressed as

$$P_{net}(t) = \begin{cases} \eta_c P_c(t) & \text{if } S(t)C \geq D(t) \\ -\eta_d P_d(t) & \text{if } S(t)C < D(t) \end{cases} \tag{25}$$

Please also note that while $P_c(t), P_d(t) \geq 0$ at any time t, $P_{net}(t)$ can be both positive and negative, and represents the rate at which the buffer changes over time.

According to the battery model in Reference [25], the instantaneous available battery capacity is a function of charge/discharge power to/from the battery. The larger the charge/discharge power the lower the instantaneous available battery capacity. This means that apart from the power constraints discussed above, we also have energy constraints in battery operations. To be more precise, the battery state of charge $E(t)$ at any time t must satisfy $B_1(t) \leq E(t) \leq B_2(t)$, where

$$B_1(t) = u_1 P_d(t) + v_1 B \tag{26}$$
$$B_2(t) = u_2 P_c(t) + v_2 B \tag{27}$$

With the above notation, the state of charge of a battery $E(t)$ at any time t can be, recursively, expressed by

$$E(t) = [E(t-1) + P_{net}(t)T_u]_{B_1(t)}^{B_2(t)} \tag{28}$$

where $[\cdot]_{B_1(t)}^{B_2(t)}$ truncates the inner expression to a lower bound $B_1(t)$ and an upper bound $B_2(t)$, or equivalently for any y

$$[y]_{B_1(t)}^{B_2(t)} = \begin{cases} B_1(t) & \text{if } y < B_1(t) \\ B_2(t) & \text{if } y > B_2(t) \\ y & \text{otherwise} \end{cases} \tag{29}$$

Recall that LOLP is the probability that at time t the energy to be withdrawn from the battery reaches the lower battery capacity boundary (i.e., $B_1(t)$) and hence the demand cannot be met at that time. Mathematically speaking, this means that

$$LOLP = \mathbb{P}\{E(t-1) + P_{net}(t)T_u < B_1(t)\} \ \forall t \tag{30}$$

where $E(t-1)$ can be computed recursively according to Eq. 28. Please note that Eq. 30 is equivalent to Eq. 11 by setting $LOLP \leq \epsilon$, both of them enforcing LOLP constraint in the steady-state regime. In the steady-state, all time instants t within T have probabilistically identical weights in violating LOLP restriction. This is attained by averaging over all time instants within T with equal weights in Eq. 11 and by formulating the terms in the probability expression in Eq. 30 in a way to represent and model any time instant in T.

The recursive equation in Eq. 30 can be turned into a complicated min-max non-recursive equation. At any time t, the min-operand searches, in the range of $[0, t-1]$, for the last reset time before t, which is the last occurrence of a loss of load event. As shown in Reference [15], instead of applying the min-operand to t scenarios in $[0, t-1]$, we can highly accurately approximate LOLP by only accounting for only two scenarios: (I) the reset time occurs at the last time slot $t-1$ and (II) there has been no reset time since $t = 0$.

Hence, define $LOLP^{\text{I}}$ and $LOLP^{\text{II}}$ representing LOLP, respectively under the two scenarios mentioned above and LOLP can be approximated by

$$LOLP \approx \min\left(LOLP^{\text{I}}, LOLP^{\text{II}}\right) \tag{31}$$

Under scenario I, the last reset time always happens at the previous time slot. The LOLP under this scenario can be closely approximated by a battery-less scenario. This means that $LOLP^{\mathrm{I}}$ can be approximated by the likelihood that the instantaneous demand is larger than the instantaneous supply, or mathematically speaking:

$$LOLP^{\mathrm{I}} \approx \mathbb{P}\{D(t) > S(t)C\} \tag{32}$$

Under scenario II, there is no reset time until time t. This means that the battery state of charge has never reached its lower boundary. Assuming that the battery is initially full ($E(0) = \nu_2 B$), $LOLP^{\mathrm{II}}$ is given by:

$$\begin{aligned} &LOLP^{\mathrm{II}} \\ &= \mathbb{P}\left\{E(t) < B_1(t)\right\} \ \forall t && (33)\\ &= \mathbb{P}\left\{v_2 B - \sup_{0\le s\le t}(-P_{net}(s,t))T_u < B_1(t)\right\} \ \forall t && (34)\\ &= \mathbb{P}\left\{\sup_{0\le s\le t}\left(\frac{u_1 P_d(t)}{T_u} - P_{net}(s,t)\right) > \frac{v_2 - v_1}{T_u}B\right\} \ \forall t && (35)\end{aligned}$$

where $P_{net}(s,t)$ is defined as

$$P_{net}(s,t) := \sum_{k=s+1}^{t} P_{net}(k) \tag{36}$$

We model the tail-bound in Eq. 35 with an exponential distribution. This means that we compute $p_l^{\mathrm{II}}, \lambda_l^{\mathrm{II}} \ge 0$ such that for any $\delta \ge 0$, the following condition holds:

$$\mathbb{P}\left(\underbrace{\sup_{0\le s\le t}\left(\frac{u_1 P_d(t)}{T_u} - P_{net}(s,t)\right)}_{:=Y^t} > \delta\right) \approx p_l^{\mathrm{II}} e^{-\lambda_l^{\mathrm{II}}\delta} \tag{37}$$

where Y^t is defined to simplify notation for the rest of derivations. Combining Eq. 35 with Eq. 37, we have

$$LOLP^{\mathrm{II}} = \mathbb{P}\left\{Y^t > \frac{v_2 - v_1}{T_u}B\right\} \approx p_l^{\mathrm{II}} e^{-\lambda_l^{\mathrm{II}}(\frac{v_2-v_1}{T_u}B)} \tag{38}$$

and finally LOLP can be computed, by inserting Eq. 32 and Eq. 38 into Eq. 31. There are three unknowns in Eq. 31 to be evaluated: $LOLP^{\mathrm{I}}$, p_l^{II}, and λ_l^{II} that can be computed as discussed next.

We can translate this mathematical presentation into an algorithm: given n scenarios, we compute n different sample paths of the stochastic processes P_d^i and P_{net}^i for a time horizon of length T. We can compute LOLP, using stochastic network calculus, following these steps in turn:

Step 1: Compute $LOLP^{\mathrm{I}}$ for $i = [1, n]$: This is a point-wise probability, expressed in Eq. 32 and can be computed as

$$LOLP^{i,\mathrm{I}} = \frac{\sum_{t=0}^{T}\mathbb{I}\left(D^i(t) > S^i(t)C\right)}{T} \tag{39}$$

where $\mathbb{I}(x)$ is the indicator function, which is 1 if x is true and 0, otherwise.

Step 2: Construct $Y^{i,t}$: To compute p_l^{II} and λ_l^{II}, we first construct the set of all $Y^{i,t}$ for any ensemble trace $i \in [1, n]$ and any time $t \le T$, defined as:

$$Y^{i,t} = \sup_{0\le s\le t}\left(\frac{u_1 P_d^i(t)}{T_u} - P_{net}^i(s,t)\right) \tag{40}$$

It can be shown that $Y^{i,t}$ can be expressed, recursively, by

$$\begin{aligned} Y^{i,1} &= \frac{u_1}{T_u}P_d^i(1) - P_{net}^i(1) && (41)\\ Y^{i,t} &= \frac{u_1}{T_u}P_d^i(t) - P_{net}^i(t) \\ &\quad + \max\left(Y^{i,t-1} - \frac{u_1}{T_u}P_d^i(t-1),\ 0\right) && (42)\end{aligned}$$

Step 3: Compute $p_l^{i,\mathrm{II}}$ and $\lambda_l^{i,\mathrm{II}}$ for $i = [1, n]$: Using $Y^{i,t}$ from the previous step, $p_l^{i,\mathrm{II}}$ is the likelihood of $Y^{i,t}$, being positive, or

$$p_l^{i,\mathrm{II}} = \frac{\sum_{t=1}^{T}\mathbb{I}\left(Y^{i,t} > 0\right)}{T} \tag{43}$$

and $\lambda_l^{i,\mathrm{II}}$ can be obtained as the exponent of fitting an exponential distribution to the following set

$$\lambda_l^{i,\mathrm{II}} \sim \text{Exponential}\left(\left\{Y^{i,t} \mid Y^{i,t} > 0\right\}\right) \tag{44}$$

Step 4: Compute LOLP: Compute $LOLP^{i,\mathrm{II}}$, according to Eq. 38. Then $LOLP = \min(LOLP^{i,\mathrm{I}}, LOLP^{i,\mathrm{II}})$. The sizing is valid if the following condition holds:

$$\frac{\sum_{i=1}^{n}\mathbb{I}(LOLP^i \le \epsilon)}{n} \ge \gamma \tag{45}$$

where γ is the confidence parameter.

2) SNC for EUE: In this section, we first formulate the value of unserved energy (UE) at any time instant t. The UE at any time t is given by

$$UE(t) = [E(t-1) + P_{net}(t)T_u - B_1(t)]_+ \tag{46}$$

Moreover, define the probability of unserved energy (PUE) as the complement cumulative distribution function (CCDF) of the unmet load, which is

$$PUE(y) = \mathbb{P}[UE(t) > y] \tag{47}$$

The expected value of the unserved energy (EUE) can be expressed as a function of the unmet load probability (PUE), as follows

$$EUE = \int_0^\infty PUE(y)dy \tag{48}$$

We use a similar strategy as used for LOLP formulation, to compute PUE. To be more precise, we consider two scenarios: (I) the reset time being the last time slot $t-1$ and (II) there has been no reset time since the beginning. Let us define PUE^{I} and PUE^{II}, representing the unmet load probability, respectively under Scenarios I and II. We approximate PUE with the minimum of what we observe in scenarios I and II; i.e.,

$$PUE(y) \approx \min\left(PUE^{\mathrm{I}}(y), PUE^{\mathrm{II}}(y)\right) \tag{49}$$

Under scenario I

$$PUE^{\mathrm{I}}(y) = \mathbb{P}\left\{D(t) - S(t)C - y > 0\right\} \approx p_u^{\mathrm{I}} e^{-\lambda_u^{\mathrm{I}} y} \quad (50)$$

where we assume that the tail bound of $D(t) - S(t)$ can be well approximated by an exponential distribution to obtain the right-hand-side in Eq. 50.

Under scenario II

$$\begin{aligned} &PUE^{\mathrm{II}}(y) \\ &= \mathbb{P}\left\{E(t-1) + P_{net}(t)T_u - B_1(t) - y > 0\right\} && (51)\\ &= \mathbb{P}\left\{\sup_{0\le s\le t}\left(\frac{u_1 P_d(t)}{T_u} - P_{net}(s,t)\right) > \frac{(\nu_2-\nu_1)B + y}{T_u}\right\} && (52)\\ &\approx p_l^{\mathrm{II}} e^{-\lambda_l^{\mathrm{II}}\left(\frac{(\nu_2-\nu_1)B+y}{T_u}\right)} && (53)\\ &= p_u^{\mathrm{II}} e^{-\lambda_u^{\mathrm{II}} y} && (54)\end{aligned}$$

where

$$p_u^{\mathrm{II}} = p_l^{\mathrm{II}} e^{\frac{-\lambda_l^{\mathrm{II}}(\nu_2-\nu_1)B}{T_u}}; \qquad \lambda_u^{\mathrm{II}} = \frac{\lambda_l^{\mathrm{II}}}{T_u} \quad (55)$$

Inserting Eq. 50 and Eq. 54 in Eq. 49, we have

$$PUE(y) \approx \left(p_u^{\mathrm{I}} e^{-\lambda_u^{\mathrm{I}} y}, p_u^{\mathrm{II}} e^{-\lambda_u^{\mathrm{II}} y}\right) \quad (56)$$

Inserting Eq. 56 to Eq. 48, yields

$$EUE = \begin{cases} \frac{p_u^{\mathrm{II}}}{\lambda_u^{\mathrm{II}}} - \frac{p_u^{\mathrm{II}}}{\lambda_u^{\mathrm{II}}}\left(\frac{p_u^{\mathrm{II}}}{p_u^{\mathrm{I}}}\right)^{\frac{\lambda_u^{\mathrm{II}}}{\lambda_u^{\mathrm{I}}-\lambda_u^{\mathrm{II}}}} + \frac{p_u^{\mathrm{I}}}{\lambda_u^{\mathrm{I}}}\left(\frac{p_u^{\mathrm{II}}}{p_u^{\mathrm{I}}}\right)^{\frac{\lambda_u^{\mathrm{I}}}{\lambda_u^{\mathrm{I}}-\lambda_u^{\mathrm{II}}}} & \text{if } p_u^{\mathrm{I,II}}, \lambda_u^{\mathrm{I,II}} > 0 \\ \frac{p_u^{\mathrm{I}}}{\lambda_u^{\mathrm{I}}} - \frac{p_u^{\mathrm{I}}}{\lambda_u^{\mathrm{I}}}\left(\frac{p_u^{\mathrm{I}}}{p_u^{\mathrm{II}}}\right)^{\frac{\lambda_u^{\mathrm{I}}}{\lambda_u^{\mathrm{II}}-\lambda_u^{\mathrm{I}}}} + \frac{p_u^{\mathrm{II}}}{\lambda_u^{\mathrm{II}}}\left(\frac{p_u^{\mathrm{I}}}{p_u^{\mathrm{II}}}\right)^{\frac{\lambda_u^{\mathrm{II}}}{\lambda_u^{\mathrm{II}}-\lambda_3}} & \text{if } p_u^{\mathrm{I,II}}, \lambda_u^{\mathrm{I,II}} < 0 \\ \frac{p_u^{\mathrm{II}}}{\lambda_u^{\mathrm{II}}} & \text{if } p_u^{\mathrm{I,II}} > 0, \lambda_u^{\mathrm{I,II}} < 0 \\ \frac{p_u^{\mathrm{I}}}{\lambda_u^{\mathrm{I}}} & \text{if } p_u^{\mathrm{I,II}} < 0, \lambda_u^{\mathrm{I,II}} > 0 \end{cases} \quad (57)$$

where $p_u^{\mathrm{I,II}} = p_u^{\mathrm{I}} - p_u^{\mathrm{II}}$ and $\lambda_u^{\mathrm{I,II}} = \lambda_u^{\mathrm{I}} - \lambda_u^{\mathrm{II}}$. Finally, for the given tolerable unmet demand ratio θ, we must ensure that

$$\frac{EUE}{\mathbb{E}[D(t)]} \le \theta \qquad \forall t \quad (58)$$

We translate this mathematical presentation into an algorithm: given n scenarios, we compute n different sample paths of the stochastic processes P_d^i and P_{net}^i for a time horizon of length T. We can compute the EUE bound using stochastic network calculus by following these steps in turn:

Step 1: Compute $p_u^{\mathbf{I}}$ and $\lambda_u^{\mathbf{I}}$: We first construct the set of all $Z^{i,t}$ for all sample paths i and all time $t \le T$ as

$$Z^{i,t} = D^i(t) - S^i(t)C \quad (59)$$

Then, $p_u^{i,\mathrm{I}}$ is given by

$$p_u^{i,\mathrm{I}} = \frac{\sum_{t=0}^{T} \mathbb{I}\left(Z^{i,t} > 0\right)}{T} \quad (60)$$

and $\lambda_u^{i,\mathrm{I}}$ can be obtained as the exponent of fitting an exponential distribution to the following set

$$\lambda_u^{i,\mathrm{I}} \sim \text{Exponential}\left(\left\{Z^{i,t} \mid Z^{i,t} > 0\right\}\right) \quad (61)$$

Table I: Battery model parameters

Parameter	α_c	α_d	u_1	u_2	v_1	v_2	η_c	η_d
Value	1	1	0.053	-0.125	0	1	0.99	1.11*

*includes inverter inefficiencies of ∼10%

Step 2: Compute $p_u^{i,\mathbf{II}}$ and $\lambda_u^{i,\mathbf{II}}$: To do so, we should first compute $p^{i,\mathrm{II}}$ and $\lambda^{i,\mathrm{II}}$ according to Step 3 in LOLP computation algorithm. Inserting these values in Eq. 55 yields the corresponding $p_u^{i,\mathrm{II}}$ and $\lambda_u^{i,\mathrm{II}}$.

Step 3: Compute the EUE restriction: By inserting the values obtained in previous steps in Eq. 57 and Eq. 58. The sizing is valid if the percentage of scenarios whose loss bound is under θ satisfies the confidence measure γ.

3) Computational complexity: For both LOLP and EUE formulations, each test requires $O(T)$ time to construct the set Y and calculate the parameters of $LOLP^I$ and $LOLP^{II}$ for each of the n scenarios. This needs to be repeated over every tested combination of (B, C), of which there are at most $(b + c)$ by using the search algorithm in Section V-B. Hence the total complexity is $O(nT(b + c))$.

VI. Numerical Evaluation

For concreteness, we numerically evaluate our approaches using four years of PV generation and load data collected from a number of homes in the Pecan Street Dataport [31]. We present a detailed view of results for three homes in the dataset, representing homes with low, mid, and high levels of consumption. We also present the results of an aggregated sizing evaluation across 52 homes from this dataset.

To evaluate the cost of a particular sizing, we set π_C, the installed cost of solar panels, to be USD 2.50/W and π_B, the cost of storage, to be USD 460/kWh[7], with battery parameters corresponding to a Lithium-Nickel-Manganese-Cobalt battery chemistry [32], [33] as summarized in Table I. The battery model is Model 1* in Reference [25] and we use the simple operating policy of charging the battery when solar generation exceeds the load, and discharging the battery when load exceeds solar generation. The battery starts with a full state of charge.

Although our approach to optimization-based sizing is robust, in that it is insensitive to small perturbations in the input trace, it is often not possible to use optimization-based sizing in real life, because it relies on optimal operation of the storage system, using a policy that cannot be determined in advance. Thus, we only evaluate our Chebyshev-curve sizing approach using sizings curves computed via simulation, and compare this sizing to the SNC approaches.

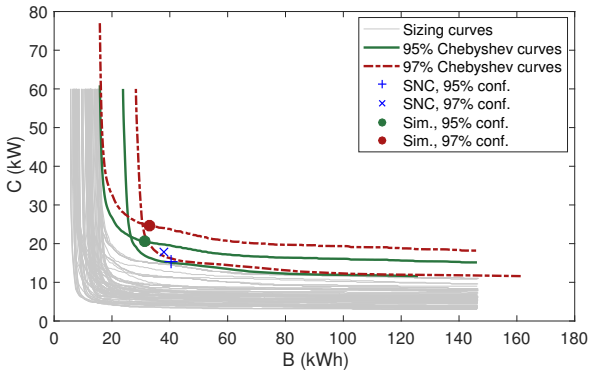
A. Convergence

In accordance with the simulation and optimization approaches, the recommended system size is based on a statistical measure of the underlying samples of computed sizing curves. With SNC, we compute a probabilistic upper-bound on the number of loss-of-load events for each scenario in the ensemble. With all three approaches, the recommended size is progressively refined as more scenarios are evaluated.

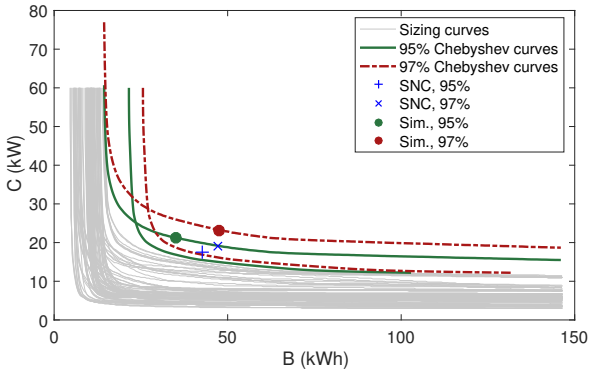
[7]Source: https://www.tesla.com/powerwall

We find that in all three approaches, B and C values converge after about 100 randomly sampled scenarios for ϵ or θ targets of 0.05. For smaller targets, more scenarios are required for convergence.

B. Sizing



(a)



(b)

Figure 3: Comparing sizings at different confidences with SNC and simulation approaches, for LOLP with $\epsilon = 0.05$ (3a) and EUE with $\theta = 0.05$ (3b).

Figure 3 shows sizings from SNC and simulation approaches for $\epsilon = 0.05$ and $\theta = 0.05$, over 100-day periods with a confidence of 95% and 97%. The figures include the sizing curves and the Chebyshev curves computed from them. The sizing obtained for all approaches with 95% confidence lie above the sizing curves. Using 97% confidence gives a more conservative sizing, which is more robust to variations in scenarios that might be observed in the future and are not fully captured in the historical data.

We note that sizing obtained using the optimization approach (not included in Figure 3) is identical to the sizing obtained using simulations for an EUE target, but is always much smaller than with the other two approaches for an LOLP target. This is not surprising, given that the optimization approach chooses the optimal operating strategy, rather than the basic strategy used by the other approaches; the basic strategy happens to be optimal for minimizing EUE.

Table II: Computation time (Linux user time) mean and standard error

Method	**Mean CPU time per 100 scenarios (h:m:s)**			
	LOLP		EUE	
	μ	std. error	μ	std. error
Simulation	0:0:38	< 0:0:01	0:0:38	< 0:0:01
Optimization	46453:20:05	896:6:55	277:45:21	1:01:22
SNC	0:0:24	< 0:0:01	0:0:16	< 0:01

C. Robustness

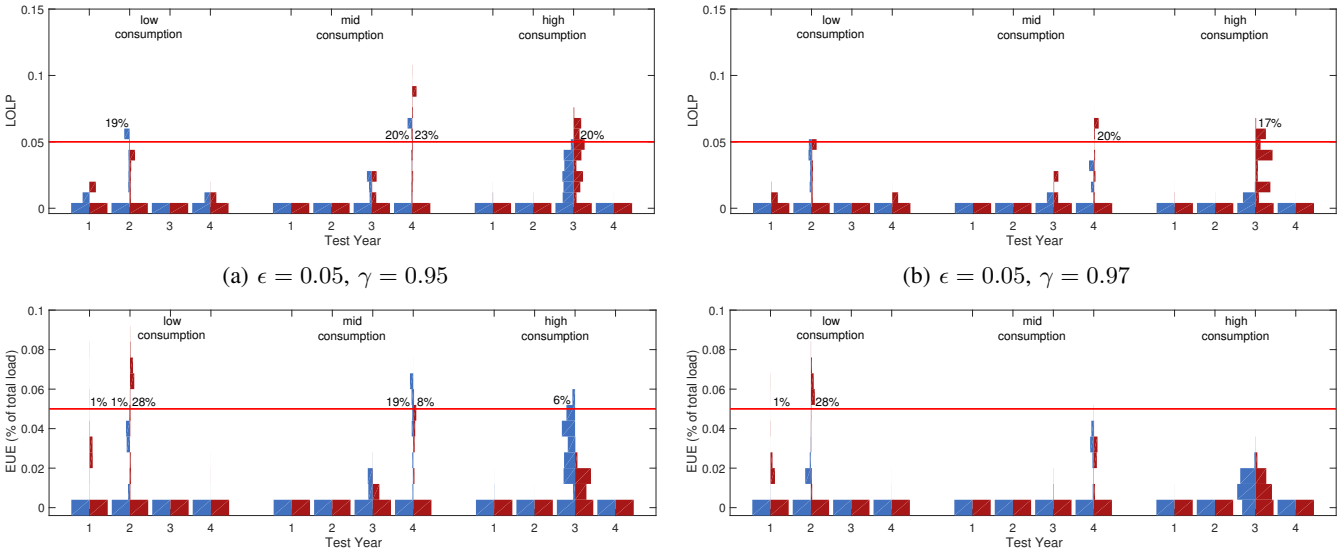
We compare robustness of the sizing that results from the different sizing approaches in Figure 4, which summarizes the results of leave-one-year-out analysis for a sample low, mid, and high electricity consumption home in the dataset. For each of the 4 years in the dataset, we compute a sizing by randomly sampling 100-day scenarios from the other 3 years. We then use the simulation and SNC approaches to compute a robust system size using either 95% or 97% confidence bounds for $\epsilon = 0.05$ (Figures 4a,4b) or $\theta = 0.05$ (Figures 4c,4d). Each size is then tested on scenarios from the test year, and the distribution of resulting LOLP and EUE values for each scenario is presented as a histogram.

Note also that each of the four subsets of three years of data can result in a substantially different sizing. If a year with particularly high load is left out, such as the 3rd year in the high-consumption household, the sizing results in a violation of the QoS, indicating that solar generation and load may highly variable across years. This variability can be accounted for by sizing conservatively, which can be achieved by using a higher confidence bound. Specifically, note that with γ = 0.95, there are several instances where both simulation- and SNC-based sizings fail to meet the performance bound. This is because of atypical behaviour in one of the year compared to the other years. When γ is increased to 0.97, the number of violations decreases for both approaches. Compared to the SNC approach, the simulation approach is more sensitive to increases in γ, since the Chebyshev curves give very loose bounds at high confidence, while the SNC approach uses an empirical confidence measure. In practice, we expect γ to be a user-supplied parameter that reflects their level of optimism.

The aggregated results for 52 Austin, Texas houses with data through years 2014-2017 are shown in Figure 5. For each house, we compute a leave-one-year-out sizing for an LOLP or EUE target of 0.05 over a period of 100 days with 95% confidence. We then test this size on 200 randomly selected 100-day periods from the test year, for a total of $52 \times 4 \times 200 = 41600$ tests for each sizing approach. The results are presented in histogram form. Notably, the fraction of values that are within the 5% LOLP and EUE target are well within $\gamma = 0.95$.

D. Computation time

Recall that the asymptotic complexity of the optimization approach is lower bounded by $O(nbT^{5.5}logT)$. The computational complexity of simulation and SNC is $O(nT(b+c))$. Thus, for large values of T, which is typical, the best approaches are simulation and SNC (with SNC up to a factor of 2.5 faster than simulation).



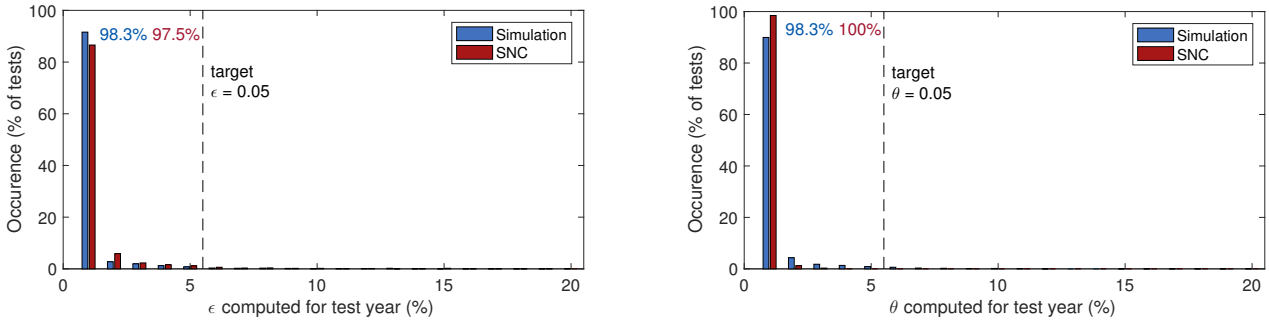
(a) $\epsilon = 0.05$, $\gamma = 0.95$

(b) $\epsilon = 0.05$, $\gamma = 0.97$

(c) $\theta = 0.05$, $\gamma = 0.95$

(d) $\theta = 0.05$, $\gamma = 0.97$

Figure 4: The vertical histograms show the leave-one-year-out sizing performance on scenarios sampled from the test year for three households, four years for each. The resulting LOLP or EUE of the system with size computed using the simulation-based approach is shown in the blue histograms extending to the left, while red histograms extending to the right show comparable results with the SNC approach. Figures 4a and 4b are for an LOLP target of 0.05, with 95% and 97% confidence respectively, while Figures 4c and 4d are for an EUE target of 0.05, with 95% and 97% confidence respectively. For test years where the resulting LOLP/EUE sometimes exceeds the target, there is an annotation showing percentage of scenarios that exceed the target.



(a) LOLP, 95% confidence target

(b) EUE, 95% confidence target

Figure 5: Aggregated leave-one-year-out test results on 52 houses. Percentage of tests that land within the LOLP (Figure 5a) or EUE (Figure 5b) target for simulation (left) and SNC (right) sizing approaches

Table II shows the CPU time required to process 100 scenarios, each of which has 365 days of data, on a 2.7GHz Intel Xeon CPU. It is clear that simulations and SNC take the least time, with optimization taking five to seven orders of magnitude more[8].

VII. Discussion

A. Comparison of the three approaches

Our work evaluates three distinct approaches to robust and practical sizing of solar and storage systems. Over and above the numerical comparison in Section VI, we now make some qualitative observations about their relative merits.

Unlike some prior work [34], [35], [9] which solve the joint problem of optimal sizing and optimal operation, in this work, we study only sizing. However, with optimization, the operation rules are a free variable, in that the output of the optimization program is also the optimal charge/discharge schedule. Note that these operation rules cannot be used in practice, because the rules depend in detail on the traces, and the details of the future are unknown. If we could encode operation rules into the optimization program, we would be able to come up with a sizing that did not have this coupling. Unfortunately, it is non-trivial, perhaps impossible, to encode arbitrary operation rules in an optimization program. For instance, consider the operation rule "Charge the store from the grid if the battery charge is below 30% and the grid price

[8]We used CPLEX 12.6.3, which has highly parallelized LP and ILP solvers

is lower than $0.12/kWh." This rule defines a dependency between the charging power and the battery charge, which complicates the formulation of the program and also makes it non-linear and hence difficult to solve efficiently.

There is a similar problem with stochastic network calculus, where encoding complex charge/discharge operations rules into Equations (22) and (23) may result in greatly complicating the subsequent analysis. In contrast, the simulation approach can be used with any operating strategy. Moreover, it has acceptable compute speed (though slower than stochastic network calculus). Thus, from a qualitative perspective, the simulation approach is perhaps the best one, especially when combined with a *post hoc* Chebyshev bound.

B. Contributions of our work

Our work makes multiple contributions. To begin with, it is the first work, to our knowledge, that provides robust and practical advice on sizing by comparing multiple approaches.

Second, our use of a univariate Chebyshev curve in combination with optimal sizing for multiple scenarios is innovative, and can be generalized to other robust optimization problems.

Third, the LOLP and EUE formulations using SNC in Section V-C considerably advance the state-of-the-art in SNC analysis of battery-equipped systems, such as in References [21], [18]. This is because the battery model used in this work is more realistic (and more complicated). Additionally, we take a different approach in characterizing energy profiles: We model net energy demand directly, instead of modelling supply and demand separately. We further characterize the tail bounds of the net load directly instead of defining envelopes and characterizing the residual processes with respect to the envelopes as done in prior work. This has substantial implications on the time complexity and also accuracy of the model. Finally, the derivations of EUE here are new and advancing the state-of-the-art in SNC and also its application in energy systems.

C. Limitations and future work

Studying the impact of energy consumption and generation patterns on the system size, especially from homes in different geographical regions, is an interesting avenue for future work. To facilitate this line of research, we have made the code for computing robust system sizes with simulation and SNC methods publicly available [7].

Our work suffers from some limitations, as discussed next.

First, we have assumed that the load is not under our control. In some cases, it is possible to ask the energy consumers to modify their behaviour, using a control signal. Thus, for example, a home owner may be asked to defer a heavy load if the state of charge of the storage was particularly low. In this situation, it is obvious that the system sizing can be much smaller. However, sizing a system in the presence of load control is a much more complex problem, in that it requires jointly optimizing the storage operation as well as the load control actions. We intend to explore this in future work.

Second, the computation times presented in this paper are only indicative. For example, both simulations and the stochastic network calculus algorithm can be tuned, or re-coded in a more efficient low-level language to improve computation times. Similarly, it is well known that choice of optimization meta-parameters can also significantly impact the computation time. Nevertheless, given the substantial differences in performance, we believe our results are representative of expected outcomes in practical scenarios.

Finally, our approaches are scenario-based, and we treat the length of each scenario as being specified by the user in their QoS target. There are challenges in this setup that we have not thoroughly explored. First, we have made an implicit assumption that each scenario starts with a particular (full) battery state of charge. When the scenarios are short and the QoS targets are tight, the initial charge state may have a significant impact on the frequency of loss events in the system; to sidestep this issue, we experimentally determined that the time duration in the QoS target must be at least 60 days for the QoS we presented in this paper. In future work, we intend to explore how to satisfy QoS targets with shorter time periods. Second, sampling long scenarios from a short dataset can result in a sample distribution that does not accurately reflect the solar and load processes due to overfitting.

VIII. Conclusion

We evaluate and compare three state-of-the-art approaches to size solar generation and storage in a realistic setting. Unlike prior work, which evaluates a single approach and does not evaluate the robustness of the resulting solution, we compare the sizing results from these approaches on identical inputs, permitting a fair comparison. We find that, due to both qualitative and quantitative reasons, simulation appears to be the best tool for sizing in a realistic setting. In carrying out our work, we have made contributions to the state of the art both in the area of stochastic network calculus and in the use of sample Chebyshev bounds to obtain a novel technique for robust optimization and simulation.

References

[1] H. Yang, W. Zhou, L. Lu, and Z. Fang, "Optimal sizing method for stand-alone hybrid solar–wind system with lpsp technology by using genetic algorithm," *Solar energy*, vol. 82, no. 4, pp. 354–367, 2008.

[2] E. D. Castronuovo and J. A. P. Lopes, "Optimal operation and hydro storage sizing of a wind–hydro power plant," *International Journal of Electrical Power & Energy Systems*, vol. 26, no. 10, pp. 771–778, 2004.

[3] M. Korpaas, A. T. Holen, and R. Hildrum, "Operation and sizing of energy storage for wind power plants in a market system," *International Journal of Electrical Power & Energy Systems*, vol. 25, no. 8, pp. 599–606, 2003.

[4] T. K. Brekken, A. Yokochi, A. Von Jouanne, Z. Z. Yen, H. M. Hapke, and D. A. Halamay, "Optimal energy storage sizing and control for wind power applications," *IEEE Transactions on Sustainable Energy*, vol. 2, no. 1, pp. 69–77, 2011.

[5] J. Mitra, "Reliability-based sizing of backup storage," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 1198–1199, 2010.

[6] F. Kazhamiaka, C. Rosenberg, and S. Keshav, "Practical strategies for storage operation in energy systems: design and evaluation," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 4, pp. 1602–1610, 2016.

[7] F. Kazhamiaka, "Robust sizing," https://github.com/iss4e/Robust_Sizing, 2019.

[8] S. Chen, H. B. Gooi, and M. Wang, "Sizing of energy storage for microgrids," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 142–151, 2012.

[9] Y. Ghiassi-Farrokhfal, F. Kazhamiaka, C. Rosenberg, and S. Keshav, "Optimal design of solar pv farms with storage," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1586–1593, 2015.

[10] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM review*, vol. 53, no. 3, pp. 464–501, 2011.

[11] A. Aichhorn, M. Greenleaf, H. Li, and J. Zheng, "A cost effective battery sizing strategy based on a detailed battery lifetime model and an economic energy management strategy," in *Power and Energy Society General Meeting, 2012 IEEE*. IEEE, 2012, pp. 1–8.

[12] J. Dong, F. Gao, X. Guan, Q. Zhai, and J. Wu, "Storage sizing with peak-shaving policy for wind farm based on cyclic markov chain model," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 3, pp. 978–989, 2017.

[13] O. Ardakanian, C. Rosenberg, and S. Keshav, "On the impact of storage in residential power distribution systems," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 3, pp. 43–47, Jan. 2012.

[14] Y. Jiang and Y. Liu, *Stochastic network calculus*. Springer, 2008, vol. 1.

[15] S. Singla, Y. Ghiassi-Farrokhfal, and S. Keshav, "Using storage to minimize carbon footprint of diesel generators for unreliable grids," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1270–1277, Oct 2014.

[16] P. Kumar, "A network calculus foundation for smart-grids where demand and supply vary in space and time," in *Proceedings of the Seventh International Conference on Future Energy Systems*, ser. e-Energy '16. New York, NY, USA: ACM, 2016, pp. 1–11.

[17] R. Basmadjian, Y. Ghiassi-Farrokhfal, and A. A. Keshav, "Hidden storage in data centers: Gaining flexibility through cooling systems," in *19th International GI/ITG Conference on "Measurement, Modelling and Evaluation of Computing Systems*, 2018.

[18] Y. Ghiassi-Farrokhfal, S. Keshav, and C. Rosenberg, "Toward a realistic performance analysis of storage systems in smart grids," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 402–410, 2015.

[19] K. Wang, F. Ciucu, C. Lin, and S. H. Low, "A stochastic power network calculus for integrating renewable energy sources into the power grid," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1037–1048, July 2012.

[20] M. Raeis, A. Burchard, and J. Liebeherr, "Analysis of the leakage queue: A queuing model for energy storage systems with self-discharge," *CoRR*, vol. abs/1710.09506, 2017. [Online]. Available: http://arxiv.org/abs/1710.09506

[21] Y. Ghiassi-Farrokhfal, S. Keshav, C. Rosenberg, and F. Ciucu, "Solar power shaping: An analytical approach," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 162–170, 2015.

[22] P. Gilman, N. Blair, M. Mehos, C. Christensen, S. Janzou, and C. Cameron, "Solar advisor model user guide for version 2.0," *National Renewable Energy Laboratory, Golden, CO, Technical Report No. NREL/TP-670-43704*, 2008.

[23] P. Attaviriyanupap, K. Tokuhara, N. Itaya, M. Marmiroli, Y. Tsukamoto, and Y. Kojima, "Estimation of photovoltaic power generation output based on solar irradiation and frequency classification," in *2011 IEEE PES Innovative Smart Grid Technologies*. IEEE, 2011, pp. 1–7.

[24] A. Bhattacharya, J. Kharoufeh, and B. Zeng, "Managing energy storage in microgrids: A multistage stochastic programming approach," *IEEE Transactions on Smart Grid*, 2016.

[25] F. Kazhamiaka, C. Rosenberg, S. Keshav, and K.-H. Pettinger, "Li-ion storage models for energy system optimization: the accuracy-tractability tradeoff," in *Proceedings of the Seventh International Conference on Future Energy Systems*. ACM, 2016, p. 17.

[26] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations research*, vol. 43, no. 2, pp. 264–281, 1995.

[27] A. Von Meier, *Electric power systems: a conceptual introduction*. John Wiley & Sons, 2006.

[28] J. G. Saw, M. C. Yang, and T. C. Mo, "Chebyshev inequality with estimated mean and variance," *The American Statistician*, vol. 38, no. 2, pp. 130–132, 1984.

[29] B. Stellato, B. P. Van Parys, and P. J. Goulart, "Multivariate chebyshev inequality with estimated mean and variance," *The American Statistician*, vol. 71, no. 2, pp. 123–127, 2017.

[30] P. M. Vaidya, "Speeding-up linear programming using fast matrix multiplication," in *Foundations of Computer Science, 1989., 30th Annual Symposium on*. IEEE, 1989, pp. 332–337.

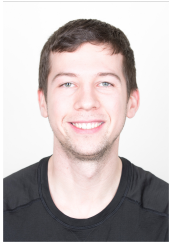
[31] "Pecan Street Inc. Dataport," 2018.

[32] *Li-ion Polymer Cell*, Kokam, 2016, li-NMC cell specifications.

[33] *LIR18650 cell*, EEMB, 2016, li-NMC cell specifications.

[34] S. Fazlollahi, P. Mandel, G. Becker, and F. Maréchal, "Methods for multi-objective investment and operating optimization of complex energy systems," *Energy*, vol. 45, no. 1, pp. 12–22, 2012.

[35] N. Kaushika, N. K. Gautam, and K. Kaushik, "Simulation model for sizing of stand-alone solar pv system with interconnected array," *Solar Energy Materials and Solar Cells*, vol. 85, no. 4, pp. 499–519, 2005.



Fiodar Kazhamiaka is a Ph.D. candidate in computer science at the David Cheriton School of Computer Science, University of Waterloo. His research interests are in the design and control of systems with energy storage and renewable energy sources.

Yashar Ghiassi-Farrokhfal obtained his PhD degree in Electrical and Computer Engineering at University of Toronto. He is currently an Assistant Professor at the Department of Technology and Operations Management at Rotterdam School of Management, Erasmus University. He is also the academic director of smart cities at Erasmus Center for Data Analytics.

Srinivasan Keshav is a Professor of Computer Science at the University of Waterloo. He received a B.Tech in Computer Science and Engineering from IIT Delhi in 1986 and Ph.D. in Computer Science from the University of California, Berkeley in 1991. He was subsequently an MTS at AT&T Bell Laboratories and an Associate Professor at Cornell. In 1999 he left academia to co-found Ensim Corporation and GreenBorder Technologies Inc. He has been at the University of Waterloo since 2003.

Catherine Rosenberg , FIEEE, is a Professor with the Department of Electrical and Computer Engineering at the University of Waterloo and the Canada Research Chair in the Future Internet. Since April 2018, she is also the Cisco Research Chair in 5G Systems. She was elected a Fellow of the Canadian Academy of Engineering in 2013. Her research interests are in networking, wireless, and energy systems. More information is available at https://ece.uwaterloo.ca/~cath/.