Collaborative Review of Planning Models

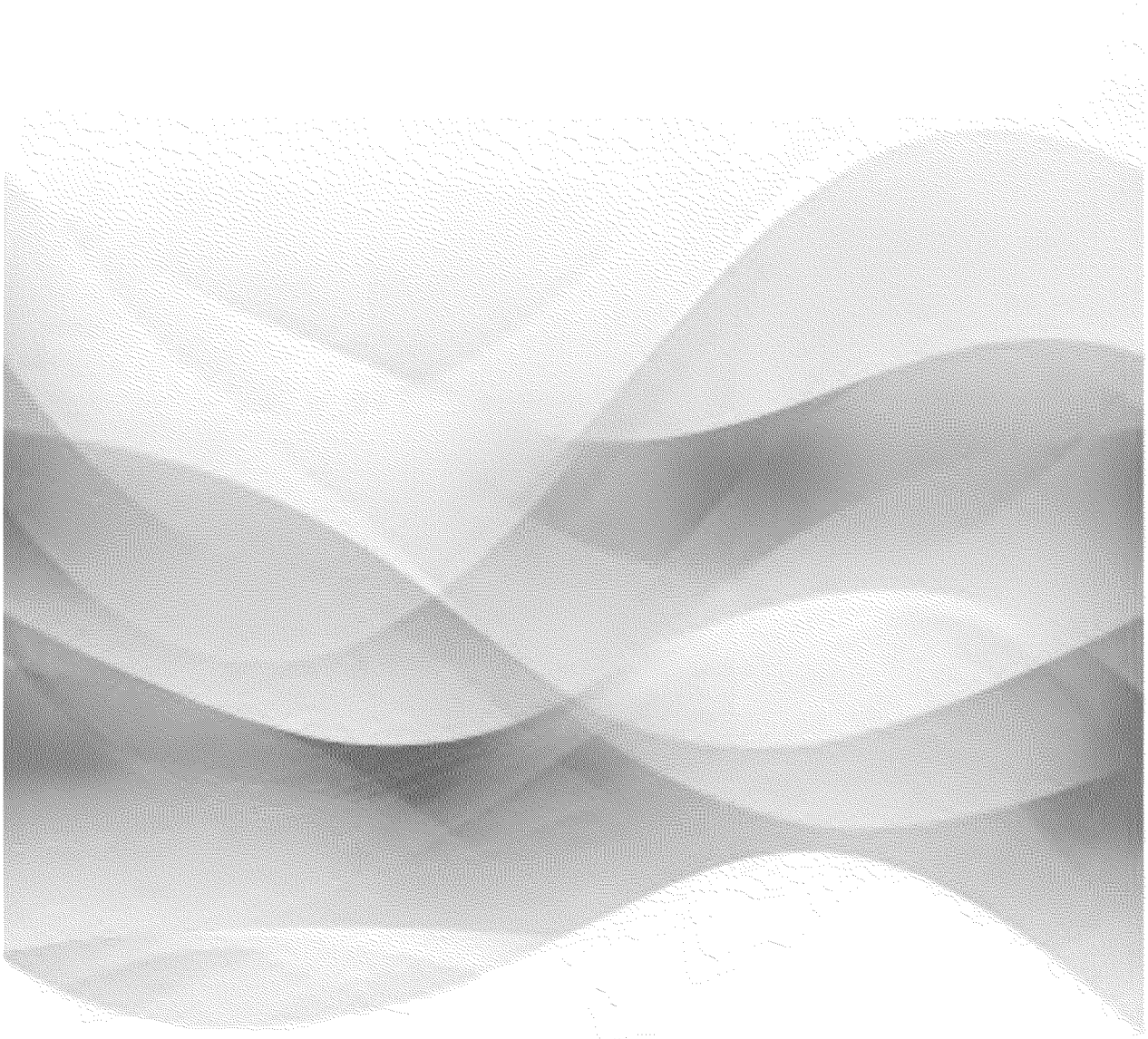


Table of Contents

Executive Summary

1. Introduction

This report is intended to improve the understanding of resource planning models under development or previously used in California to answer questions about system reliability and operating flexibility needs.[1] This effort is not designed to select a model, but rather to improve understanding of how models can be used to evaluate a system's performance, and inform future policy and planning decisions.[2]

1.1 Questions That Planning Models Should Help Answer

With the increase in weather-dependent renewable generation, the system is increasingly challenged to provide operational flexibility to accommodate increased variability and forecast uncertainty. As a starting point, the report identifies questions listed in Table 1 that planning models should help answer when evaluating the adequacy of a system. California has a number of planning processes and proceedings that are addressing some of these questions, including the CPUC's LTPP and Resource Adequacy (RA) proceedings, the CAISO's Flexible Resource Adequacy Criteria and Must-Offer Obligation (FRAC-MOO) stakeholder initiative, and the CPUC-CAISO Joint Reliability Plan (JRP) proceeding.

Table 1: Questions Models Should Help Answer

1. How to evaluate the future performance of a system.
2. What is the frequency, duration, and magnitude of projected shortfalls or deficiencies[3] in a given system, if any?
3. What is causing any projected shortfalls or deficiencies?

[1] Some of these models have not yet been used to evaluate the electric system's performance in California. Others have been used in the past, but are in process of modification for future work. For example, Southern California Edison Company's (SCE) stochastic model, as described herein, is different than the model that SCE plans to use in the California Public Utilities Commission's (CPUC or Commission) 2014 Long Term Procurement Plan (LTPP) proceeding.

[2] Parties that participated in this collaborative effort include: California Independent System Operator Corporation (CAISO), The Utility Reform Network (TURN), Pacific Gas and Electric Company (PG&E), National Renewable Energy Laboratory (NREL), Lawrence Livermore National Laboratory (LLNL), Electric Power Research Institute (EPRI), Energy Exemplar, Astrape Consulting, E3, and Gary Schultz, an independent consultant. **No party's participation should be construed as an endorsement or criticism of any particular model, modeling approach, or modeling results presented herein.** We are deeply grateful to all the participants for their contributions of time and effort, especially the model developers.

[3] The terms shortfall and deficiency are used throughout this report, these should be interpreted to mean deficiency or shortage of resources (generic capacity or upward/downward flexibility) that reaches a threshold where mitigating action is desired.

4. What is the cost and effectiveness of alternatives available to remedy any projected shortfalls or deficiencies?
5. What metrics, standards, and system requirements should be adopted from the evaluation of the system's performance and alternatives to remedy any shortfalls or deficiencies?

Evaluating the performance of a system is an iterative process.[4] First, initial assumptions about a system's loads, resources, and other characteristics are made, including amounts of variable energy generation and their potential impacts on system operations. The demands or requirements for reliability and operating flexibility[5] are included as an input assumption or estimated by modeling the system. Then, initial estimates of system shortfalls or deficiencies relative to the assumed or computed requirements are calculated, and after evaluating the cost and benefit of alternative solutions to reduce shortfalls or deficiencies, a final action plan or strategy is adopted to implement possible solutions.

2. Basic Model Descriptions

This report reviews the following five models or modeling approaches:

CAISO Deterministic Approach: The CAISO developed a modeling approach for use in the LTPP proceeding using Energy Exemplar's PLEXOS production simulation model to study the need for resources and flexibility to integrate 33 percent renewable energy. The CAISO's Deterministic Approach evaluates one scenario at a time assuming perfect foresight.

E3's REFLEX Model: E3 developed the Renewable Energy Flexibility (REFLEX) model to calculate the need for power system flexibility under high renewable penetration and to evaluate alternative strategies for meeting power system flexibility needs. REFLEX performs stochastic production simulation through Monte Carlo draws of load, wind and solar production.

SCE's Approach: SCE developed an approach that assesses the needs of an electrical system through stochastic production simulation modeling using PLEXOS. SCE initially developed this approach for analysis in the 2012 LTPP, and is continuing to improve it for future work.

SERVM Model: The Strategic Energy and Risk Valuation Model (SERVM) developed by Astrape Consulting is a hybrid resource adequacy and production cost model that stochastically evaluates a system's performance. The CPUC's Energy Division selected SERVM to estimate the reliability contribution of wind and solar in the CPUC's RA proceeding.

LLNL- California Energy Commission (CEC) Model: Lawrence Livermore National Laboratory (LLNL) developed an atmospheric physics-based stochastic weather model to represent the day-ahead uncertainty in renewable generation and load. The weather and renewable generator models provide an ensemble of potential net loads that are passed to the PLEXOS model.

[4] For purposes of this report, we will assume that the system's local reliability needs have been satisfied given that the models reviewed in this report are not designed to evaluate a system's local reliability.

[5] Defined in Appendix B.

3. Comparison of Modeling Approaches

The report compares the methodologies and inputs used by the models, and sample results for the 2012 LTPP Base Scenario without San Onofre Nuclear Generating Station (SONGS), or the Replicating Transmission Planning Process (TPP) Scenario without SONGS and low demand response. Key features of the models are summarized in Figure 1 below.

Figure 1: Key Model Features

	Scenario(s) Considered	Simulating Operating Decisions
Deterministic (CAISO) Deterministic)	A single "base case" or "stress" scenario at a time	Assumes perfect foresight, considers operating cost
Stochastic, statistical model (SCE)	Many scenarios, enables calculation of probability metrics (e.g. LOLE)	Assumes perfect foresight
Stochastic +uncertainty + recourse (REFLEX, SERVM)	Many scenarios, enables calculation of probability metrics (e.g. LOLE)	Considers uncertainty, operating costs, and ability to adjust decisions (recourse)

3.1 Approaches for Considering Uncertainty in Operating Decisions

All models have some representation of uncertainty in their inputs, and simulate economic commitment and dispatch decisions, but only REFLEX and SERVM consider changes in uncertainty as operating decisions are made, and allow for adjustments to these decisions as uncertainty is resolved (i.e., recourse).

Deterministic models simulate a scenario with perfect foresight. The CAISO's Deterministic Approach is an example of this method. This approach utilizes a single year of load, wind, solar, and hydro conditions. However, a single weather year provides no information about the system performance in other more or less stressful conditions. This approach incorporates regulation and load following requirements in commitment and dispatch decisions to accommodate the realization of uncertainty; however, this approach does not adjust early decisions to realized conditions, and may commit and dispatch more or less than is actually needed.

Statistical models stochastically simulate different conditions with a set of scenarios with different weather years, also assuming perfect foresight in unit commitment decisions.[4] The SCE Approach is an example of this modeling method. System performance can be summarized across multiple scenarios to compare to reliability standards such as a 1-day-in-10-year Loss of Load Expectation[6] (LOLE).

Stochastic models with uncertainty and recourse decisions develop an initial commitment considering the uncertainty at that time and, similar to how operators do as conditions change during the day, adjust commitment or dispatch of resources as needed. These later decisions are

[6] Defined in Appendix B: Glossary.

referred to as "recourse" decisions. REFLEX and SERVM are examples of this approach, but they do so in different ways – REFLEX with a value function incorporated in the optimization, and SERVM with updates to operating decisions as it walks through time.

Models of weather derived from models of atmospheric physics take into account the atmospheric information available in the day-ahead and recreate the weather forecasts and uncertainty as seen by the system operator. These models can be represented as possible scenarios over the next day's weather. The LLNL model is an example of this approach. LLNL samples the distribution of inputs to achieve a number of weather scenarios, and then makes a stochastic unit commitment decision that best fits all of the input scenarios simultaneously.

3.2 Other Differences Among Models

All models reviewed except for SERVM use PLEXOS.[7] PLEXOS uses Mixed-Integer Programming (MIP) techniques to perform optimal unit commitment and economic dispatch, starting with the day-ahead time commitment window and incorporating additional commitment windows during the operating day. However, only the CAISO's Deterministic Approach uses PLEXOS for a complete 365 day Western Electricity Coordinating Council (WECC)-wide simulation. Other approaches use simplifications to reduce run time. For example, REFLEX samples multiple three-day periods and assumes no transfer constraints within CAISO. SCE samples individual stressful days.

SERVM does not use a MIP optimization; instead, to reduce run time, SERVM breaks the large-scale optimization into a number of sub-problems using an evolutionary algorithmic approach. SERVM runs all 365 days for multiple scenarios and can incorporate transfer limitations within the CAISO and the rest of WECC, although for the sample analysis presented in this report only a simplified modeling of the rest of WECC was used. REFLEX and SCE's Approach also use a simplified representation of the rest of WECC.

3.3 Discussion of Sample Results for Each Model

The results show some similarities and some differences. The similarities are that all models show some type of reserve shortfalls or unserved energy in 2022. The amounts and types of shortfalls produced by a model are influenced by user-defined cost penalty assumptions. Whether shortfalls require new resources depends on the decision-makers' risk preferences and willingness to experience those deficiencies, and the trade-offs between the cost of deficiencies and the cost of new resources. The following discusses the sample results presented in this report.

Peak and Upward Flexibility Deficiencies

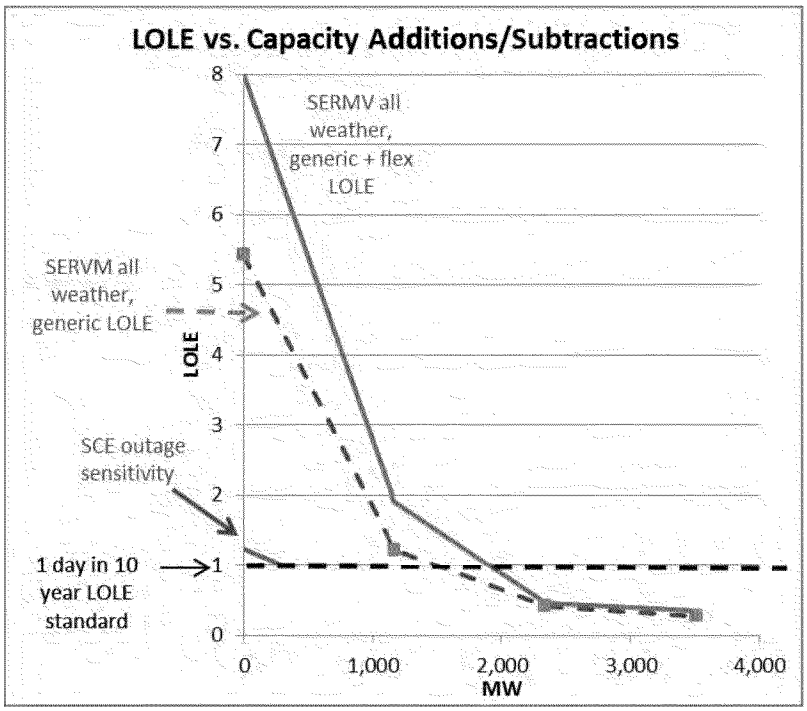
Traditionally, reliability modeling has been focused on the adequacy of the system to meet peak load using probabilistic methodologies. These methodologies measure Loss of Load Probability (LOLP) ignoring operational details, such as ramping needs and economic dispatch. These methodologies are the basis for California's 15 percent planning reserve margin requirement. This report, however, focuses on a recent class of models that bring additional operational detail to the evaluation of the system's operating flexibility adequacy. Although both traditional and

[7] E3's REFLEX methodology has also been implemented by ECCO International on the ProMaxLT™ platform and used in the study "Investigating a Higher Renewables Portfolio Standard in California": https://ethree.com/public_projects/renewables_portfolio_standard.php.

new models use loss of load metrics measured in hours/year and unserved energy megawatt-hours (MWh), the loss of load metrics calculated with REFLEX use the traditional approach where the capacity of a resource is counted to serve load whether or not it is committed as long it is available, while other models calculate loss of load metrics based on their unit commitment, which may account for the REFLEX's loss of load metrics being lower than those calculated from other models.

The CAISO's Deterministic Approach shows a maximum 2,709 megawatts (MW) reserve deficiency under 2005 weather in the 2012 LTPP Base Scenario without SONGS. SCE's Approach shows a maximum of 3,500 MW reserve deficiency in its benchmark analysis. SCE's model estimates an expected 1.24 Stage 3[8] events in 10 years assuming resource outages are capped at 1,000 MW. SERVM shows 1,908 MW are needed to reach a 1-day-in-10-year LOLE standard. SERVM estimates an expected 8.03 Stage 3 events in 10 years, excluding load growth uncertainty. SERVM LOLE is higher in part because SERVM assumes demand response availability is limited after 6 p.m. This is significant since Expected Unserved Energy[9] (EUE) occurs almost exclusively in hours after 6 p.m. The LOLE is driven by Stage 3 events in southern California. As noted in reviewing SCE's results, this need may be satisfied by new resources recently authorized in Track 1 or Track 4 of the 2012 LTPP. Figure 2 shows the LOLE to resource additions/subtractions relative to the resources available in the scenario. All the LOLE metrics shown in this figure are calculated based on each models' unit commitment.

Figure 2: Relationship between LOLE and Capacity Additions/Subtractions



[8] Defined in Appendix B: Glossary.

[9] Defined in Appendix B: Glossary.

Downward Flexibility Deficiencies

The CAISO's Deterministic Approach shows no need to dump or spill energy to balance the system most likely because of the assumption that the CAISO is able to export surplus generation and rely on the ramping capacity of its neighbors to meet its own net load ramping requirement. Wet hydro conditions may show some dump energy. SCE's Approach found no over-generation. REFLEX shows close to 60 gigawatt-hours (GWh) of expected over-generation, most likely the result of limited exports assumptions. SERVM shows close to 380 GWh of expected over-generation, most likely the result of no exports assumptions, less flexibility in hydro generation and imports compared to REFLEX, and forcing all dedicated imports into CAISO as must-take energy.

3.4 Answers to Planning Questions Posed in the Introduction Section

The following discusses how the models reviewed in this report answer or help answer the questions identified in the Introduction section.

Question 1: How to evaluate the future performance of a system.

All models can evaluate the performance of the system. However, they differ in the type of information they provide. Models that consider multiple scenarios provide a more complete picture of a system's performance. However, none of models reviewed offer a direct way to determine whether the upward flexibility deficiencies can be satisfied with flexible or inflexible resources. Additional sensitivities are necessary to determine the effectiveness of alternative solutions and the minimum flexible capacity the system needs. REFLEX and SERVM provide more system performance indicators about whether system deficiencies are associated with flexible or non-flexible requirements.

Question 2: What is the frequency, duration, and magnitude of shortfalls or deficiencies in a given system?

The CAISO's Deterministic Approach estimates the magnitude of deficiencies, but does not provide reliability metrics to compare against a 1-day-in-10-year LOLE standard. SCE's Approach provides reliability metrics. REFLEX and SERVM can also provide similar system performance information. In addition, these models consider intra-hour uncertainty and resources' variable costs, which is necessary for operating decisions, and resource evaluations.

Question 3: What is causing these shortfalls or deficiencies?

The REFLEX model incorporates an explicit step to test for pure capacity deficiencies prior to conducting flexibility analysis. SERVM calculates deficiencies before and after accounting for the operating constraints of committed resources. These metrics may be useful to provide an initial indication of the deficiency drivers. However, with both of these models, and any of the other models reviewed, additional sensitivities need to be run to test the sensitivity of the results to changes in input assumptions.

Question 4: What is the cost and effectiveness of alternatives available to remedy these shortfalls or deficiencies?

All models with the exception of SCE's Approach consider the variable cost of different alternatives. SCE is currently in the process of adding the capability of considering costs in its

system evaluations. However, because the models only consider variable or productions costs, the fixed cost of resource needs to be added outside of the model to complete a cost–benefit assessment of alternatives.

Question 5: What metrics, standards, and system requirements should be adopted from the evaluation of the system's performance and alternatives to remedy shortfalls or deficiencies?

Traditional loss of load reliability metrics and standards are used throughout the industry in system evaluations. The 1-day-in-10-year LOLE standard is the most widely used standard, but does not traditionally include flexibility driven outages. All stochastic models reviewed here calculate LOLE metrics. REFLEX calculates LOLE metrics excluding resources' flexibility limitations. SERVM calculates LOLE metrics with and without resources' flexibility limitations, and SCE calculates LOLE with flexibility limitations.

Today, there are no upward or downward operating flexibility standards. Some of the models reviewed here (SCE's Approach and SERVM) embed upward flexibility into the calculation of a traditional LOLE metric. This may be an acceptable approach. However, more work is needed to determine if there is a minimum amount of flexibility required to ensure the performance of a system and what the desired level of flexibility is given the cost of providing such flexibility.

4. Conclusions and Recommendations for Future Work

This report offers: (1) conclusions on model features useful to evaluate the performance of a system and possible solutions to deficiencies; and (2) recommended model improvements for future system evaluations.

4.1 Important Model Features to Perform System Evaluations

Ability to run multiple scenarios to capture the range of potential conditions: Evaluations that consider multiple scenarios rather than a single scenario provide robust results.

Modeling operating uncertainty: Modeling uncertainty that affects operating decisions is useful to estimate the amount and operating attributes of resources that the system needs. An accurate representation of system operating decisions is also useful to determine whether the system is flexible enough to accommodate increased variability and uncertainty and to evaluate alternate solutions to remedy any shortfalls.

Finding the best solution: The models reviewed in this report do not directly answer the question whether system deficiencies can be remedied with flexible or inflexible alternatives. Additional simulations are required to determine the effectiveness of different solutions.

Consideration of production costs: Considering costs is important to evaluate the system's performance and alternative solutions to deficiencies.

Consideration of transmission constraints within the CAISO: Transmission constraints may be relevant to the evaluation of the performance of the system and evaluation of possible solutions to deficiencies. Ignoring transmission can mask deficiencies that might arise in transmission-constrained areas.

Modeling interactions between the CAISO and the rest of WECC: Modeling the interactions between the CAISO and the rest of WECC is challenging not just because it adds computing

time, but because it raises questions about the ability to rely on the flexibility of other regions and the cost of flexibility services, compared to in-area alternatives.

Transparency: Transparency of the workings of the models and inputs is essential to get parties and decision-makers comfortable with the evaluation results.

Run time: Running more complex simulations requires computational resources. Regardless of which model is used (whether deterministic or stochastic), sensitivities are needed to examine the performance of a system, evaluate alternative solutions, and ensure results are robust.

4.2 Recommendations for Future Work

This report identifies the following three main areas where additional work is desired.

Flexibility metrics and standards: Traditional reliability metrics such as LOLE and EUE, and standards such as a not to exceed 1-day-in-10-year LOLE, are well understood and generally used in the industry. However, there are no standardized flexibility metrics or standards generally accepted in the industry to guide investment and procurement decisions for operating flexibility.

Modeling the rest of WECC is a significant challenge in planning studies because of the increase in computing time, but also because of the implicit assumption that the system can rely on neighboring areas to provide reliable operating flexibility at variable costs, ignoring fixed costs and/or the premium paid for these services in the market. At a minimum, additional sensitivities are needed to determine the robustness of the future decisions to these assumptions.

Reducing the run time of planning models to enable sensitivity runs is needed to develop confidence about the robustness of system performance evaluations and to compare the effectiveness of possible solutions.

Collaborative Review of Planning Models Report

Section 1 – Introduction

1.1 Scope and Objectives

This report reviews resource planning models that have been previously used in California or that are under development to evaluate the electric system's performance and study changes in infrastructure and operating practices that could facilitate the integration of large amounts of renewable variable generation. While other jurisdictions are facing similar needs for power system flexibility, this effort focuses on models that have been developed for or have been used in California.[10] The report is the result of a collaborative effort of various parties familiar with these models.[11]

This effort is not designed to select a model, but rather to improve understanding of how existing models can be used to address renewable integration questions and inform policy and planning decisions.

Traditional resource planning practice focuses on sufficiency of capacity to meet peak loads, and does not consider the system's requirement for operating flexibility.[12] However, given the increased reliance on highly weather-dependent renewable generation, the system's operating flexibility has become an increasingly important concern that should be addressed by planning models. Models now need to simulate the system's operating flexibility by optimizing unit commitment and dispatch decisions taking into account resource operating characteristics (e.g., start times, minimum/maximum generation levels, and ramp rates) and inter-zonal electricity transport limits. Advances in computing have allowed planning models to better represent the operating details and inherent uncertainties of the system, although there is always a trade-off between computational speed and level of detail.

This report does not consider modeling approaches used to answer local reliability or intra-zonal congestion questions, which require a more detailed representation of the transmission system than is used in the planning approaches considered here. Nor do these models resolve frequency response and system stability concerns at very short time-scales.

1.2 Background

The electric system is continually changing in response to evolving policy preferences and customer needs. In the United States (U.S.), California is leading the trend toward energy policy

[10] Some of these models have not yet been used to evaluate the electric system's performance in California. Others have been used in the past, but are in process of modification for future work. For example, SCE's stochastic model, as described herein, is different than the model that SCE plans to use in the CPUC's 2014 LTPP proceeding.

[11] Parties that participated in this collaborative effort include: CAISO, TURN, PG&E, NREL, LLNL, EPRI, Energy Exemplar, Astrape Consulting, E3, and Gary Schultz, an independent consultant. **No party's participation should be construed as an endorsement or criticism of any particular model, modeling approach or modeling results presented herein.** We are deeply grateful to all the participants for their contributions of time and effort, especially the model developers.

[12] Defined in Appendix B: Glossary.

goals aimed at significantly reduce greenhouse gas emissions. Large amounts of wind and solar generation are being added, such that at least 33 percent of electric energy needs in California is expected to be supplied from Renewable Portfolio Standard-eligible generation by 2020.[13] At the same time, the State is requiring the retirement or retrofit of about 13,000 MW of operationally flexible generation that currently uses once-through cooling processes.[14] These large infrastructure changes require an examination of the electric system's capabilities and operating practices,[15] to ensure that the system can meet the new demands for operational flexibility in a reliable and economic manner.

1.3 Questions That Planning Models Should Help Answer

With the increase in weather-dependent renewable generation, the system is increasingly challenged to provide sufficient operational flexibility to accommodate the associated variability and forecast uncertainty. In order to achieve a certain level of reliability, it is insufficient to simply plan to procure enough dependable capacity to cover the forecasted annual peak and a planning reserve margin, as has been the practice in California and elsewhere. It is now also necessary to understand what types of system capacity and specific operating attributes are required, and what changes in operating practices can also effectively reduce flexibility requirements or enable access to more operating flexibility. These and other questions provide decisions makers, utility planners, and model developers with a challenge as well as an opportunity to use existing models in new ways, or to modify or improve existing models.[16]

Answering such questions in a planning study involves not selecting a modeling tool, but also establishing how that tool should be used—the *modeling approach*. Any modeling tool can be used in a variety of ways: different input assumptions can be used, functionalities can be turned on or off, and results can be interpreted differently. This report attempts to capture intrinsic differences in the modeling tools themselves as well as differences in the modeling approaches used in recent California planning studies.

As a starting point, this report identifies five questions listed in Table 1.1 that planning models or modeling approaches should help answer.

---

[13] For further information on California's Renewable Portfolio Standard, please visit the following webpages: http://www.cpuc.ca.gov/PUC/energy/Renewables/ and http://www.energy.ca.gov/portfolio/.

[14] For further information on California's once-through cooling regulation, please visit the following webpage: http://www.swrcb.ca.gov/water_issues/programs/ocean/cwa316/.

[15] Changing operational practices can better use existing flexibility rather than provide new flexibility in and of itself.

[16] This description of coming changes to the traditional planning tools should not be taken to mean that the state's current amount of existing capacity—or the subset of existing flexible capacity—is not sufficient to meet such operating challenges in the coming years. A similar caveat applies to current and planned future operating practices designed to integrate variable generation.

Table 1.1 – Questions Models Should Help Answer

1. How to evaluate the future performance of a system.
2. What is the frequency, duration, and magnitude of projected shortfalls or deficiencies in a given system, if any?
3. What is causing any projected shortfalls or deficiencies?
4. What is the cost and effectiveness of alternatives available to remedy any projected shortfalls or deficiencies?
5. What metrics, standards, and system requirements should be adopted from the evaluation of the system's performance and alternatives to remedy any shortfalls or deficiencies?

(a) The terms shortfall and deficiency are used throughout this report. These terms should be interpreted to mean deficiency or shortage of resources (generic capacity or upward/downward flexibility) that reaches a threshold where mitigating action is desired.

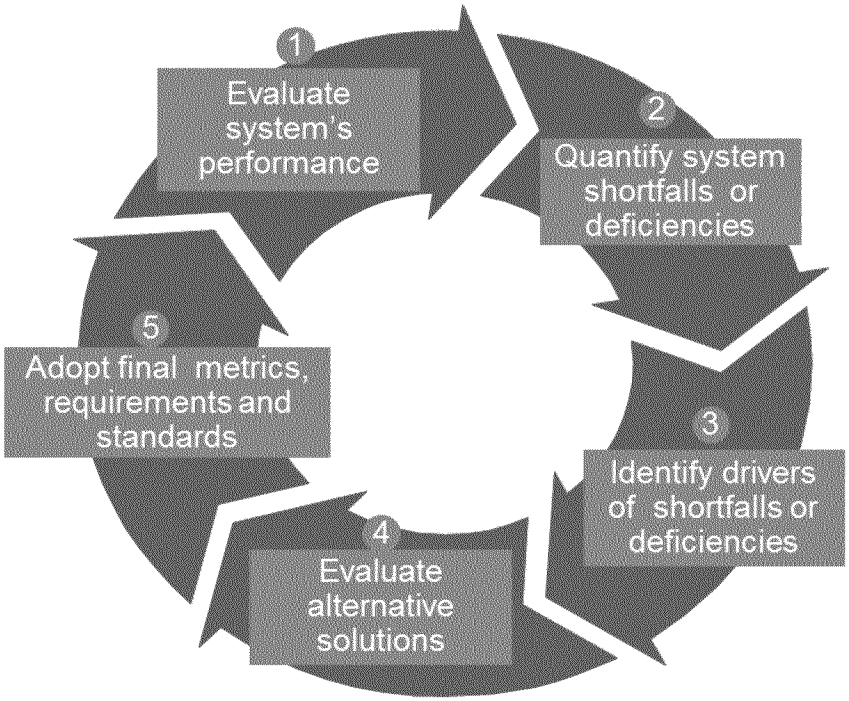
1. How to evaluate the future performance of a system.

The challenges posed by high levels of variable generation are the increase in variability and uncertainty of the net load—the residual load remaining after accounting for variable renewable generation—and the increase in the frequency and magnitude of net load ramps, relative to prior experience. Evaluating the performance of a system therefore requires a model that can characterize the ability of the power system to meet changing net load conditions across various timescales (e.g., 5 minutes, 20 minutes, 1 hour, 3 hours, and 5 hours). Important characteristics to evaluate include resource start times, minimum up times, minimum down times, minimum stable generation levels, and ramp rates in both the upward and downward directions. Restrictions on operations imposed by air permits, or energy budgets for energy limited resources such as storage or hydroelectric power, should also be considered.

Evaluating the performance[17] of a system is an iterative process meaning that initial assumptions about the demands or requirements for reliability and operating flexibility are made at the beginning of the process. This iterative process is depicted in Figure 1.1. Then, initial estimates of system shortfalls or deficiencies relative to the assumed demands and requirements are calculated, and, after evaluating the cost and benefit of alternative solutions to reduce shortfalls or deficiencies, a final action plan or strategy is adopted to implement possible solutions.

[17] For purposes of this report, it is assumed that the system's local reliability needs have been satisfied given that the models reviewed in this report are not designed to evaluate a system's local reliability.

Figure 1.1: Iterative Process Used to Evaluation a System's Performance



2. What is the frequency, duration, and magnitude of projected shortfalls or deficiencies in a given system, if any?

Planning models should be able to determine the frequency, duration, and magnitude of any projected shortfalls or deficiencies relative to reliability and flexibility requirements. This means that the planning models must have quantitative metrics that characterize these shortfalls or deficiencies. Examples of metrics[18] that have been used by the models considered in this report include:

- Loss of Load Expectation (LOLE), also referred to as Loss of Load Frequency (LOLF), measured as the number of loss of load events
- Expected Unserved Energy (EUE), measured in MWh
- Load following or ramping shortfalls, measured in MW, or ramping rate shortfalls measured in MW/minute
- Expected Over-Generation (EOG), measured in MWh

In some models, the demand for reliability and flexibility services is represented as a downward-sloping curve, where the demand for these services is inversely proportional to their cost. In other models, reliability and flexibility services are treated as fixed requirements (i.e., hourly MW of different types of capacity, such as contingency reserves, regulation, or load following capability).

[18] See Appendix B: Glossary for Definitions of LOLE/LOLF, EUE, and EOG.

3. What is causing any projected shortfalls or deficiencies?

There are many constraints on power system operations that could prevent a system from meeting all potential net load conditions. These include insufficient capacity to meet peak loads, a shortfall of upward or downward ramping capability, insufficient minimum generation flexibility to accommodate all of the variable generation, or a combination of the above (e.g., conventional generators may not be able to turn off during the middle of the day to accommodate an influx of solar energy and turn back on in time to help meet an upward ramping event that occurs after sundown). A planning model should be able to provide information about which of the potential constraints is the most binding in order to inform the search for potential solutions.

Sensitivity analysis can be used to understand the drivers of shortfalls or deficiencies. Care should be taken in this step regarding the priority in which available resources are used to meet different system requirements and how the assumed cost functions representing the demand for different types of capacity can influence the resulting deficiencies. As explained later in this report, the priority with which a model allocates resources to different functions determines the type of deficiency it calculates. This step can help identify alternative solutions, and, together with the next step's evaluation of possible solutions, should confirm the cause of system deficiencies.

4. What is the cost and effectiveness of alternatives available to remedy any projected shortfalls or deficiencies?

Just as there are many constraints on power system flexibility, there are many ways to add flexibility to a power system (or reduce the need for physical operational flexibility). Examples include:

- Operational changes: Improved forecasting, improved unit commitment processes, shorter scheduling timelines, new market instruments such as the proposed Energy Imbalance Market, etc.
- Demand-side investments: Development of new demand-side programs such as demand response or flexible loads.
- Renewable dispatch: Providing the system operator with the ability to dispatch or curtail variable generation in order to avoid immediate or potential future reliability challenges.
- Supply-side investments: Addition/modification of resources such as fast-starting or fast-ramping combustion turbines, advanced combined-cycle generators, energy storage, and transmission to allow greater use of remote flexible resources.

Alternative solutions have varied costs and degrees of effectiveness in remedying shortfalls, which can be estimated. The benefits of alternatives can be estimated in some models by assuming a penalty or price for system shortfalls or deficiencies. For example, $/MWh penalties can be assumed for unserved energy, over-generation, and load following deficiencies. This provides an economic signal to inform a potential investment decision. In other cases, benefits may require the decision-maker's cost vs. risk trade-off. For example, a decision-maker may not want the frequency of unserved energy to exceed the typical 1-day-in-10-year expectation. In other cases, the system operator may be required to meet specific North American Electric Reliability Corporation (NERC) or WECC metrics or performance standards. In these situations, the benefit of an alternative solution will be in the form of avoiding unacceptable frequency or magnitude of certain events.

5. What metrics, standards, and tem requirements should be adopted from the evaluation of the system's performance and alternatives to remedy any shortfalls or deficiencies?

After evaluating the system's performance and possible alternative solutions, a final action plan or strategy can be developed, as mentioned above. Such an action plan could take the form of a regulatory decision, depending on the jurisdiction. The action plan or strategy could be based on adopted metrics and standards, and could take the form of new forward procurement requirements, long-term investments or procurement plans, and/or changes in resource operating characteristics or system/market operations.

As explained later, the modeling approaches considered in this report can address some or of all the questions identified above. In some cases, a model might directly address several or all of these questions in one step. In other cases, a model may not address all of these questions directly, but can be used within a larger analytic process to address a broader range of questions.

1.4 Proceedings and Stakeholder Processes Considering Planning Questions in California

California has a number of planning processes and proceedings currently taking place that are attempting to answer some or all of the questions above. The purpose of this report is to improve the understanding of models or modeling approaches that have been used in past proceedings and/or are under development for future use in California.

CPUC LTPP Proceeding:[19] The LTPP proceeding is a biennial proceeding intended to ensure that there are sufficient resources to meet the long-term future energy and capacity needs of CPUC-jurisdictional entities serving customers in California. The LTPP proceeding typically concludes with a determination of need for new resources to meet system and local area reliability standards. Recent LTPP cycles have attempted to investigate the need for incremental (to the existing resource base) flexible capacity, especially due to increasing renewable generation. Several planning models discussed in this report have been used in prior LTPP cycles and/or may be used in future LTPP cycles. The LTPP is also the venue in which the utilities submit procurement plans to serve bundled customers over a ten-year horizon.

CPUC RA Proceeding:[20] The RA proceeding is an annual proceeding that establishes one-year-forward procurement requirements for all load serving entities under the CPUC's jurisdiction, namely capacity procurement responsibilities for local and system reliability, and addresses other RA program policies. In its 2013 RA decision, the CPUC adopted a framework for flexible capacity procurement requirements with targets set for 2014 and the intention to begin implementing procurement obligations in the 2015 RA compliance year. Additionally, in the current RA cycle, the CPUC is considering revising the methodology to determine the qualifying capacity of wind and solar resources using the Effective Load Carrying Capability (ELCC) methodology. The CPUC's Energy Division is using the SERVM model discussed in this to complete the ELCC analysis.

[19] For further information on the CPUC's LTPP process, please visit the following webpage: http://www.cpuc.ca.gov/PUC/energy/Procurement/LTPP/.

[20] For further information on the CPUC's RA process, please visit the following webpage: http://www.cpuc.ca.gov/PUC/energy/Procurement/RA/.

CAISO FRAC-MOO:[21] The CAISO's FRAC-MOO stakeholder initiative is intended to operationalize the flexible capacity requirements and capabilities in its market. The initiative is expected to establish the operational obligations for different flexible capacity categories that can be used to meet flexible RA requirements. The CAISO will implement a methodology to calculate the systemwide flexible capacity requirement for the upcoming year and allocate that requirement to the local regulatory authorities in its territory in a separate stakeholder process.[22]

CPUC-CAISO JRP:[23] The JRP continues the ongoing cooperation of the CPUC and CAISO to ensure the reliability of California's electric system. It includes three initiatives: (1) establish multi-year forward RA requirements for system, local and flexible capacity; (2) consider development of a market-based CAISO backstop procurement mechanism to replace or augment the existing CAISO Capacity Procurement Mechanism; and (3) develop a unified long-term reliability planning assessment. The CPUC and CAISO jointly launched the JRP process at the end of 2013 to address the previously described changes to the electric system, which "pose new operational and market challenges that resource planners and transmission operators must be aware of, and responsive to, in order to ensure reliable electricity supplies."[24] This process includes a dedicated CPUC proceeding[25] and a dedicated CAISO stakeholder initiative.[26]

Integration Cost Adders (proceeding to be determined): One key policy issue that is expected to be addressed in a CPUC proceeding in the near future is the development of integration cost adders to inform the least-cost, best-fit procurement of renewable generation. Development of such adders requires assessing the system's reliability and flexibility requirements (increase or decrease) and associated costs, not simply overall system reliability and flexibility requirements and deficiencies.

Section 2 – Basic Model Descriptions

This section introduces the models or modeling approaches considered in this report. Additional information on these models and modeling approaches is provided in the descriptions prepared by developers or expert users of these models in Appendix A.

---

[21] For further information on the CAISO's FRAC-MOO initiative, please visit the following webpage: https://www.caiso.com/informed/Pages/StakeholderProcesses/FlexibleResourceAdequacyCriteria-MustOfferObligations.aspx.

[22] For further information on the CAISO's Flexible Capacity Requirements stakeholder process, please visit the following webpage: http://www.caiso.com/informed/Pages/StakeholderProcesses/FlexibleCapacityRequirements.aspx.

[23] The JRP is provided as an attachment to the CPUC's Order Instituting Rulemaking for the JRP proceeding located here: http://docs.cpuc.ca.gov/PublishedDocs/Published/G000/M087/K779/87779434.PDF.

[24] Joint Reliability Plan of the CPUC and the CAISO, November 8, 2013, p. 1.

[25] For further information on the CPUC's Joint Reliability Plan proceeding, please visit the following webpage: http://delaps1.cpuc.ca.gov/CPUCProceedingLookup/f?p=401:56:4872093540992::NO:RP,57,RIR:P5_PROCEEDING_SELECT:R1402001.

[26] For further information on the CAISO's Multi-Year Reliability Framework initiative, please visit the following webpage: https://www.caiso.com/informed/Pages/StakeholderProcesses/Multi-YearReliabilityFramework.aspx.

CAISO Deterministic Approach

The CAISO developed a modeling approach for use in the LTPP proceeding using Energy Exemplar's PLEXOS production simulation modeling tool to study the need for resources and flexibility to integrate 33 percent renewable energy. The CAISO Deterministic Approach consists of two steps: Step 1 uses a Pacific Northwest National Laboratory (PNNL) statistical model to calculate operating and flexibility reserve requirements (regulation and load following) for every hour and Step 2 uses PLEXOS to calculate capacity and flexibility deficiencies with deterministic assumptions for hourly load, renewable generation, and reserves from Step 1. The model uses the MIP method for unit commitment and dispatch. Simulations run chronologically to co-optimize energy dispatch, ancillary services (regulation and spin/nonspin contingency reserves), and load following provisions.

E3's REFLEX Model

E3 developed the Renewable Energy Flexibility (REFLEX) model to calculate the need for power system flexibility under high renewable penetration and to evaluate alternative strategies for meeting power system flexibility needs. REFLEX performs stochastic production simulation after first making Monte Carlo draws of load, wind, and solar production. REFLEX then uses the MIP method to perform optimal unit commitment and economic dispatch for the operating conditions drawn. REFLEX is implemented on the PLEXOS and ProMaxLT MIP production simulation modeling platforms. Demand curves for load following reserves are developed exogenously for each of the draws and incorporated into the cost-minimizing objective function. REFLEX estimates the likelihood, magnitude, duration, and cost of flexibility violations at both the hourly and sub-hourly level to characterize flexibility constraints and inform potential solutions. REFLEX provides an economic framework for determining optimal flexible capacity investments by trading off the cost of investments in new flexible resources against the value of avoided flexibility violations. REFLEX has been used to investigate flexibility constraints under high renewable penetration by the CAISO[27] and the five largest California utilities.[28]

SCE's Approach

SCE developed an approach that assesses the needs of an electrical system through stochastic production simulation modeling using PLEXOS. SCE's method incorporates key drivers such as load, wind, and solar production, and available capacity to estimate the likelihood of a loss of reliability event, either insufficiency of generation or reserve provision. Analysis of load and intermittent generation at a 5-minute granularity is employed to evaluate the flexibility of resources to manage intra-hour variability. Similar to the CAISO's Deterministic Approach, this model assesses the different load serving entities in California as part of a zonal network within the WECC electrical system.

SCE's model was initially developed for analysis of the CAISO system area as part of the 2012 LTPP, but is a part of ongoing modeling efforts and will continue to be developed and expanded for use in future work.

---

[27] Results from E3's study performed on behalf of the CAISO can be found at http://www.caiso.com/Documents/RenewableEnergyFlexibilityResults-Final_2013.pdf.

[28] Results from E3's study performed on behalf of the five largest California utilities can be found at https://ethree.com/documents/E3_Final_RPS_Report_2014_01_06_with_appendices.pdf.

SERVM Model

SERVM is a hybrid resource adequacy and production cost model developed by Astrape Consulting. The model stochastically simulates unit performance, weather conditions, and other stochastic variables representing load growth uncertainty, and resource outages. It simulates thousands of iterations representing full years with a 5-minute granularity taking into account the short-term uncertainties introduced by load, wind, and solar generation. SERVM also evaluates the trade-off between reliability and costs as it estimates the frequency and duration of deficiencies, as well as the costs of unserved energy, operating reserve deficiencies, and generation curtailment.

The CPUC's Energy Division has selected SERVM for building a WECC-wide model to estimate the contribution of wind and solar generation towards reliability requirements in the ongoing CPUC RA proceeding.

LLNL-CEC Model

LLNL developed an atmospheric physics-based stochastic weather model to represent the day-ahead uncertainty in renewable generation and load. The weather and renewable generator models provide an ensemble of potential net loads that are transferred to a PLEXOS production simulation model. LLNL used this model to study the value of energy storage and demand response for renewable integration for the CEC.

The PLEXOS model executes an hourly day-ahead stochastic unit commitment and a 5-minute economic dispatch using the realized net load.

The model has been applied along with KERMIT, DNV GL's electromechanical simulation model, to estimate the value of storage and demand response for energy arbitrage and regulation given load following and regulation requirements.

Models Not Considered

Other tools and approaches beyond those described here are also available. Some are well-known commercial production cost tools such as PROMOD IV, GE MAPS, GridView, Aurora, UPLAN, and others. These tools could be used in a similar manner to the tools used in this study, and with some adaptation, could potentially be used in future studies of flexibility in California.

Another set of tools are those which have been designed to look at issues relating to system flexibility, but are still at the point where they have not been used extensively on a commercial basis. Two particularly advanced models in this category are described below:

- Polaris Systems Optimizer (PSO):[29] This tool uses a number of advanced features, chief among them being a multi-cycle approach which allows users to accurately represent the various operational decision-making stages to simulate the impact of imperfect information in the decision-making process. This allows for examination of how decisions made at different cycles (e.g., day-ahead, hour-ahead, "real-time") impact each other. The EPRI is using PSO to examine the use of stochastic optimization techniques to determine optimal levels of load following reserves. Through a Department of Energy-funded project, they are examining the use of these methods in the CAISO

[29] For more information, see http://www.psopt.com.

system with high levels of wind and solar. Approaches such as this could be utilized in future studies for the California system, once they have been proven valuable in the demonstration studies.

- Flexible Energy Scheduling Tool for Integration of Variable Generation (FESTIV):[30] This tool, developed by NREL, was designed to understand the impacts of variability and uncertainty on operating reserve requirements. Using nested models of security-constrained unit commitment, security-constrained economic dispatch, and automatic generation control, the impact of load, wind, and solar variability and uncertainty on the Area Control Error can be examined in conjunction with the system operating costs. FESTIV has not been used to date in California-specific studies.

Other advanced simulation tools are available or are under development at research labs, consultancies, and universities. Many may be able to address issues identified as weaknesses of the existing tools described in this report, though the tools described here will also continue to evolve and adopt new features.

Section 3 – Comparison of Modeling Approaches

This section provides a more in-depth comparison of the methodologies, inputs, and the results from the various modeling approaches considered in the report; it includes the following subsections:

3.1: Methods for modeling uncertainty, operating decisions and resolution of uncertainty

3.1.1: Consideration of variations in system conditions and uncertainty in each model

3.1.2: Discussion of model features

3.2: Reliability and flexibility metrics used by the models

3.3: Method for determining system's demand (or requirements) for flexibility

3.3.1: Demand for "pure capacity" reserves

3.3.2: Demand for upward flexibility

3.3.3: Demand for downward flexibility

3.4: Methods for determining system deficiencies

3.4.1: Methods for determining pure capacity deficiencies

3.4.2: Methods for determining flexib le resource deficiencies, both upward and downward

3.5: Methods for evaluating alternatives to meet deficiencies

3.6: Comparison of Sample Results

3.7. Discussion of Sample Results for Each Model

3.8: Answers to planning questions posed in the introduction section

3.9: Other key inputs and solving approaches used by models

[30] For more information, see http://www.nrel.gov/electricity/transmission/festiv.html.

3.1 Methods for Modeling Uncertainty, Operating Decisions and Resolution of Uncertainty

All of the models and modeling approaches considered have some representation of uncertainty in their inputs. These are done with "random variables" that have an associated probability distribution. This section describes how real uncertainties such as weather are related to representations of load, wind and solar uncertainty in each model. It also describes the relationships among uncertainties.

When modeling systems with large volumes of intermittent renewable generation, we must assess the performance of the system taking into account the following factors: (1) the wide variety of possible conditions that the system must deal with (i.e., long-term statistical variance); (2) the fact that the system operator must make operating decisions (unit commitment and dispatch) using imperfect forecasts of conditions in the following hours or days, (i.e., operating uncertainty); and (3) the resolution of this operating uncertainty as the actual or realized conditions occur. Below we consider how each of the models or modeling approaches considers each factor.

3.1.1 Consideration of Variations in System Conditions and Uncertainty in Each Model

CAISO Deterministic Approach

Representing the Statistical Variance of Possible Conditions Over Long Time Horizon

The analysis models the conditions and system performance over all days in a study year (e.g., 2022) based on a single scenario. The scenario includes the renewable generation and load profiles based on one year of historical conditions (e.g., 2005 weather conditions) with adjustments for growth in loads and renewable generation for the study year.

Modeling Operating Uncertainty

The CAISO's Deterministic Approach considers two types of operating uncertainties. For the first intra-hour variability and forecast uncertainty, in "Step 1," the PNNL model calculates hourly load following and regulation requirements to cover the hour-ahead and real-time forecast errors of load and wind and solar generation. For the second type, forced outage uncertainty, the PLEXOS production simulation model generates resource forced outages randomly using uniform distribution functions and converged Monte Carlo method. Planned outages are also generated randomly, but with some limits during times when system supply is short.

Modeling Commitment Decisions

For each day, an optimal hourly unit commitment schedule is computed in PLEXOS. Uncertainty in the forecast of variable generation and load is accounted for by adding the hourly load following and regulation requirements described above, in addition to contingency reserve requirements to manage unit forced outages.

Modeling the Realization and Dispatch Over the Operating Day

The results of the hourly performance computed in the hourly unit commitment and dispatch are the system's performance over the day. The model's results include hourly operation costs,

reserve costs, and possible violations of non-spinning, load following requirements, and load not served when the system's total reserves drop below 3 percent of load. In the 2010 LTPP study, simulations on 5-minute interval were conducted for selected days. The 5-minute simulations were for economic dispatch only, unit commitments were inherited from the hourly simulations.

E3's REFLEX Model

Representing the Statistical Variance of Possible Conditions Over Long Time Horizon

REFLEX models the dispatch and performance of the system over a large number of days. REFLEX takes 30 years or more of historical weather and synthesizes multiple weather-years of load for the future study year. These are binned by month, weekend versus weekday, and load type (high, medium or low). Upon drawing a load for a given day, the wind and solar generation profiles are drawn from the corresponding bin (i.e., same month and load type). Therefore, REFLEX implicitly uses the bins to capture the relationship between load, wind and solar.

Each day is modeled by randomly selecting a scenario consisting of three consecutive days, each with loads, renewable generation, hydro conditions, and outages. REFLEX uses the middle day's performance and discards the first and last day. The statistical model has been structured to maintain the correlation between these scenarios via the bins.

Modeling Operating Uncertainty

Forecast error and variability are incorporated into day-ahead and hour-ahead unit commitment and optimization process using exogenous demand surfaces for sub-hourly upward and downward flexibility. The demand surfaces express the volume of within-hour EUE and EOG as functions of the load, wind and solar forecast error and variability and the quantity (in MW) and speed (in MW/minute) of load following reserves carried. The demand surfaces are incorporated into the dispatch optimization by two terms that account for the per-MWh value of unserved energy (VUE) and of the value of over-generation (VOG). VUE and VOG are constant and independent from the volumes of flexibility services the system might dispatch. REFLEX calculates the final volumes of expected hourly and, potentially, sub-hourly.[31] EUE and EOG through the optimal unit commitment and economic dispatch process performed by PLEXOS or ProMaxLT.

Modeling Commitment Decisions

The unit commitment algorithm minimizes the total operating cost over the day, including the effects of load, wind and solar variability and forecast error. Demand surfaces similar to those described above are utilized to express uncertainty at the day-ahead commitment window to inform the commitment decision. The optimal commitment and dispatch trades off the fuel and operating cost of committing or decommitting resources against the change in the expected costs of reliability and flexibility violations.

[31] System dispatch in REFLEX can be run at the hourly or sub-hourly (e.g., 5-minute) level. If hourly dispatch is used, sub-hourly EUE and EOG is accounted for using the demand surfaces. If 5-minute dispatch is used, the demand surfaces are used only to inform hourly unit commitment decisions and the volumes of sub-hourly EUE and EOG are calculated through the 5-minute dispatch.

Modeling the Realization and Dispatch Over The Operating Day

REFLEX models the dispatch over the operating day using hourly steps and uses intra-hour surfaces to represent the cost trade-off of not meeting all the reliability and flexibility requirements in the optimization process as described above. The scenario of loads and renewable generation over the day is the scenario used for the day-ahead unit commitment analysis, subject to exogenously-determined forecast error parameters represented as the demand surfaces for reliability and flexibility services over the operating day. The model calculates the four primary flexibility metrics: EUE, EOG, Within-Hour Expected Unserved Energy (EUE_{WH}) and Within-Hour Expected Over-Generation (EOG_{WH}), over each interval in the operating day.

REFLEX explicitly models the recourse decisions by pre-computing a value function and embedding it within the unit commitment optimization. This value function is the assumed cost of unserved energy, and over-generation, multiplied by the amount of unserved energy, and over-generation as represented by the demand surface, at all time frames. E3 estimates the functional relationship between amount (MW) and ramp rate (MW/minute) of reserves as inputs, and how they map to the amount of expected unserved energy (MWh) and expected over-generation (MWh). Each surface may be considered as a mapping with three inputs to one output. The three inputs are: (1) load net of inflexible renewable generation; (2) MW of available reserves, either up or down as appropriate; and (3) the maximum ramp rate of the reserves. The outputs are the amount of load energy not served or must-run energy not delivered. A different surface is computed for upward deficiencies (unserved energy) and for downward deficiencies (over-generation), so these may be given different penalties.

Both unserved energy and over-generation surfaces are computed for day-ahead and hour-ahead, yielding four surfaces. The day-ahead surface models intra-daily recourse, and the hour-ahead surface accounts for intra-hourly recourse. Distributions of forecast error are used in the surface calculation to calculate the expected values. E3 finds that the correlations between load, wind and solar forecast error are near enough to zero to be ignored within the day-type bins described above. This allows them to add the variances of load, wind and solar forecast error together to produce a net load variance.

These "surfaces" are then approximated by a finite number of linear surfaces, or "cutting planes." Since the surface is convex, these cutting planes are suitable for inclusion into a unit commitment optimization. No intra-hourly uncertainty is modeled, but minute-by-minute load is run through the hourly dispatch, and any unserved load or over-generation is counted. REFLEX counts intra-daily and intra-hourly deficiencies separately, so that small short-term deficiencies that are likely to be absorbed by the Alternating Current transmission grid may be discounted or removed.

One limitation of E3's approach is in the trade-off between the accuracy of the future value function versus the computation time needed to do it. To keep computational times low, the flexibility in the hydro system and the tie lines outside of the CAISO have been modeled using relatively simple minimum and maximum generating capacities, and maximum ramp rate parameters.

SCE's Approach

Representing the Statistical Variance of Possible Conditions Over Long Time Horizon

SCE models many individual days at a time and records flexibility or generation shortfalls for each day modeled. To develop useful violations statistics, a large number of days must be modeled. To reduce the total computational burden, stratified sampling is used to concentrate the computational effort on those types of days that have the highest net load and ramping requirements, therefore, the most difficult to quantify.

SCE uses 30 years of historical weather data to create 30 years of load data for the study year. As much wind and solar data as is available is used to create generation profiles. All load, wind, and solar data is produced at 5-minute granularity in order to capture intra-hour flexibility needs. The different forecasts are combined to create over 10 million potential net load days that could exist in the study year. Correlation between load, wind generation and solar generation is studied and applied to the analysis at present (data used for the 2012 LTPP analysis did not show any strong correlations between load, wind, and solar, so independence was assumed).

Outage uncertainty is represented outside of the production simulation. Prior to simulation, many PLEXOS unit outage samples are drawn to build up an outage cumulative distribution function (both planned and forced) for each season. Simulations are then run assuming the highest outage level drawn, and wherever the simulation indicates a shortfall, the distribution functions are used in a post-processing step to estimate the probability of outages reaching a level that would cause a Stage 3[32] emergency.

Modeling Operating Uncertainty

SCE's Approach does not account for day-ahead uncertainty, but it accounts for intra-day and intra-hour variability by using 5-minute profiles of load, wind, and solar generation to estimate load following requirements. Load following reserves are set based on the difference between the hourly average and the 5-minute net load values, thus capturing the 5-minute net load variability. For each day modeled, SCE's Approach selects a scenario over loads and renewable generation and uses a deterministic optimization to model day-ahead and operating-day decisions.

Modeling Commitment Decisions

Unit commitment is determined for each randomly drawn day using PLEXOS. For each day draw, a different unit commitment is modeled. Regulation and contingency reserves are set using a fixed percentage of load. Load following requirements are calculated as explained above. Unit commitment and dispatch decisions are made with perfect foresight by the model.

Modeling the realization and Dispatch Over the Operating Day

The SCE model computes the day-ahead unit commitment and calculates reliability and upward flexibility deficiencies on a 5-minutes basis since it represents load, wind, and solar via 5-minute profiles.

---

[32] Stage 3 is initiated by the CAISO when operating reserves are forecasted to be less than 3 percent. See: http://www.caiso.com/Documents/EmergencyFactSheet.pdf.

LLNL-CEC Model

Representing the Statistical Variance of Possible Conditions Over Long Time Horizon

The LLNL analysis modeled 361 of the days in 2005. Four days were omitted due to insufficient data. The LLNL-CEC model can provide an assessment of the probabilities of ramping shortfalls for that year. If needed, the analysis can be extended to other weather years for a more accurate estimate of the probabilities of flexibility shortfalls.

To reduce the computational burden, the days in 2005 were statistically clustered so that similar days were grouped together. Using this approach, a much smaller set of runs would be needed to represent a full year (about 24). The results from the full set of days were compared with the results from the smaller set, and the results varied by less than 10 percent.

Modeling Operating Uncertainty

The day-ahead uncertainty was modeled to capture information similar to that which a system operator would have in the day before the operating day. The atmospheric conditions over the Western U.S. are available in meteorological archives for each day in 2005. Using the Weather Research and Forecasting (WRF) model, 30 forecasts over the weather for the operating day were computed. Each of the forecasts used a different set of physics parameters in the WRF model. Consequently, the uncertainty over the atmospheric scenarios is derived from the uncertainty over the set of modeling parameters for each day. This set of 30 was clustered into six groups and one scenario was chosen from each scenario. Each selected scenario was weighted according to the number of scenarios in its cluster.

The renewable generation for each of the six scenarios was computed from the modeled wind speed and cloud cover. The loads for each scenario were also adjusted based of the temperatures in each of the scenarios.

Modeling Commitment Decisions

The day-ahead unit commitment was determined using the stochastic unit commitment capability in the PLEXOS modeling system. This finds the unit commitment schedule that minimizes the expected cost over the operating day given the six scenarios and associated weightings. In executing the analysis, the algorithm explicitly computes the operating costs for each candidate commitment schedule against the each of the scenarios over the atmospheric conditions. It then determines the schedule with the minimum expected operating cost.

The day-ahead analysis was done on a one hour time step. To account for sub-hourly ramps, additional ramping reserves were required in the model. The required ramping reserves were computed based on a rule that there should be enough ramping reserves each hour to meet a specified percentile of the ramping requirements observed in the set of scenarios.

Modeling the Realization and Dispatch Over the Operating Day

The actual, realized conditions that prevailed over each day in 2005 were used to model the dispatch of the system over the day using a 5-minute time step in the PLEXOS model.

SERVM Model

Representing the Statistical Variance of Possible Conditions Over Long Time Horizon

SERVM takes 30 years of historical weather and six different load growth forecast errors (load growth accounts for economic and energy efficiency factors but excludes weather effects). Loads, wind and solar generation—as well as ambient temperature derates for thermal generators—are developed for each of those years' weather profiles. Instead of drawing cases, SERVM simulates all 30 weather years which includes the load and generation profiles associated with that weather year. Each of the 30 weather years is simulated with 6 different load growth forecast errors resulting in 180 actual years simulated at 5 minute intervals. Then each of the 180 years is simulated with 100 iterations of unit outage draws. This ensures convergence as it results in 18,000 yearly simulations at 5 minute intervals. This representation of the probabilistic dependency between variables reproduces the relationships seen in the historical weather data.

Modeling Operating Uncertainty

Additional regulation and load following requirements are added to represent intra-hour forecast uncertainty and variability. Both regulation and load following requirements can be defined as percent of load, or a MW value by year, month, or hour.

Modeling Commitment Decisions

SERVM commits and dispatches resources economically. It performs a weekly commitment, and then in each hour looks over the next four hours to identify needed changes to commitments. An economic dispatch routine is used every hour (and intra-hour if desired) to identify the operating point, ancillary service contribution, and ramping capability of each unit.

Modeling the Realization and Dispatch Over the Operating Day

SERVM uses a chronological simulation, where random variables such as load, wind and solar generation at time t are drawn, and unit commitment and dispatch decisions made before moving on to time t+1, etc. Because the model does not have a perfect foresight at time t, SERVM updates these decisions as it walks through time in hourly time steps where it draws outcomes for the random load forecast error, and forecast errors for solar and wind generation. SERVM uses separate but correlated 4-hour, 3-hour, 2-hour, and 1-hour forecast errors updated and applied every hour. The magnitude of the error generally decreases as the operating hour approaches. There is also a day-ahead error that is used in the initial daily commitment.

SERVM also simulates unit outages chronologically. Unit outages are modeled with distributions of time-to-failure and time-to-repair values, and random draws from these distributions are independent of load, wind and solar generation. Unit performance is also modeled using start-failures, maintenance outages, single contingency outages, and planned outages to be consistent with the impact that unit performance has on actual operations.

SERVM uses historical weather years to generate a year of loads, and wind, solar and hydro generation amounts. It then simulates each week separately by walking along the time axis, drawing random load, wind, and solar error amounts based on forecast error distributions around a number of parameters for each component, and drawing outage states for each thermal unit. Hydro dispatch is based on the available energy and operational constraints defined by the

historical weather year drawn. At each hourly time step it updates the unit commitment and dispatch for the rest of the week. These decision variables are recourse for the outcome of the stochastic processes up to that point in the time axis.

Like E3, Astrape Consulting finds that the correlations in forecast error distributions of load, wind and solar are low enough to warrant independent draws of each error component during simulations.[33] However, they believe that the forecast distributions themselves of load, wind and solar generation amounts are not independent. This kind of a chronological algorithm—drawing outcomes for variables at one time step, then having recourse for one time step, and proceeding on in the same way for the subsequent time step—is a way of modeling recourse. The results (prices, transfers) from prior iterations also feed into future iterations of the same weather year so that the commitment algorithm learns the relationship during the course of the simulations. The information that is fed into subsequent iterations does not represent explicit requirements that must be met by the commitment, but rather provide indicative shaping to guide the commitment. For example, if during a particular week the model commits capacity outside of CAISO, in subsequent iterations a similar amount of capacity at a similar purchase price would be made available during commitment. The amounts and prices across all regions are adjusted each iteration such that they converge to an optimal level.

Ultimately, prices and transfers are determined separately for each iteration, but the evolutionary approach allows them to converge to the most efficient systemwide commitment and dispatch. Doing so loses some of the characteristics of an optimization-based method (like prices associated with constraints, and consistency with the CAISO's models), but adds the important feature of intra-daily recourse.

3.1.2 Discussion of Model Features

Accounting for Imperfect Forecasts and Uncertainty in Operational Decisions

The following discusses the various approaches to incorporate variations in system conditions and uncertainty which were presented in the prior section. The performance of a system depends on the system's ability to handle a variety of system conditions such as resource outages, different loads and hydro conditions, and the uncertainty of weather dependent variables such as load, and wind and solar generation. To accurately evaluate the performance of the system and determine if there is any unserved load or flexibility shortages, if any, it is essential to model the probabilistic dependence between these conditions. Various models consider them in different ways. The way that models account for uncertainty in forecasts can be described along three dimensions:

1. The method for modeling uncertainty before decisions are made.
2. The method for making commitment decisions (e.g., for day-ahead commitment), given the uncertainty.
3. Modeling the actual realizations of conditions and dispatch over the operating day.

[33] However, there is some implicit correlation in the variables used for drawing load and solar error (and possibly even wind). The solar error drawn is a function of the magnitude of solar output, (e.g., the forecast error on days with high solar output will always come from the same distribution). The load forecast error is treated similarly. To the extent load and solar actuals are correlated (i.e., high solar coincides with high load), the error will also be correlated.

The modeled performance of the system depends on all three of these dimensions. More important, the fidelity of the model results is determined by the way that these three dimensions are handled in the model. To achieve maximum fidelity, the uncertainty in the model should represent the uncertainty as seen by the operator each day, the decision processes should mimic the actual decision processes that the operator uses (or could use), and the realizations should be consistent with the forecasts that have been made ahead of time.

By doing a large number of scenarios with different (i.e., statistically variant) hourly or intra-hour profiles of load, wind and solar generation, a model can use this statistical variance in inputs to produce expected values of a system's performance metrics, such as loss of load probability. This can capture uncertainty over longer time periods (e.g., seasons or years). However, if the input profiles are certain (i.e., deterministic), then the model's unit commitment and dispatch decisions will not account for uncertainty at the time these decisions are made (i.e., operating uncertainty), which in actual operation would require different commitment and dispatch decisions to cover the uncertainty.

Combining the variability and uncertainty of inputs, for example, by modeling unit commitment decisions subject to forecast error, a model can more closely resemble the environment and decision-making process when operating the system, and therefore produce more accurate performance metrics.

Methods for Modeling Uncertainty

There are different methods for modeling different conditions and uncertainty. Figure 3.1 summarizes the differences between stochastic and deterministic models.

Figure 3.1: Stochastic Versus Deterministic Models

	Scenario (s)	Simulating Operations
Deterministic	A single "base case" or "stress" scenario	Optimize with perfect foresight
Stochastic	Many possible scenarios (enables calculation of probability metrics (e.g. LOLE)	Optimize with uncertainty (more realistic representation of system operations)

Differences between the deterministic and stochastic methods are described further below.

- Single deterministic scenario that is known ahead of time to the operator. In a sense, this represents the best that could be done with perfect forecasts. The CAISO's Deterministic Approach is an example of this method. This approach utilizes a single year of load, wind, solar and hydro conditions. An "average year" like 2005 in the WECC region is

typically chosen because it represents a year with generally normal conditions across much of the West, and because data is available. However, a single weather year provides no information as to the system performance in other more or less stressful conditions than the chosen year.

- Statistical models of different system conditions and the weather dependent variables. This method relies on a set of scenarios with different weather years, and can also be supplemented with probability distributions of forecast errors for weather dependent variables over different time frames. SCE's representation of load, wind and solar uncertainty is an example of this modeling method. REFLEX and SERVM also use weather year scenarios with probability distributions of forecast errors at different time frames when unit commitment and dispatch decisions are made.
- Models of weather derived from models of atmospheric physics: these attempt to take into account the atmospheric information available in the day-ahead and recreate the weather forecasts and uncertainty as seen by the system operator. These models can be represented as possible scenarios over the next day's weather. The LLNL-CEC model is an example of this approach. LLNL samples the distribution of inputs to achieve a number of scenarios of weather inputs, and then makes a stochastic unit commitment decision that best accommodates all of the input scenarios simultaneously.

Approaches for Incorporating Uncertainty in Commitment Decisions

Some models rely on explicit optimization methods to model operating decisions. In the case of deterministic representation of the forecasts, this consists of finding a commitment schedule that minimizes costs over the deterministic simulated operating day.

Uncertainty in the day-ahead or intra-day forecasts can be accounted for in several ways. Even in cases where a deterministic scenario is used for day-ahead scheduling, the operator can require that additional resources be scheduled to allow for errors in the forecast, as in the CAISO Deterministic Approach. In the cases where the uncertainty is represented as a probability distribution over the weather or over scenarios, some approaches like REFLEX or SERVM incorporate a demand for capacity cost function into their commitment and dispatch decisions as explained in the previous section. The LLNL-CEC model uses stochastic optimization to find the commitment schedule that minimizes the expected cost over the operating day.

Approaches for Modeling the Realization of Conditions and Dispatch Decisions Over the Operating Day

A system's performance depends on the actual weather pattern that is realized over the operating day and the dispatching decisions made during the day. In the case where the forecasts are modeled as deterministic, this is straightforward: the realized scenario is the same as the scenario that was used for day-ahead planning, and the operating decisions made for that scenario. With statistical methods, the realization is derived from statistically compatible weather dependent variables selected for each scenario used to evaluate the system's performance. Finally, in the case of physical modeling of the forecasts using the recorded day-ahead conditions, the model can use the actual, recorded realization over the operating day.

The CAISO Deterministic Approach, SCE, LLNL and REFLEX all utilize the PLEXOS production simulation model. PLEXOS uses MIP techniques to perform optimal unit commitment and economic dispatch, starting with the day-ahead time commitment window and

incorporating additional commitment windows during the operating day. As explained before, these approaches use PLEXOS differently from a complete 365-day WECC-wide simulation in CAISO's Deterministic Approach and the LLNL-CEC model, to three-sequential sample days in REFLEX with no transfer constraints within CAISO, to sampling of stressful days in SCE's Approach. REFLEX and SCE's Approach also use a simplified representation of the rest of WECC. Also, each approach has a different way of representing operational variability and uncertainty and long-term statistical variance in system conditions as explained in the previous section. REFLEX is also available on the ProMaxLT platform which utilizes similar MIP optimization techniques. REFLEX also uses cost surfaces to allow the model to choose whether to incur the cost of shortages or change unit commitment and dispatch to prevent or reduce those shortages.

All these techniques are processor-intensive and require significant run times. SCE's Approach and REFLEX address these computing challenges by simulating sample operating days, rather than modeling all days in a year or entire weeks of time-sequential operations. This enables operations to be simulated over a broader range of system conditions, as described below. The CAISO Deterministic Approach, LLNL's model, and SERVM simulate a complete year for each scenario.

SERVM does not use a MIP optimization; instead, to reduce run time, SERVM breaks the large-scale optimization into a number of sub-problems using an evolutionary algorithmic approach. The first sub-problem includes meeting load for every hour up to the minimum load of the week. The next sub-problems are then set up to meet remaining unserved load. For each subsequent sub-problem up to the final sub-problem which fully meets load plus operating reserve requirements, the unit constraints become more critical and all relaxations are progressively dismissed. The selection of resources to optimally meet the need in each sub-problem is performed using a proprietary indexing technique. Results are saved from each weekly commitment for use in an evolutionary algorithm to adjust the commitment for subsequent iterations, where it is re-optimized for the conditions drawn. SERVM can incorporate transfer limitations within the CAISO and the rest of WECC, although for the sample analysis presented in this report only transfer constraints within the CAISO and a simplified bubbles and pipes modeling of the rest of WECC were used.

All of the methods develop a commitment schedule. However, they differ in the way they handle subsequent decisions over the operating day. In reality, the operator observes the developments over the day and then adjusts the commitment or dispatch of resources as needed. These later decisions are referred to as "recourse" decisions. Some models may be relatively rigid in the way they handle recourse decisions. For example, the CAISO Deterministic Approach, and SCE's Approach do not allow the operations to adjust to realized conditions. These two approaches, however, incorporate regulation and load following requirements in prior commitment and dispatch decisions to allow enough flexibility in the system to accommodate the realization of uncertainty; however, these approaches do not adjust unit commitment or dispatch to realized conditions, and therefore prior commitment and dispatch may be in excess or short of what was needed to satisfy those conditions.

As explained below, of the models reviewed, two incorporate recourse decisions: REFLEX and SERVM. They do so in different ways—REFLEX with a value function incorporated in the optimization, and SERVM with a non-anticipative forward walk along the time domain. LLNL's approach has daily re-commitment, which acts as recourse to the realization of the

randomness for that day, and tries not to over-fit the unit commitment by using a stochastic unit commitment that incorporates the weather uncertainty operators see at the time unit commitment decisions are made. SCE's Approach uses a scenario approach and optimizes commitment and dispatch for each simulated scenario assuming perfect foresight. SCE's Approach provides a much better assessment of the system performance by considering more scenarios than the CAISO's single year deterministic approach. The CAISO's Deterministic Approach however, offers the more detailed representation of the system and the rest of WECC, compared to other modeling approaches.

3.2 Reliability and Flexibility Metrics Used by the Models

Metrics are used to measure the performance of a system. Where an accepted standard exists for a given metric, model results can be easily interpreted by decision-makers, for example the standard for planning reserve margin in California is 15 to 17 percent. However, it is important that a standard applied to any metric be developed with consideration for the cost and risk trade-offs inherent in that level of performance, or assumed acceptable deficiencies. Table 3.1 provides a list of the metrics produced by each of the models considered.

As can be seen in the table, all models produce a measure of unserved energy, either by scenario or by a compilation of statistics about unserved energy events for multiple simulated scenarios. These are expressed as expected Stage 3 events or expected LOLF, LOLP and EUE.

All models also produce a measure of over-supply conditions, and report different metrics including amounts of energy that the model needs to dump or spill in order to balance loads and resources (dump energy). Depending on whether the model simulates one or multiple scenarios, this metric is expressed as MWh of dump energy for a scenario or expected MWh of over-generation (EOG).

Models that simulate intra-hourly operating conditions can also produce within hour flexibility deficiencies metrics such as EUE_{WH}, and EOG_{WH}.

Table 3.1: Reliability and Flexibility Metrics Produced by Models

Modeling Approach	Reliability Metrics	Flexibility Metrics
CAISO Deterministic Approach	Unserved energy amount	Shortage of contingency reserves, regulation, and load following reserves Dump energy amount
E3's REFLEX	Traditional LOLP, LOLE/LOLF, EUE via RECAP module	EUE, EOG, EUE_{WH}, EOG_{WH}
SCE's Approach	Expected Stage 3 Events or LOLE	Expected Stage 3 Events or LOLE, EOG
LLNL-CEC Project	Same as CAISO Deterministic Approach	Same as CAISO Deterministic Approach

Modeling Approach	Reliability Metrics	Flexibility Metrics
SERVM	LOLP, LOLE/LOLF, EUE	Shortage of contingency reserves, regulation, and load following reserves, EUE, EOG

3.3 Method for Determining System 's Demand (or Requirements) for Flexibility

There are multiple demands for different types of capacity requirements considered by the models. Some are simply input assumptions; some are calculated by the models as part of simulations. The main types are contingency reserves (spinning and non-spinning), regulating capacity, upward flexibility, and downward flexibility and avoidance of over-generation. The requirements are a function of the composition and characteristics of the system's load and resources, including their respective uncertainty.

Some of the models consider these requirements for different types of capacity as a demand function consisting of a quantity of capacity and a price or cost that the model chooses in the optimization process. REFLEX and SERVM are examples of these models. Other models consider these requirements as absolute requirements for capacity that the system is required to have regardless of its costs. The CAISO's Deterministic Approach, the LLNL-CEC approach, and SCE's Approach are examples of these models.

As explained below, the requirements may be endogenous, exogenous, or some combination of both. For example, the requirement for load following in the CAISO's Deterministic Approach is calculated using PNNL's Monte Carlo model in "Step 1" and used as an exogenous input to PLEXOS, where it is treated as an absolute requirement. In REFLEX, load following adequacy in the system is assessed by examining the range of possible variability in the model and determining whether there is sufficient supply at different costs to meet this requirement. No matter which way this is calculated or assumed, the approaches will somehow consider whether the requirements can be met, how much the shortfall there is, if any. For the upward flexibility and peaking capacity, all the approaches consider the requirements and determine system deficiencies to some extent. Some explicitly calculate each requirement and the associated deficiency separately, generally considering flexibility requirements as an increment to traditional contingency reserves. For example, the flexibility requirements for load following and regulation would be incremental to the traditional contingency reserve requirement. Other approaches calculate these as an overall requirement.

An important aspect which will drive many of the results in terms of requirements and needs is the cost assumed when not meeting requirements, which may be a value of lost load, value of reserve deficiency, curtailment cost, etc. These are exogenous assumptions in all approaches, but these inputs should be compared.

Table 3.2 summarizes how each modeling approach calculates system requirements either calculated separately and used as an input to the model or calculated intrinsically as part of the model's simulation.

Table 3.2: System's Demand (or Requirements) for Flexibility Considered by the Models

Modeling Approach	**Contingency Reserves**	**Regulation Reserves** (up and down)	**Load Following Reserves** (up and down)
CAISO Deterministic Approach	6% percent of load. 50% of the contingency reserves is spinning and the other 50% is non-spinning.	Exogenous input calculated using the CAISO-PNNL Monte Carlo simulations based on 1-minute load, and wind/solar generation profiles and historical forecast errors. The hourly requirements cover 95% of deviation between actual load and 5-minute forecast of net load within the hour.	Exogenous input calculated using the CAISO-PNNL Monte Carlo simulations based on 1-minute load, and wind/solar generation profiles and historical forecast errors. The hourly requirements cover 95% of deviation between 5-minute and hourly forecasts of net load within the hour.
E3's REFLEX	Same as CAISO Deterministic Approach with a minimum 3% spinning reserves to avoid loss of load events.	Same as CAISO Deterministic Approach	Calculated intrinsically using demand surfaces to incorporate the value of unserved energy and of over-generation in the optimized commitment and dispatch.
SCE's Approach	6% percent of load, but 3% minimum before loss of load event (Stage 3 firm load curtailments).	1.5% percent of load.	For each of the 2,400 days drawn, estimated as the largest difference between the hourly average and the highest/lowest 5-minute interval value of net load.
LLNL-CEC Model	Same as CAISO Deterministic Approach	Same as CAISO Deterministic Approach	Uses 95% confidence intervals from ensemble weather model.

Modeling Approach	**Contingency Reserves**	**Regulation Reserves** (up and down)	**Load Following Reserves** (up and down)
SERVM	Contingency minimums for spinning and non-spinning are considered. These can be input as % of load or MW values. by year/month/hour of day. Model allows users to set the point at which firm load is shed (i.e., 3% spinning). SERVM explicitly considers contingency events and allows contingency reserves to dip below the minimums only during such events. The reserves must be restored within 90 minutes.	Can be input as % of load or MW values by year/month/hour of day.	Can be input as % of load or MW values by year/month/ hour of day. The model uses a reserve duration price curve to determine whether or not to fully procure the load following reserve target.

3.3.1 Demand for "Pure Capacity" Reserves

This requirement is concerned with having enough capacity to meet peak demand, (i.e., a traditional resource adequacy problem). This problem is well understood and is normally covered in one of two methods. The first method is to ensure reserve margin is sufficient; in California, generally 15-17 percent planning reserve margin. All modeling approaches should be able to assess this 15-17 percent margin requirement if supplemented with the "dependable" or RA qualifying capacity of each resource modeled.

The second method is to use detailed modeling of outages (scheduled and maintenance) and peak demand in a probabilistic fashion, covering many possible scenarios. This is done in SCE's Approach and the SERVM and REFLEX models. SERVM and REFLEX calculate traditional LOLE and EUE before operational constraints are imposed. In addition, REFLEX and SERVM models recalculate these metrics after operational flexibility constraints are imposed and allow for trade-off between reduced LOLE and EUE and the costs of meeting reliability and flexibility requirements. The LLNL and CAISO approaches, as they only sample one year, do not calculate these probabilistic resource adequacy metrics, but they calculate capacity deficiencies to meet non-spinning, regulation, load following and unserved energy for the scenarios simulated.

Analysis of "pure capacity" needs can be an important component of flexibility analysis, as it prevents a capacity shortfall from being misidentified as a flexibility deficiency. For example, if a system does not have sufficient capacity to meet loads, a flexibility analysis may show a lack of capability to meet the demand for flexibility. However, adding inflexible capacity may remedy the shortfall by freeing up flexible capacity to provide flexibility products. Sequential or

concurrent analysis of pure capacity and flexibility needs can avoid this issue by correctly identifying the binding constraint.

3.3.2 Demand for Upward Flexibility

The demand for upward flexibility is considered in all of the approaches, in some fashion or another. Again, there are multiple methods to consider demand for upward flexibility. Upward flexibility is needed to cover both variability and uncertainty due to forecast error, and is needed on multiple time scales: 1-minute up to several hours. In each of the models, demand for upward flexibility on the hourly or longer timeframes is assessed through the time-sequential simulations of system operations.

The models have different approaches for assessing upward flexibility on the sub-hourly timescale. The CAISO Deterministic Approach uses the PNNL tool to calculate demand for regulation and load following reserves (both upward and downward) to manage load, wind and solar variability and uncertainty in the within-hour time frame. These quantities are treated as absolute requirements in the production simulation modeling. This approach does not consider uncertainty in the day-ahead, intra-day; only in intra-hour time frames.

In SCE's Approach, the stratified sampling methodology and choosing of particularly strenuous days ensures that upward flexibility requirements and associated needs are considered. SCE's Approach does not consider uncertainty explicitly, but through the sampling approach a wide range of sampled day ensures that the most difficult conditions are covered. Choosing days based on 3-hour ramping and then modeling in a 5-minute dispatch calculates whether there is enough flexibility to cover the intra-hour variability of net load. Therefore, SCE's Approach considers day-ahead uncertainty, and intra-hour variability only, but no intra-hour uncertainty.

SERVM explicitly represents the additional load following required for load, wind and solar variability and intra-hour forecast uncertainty with a combination of load following requirements plus an intra-hour forecast error. Upward and downward flexibility shortages are reported separately.

In REFLEX, demand for upward flexibility on a sub-hourly timescale is assessed not through a hard requirement, but through a demand surface which is brought into the objective function. The upward load following reserve carried thus varies every hour as a function of the demand and the cost of carrying it.

The LLNL approach examines the day-ahead scenarios of wind and solar and estimates the upward and downward flexibility requirements each hour, for each forecasted scenario. It estimates the 95th percentile of flexibility requirements for each hour. This is applied to the model as a ramping capacity constraint for each hour in the day-ahead unit commitment algorithm.

It is clear therefore that all models to a certain extent cover the upward flexibility requirements and needs, but the way they do so may differ significantly. Models which explicitly account for within-hour uncertainty and variability will more accurately represent within-hour flexibility requirements. The other main difference is in how they investigate deficiencies; either as a combined LOLE-type metric which covers both pure capacity and flexibility shortfalls, or as a separate flexibility deficiency metric. These differences result in different performance metrics and estimates of various types and amounts of capacity need if and when deficiencies are found.

3.3.3 Demand for Downward Flexibility

Downward flexibility is important in order to ensure that the system can manage the ramps as net load decreases, and balance loads and resources at different time intervals. Downward flexibility is related to over-generation, which occurs when must-run and variable generation has to be curtailed to balance load. Depending on the severity of the conditions, curtailments of must-run and variable generation may occur before or after the CAISO short-term markets pay negative prices to dispose of excess energy.

Downward flexibility may be needed on multiple time scales: one-minute up to several hours. In each of the models, demand for downward flexibility on the hourly or longer timeframes is assessed through the time-sequential simulations of system operations.

The models have different approaches for assessing downward flexibility on the sub-hourly timescale. All modeling approaches consider some aspect of downward flexibility by including regulation and load following down requirements. The CAISO Deterministic Approach focuses mainly on whether there is enough capacity to meet regulation and load following down, and reports down capacity deficiencies and dump energy (i.e., energy the system needs to curtail or dump out of market to balance its net load). SCE's Approach calculates a load following down requirement, and tracks violations as well as other over-generation metrics (over-generation, dump energy, and exports). The LLNL approach will ensure there is also enough downward flexibility to manage all potential scenarios in the stochastic unit commitment or incur a penalty. The SERVM approach also includes downward flexibility requirements during both commitment and dispatch. During commitment, the production cost penalty of curtailed generation equals the contract price plus grossed up production tax credits, though this input can be modified. Therefore, the commitment should limit the magnitude of expected curtailment based on economic considerations, but some may still exist after commitment but prior to the hourly dispatch. During the hourly simulation, SERVM will attempt to sell power in hours in which curtailment is a potential, and will attempt to adjust the commitment of short-lead resources. If neither of these mitigation procedures completely eliminates curtailment, SERVM will accumulate over-generation as a metric or output.

In REFLEX, demand for downward flexibility on a sub-hourly timescale is assessed not through a hard requirement, but through an exogenous demand surface, similar to upward flexibility as described above, and the downward load following reserve carried thus varies every hour as a function of the demand and the cost of carrying it. This approach sets up an implicit trade-off between upward and downward flexibility: carrying higher levels of load following up reserves may allow the system to avoid EUE_{WH}, however, this may result in more generators running at minimum levels, potentially resulting in higher levels of EOG_{WH}. The model optimizes its decisions to result in the least-cost outcome given the penalty and cost inputs. Downward flexibility at the hourly level is accommodated through the unit commitment and economic dispatch decisions that are optimized over the three-day operating horizon. If sufficient downward flexibility exists, either on internal resources or on interties to neighboring systems, then over-generation is avoided. In the presence of flexibility constraints, the cost of incurring over-generation at the hourly level is traded off against the fuel costs, operating costs, emissions costs, and the potential cost of unserved energy. Hourly EOG will be logged to the extent that it represents the least-cost solution to the flexibility-constrained operating problem.

3.4 Methods for Determining System Deficiencies

There are different methods to calculate the amount of system deficiency, if any, when evaluating the performance of a system. The CAISO Deterministic Approach counts as a deficiency when the full amount of spinning reserve, regulation or load following requirements are not met, and the system is deemed to be deficient. This may be a reasonable approach to ensure reliability since only one weather year is studied deterministically; however it may exaggerate at times the need for additional capacity. LLNL uses a similar approach.

SCE's Approach calculates system deficiencies based on loss of load events or occurrences of total reserve falling below 3 percent, after which point CAISO is permitted to initiate rolling blackouts to preserve system fidelity. For the 2012 LTPP, these expected outages are then compared to the industry standard of one expected outage in 10 years LOLE. Although this standard is not a NERC or Federal Energy Regulatory Commission requirement, it is used throughout the energy industry to assess resource adequacy. If the system is expected to experience more than one outage event in 10 years, additional resources are needed to ensure sufficient system reliability. REFLEX and SERVM use a similar approach, and the resulted deficiencies can be compared to a selected LOLE standard to determine when additional capacity is needed.

A major difference between the CAISO Deterministic Approach and the three stochastic modeling approaches reviewed in this report (SCE's Approach, REFLEX and SERVM) is that stochastic approaches incorporate extreme scenarios, thus not all deficiencies calculated with these approaches require new capacity be added; only when deficiencies exceed a given reliability or flexibility standard. In the case of traditional flexibility deficiencies, new capacity is needed only when the calculated LOLE metric exceeds a 1-in-10-year LOLE standard. For flexibility deficiencies, however, there is no standard to measure system deficiencies. However, REFLEX and SERVM provide the means to compare the cost of adding upward or downward capacity against the cost of an assumed penalty for not meeting flexibility requirements.

3.4.1 Methods for Determining Pure Capacity Deficiencies

Some of the models determine "pure capacity" deficiencies. REFLEX and SERVM both separate the deficiencies caused by flexibility and those caused by pure capacity. They do so by first calculating traditional LOLE and EUE metrics without considering the operational limitations of resources such as ramp rates, and minimum up and down times. If the deficiencies exceed the traditional 1-day-in-10-year LOLE standard, the need may be satisfied with operationally flexible or inflexible resources. To determine whether flexible or inflexible solutions are useful, the models re-calculate LOLE and EUE metrics a second time considering the operational limitations of resources. If the LOLE metric does not increase beyond the selected 1-day-in-10-year standard and the flexibility metrics are reasonable[34] values, then pure capacity solutions can be added. Additional sensitivities adding incremental inflexible resources may be needed to confirm that inflexible resources are truly effective in meeting the "pure capacity" deficiencies.

[34] The reasonableness of flexibility metrics needs to be determined since there are no operating flexibility standards today.

The CAISO Deterministic and LLNL approaches do not calculate LOLE or EUE metrics because they use a single weather year and perhaps one or more stress scenarios. Also, the results provide no information as to whether the deficiencies can be met with operational flexible or inflexible resources because the type and amount of the deficiencies found, if any, are a function of priority with which resources are used to meet peak and flexibility requirements. SCE's Approach calculates a single set of deficiencies based on the minimum total reserve threshold at 3 percent of load. Additional sensitivities adding incremental inflexible resources are needed to confirm that inflexible resources are truly effective in meeting the "pure capacity" deficiencies.

3.4.2 Method for Determining Flexible Resource Deficiencies, Both Upward and Downward

Broadly speaking, there are two approaches for determining resource deficiencies. The CAISO, SCE and LLNL approaches determine requirements for upwardly flexible resources. If these requirements are not met during production simulation, shortages are logged, denominated in MW. Generic resources are added until the violations are "cleared" (i.e., until the simulation can meet all upward requirements), or the LOLE metrics are below a selected standard such a 1-day-in-10-year LOLE. The quantity of MW added is said to represent the need for new resources.

REFLEX and SERVM take a different approach. These models do not identify an absolute "need" for new resources. Rather, they posit the investment decision as an economic trade-off between the cost of new resources added to avoid or reduce deficiencies, against the value that they provide, in the form of reduced flexibility deficiency costs. If the value of avoided flexibility violations is high, or the cost of flexibility solutions is low, more resources are added and a more flexible system results. Conversely, if new resources are expensive or flexibility violations are inexpensive, fewer additions will be indicated. This approach requires robust penalty values for upward and downward flexibility violations, as well as information about the cost of resource or flexible capacity additions.

All modeling approaches, however, do not have a direct method to calculate what portion of the need is for flexible versus non-flexible capacity as explained below. Additional sensitivities are likely needed to estimate the minimum amount and type of flexible capacity the system needs.

CAISO, SCE and LLNL rely on user-input cost penalties to prioritize the type of capacity or reserve requirements that the model shortfalls. For example, if the user assigns a lower cost to load following (or flexible capacity) shortfalls than to unserved energy, the models will produce larger and more frequent load following shortfalls than unserved energy. However, it is possible that these load following deficiencies can be reduced by adding non-flexible capacity; therefore, it is not possible to conclude whether the need can only be satisfied by flexible capacity without doing additional sensitivities. Although not done, it seems possible to replace different amounts of flexible with non-flexible capacity to test the impact of decreasing operating flexibility without reducing overall dependable capacity to determine the minimum amount of flexible capacity that the system needs without increasing flexible capacity deficiencies found in the performance evaluation of the system with the user-input cost penalties.

REFLEX and SERVM also use a system of user-input cost penalties implemented via a demand curve or surface to prioritize different types of capacity shortfalls. As noted earlier, in addition, these two models take an additional first step to determine the system's pure capacity shortfalls

without considering flexibility requirements. Any capacity shortages found at this stage may be satisfied with non-flexible resources. However, additional sensitivities are needed to confirm whether inflexible solutions are effective.

Table 3.3 summarizes how each modeling approach calculates deficiencies for different system requirements.

Table 3.3: Approaches Used to Calculate System Deficiencies

Methodology or Model	How are deficiencies calculated?
CAISO Deterministic Approach	Deficiencies are calculated hourly (and can be sub-hourly). Peak/upward deficiencies can be in the form of deficiencies in unserved load, contingency reserves or net load following. A deficiency occurs when contingency, regulation, or load following falls below required levels (i.e., any use of reserves is a deficiency). Downward deficiency can be in the form of excessive exports, and load following-down deficiencies. Uses penalty prices to prioritize use of resources and deficiencies. • Low to high priority order for peak/upward deficiencies: (1) load following-up; (2) non-spinning; (3) spinning; (4) regulation-up; and (5) unserved energy. • Low to high priority order for downward deficiencies: (1) excessive net exports; (2) load following-down; (3) regulation-down; and (4) dump energy.

Methodology or Model	How are deficiencies calculated?
E3's REFLEX	Probabilistic measures of deficiencies via correlated Monte Carlo draws of load, wind, solar and hydro shapes (load, wind, solar 1-minute profiles; hydro hourly) First, calculates capacity needed to meet a 1-event-in-10-year standard based on LOLF (RECAP Module) Unit commitment algorithm minimizes cost considering the assumed penalty of deficiencies (unserved energy, over-generation, contingency reserves, and up/down flexibility requirements) Makes unit commitment decisions day-ahead, 4 hour-ahead, 1 hour-ahead. Incorporates ramping policy functions into commitment decisions to account for forecast error and net load variability (i.e., willingness to pay to add capacity to manage forecast error and variability). Sub-hourly deficiencies are read from the demand surfaces as a function of the endogenous commitment decisions or from sub-hourly dispatch. Hourly deficiencies are determined directly through the optimal commitment and economic dispatch process. Calculates EUE, EOG, EUE_{WH}, and EOG_{WH}.
SCE's Approach	Probabilistic measures of deficiencies via stochastic representation of weather uncertainty drawing daily load, wind, and solar from historic 5-minute profiles adjusted to reflect a simulated future year's load/wind/ solar assumptions. (One day simulated per season, but many draws.) SCE is currently exploring ways to maintain historic correlations of load/wind/solar. Calculates deficiencies on a 5-minutes basis. Peak/upward deficiencies can be in the form of deficiencies in unserved load, contingency reserve or regulation. Calculates loss of reserve probability and LOLP. A loss of load event is assumed when contingency reserves drop below 3 percent. Can calculate over-generation or downward ramping deficiencies.
LLNL-CEC Model	Observe which constraints are violated in PLEXOS model

Methodology or Model	How are deficiencies calculated?
SERVM	Typical study will calculate probabilistic measures of deficiency by doing full hourly chronological simulation of 30 distinct load shapes with corresponding solar/wind/temperature shapes and hydro constraints combined with six different estimates of load forecast error (these reflect that the realized reserve margin will be higher or lower than the 15% target most years) and combined with 100 iterations of unit performance draws (for a total of 18,000 annual hourly simulations). Calculates deficiencies on an intra-hour basis and rolling it up on an hourly basis using economic commitment and dispatch of resources to load, and Monte Carlo techniques to model generation outages, and weather impacts on load and generation. Model reports unserved energy (and events) due to peak capacity deficiencies and ramping capability deficiencies separately. SERVM constructs weekly (and multi-hour) commitment based on inclusion of all ancillary service requirements. Commitment to meet downward requirements considers a curtailment penalty input to minimize curtailment energy.

3.5 Methods for Evaluating Alternatives to Meet Deficiencies

While the previous section shows there are differences in how the requirements and deficiencies are assessed in each approach, there is a greater difference when it comes to how individual resources, either new or existing, and costs are assessed. Generally, all models can be used to evaluate resource alternatives in the same way, by adding resources, and measuring their impact in terms of reduced deficiencies, and changes in cost. The differences are in the way models have been used to evaluate alternative solutions to deficiencies. For example, the CAISO Deterministic and SCE approaches, have been used until now to calculate system deficiencies, rather than to estimate the effectiveness (physical and net cost or benefit) of resource additions. It is feasible that in the future both approaches could be used to test the impact of possible solutions.

The LLNL approach currently is used to assess the contribution of demand response and energy storage resources. This is done by adding them to the CAISO system and assessing how various metrics are improved. In particular, the focus is on cost and price metrics, but prices would also indicate when there is a reduction in needs, due to reduction in scarcity prices.

REFLEX is used to assess the various options and resources available to manage violations; this examines the reduction in costs and reliability or flexibility violations when different resources are considered. A similar approach is used in the SERVM tool, where the overall system production cost benefits of different resources are considered.

It should be noted that all models *could* be used to assess how resources are used to meet requirements and/or needs; however, REFLEX and SERVM have been designed and used in the past to evaluate the cost and benefits of resource additions, and provide additional metrics to evaluate resources, as explained in prior sections. Models which represent uncertainty seen by

the system operator can evaluate solutions associated with improving operations, such as improved forecasting.

3.6 Comparison of Sample Results [35]

To the extent possible, this section provides results from different modeling approaches that are applied to a common scenario, the 2012 LTPP Base scenario without SONGS, defined in CPUC Decision 12-12-010 (referred to as the "Early SONGS Retirement Sensitivity"), issued on December 24, 2012. This scenario, however, is no longer used in the 2012 LTPP, will not be used in the new 2014 LTPP proceeding, and is used in this report simply to provide a common scenario to compare model results. The information presented in this section was primarily obtained through the documentation provided by the model developers in Appendix A and through public documents from the 2012 LTPP.[36]

A summary of the key assumptions used is presented in Table 3.4. The LLNL-CEC model is not shown in this comparison because this model was used to study the 2010 LTPP scenario so its inputs and results are difficult to compare with the other models.

Table 3.4: Comparison of Key Inputs for 2022 in the 2012 LTPP

Model	CAISO Deterministic Approach	E3's REFLEX	SCE's Approach	SERVM
Starting Set of Assumptions or Scenario	2012 LTPP Base Scenario Without SONGS	2012 LTPP Replicating TPP Scenario SONGS out	2012 LTPP Base Scenario Without SONGS	2012 LTPP Base Scenario Without SONGS
Other key assumptions used by model or modeling approach				
Contingency Reserves	6% of load	Same as CAISO Deterministic Approach with a minimum 3% spinning reserves	Same as CAISO Deterministic Approach with a minimum 3% spinning reserves	Same as CAISO Deterministic Approach with a minimum 3% spinning reserves

[35] These results are being presented for the purposes of making this report as complete as possible. None of these results have been reviewed adequately to form the basis of a CPUC finding of deficiency or need, and their inclusion in this report should not be interpreted as a statement of such need.

[36] SCE's presentation of their results from a September 18, 2013 workshop is available at the following link: http://www.cpuc.ca.gov/NR/rdonlyres/8A04F5B8-4990-4089-9E40-B7A064387C67/0/CPUC_ED_SCE_Workshop_StochasticModeling.pdf.

E3's final presentation of their results from a December 9, 2013 webinar available at the following link: http://www.caiso.com/Documents/RenewableEnergyFlexibilityResults-Final_2013.pdf.

Model	CAISO Deterministic Approach	E3's REFLEX	SCE's Approach	SERVM
Regulation Up	612 MW (1.2% of load on July 22, HE 19)(a)	Same as CAISO Deterministic Approach	1.5% of load	1.5% of load
Load Following Up	1,525 MW (3% of load on July 22, HE 19)	Calculates requirements based on statistical analysis of 5-minute deviations within the hour, up to +/- 2,100 MW	Max difference between average hourly and 5-minute net loads	Difference between forecasted average hourly and maximum and minimum 5-minute interval of load, calculated separately for each hour plus 1% of hourly load
Imports	11,197 MW (22% of load on July 22, HE 19)	13,308 MW max import; ramping available across the ties is limited to historical ramping rates	Seasonal limits from 11,400 MW (summer & fall) to 7,522 MW in winter	13,000 MW Max
Exports	Exports allowed subject to tie line limitations	Not allowed	Exports allowed with seasonal limits from 3,568 MW in Summer to 1,287 MW in Fall	Not allowed
Minimum In-area Generation	LA Basin: 40% SDG&E: 25%	LA Basin: 40% SDG&E: 25%	LA Basin: 40% SDG&E: 25%	LA Basin: 40% SDG&E: 25%
Within CAISO Path Constraints	Modeled	Not modeled	Modeled	Modeled

Model	CAISO Deterministic Approach	E3's REFLEX	SCE's Approach	SERVM
Resource Portfolio Differences From Original Base Scenario	n/a	Replicating TPP has a 1,887 MW higher managed load and 1,511 MW higher net supply than the Base scenario	Some DR programs extended past hour-ending 18	None
Resource Planned Outage Cap	None	None	1,000 MW	None

(a) July 22 at hour ending (HE) 19 was the largest shortfall hour.

Figure 3.2 depicts the relationship between the LOLE and resource additions/subtractions that were simulated by the models for the 2012 LTPP Base Scenario without SONGS. Only two models estimated the relationship between LOLE and different amounts of resource additions and subtractions. In SCE's case, SCE estimated the sensitivity of LOLE for different assumptions around the resource 1,000 MW outage cap initially assumed. With a 1,000 MW planned outage cap, SCE found the LOLE slightly above 1-day-in-10 year. SERVM found a LOLE of about 8 days in 10 years with the Base scenario as is, and estimated a generic CT resource need of about 1,900 MW was needed to achieve a 1-day-in-10-year LOLE standard.

Figure 3.2: Relationship Between LOLE and Capacity Addition/Subtraction

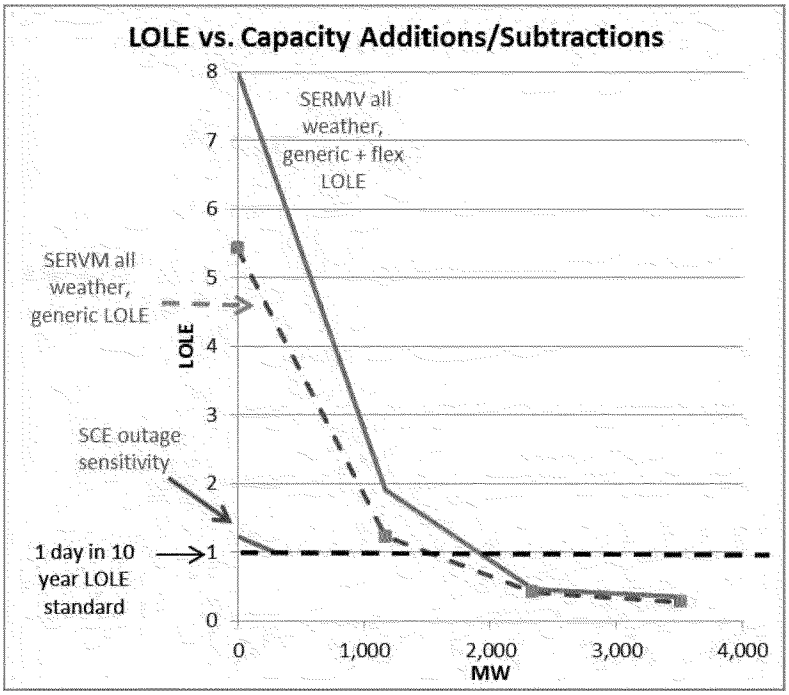


Table 3.5 compares the model results. The results show some similarities and some differences. The similarities are that all models show some type of reserve shortfalls or unserved energy in 2022. The amounts and types of shortfalls produced by a model are influenced by user-defined cost penalty assumptions. Whether shortfalls require new resources depends on the decision-makers risk preferences and willingness to experience those deficiencies, and the trade-offs between the cost of deficiencies and the cost of new resources. The differences are that all models produce different results, some which may be explained by differences in modeling approaches and assumptions used. Other differences are difficult to explain.

Table 3.5: Comparison of Results for 2022 in the 2012 LTPP

Methodology or Model	CAISO Deterministic Approach	E3's REFLEX	SCE's Approach	SERVM
Loss of load metrics excluding flexibility constraints[37]				
Loss of Load Frequency (Expected Stage 3 event in 10 yrs)	n/a	0.3 in 10 years	n/a	5.43 in 10 years 1,500 MW to meet 1-in-10-year LOLE
Loss of Load Hours (Expected Stage 3 hours per year)	n/a	0.07	n/a	0.74
Unserved Energy (MWh/year)	n/a	100 MWh/yr expected	n/a	69 MWh/yr expected
Loss of load metrics including flexibility constraints				
Loss of Load Frequency (Expected Stage 3 events in 10 yrs)	n/a	Not Reported (n/r)	1.3 in 10 years	8.03 in 10 years 1,908 MW of capacity needed to meet 1-in-10-year LOLE; Other methods of achieving 0.1 LOLE include adding flexibility, or changing operating strategies
Loss of Load Hours (Expected Stage 3 hours per year)	n/a	n/r	n/r	1.40 hours
Unserved Energy (MWh/year)	0 MWh	0 MWh expected	n/r	100 MWh expected

[37] The reliability metrics calculated with RECAP, as part of the REFLEX analysis, reflect a traditional reliability analysis without accounting for the system's flexibility requirements or the resource's flexibility constraints.

Methodology or Model	CAISO Deterministic Approach	E3's REFLEX	SCE's Approach	SERVM
Upward flexibility or peak shortfall				
Maximum Reserve Shortfall	2,709 MW	n/r	3,500 MW	2,900 MW
Regulation Up (MWh/yr)	0	1,124 MWh expected	n/r	0 MWh expected
Unserved Energy Within Hour (MWh/yr)	n/a	44 MWh/yr expected	n/r	31 MWh/yr expected - Not reported distinctly as within hour, so some events could have extended to inter-hour
Unserved Energy Intra-Day (MWh/yr)	0 MWh	0 MWh/yr expected	n/r	0 MWh/yr expected
Downward flexibility shortfall				
Regulation Down (MWh/yr)	0 MWh	0 MWh/yr expected	0 MWh/yr expected	0 MWh/yr expected
Over-Generation Within Hour (MWh/yr)	0 MWh	1,908 MWh/yr expected	0 MWh/yr expected	Included in Intra-Day Results
Over-Generation Intra-Day (MWh/yr)	0 MWh	61,372 MWh/yr expected	0 MWh/yr expected	377,795 MWh/yr expected (Assumes Inflexible Hydro energy budget and DI across peak hours.)
Net Exports	85,801 MWh 103 hours 2,798 MW Max	Not allowed	n/r	Not allowed

Methodology or Model	CAISO Deterministic Approach	E3's REFLEX	SCE's Approach	SERVM
Information available to inform operating characteristics of resource needed for shortfalls				
Peak	Deficiency occurs in the summer peak hours			Deficiency occurs in the summer peak hours
Upward Flexibility				Max Duration is ~1 hour
Downward Flexibility	All net exports occur in the spring solar hours			Overgen occurs mostly in shoulder months. Highly dependent on assumptions of hydro and import flexibility.

3.7 Discussion of Sample Results for Each Model

CAISO Deterministic Approach

Peak and Upward Flexibility Deficiencies

The CAISO's Deterministic Approach shows a maximum 2,709 MW reserve deficiency under 2005 weather in the 2012 LTPP Base Scenario without SONGS. All deficiencies occur in the peak month of July. It is difficult, however, to tell whether deficiencies can be satisfied with flexible or inflexible capacity without additional simulations to test the effectiveness of different alternatives.

The CAISO's Deterministic Approach does not show any unserved energy. The absence of unserved energy in the analysis is likely the result of the high priority given to resources serving energy, rather than load following or other reserve deficiency. The average or normal weather assumption can also contribute to the absence of unserved energy, compared to stochastic models that account for more stressful weather conditions. However, it is not possible to confirm this conclusion without running additional sensitivities.

Downward Flexibility Deficiencies

The CAISO's Deterministic Approach also shows no need to dump or spill energy to balance the system, most likely because of the assumption that the CAISO is able to export surplus generation and rely on the ramping capacity of its neighbors to meet its own net load ramping requirement. Wet hydro conditions could show some dump energy. CAISO's Deterministic Approach shows 86 GWh and 103 hours of net exports. Other models, as explained below, limit the exports or the ramping over the ties reflecting operating history, and therefore show higher over-generation levels.

SCE's Approach

Peak and Upward Flexibility Deficiencies

SCE's Approach shows a maximum of 3,500 MW reserve deficiency in its benchmark analysis.[38] SCE's model estimates an expected 1.3 Stage 3 events in 10 years. SCE notes that this reserve shortage, or the above 1.0 expected Stage 3 frequency, does not mean that the system needs resources because the scenario did not include all the capacity recently authorized in Track 1 or Track 4 of the 2012 LTPP, and that there is more existing capacity that was assumed retired in the scenario that could continue to operate.

Downward Flexibility Deficiencies

SCE's model description did not contain information about the downward flexibility deficiencies or CAISO net export, however, SCE confirmed that no over-generation or downward ramping violations were observed in their 2012 LTPP analysis.

E3's REFLEX Model

REFLEX used the Replicating TPP scenario without SONGS from the 2012 LTPP, rather than the Base scenario without SONGS. The Replicating TPP scenario is slightly more stressful than the Base scenario because it assumes less incremental energy efficiency and solar photovoltaics, effectively increasing the resulting CAISO load. This is partly offset because it has more resources than the Base scenario.[39]

Peak and Upward Flexibility Deficiencies

Even considering the more stressful conditions of the Replicating TPP scenario, the traditional reliability loss of load metric calculated with RECAP shows the system has adequate resources to meet a traditional 1-day-in-10-year LOLE metric that excludes the flexibility limitations of resources and does not account for limitations associated with economic commitment and dispatch decisions. However, including the flexibility requirements, REFLEX shows some within hour regulation-up deficiencies, rather than intra-day unserved energy. This is likely the result of a lower penalty assumed in REFLEX's optimization for regulation-up deficiencies ($1,000/MWh) compared to unserved energy ($50,000/MWh).

Downward Flexibility Deficiencies

REFLEX also shows close to 60 GWh of expected over-generation, most likely driven by the no export constraint.

SERVM Model

SERVM was run using the 2012 LTPP Base scenario without SONGS. The simulation includes a +/-4 percent additional load growth forecast error to account for economic factors, which was not included in other models. However, for the comparison results in Table 3.5, that impact was

[38] SCE's model description included in Appendix A. The 3,500 MW value is read from Figure 13.

[39] The differences in load and resources between the Replicating TPP without SONGS and the Base scenario without SONGS are shown in Table 3.4 in the Differences between Replicating TPP and Base scenarios row.

removed. Exports were not allowed in the simulation run to compare results to REFLEX, which also had no exports.

Peak and Upward Flexibility Deficiencies

SERVM shows 2,271 MW are needed to reach a 1-day-in-10-year LOLE standard. SERVM estimates an expected 10.6 Stage 3 events in 10 years, including a +/-4 percent load growth uncertainty. Excluding this load growth uncertainty that other models did not consider in the analysis reduces the need from 2,271 MW to 1,900 MW. SERVM LOLE is higher in part because SERVM assumes that demand response availability is very limited after 6 p.m. This is significant since unserved energy occurs almost exclusively in hours after 6 p.m. The LOLE is driven by Stage 3 events in southern California. As noted in reviewing SCE's results, this need may be satisfied by new resources recently authorized in Track 1 or Track 4 of the 2012 LTPP.

Downward Flexibility Deficiencies

SERVM shows close to 380,000 MWh of expected over-generation, most likely the result of no exports assumptions and limited flexibility in hydro generation and imports. Both REFLEX and SERVM did not allow exports. However, SERVM used a more limited hydro flexibility than REFLEX. The weekly hydro constraints imposed in SERVM prevented hydro dispatch from dropping below several thousand MW during several minimum load conditions. This minimum dispatch level was significantly higher than the minimum dispatch input used in the REFLEX simulations. SERVM also limited the flexibility of imports by forcing all direct imports into CAISO as must-take energy. These imports averaged several thousand MW during minimum load conditions.

3.8 Answers to Planning Questions Posed in the Introduction Secti on

The following discusses how the models reviewed in this report answer or help answer the questions identified in the Introduction section.

Question 1: How to evaluate the future performance of a system.

As noted before, all models can evaluate the performance of the system. However, they differ in the type of information they provide, and the range of scenarios and uncertainty they can consider in the evaluation. Considering multiple scenarios provides a more complete picture of a system's ability to perform under different conditions. Considering uncertainty—weather-related uncertainty impacting load, and variable generation, and resource outage uncertainty—similar to what operators experience provides insight into potential system deficiencies and resource need, if any.

The CAISO's Deterministic Approach can estimate capacity and any upward or downward flexibility deficiencies for one scenario at a time. Several scenarios can be run sequentially to understand the range of these deficiencies. At this point, the CAISO's Deterministic Approach provides the most accurate representation of the rest of WECC system, compared to other models' simplified representation of the WECC system.

SCE's Approach can estimate capacity or upward flexibility deficiencies for multiple sampled scenarios. SCE's Approach can also estimate downward flexibility deficiencies, although this wasn't the focus of the analysis presented here.

REFLEX and SERVM provide system performance indicators regarding system deficiencies, if any, of pure capacity and upward or downward flexibility for both intra-day and within hour, which can provide some indication of whether system deficiencies are associated with flexible or non-flexible requirements. However, as noted before, none of the models reviewed offer a direct way to determine whether the upward flexibility deficiencies can be satisfied with flexible or inflexible resources. Additional sensitivities are necessary with all models to determine the effectiveness of alternative solutions and the minimum flexible capacity the system needs.

Question 2: What is the frequency, duration, and magnitude of shortfalls or deficiencies in a given system, if any?

The CAISO's Deterministic Approach estimates the magnitude of deficiencies for a particular scenario, but it is not designed to provide loss of load metrics that can be compared to a loss of load reliability standard such as 1-day-in-10-year LOLE.

SCE's Approach can provide the following reliability metrics: frequency, duration, and magnitude of shortfalls or deficiencies, if any. It can also aggregate deficiency statistics about Stage 3 events from multiple sampled simulations to compare against a 1-day-in-10-year LOLE standard.

REFLEX and SERVM can provide similar system performance information. In addition, these models consider intra-hour uncertainty when calculating the various capacity or flexibility deficiencies, and variable resource costs when making commitment and dispatch decisions. Considering intra-hour uncertainty is necessary to represent the uncertainty operators experience when making unit commitment and dispatch decisions. Considering the variable costs of each resource is also necessary when making operating decisions, and to evaluate resource alternatives.

Question 3: What is causing these shortfalls or deficiencies?

The REFLEX model incorporates an explicit step to test for pure capacity deficiencies prior to conducting flexibility analysis. SERVM calculates deficiencies before accounting for the operating constraints of committed resources. However, with both of these models and any of the other models reviewed, additional sensitivities need to be run to test the sensitivity of the results to changes in input assumptions.

Question 4: What is the cost and effectiveness of alternatives available to remedy any shortfalls or deficiencies?

All models, with the exception of SCE's Approach, consider the variable cost of different alternatives. SCE is currently in the process of adding the capability of considering costs in its system evaluations. However, because the models only consider variable or productions costs, the fixed cost of resource needs to be added outside of the model to complete the cost–benefit assessment of alternatives.

Question 5: What metrics, standards, and system requirements should be adopted from the evaluation of the system's performance and alternatives to remedy any shortfalls or deficiencies?

Traditional loss of load reliability metrics are used throughout the industry in system evaluations. The 1-day-in-10-year LOLE standard is the most widely used standard, although it can be

measured and interpreted in different ways. Generally, the 1-day-in-10-year LOLE standard is used to measure pure RA capacity, excluding the operating flexibilities of resources. That means that the capacity of resources is available as long as they are available without considering start times, ramp rates, and minimum up and down limitations. All stochastic models reviewed for this report calculate LOLE metrics. REFLEX and SERVM calculate the traditional LOLE metrics, which don't impose a resource's flexibility limitations. SCE's Approach and SERVM calculate LOLE metrics including a resource's flexibility limitations.

Today, there are no upward or downward operating flexibility standards to guide the calculation of flexibility requirements or the need for new flexible capacity additions to the system. As noted before, SCE's Approach and SERVM embed upward flexibility into the calculation of a LOLE metric. This may be an acceptable approach. However, embedding downward flexibility into the calculation of a LOLE metric is controversial because of the lack of experience with these conditions and the inability to conclusively determine the cost of alternatives to deal with downward flexibility deficiencies.

Therefore, the decision about what flexibility metrics and standards to adopt (and in particular for downward flexibility metrics and standards) will likely require the systematic evaluation of the system's performance and of alternative solutions to remedy any shortfalls or deficiencies and their cost-effectiveness. All models can be used to some extent to perform these evaluations; however REFLEX and SERVM are better able to provide these evaluations because they consider variable costs and year-to-year weather uncertainties. Improvements to SCE's Approach can make this approach useful for this purpose in the future as well.

3.9 Other Key Inputs and Solving Approaches Used by Models

This section describes other key inputs used by the models, including the representation of the transmission system, imports and exports, and additional information about the modeling of hydro generation.

Transmission Within California

The transmission system is an important part of the electrical system, especially in large and complex areas like California. Transmission limitations could have a meaningful impact on results. As shown in Table 3.6, with the exception of REFLEX, all models represent several areas in the CAISO with bubbles and pipes (meaning that transmission constraints are imposed).

Table 3.6: Model Topology

	CAISO	SCE Model	REFLEX	SERVM
CAISO	"Pipe & bubble"	"Pipe & bubble"	"Copper plate" (no transmission constraints within CAISO)	"Pipe & bubble"
Rest of WECC	"Pipe & bubble"	A single resource representation	A single resource representation	Three resource representation

The CAISO Deterministic and SCE approaches as well as the LLNL-CEC model use the pipes and bubbles representation used in past CAISO studies performed with PLEXOS. SERVM also uses a pipe and bubble model, and can provide a similar representation of the transmission.

To date, REFLEX simulations have been performed using a "copper plate" (meaning that no transmission constraints are imposed) representation of the CAISO. However, the production simulation platforms upon which REFLEX is implemented (PLEXOS and ProMaxLT) include nodal representations of the CAISO system. Thus, it is possible to incorporate transmission constraints at a variety of levels. However, increasing the number of constraints will increase model run time. Deciding whether or not and at what level to incorporate transmission constraints involves a tradeoff between the increased accuracy of each simulated day against the potential reduction in the number of operating days that can be simulated or against other simplifications needed to achieve acceptable run time. Utilization of REFLEX in a high-performance computing environment may allow increased operating accuracy without loss of stochastic robustness.

Imports and Exports

Imports and exports are essentially other resources that may relieve or exacerbate unserved energy and over-generation problems. The CAISO Deterministic Approach and LLNL-CEC Model offer a more detailed representation of the rest of WECC than the other models considered. SCE's Approach and REFLEX use an aggregate representation of CAISO's neighbors. REFLEX uses the historical distribution of tie line flows to imply a minimal amount of flexibility available to the CAISO from without. SCE's Approach represents CAISO's interchange as a resource with limits representing the combined transfer limits of CAISO's ties, southern California and CAISO simultaneous import constraints, and flexibility limitations available from imports. SERVM also models transmission with a pipe and bubble topology.

All modeling approaches rely on CAISO's imports and exports to meet operating flexibility requirements, both in terms of ramping requirements and managing over-supply conditions. However, there are significant institutional and other non-technical constraints to accessing flexibility from neighboring systems. Local entities may also face policy decisions about the

extent to which they wish to develop internal strategies for managing flexibility, as opposed to relying on their neighbors. This is an area in need of further research to determine how much flexibility is available from neighboring systems and at what cost (e.g., the magnitude and frequency of negative prices to manage over-generation conditions, relative to other alternatives available to CAISO).

Hydro

Three of the approaches—CAISO, SCE, and the LLNL model—model hydro deterministically, as either run of river or dispatchable aggregate hydro energy. Dispatchable hydro uses energy budgets on a monthly (CAISO) or daily (SCE and LLNL) basis, along with minimum and maximum dispatch levels and maximum upward and downward ramps. We expect that these are all very similar, because they all use PLEXOS.

SERVM breaks hydro into three different categories: emergency, scheduled, and run of river. These quantities depend on the historical year sampled within the model. A portion of the hydro energy is scheduled prior to the sampling of intra-weekly uncertainty, so hydro operations is limited in the amount of recourse that it can provide to account for the intra-weekly (load, wind, solar, and forced outage) uncertainty. However, this scheduled block can provide regulation-up and regulation-down capability so it can reduce the amount of ancillary services that are required from the thermal fleet. The capacity block of hydro (represented as emergency hydro) can be used to provide recourse in the event of upward flexibility constraints.

REFLEX draws hydro minimum and maximum daily and hourly energy, maximum upward and downward ramping capability, and a daily energy budget from a historical dataset. The hydro is then scheduled within the unit commitment, so that the hydro ramping capability is combined with the scheduling of hydro to form part of the ramping capability to pass to the value function. In this way, REFLEX can implicitly use hydro ramping capability not used in the unit commitment as a recourse within the value function.

None of the approaches model the storage limits and flow topology constraints in the hydro system.

Solvers

SERVM builds the solver directly into the methodology. This means that it solves relatively fast, but should be benchmarked to the other optimization models, especially if one were to require a detailed dispatch representation or a more detailed deterministic case to test transmission system feasibility.

All of the other approaches considered use optimization models of unit commitment and dispatch as their core operational computation. Optimization models have nice features that allow adding more constraints, relaxing other constraints, and pricing constraints. However, optimization models also tend to over-fit the particular scenario, which is why recourse and robust design patterns become so important. It is less important which solver is used to find the solution to an optimization problem, so long as the solver works on large problems, and can read and write data in the right formats. All of the approaches using optimization use PLEXOS, and REFLEX uses either PLEXOS or ProMax.

Section 4 – Conclusions and Recommendations for Future Work

4.1 Introducti on

The objective of this section is to review the key features of models or approaches that are needed to evaluate the performance of a system and of alternative solutions to deficiencies identified in the evaluation, rather than to select a particular model. As noted elsewhere in this report, the models reviewed are changing. We hope that this collaborative process can help model developers identify features they would like to add to their models and help model users select their own preferred model. We leave it to the reader to reach his/her own conclusions.

4.2 Important Model Features Needed to Perform System Evaluations

The following model features and approaches are important in the evaluation of a system's performance and of possible alternatives to remedy any identified deficiencies. Some of the items noted below apply only to evaluations of the CAISO.

Ability to Run Multiple Scenarios to Capture the Range of Potential Conditions

Evaluations that consider multiple scenarios rather than a single scenario provide more robust results. Multiple scenarios can be examined by either making several deterministic model simulation runs or by using a stochastic model. For example, a single weather scenario that assumes average weather conditions is not sufficient to evaluate the true performance of a system under system stress: most loss of load events are found in hot weather scenarios where load is higher, or hydro generation is reduced, whereas most downward flexibility deficiencies are found under conditions with low load, high hydro, high renewable generation, or a combination of these conditions. This conclusion is supported by the results contained in this report for both deterministic and stochastic modeling approaches.

Modeling Operating Uncertainty

In addition to simulating multiple weather years, modeling the uncertainty that affects operating decisions—of weather, resource outages, etc.—is important to understand the capacity and necessary operating attributes of resources that the system needs. In general, more resources need to be committed (and therefore need to be procured ahead of time to be available for operators to commit and dispatch) to cover operating uncertainty. An accurate representation of system operating decisions is also useful to determine whether the system is flexible enough to accommodate increased variability and uncertainty and to evaluate alternate solutions to remedy any shortfalls, such as improved forecasts or different reserve setting methods.

Finding the Best Solution

Depending on whether flexibility deficiencies occur intra-day (from one hour to the next) or within hour, solutions may be found in changes to start times, ramp rates, and minimum up and down times. Models which represent uncertainty seen by the system operator can evaluate solutions associated with improved operations, such as forecast improvements. The model results presented in this report do not clearly answer the question of what type of resources can be effective solutions to the deficiencies found in the system evaluation (i.e., whether flexible or inflexible resources are needed). Additional simulations are necessary to determine the effectiveness of different solutions with different operating features.

Any of the models reviewed in this report can evaluate the effectiveness of alternative solutions. However, models that can evaluate the reliability and cost trade-off between increased or decreased commitment and the variable cost of the changes in resource commitment are able to provide additional insight as to the needs of the system and the effectiveness of possible solutions. These trade-offs can be evaluated with more detailed models that include recourse decisions available to the operator as he or she walks through time and considers the cost of deficiencies against procuring additional reserves, for example. In the end, regardless of the model used, multiple iterations are needed to strike the appropriate balance between the fixed cost of additional resources and penalty assumptions for deficiencies in the optimization. There is no one-stop, press-the-button-and-get-the-answer model.

Consideration of Production Costs

Considering costs is important to evaluate the system's performance and alternative solutions to deficiencies. If all requirements are not viewed as equally important and if there are different costs or penalties associated with not meeting different requirements, then costs need to be considered in commitment and dispatch decisions. Even more important, after determining that the system has shortfalls that need to be remedied, cost estimates are necessary to evaluate and select from potential solutions.

Consideration of Transmission Constraints Within the CAISO

It is important to consider constraints within the CAISO. All else being equal, models that ignore transmission constraints can mask a shortfall of capacity, or operating flexibility, because transmission constraints may prevent resources available in one part of the system from assisting other parts of the system. All models except for the current version of REFLEX consider transmission constraints within the CAISO region. Consideration of transmission constraints is essential to evaluate possible solutions or identify the preferred location of new resources when there are transmission constraints.

Modeling Interactions Between the CAISO and the Rest of WECC

Modeling the interactions between the CAISO and the rest of WECC is one of the more challenging aspects of evaluating the performance of CAISO's system. The models reviewed in this report use a range of approaches, from detailed modeling of loads and resources in the WECC and transfer limits in the CAISO Deterministic Approach, to one or more aggregated neighboring areas and the simple heat rate representation used by other models. Even the CAISO's representation assumes ramping and balancing services are provided at variable cost rather than with bids, with no capacity cost charged for energy, ancillary services, ramping, or other flexibility services. In reality, both fixed and variable costs need to be considered in evaluating the cost of services provided by the CAISO's neighboring systems. The challenge of modeling the interactions of the CAISO with the rest of WECC is not just a problem of scaling up the models, it is also a policy issue. To what extent can we rely on the flexibility of other regions and what is the true cost of the flexibility they would provide? The models underestimate the cost of flexibility services because they are based on variable costs alone, rather than bid prices, and ignore capacity payments, so decisions about resource additions should be made considering the sensitivity of these assumptions about the CAISO's interactions with its neighbors.

Transparency

Transparency of the workings of the models and inputs is essential to get parties and decision-makers comfortable with the results and recommendations. In general, more information is desirable about the workings of the models, especially where proprietary algorithms are used. We think this report should help improve understanding of the existing models, but the constant evolution of the models requires continuous education for all involved.

Run Time

Requirements for computing resources is also a major issue. Building more complex models takes additional effort and running more complex simulations requires greater computational resources. Regardless of which model is used, whether a deterministic or stochastic approach, sensitivities are needed to examine the performance of a system, evaluate alternative solutions, and ensure results are robust. If preparing a scenario takes a day or two of computer run time, and additional time is needed to validate the results and understand the drivers of the results, pretty soon time and resource costs become constraints.

Ability to Test the Costs or Penalties Assumed for Resource Deficiencies

The CAISO Deterministic Approach assumes all requirements are important and any deficiencies should be remedied. SCE's Approach assumes that as long as there is 3 percent or more operating reserves available, new capacity is not needed. REFLEX and SERVM consider the cost and benefits (or reduced cost of violations) to inform when new capacity is needed. This is an area where further research is needed. Possible definition of metrics and standards for flexibility can also be used in lieu of cost assumptions for deficiencies. The cost of alternatives has a major impact on what standards are selected. The cheaper it is to meet a standard, the higher the standard (i.e., if the cost of reducing Stage 3 outages is cheaper than the cost of outages to customers, the more resources it makes sense to add.). There is unfortunately not a single solution for electric systems given that the needs of systems are different, and the availability and cost of solutions vary as well. Trade-offs can be subjective. Some decision-makers prefer no deficiencies at any cost while others have a clear high or low cost perspective for addressing deficiencies.

Ability to Test the Model's Unit Commitment and Dispatch

All models assume ideal commitment and dispatch of resources based on variable costs rather than market bids and ignore contractual limitations or resource owners' self-scheduling preferences. On the other hand, due to confidentiality, resources may actually be more or less operationally flexible than represented in the model. It is therefore useful in a system's performance evaluation to be able to test the sensitivity of the idealized commitment and dispatch in the model, and modeling assumptions. For example:

1. Self-scheduling due to contract limitations or a resource owner's scheduling preferences may reduce the operating flexibility available to the system. Comparing actual versus simulated dispatch may help identify possible sensitivities that can be run with different levels of self-scheduling.
2. Use of variable cost versus bids will result in differences in commitment and dispatch of resources, and may distort the price forecast that comes from simulations. Again, additional sensitivities would be useful to test the robustness of the results.

4.3 Recommendations for Future Work

In addition to improvement that the models reviewed in this report are currently undergoing, or that are being planned according to the model descriptions in Appendix A, we have identified the following three main areas where additional work is desired.

Flexibility Metrics and Standards

Traditional reliability metrics such as LOLE and EUE, and standards such as a not to exceed 1-day-in-10-year LOLE, are well understood and generally used in the industry. However, there are no flexibility metrics standards generally accepted in the industry that can guide how much flexibility an electric system should have. As noted in the prior section when discussing the need to test the costs or penalties assumed for resource deficiencies, the models reviewed here use different flexibility metrics and assume different flexibility deficiency costs.[40]

Additional work is needed to answer the question of how much flexibility the CAISO's system CAISO should have, whether new flexibility metrics are needed, and whether and how to determine possible flexibility standards to guide the planning and procurement of resources. Some of the important considerations that can help answer these questions are:

1. Existing performance standards that may establish minimum flexibility standards.
2. The cost of acquiring additional flexibility for the system.
3. The risk aversion or penalty associated with not meeting all flexibility requirements.
4. The availability of tools to perform the necessary analysis to determine desired flexibility standards and requirements.

Some of the models reviewed here provide a framework to address the flexibility metrics and standards questions outlined above; however, additional improvements to these models may be necessary to comprehensively address these questions.

Representation of the Rest of WECC

Modeling of the rest of WECC is a significant challenge in planning studies, not only because of the increased run times of economically committing and dispatching thousands of resources and managing even a greater number of resource and transmission constraints, but also because of the implicit assumption that the system actually operates as modeled. With regards to operating flexibility, this becomes an important assumption because the simulations of the CAISO may show that the system relies heavily on the flexibility of its neighbors to balance loads and resource. Two questions seem important to explore in the future:

1. To what extent can neighboring areas provide reliable operating flexibility to the state or the CAISO?
2. How does the cost of flexibility services available from California's neighbors compare to in-area alternatives?

As noted before, the studies performed so far assume services are available at variable costs, and do not include the fixed cost and/or the premium that may be charged for these services when needed. Also, the simulations ignore any contractual or self-scheduling constraints of resources. At a minimum, additional sensitivities are needed to determine the robustness of the system

[40] The flexibility metrics uses by the models reviewed here are summarized in Table 3.1.

performance evaluations or the evaluation of alternatives to remedy flexibility deficiencies to these assumptions.

Reducing the Run Time of Planning Models to Enable Sensitivity Runs

Performing system evaluations is time consuming with any of the models reviewed in this report. There is always the trade-off between speed and the detail used in representing a system and its operating constraints. Some of the models reviewed require more than one day of computing time to evaluate a single year with a single scenario or set of assumptions. Multiple runs are needed to compare the effectiveness of possible solutions. Therefore, reducing the run time becomes a priority. It may be worth exploring in the future how to reduce the run time of the models without losing valuable information. It is worth noting that all models make simplifications to improve speed; however, choosing which simplifications to make without impacting the results requires testing. The results must be robust enough to make decisions.

Appendix A: Descriptions of Modeling Approaches Considered

The five documents in this appendix were provided by CAISO, SCE, E3, Astrape Consulting and LLNL for purposes of supporting this collaborative model review. The intent of these documents is to provide an accessible yet detailed description of the modeling approaches described in this report. The reader is encouraged to read these documents where additional detail on any of the models described here is desired.

Additional resources used in development of this report that readers may find useful are listed below.

- SCE's presentation of their results from a September 18, 2013 workshop is available at the following link: http://www.cpuc.ca.gov/NR/rdonlyres/8A04F5B8-4990-4089-9E40-B7A064387C67/0/CPUC_ED_SCE_Workshop_StochasticModeling.pdf.
- E3's final presentation of their results from a December 9, 2013 webinar is available at the following link: http://www.caiso.com/Documents/RenewableEnergyFlexibilityResults-Final_2013.pdf.
- E3's preliminary presentation of their results from an August 26, 2013 workshop is available at the following link: http://www.cpuc.ca.gov/NR/rdonlyres/F62DC247-8823-4861-8352-4EAC1341CB73/0/E3_REFLEX.pptx.
- Astrape's report describing their analysis of the 2012 LTPP CAISO Base Scenario and providing detailed results is provided at the following link: http://www.astrape.com/publications/

The California ISO Simulation Model Set for Renewable Integration Study

I. Introduction

The California Independent System Operator Corporation (ISO) developed a simulation model set for its renewable integration study supporting the California Public Utilities Commission (CPUC) Long-Term Procurement Plan (LTPP) proceeding.

The study evaluates the adequacy of capacity and operational flexibility of the fleet for integrating 33% renewable energy in the CA system considering the externalities of all other Balancing Authority Areas of the WECC region. The findings of the study will be used to support the CPUC decision on Bundled Procurement Plans.

The model set mimics the methodologies implemented in the ISO market and operational practices in enforcing operational constraints, including chronological minimum-cost optimization, unit commitment, time-based ramping limitations, minimum up and down time, start-up and shout-down time, random forced outages and chronological planned outages, etc. The model set also considers load and renewable (wind and solar) generation forecast errors and reserve flexible capacity for load following and regulation based on probabilistic assessments of the impacts of the forecast errors.

The model set co-optimizes generation dispatch, ancillary services and load following reserves to achieve minimum cost solutions to meet load, ancillary service and load following requirements simultaneously. The need for additional capacity or flexibility can be identified when load, ancillary service, or load following requirements are not met. The model set can be then used to evaluate effective alternatives to meet the identified need based on various criteria, such as cost, environment, etc.

II. Methodologies

The model set consists of two functional modules. One is a probabilistic model for calculating regulation and load following requirements through Monte Carlo simulations. The other is a deterministic production simulation model that takes the regulation and load following requirements as inputs. It simulates the system operation to identify need for additional capacity or flexibility in the generation resource fleet.

1. Calculation of Regulation and Load Following Requirements

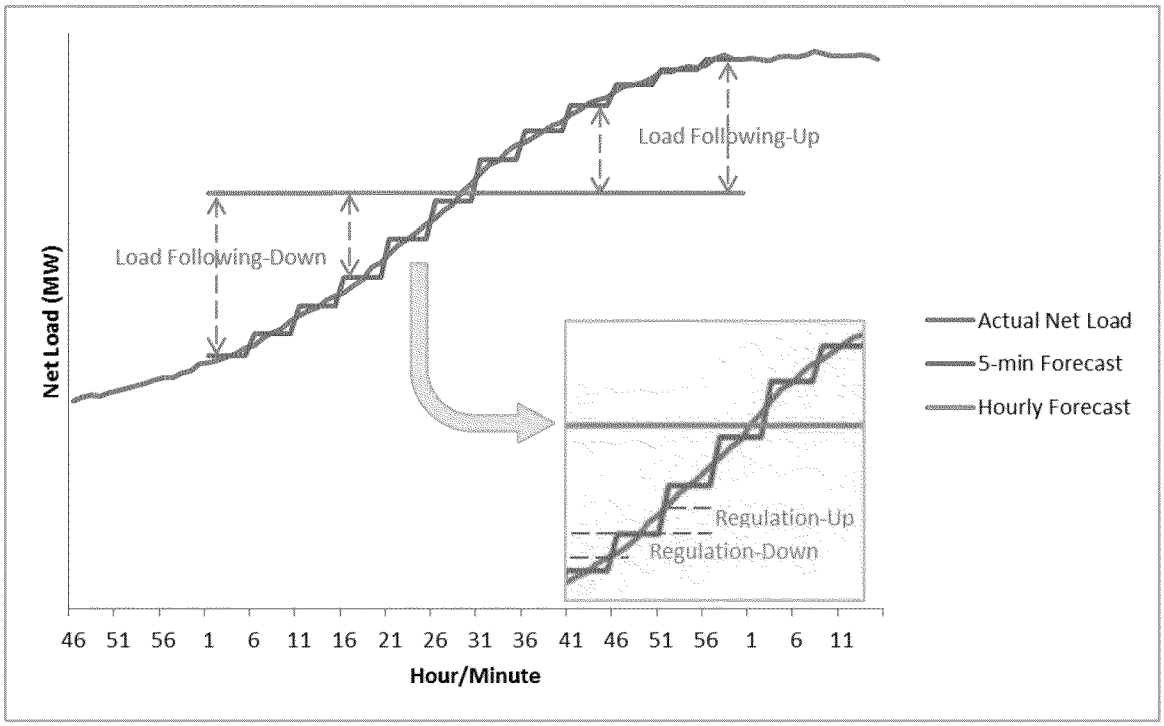
The calculation of regulation and load following requirements uses a probabilistic model developed by Pacific Northwest National Laboratory (PNNL) and the ISO. The calculation is carried out using Monte Carlo simulations.

The model simulates the ISO scheduling process, from hourly (in Day-Ahead and Hour-Ahead market) to 5-min (in Real-Time Dispatch, RTD, market), and to actual operation. In hourly scheduling the

ISO commits and dispatches generation resources economically to meet hourly average net load based on forecast. The hourly schedule should also reserve sufficient upward and downward ramping headroom to be used in RTD dispatch within the hour. In RTD the ISO dispatches generation resources economically to meet the 5-minute average forecasted net load, which are usually different than the hourly forecasted net load. The ramping headroom reserved in the hourly schedule needs to be sufficient to cover the maximum net load difference between 5-minute and hourly forecasts within the hour. The headroom in the model is called load following capacity. In operation the actual net load changes constantly. Its deviation from the 5-minute schedules needs to be balanced using regulation reserve. Regulation requirements should be able to cover the largest deviation for each 5-minute interval.

Figure 1 illustrate the relationship between actual net load, 5-minute forecast, and hourly forecast as well as the concept of regulation and load following requirements.

Figure 1. Load Forecasts and Regulation and Load Following Requirements



In an ideal situation when there is not forecast error, 5-minute forecast is the average of the actual net load of the 5 minutes and hourly forecast is the average of the actual net load of the hour. The load following requirement of each 5-minute interval is the difference between the 5-minute forecast and the hourly forecast. When 5-minute forecast is greater than hourly forecast, load following-up is required. Likewise when 5-minute forecast is less than hourly forecast, load following-down is required. For the hour the maximum of the load following-up requirements of the twelve 5-minute intervals is the hourly load following-up requirement. The maximum of load following-down requirements of the twelve 5-minute intervals is the hourly load following-down requirement. In hourly scheduling headroom equal the hourly load following-up and load following-down requirements need to be reserved to ensure that

there is sufficient online capacity to follow the 5-minute forecasted net load in RTD dispatch. In 5-minute RTD scheduling there is no load following requirement.

Similarly, regulation requirement for each minute is the difference between the actual net load and 5-minute forecast. When actual net load is greater than 5-minute forecast, regulation-up is required. When actual net load is less than 5-minute forecast, regulation-down is required. The regulation requirement of each interval (5-minute or hourly) is the maximum of the of the regulation requirements of all the minutes in the interval. Hourly regulation requirements need to be met in hourly scheduling. In 5-minute scheduling the 5-minute regulation requirements need to be met.

In actual operation there are always forecast errors. The net load forecasts are random variables. The 5-minute and hourly forecasts are not averages of actual net load, but vary independently around the average of actual net load. To calculate the regulation and load following requirements accurately the PNNL probabilistic model is used. The requirements are determined through Monte Carlo simulations using the model.

The model takes 1-minute generated actual load, wind and solar generation profiles of the target year as well as hourly forecast standard deviations of load, wind and solar generation, and RTD load forecast standard deviation as inputs.[1] It assumes that these forecast errors have truncated normal distributions.[2] 5-minute wind and solar generation forecasts use persistence method. That is[3]

$$\bar{\quad} \quad \tilde{} \ (\quad)$$

$$\bar{\quad} \quad \tilde{} \ (\quad)$$

$$\bar{\quad} \quad \tilde{} \ (\quad)$$

$$\bar{\quad} \quad \tilde{} \ (\quad)$$

Where,

- load forecast for hour

$\bar{\quad}$ - average 1-minute actual load of hour

$\tilde{}$ ( ) – hourly load forecast error represented by a truncated normal distribution with zero mean value and hourly load forecast standard deviation

- 5-minute load forecast for interval starting at minute

[1] Forecast standard deviations of load, wind and solar generation are derived from historical data.
[2] The tails beyond historical maximum and minimum forecast errors are truncated to avoid unrealistic outcomes.
[3] These formulas are simplified for discussion of the concepts. The actual formulas are described in the draft "Technical Appendices for California ISO Renewable Integration Studies" at http://www.caiso.com/282d/282d85c9391b0.pdf.

– - average 1-minute actual load of 5-minute interval starting at minute

~ ( ) – 5-minute load forecast error represented by a truncated normal distribution with zero mean value and 5-minute load forecast standard deviation

- 5-minute wind generation forecast for interval starting at minute

- actual 1-minute wind generation at minute [4]

- 5-minute solar generation forecast for interval starting at minute

- actual 1-minute solar generation at minute

Then,

( )

Where,

– load following requirement of the 5-minute interval starting at minute in hour

– regulation requirement of minute in the 5-minute interval starting at minute

In Monte Carlo simulations the forecast errors were generated randomly in 100 iterations for the whole target year. For each hour, 1,200 5-minute load following requirements of the 100 iterations (12 requirements each iteration) are generated. The 1,200 requirements form a load following requirement distribution. The 2.5th percentile value of the distribution is defined as load following-down requirement and 97.5th percentile value of the distribution is defined as the load following-up requirement of the hour, as shown in Figure 2.

Based on the definitions, when load following-up requirement is just met, there is still a 2.5% probability that there is insufficient upward headroom reserved in hourly schedule to meet the 5-minute forecasted net load within the hour. Also when load following-down is just met, there is a 2.5% probability that dispatch will be greater than 5-minute forecasted net load (over-generation). This is considered as acceptable risk. However, when one or both of the load following requirements are not met the risk of unmet demand or over-generation will be greater and may not be acceptable.[5] When that happens while there is still available capacity not fully utilized, it reflects lack of flexibility of the generation fleet. If there is no more capacity available when the load following-up requirement is not met, it is lack of capacity.[6] In both cases the maximum and expected MWh of intra-hour load not served or over-generation can be calculated based on the 100-iteration Monte Carlo simulation results.

---

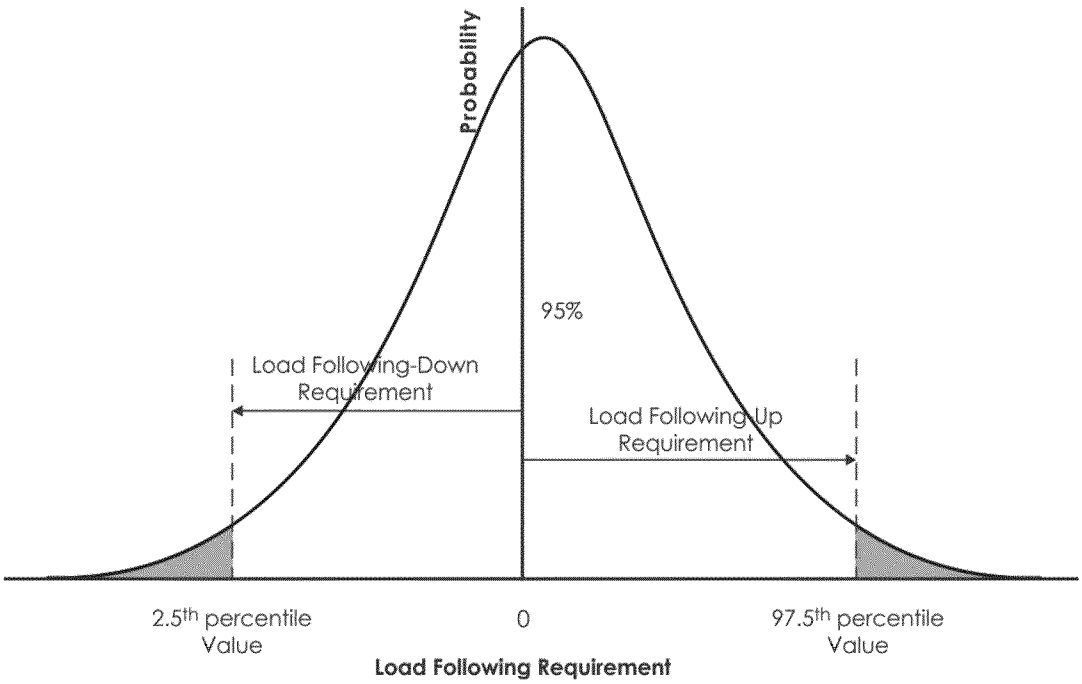
[4] The latest forecast is done at minute .

[5] There is no universal standard for this risk yet. It should be established in the proposed development of a criterion metrics for measuring stochastic modeling assumptions and results.

[6] See section III.-2-3) below about priority orders of the requirements. Insufficient capacity or flexibility in supply will cause shortage to meet load-following requirement first.

Similarly the 6,000 1-minute regulation requirements of the 100 iteration in each hour (60 requirements each iteration) form a regulation requirement distribution function. The 2.5th percentile value of the distribution is defined as regulation-down requirement and 97.5th percentile value of the distribution is defined as the regulation-up requirement of the hour. Regulation requirements can also be calculated in sub-hour (e.g., 5-minute) intervals.

Figure 2. Determination of Load Following Requirements



Regulation and load following requirements calculated in this step using the PNNL probabilistic model are inputs of the production simulation model that is discussed in the next section.

2. Production Simulation Model

The production simulation model is a deterministic zonal model developed using Plexos software. The model mimics the methodologies implemented in the ISO market and operation practices, specifically in the following aspects.

1) Optimization

The model uses Mixed-Integer Linear Programing (MIP) optimization for unit commitment and dispatch. The simulation runs chronologically to co-optimize energy dispatch, ancillary services and load following provision. The outcome of the co-optimization is a least-cost solution that meets load, ancillary service and load following requirements simultaneously. When there is insufficient capacity or flexibility to meet the requirements, the shortage is captured and reported. The chronological simulation can run in hourly or sub-hourly intervals.[7]

[7] 5-min chronological simulation was conducted using the model in the study for the CPUC 2010 LTPP proceeding.

2) Operational constraints

To ensure that the resource availability and flexibility simulated in the models can be achieved when the resources bid into the ISO market, the model enforces operational constraints similar to those enforced in the ISO market and operation practices, including

- Unit commitment and dispatch
  - The model uses MIP optimization for unit commitment in the hourly scheduling process. It does not allow partial unit commitment.
  - Commitment and dispatch decisions are made based on demand (net load, ancillary service and load following requirements), costs (start-up, VOM, fuel, and emission) and availabilities (including outage, ramping, fuel or emission limits) of generation resources, transmission limits, etc.
- Ramping limitation
  - In hourly simulation inter-hour energy ramp has a 60 minutes ramping time; load following has a 20 minutes ramping time, and ancillary services have a 10 minutes ramping time.
  - Each dispatchable generation resource is subject to a ramp rate limit between minimum and maximum capacity. In upward direction, its total provision of ancillary services cannot exceed its 10-minute ramping capability and unused capacity; total provision of ancillary services and load following cannot exceed its 20-minute ramping capability and unused capacity; and the sum of energy ramping and provision of ancillary services and load following cannot exceed its 60-minute ramping capability and unused capacity. In downward direction the dispatch above its minimum capacity limits the resource's provision of regulation-down and load following-down.
- Minimum up and down time, start-up and shut-down time
  - A generation resource may take several hours to ramp from 0 MW to minimum capacity. Before reaching minimum capacity the generation resource cannot be dispatched or provide ancillary service or load following. Similarly when a generation resource is in the shutting down process from minimum capacity it cannot be dispatched or provide ancillary service or load following.
  - Once committed, a generation resource may have to stay on for certain hours before it can be shut down. Once it is shut down the resource may not be available for commitment until certain hours later.
- Random forced outage vs. planned maintenance outage with monthly weights
  - Forced outages of each generation resource are generated randomly using uniform distribution function and the forced outage rate of the resource. The ratio of generated outage time of the year matches with the forced outage rate of the resource.
  - Maintenance outage rate is allocated to each month of the year based on maintenance outage allocation factors (a set of weights). The allocation factors are derived from the ISO historical monthly maintenance outage pattern. The allocation factors for CA gas-firing

resources are adjustments to reflect the high net load ramping need in the spring and fall months (reduced in February-May, October, and increased in January, November, and December). The summer months have lower maintenance outage rates and winter months have higher rates. In each month the planned outages are generated with consideration of load, availabilities of other resources, etc. The ratio of generated maintenance outage time of the year matches with the maintenance outage rate of each generation resource.

- In generating forced and maintenance outages, the minimum time to repair (another input that is usually longer than minimum down time) of each generation resource is enforced.

3) Others

The following methodologies implemented in the production simulation model were developed through the discussion with the Advisory Team and stakeholders of the CPUC 2010 and 2012 LTPP proceedings.

- Zonal configuration
  - The model has a zonal configuration. The transmission limits between zones are enforced. Transmission limits within the zones are not enforced.
  - The transfer capabilities between any two adjacent zones reflect the maximum simultaneous transfer capabilities between the two zones.
  - On each transmission path there is a wheeling charge for each direction. It reflects the Transmission Access Charge and transmission loss of energy (in financial term).
  - Transmission loss quantity (MWh) is not modeled explicitly, but assumed to be included in the load forecast.
- CA import and export
  - Besides the transmission limit on each path, a total CA maximum simultaneous import limit is enforced.
  - 70% of out-of-state RPS renewable generation is must-take import into CA. Some CA parties have ownership of certain out-of-state generation resources. The corresponding portion of generation of these resources is also must-take import. The must-take imports use the CA import capability.
  - Ancillary services and load following provided by out-of-state generation resources also use CA import capability.
  - Export from CA is only subject to the transmission limits of the export paths.
- Local transmission constraints
  - Southern California Import Transmission (SCIT) nomogram constraint that limits total simultaneous import into Sothern CA is enforced in the model.
- Load
  - Each zone has its own hourly chronological load profile.
- Renewable generation

- Renewable generation has fixed hourly generation profiles that are developed based on capacity, location, weather, and historical generation data.
- Renewable generation is not curtailable.
- Non-RPS distributed PV generation is modeled as supply side generation resource. Load forecast is adjusted upward by the same amount of MW and GWh to account for the forecast of distributed PV generation.

- Hydro generation
  - Hydro generation is aggregated and modeled as two types: run-of-river and dispatchable.
  - Run-of-rive hydro has a fixed generation profile derived from historical data.
  - Dispatchable hydro has monthly energy limits, which are also derived from historical data. Disptachable hydro can provide ancillary services and load following.

- Pumped storage scheduling
  - The schedules of pumping and generation are optimized subject to storage capacity, inflow and target limits, and cycling efficiency, etc.

- Demand response
  - Demand response has event based and non-event based two types.
  - Non-event based demand response is modeled as load forecast (peak load and energy) reduction.
  - Event based demand response is modeled as supply resources that can be triggered by energy price.
  - Event-based demand response resources are not dispatchable (can be either on or off, but cannot ramp). Some can provide non-spinning reserve.

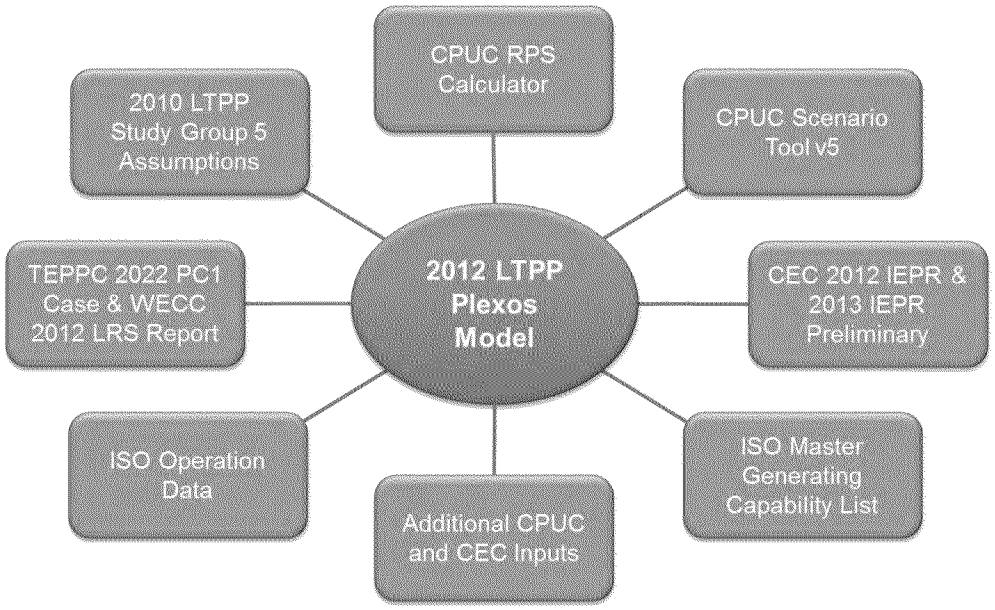
- CO_2 emission cost
  - Each CA generation resource has CO_2 emission cost included in its generation variable cost
  - Generation resources outside CA do not have CO_2 emission cost in their variable cost. The emission cost is added to wheeling charges for imports into CA. The emission cost wheeling adder on path from BPA to CA is only 20% of that on other paths.[8]

III. Input Assumptions and Data Sources

The model was last updated for the CPUC 2012 LTPP proceeding. Most of the assumptions are either defined in the CPUC LTPP scenario definitions or developed through the discussion with the LTPP proceeding stakeholders. The model also uses assumptions and data from other sources, as shown in Figure 3.

[8] Consistent with CARB rule (http://www.arb.ca.gov/regact/2010/ghg2010/ghgisoratta.pdf).

Figure 3 Data Sources of the Model



1. Model base data

The model covers the whole WECC in a zonal configuration. It has 25 zones (8 in CA) and more than 1,500 generation resources. The model was updated based on the TEPPC 2022 PC1 Case, with updates according to the WECC Loads and Resources Subcommittee (LRS) 2012 Power Supply Assessment report.[9] The model uses the TEPPC/WECC data for

- Transmission path ratings and wheeling charges;
- Fuel prices, except natural gas prices; and
- Load forecasts and profiles, generation resources and characteristics in zones outside CA.

2. Other assumptions and data

1) CA generation resources

- CPUC Scenario Tool
- The ISO "Master Control Area Generating Capability List" (Nov 26, 2012, http://www.caiso.com/participate/Pages/Generation/Default.aspx)
- SONGS nuclear power plant is retired.

2) CA load

- Load forecasts and adjustments (energy efficiency, CHP, non-event based demand response, distributed generation, etc.) are from the CEC 2012 IEPR forecast, 2013 IEPR preliminary forecast, and CPUC Scenario Tool. Load forecast is adjusted down by energy efficiency, CHP, and

[9] These are the same sources for the ISO transmission planning economic study model.

non-event based demand response. Distributed PV generation is modeled as supply resource. Load forecast is adjusted up accordingly to account for that.

- Some scenarios use 1-in-2 peak load forecast, while the others use 1-in-5 peak load forecast.
- Load profiles – developed based on load forecast and 2005 historical load shape.
- Pump load - average of 2009-2011 actual profiles, scaled to match with 12,530 GWh energy forecast in the CPUC Scenario Tool.

3) Ancillary service and Load Following Requirements

- In CA - spinning and non-spinning requirements are 3% of load each; regulation and load-following requirements are from PNNL probabilistic model calculation and are added to the model as exogenous requirements, which are scenario specific.
- Outside CA – spinning and non-spinning requirements are 3% of load each; regulation up and down requirements are 1% of load each. Load following requirements were developed based on discussion with the Advisory Team Study Group 5 in 2010 LTPP proceeding.
- The priority orders from high to low in the simulations are: upward – load, regulation-up, spinning, non-spinning, load following-up; downward – load, regulation-down, load following-down. Higher priority requirements are met before lower priority requirements when there is insufficient capacity or flexibility in supply.
- There is no reserve sharing across Balancing Authority Areas.

4) CA hydro generation

- Run-of-river hydro uses 2005 actual hydro generation profiles.
- Dispatchable hydro has 2005 actual monthly energy as limits. The dispatch is optimized subject to the monthly energy limits.

5) CA renewable generation[10]

- RPS renewable portfolios – CPUC RPS Calculator
- Wind and solar generation uses hourly profiles. Other RPS renewable resources have fixed dispatch with energy limits.
- The wind and solar generation profiles are developed based on 2005 weather data, location, installed capacity, and energy requirements.
- Solar thermal with storage (the Rice Project) has an hourly energy profile as source, a storage device, and a steam turbine. The turbine is dispatchable. It is assumed to have a solar multiple equal 1 and 6 hours of storage.

6) CA event based demand response

- The MW and availability is from the CA utility 2011 Load Impact Ex Ante Reports.
- There are two types of demand response programs. One has a triggering price $600/MWh and is available from hour 14 to 21 with monthly energy limits. The other has a triggering price $1,000/MWh and is available from hour 14 to 18 without energy limit.

[10] Some CA RPS renewable resources are located outside CA.

7) Outage rates

- Forced and maintenance outage rates by technology are calculated based on the ISO 2006-2010 operation data.
- The outage rates are applied to all applicable generation resources in the WECC.

8) Ramp rates

- Average ramp rates by technology and by size are calculated based on the ISO Master File data.
- The ramp rates are applied to all applicable generation resources in the WECC.

9) Natural gas price forecast

- WECC-wide natural gas forecast is from the CEC staff forecast.

10) CO_2 emission cost forecast

- 2013 IEPR preliminary forecast

11) CA and SCIT import limits

- CA total import limit and the SCIT limits are calculated using a tool developed by the ISO based on the SCIT nomogram limit.[11]

12) Forecast errors in the PNNL model calculation

- The ISO 2012 actual and forecast errors are used in the calculation of regulation and load following requirements with the PNNL model.

IV. Simulations

The production simulation model was used for the CPUC 2012 LTPP study. The simulations were run on hourly interval chronologically for the whole year of 2022. Each optimization has a horizon of 24 hours.[12] The end conditions of one optimization (one day) are used as initial condition of the next optimization (the next day) so that the operational constraints and maintenance outage schedules can be enforced correctly.

The model simulation generates various results at different granularity level. It can report generation, ancillary service and load following provision, import, export, resource utilization, etc. in annual or monthly total or average. It can also report hourly results such as energy price of each zone, ancillary service and load following prices by product, load and dispatch of generation resources, ancillary service and load following requirements and provision by individual resources, etc.

The LTPP study focuses on if there is sufficient capacity and flexibility in the fleet to integrate 33% renewable energy in the CA system. It also looks at the impacts of 33% renewable energy on the system,

[11] Prior study work also applied at 60/40 import to generation requirements for SCE area and a 75/25 for SDGE. Going forward these requirements are no longer applied as a result of changes to the ISO operating practice.

[12] The interval length and optimization horizon can be set differently. In 2010 LTPP study the model was configured to run 5-minute interval with an optimization horizon of 3 days.

such as how differently the conventional resources will operate, how production cost will change, and how much CO_2 emission will be reduced. The model produced all the results we need.

For sufficiency of capacity and flexibility, we examine if load, ancillary service and load following requirements are all met. If load following or ancillary service requirements are not met while there is still capacity available (not in outage mode), it indicates lack of flexibility. If there is no capacity available it is lack of capacity.

When there is a shortage of flexibility or capacity, various types or combinations of types of new resources can be added to the model. Re-running the simulation can tell us how much of each type or combination of resources will need to address the shortage. Based on that the effective solutions can be identified based certain criteria, such as cost, emission, etc.

V. Sample Results

The following are some preliminary results from the 2012 LTPP study.

Figure 4 Maximum Upward Ancillary Service and Load Following Shortage

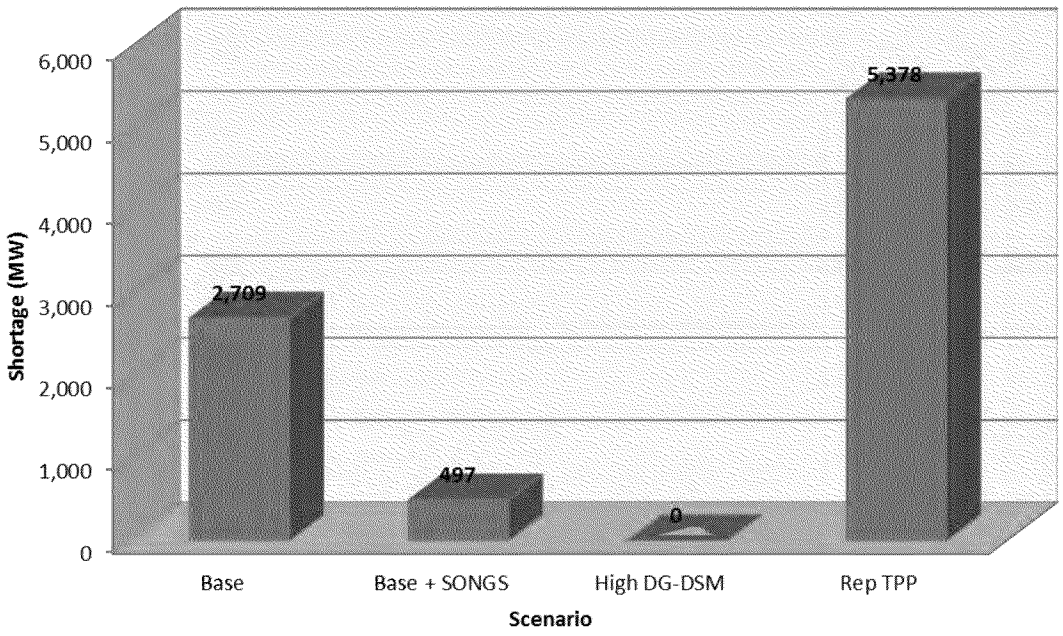


Figure 5 Number of Hours of Upward Shortage

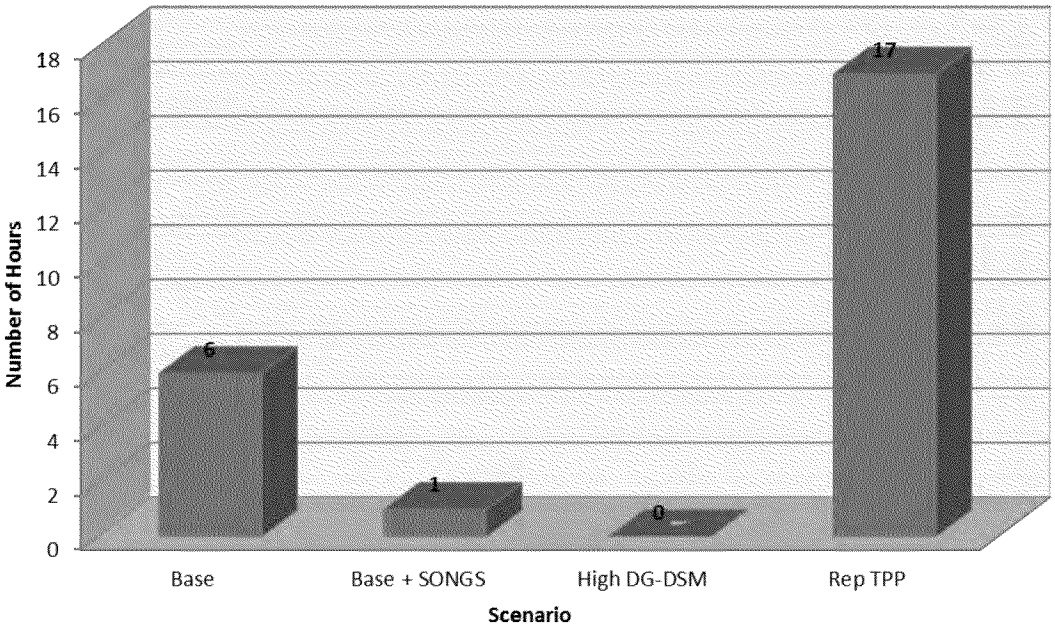


Figure 6 CO_2 Emission Attributed to Meeting CA Load

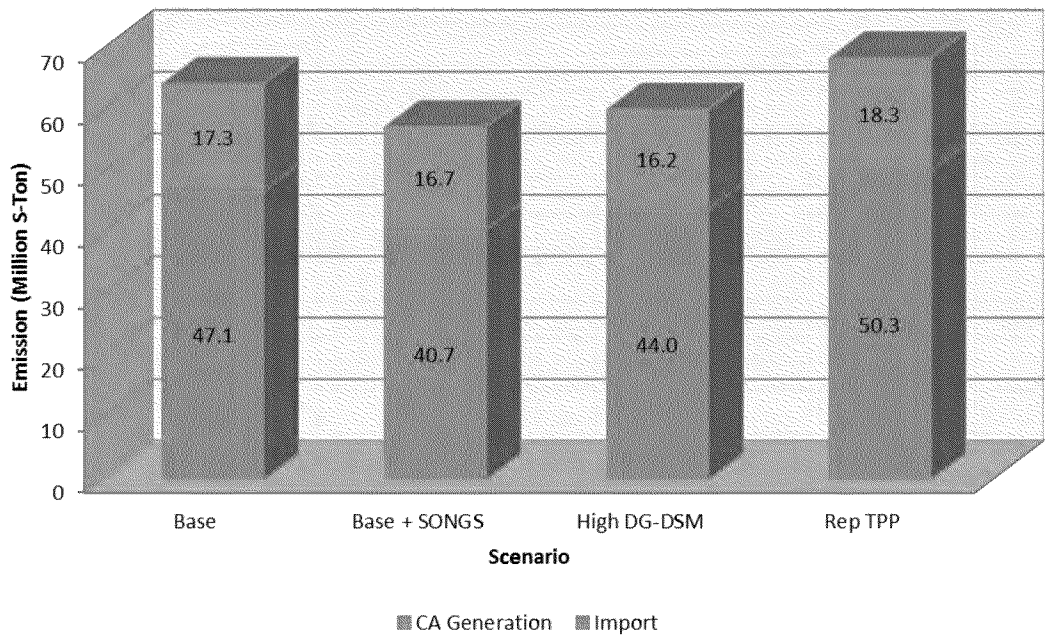


Figure 7 Production Cost Attributed to Meeting CA Load

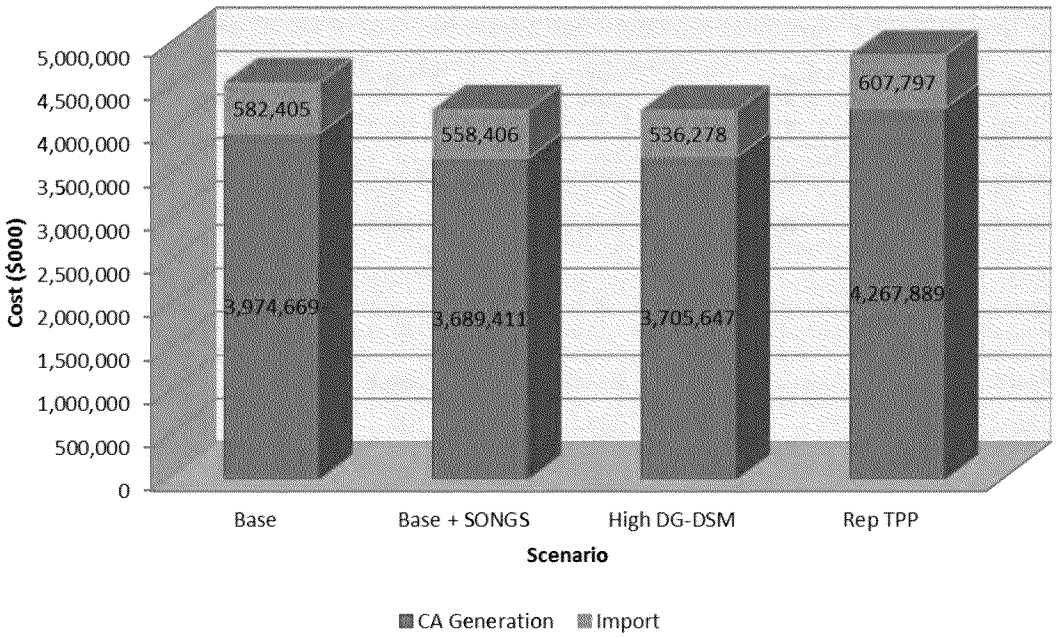


Figure 8 Histogram of CA Net Import

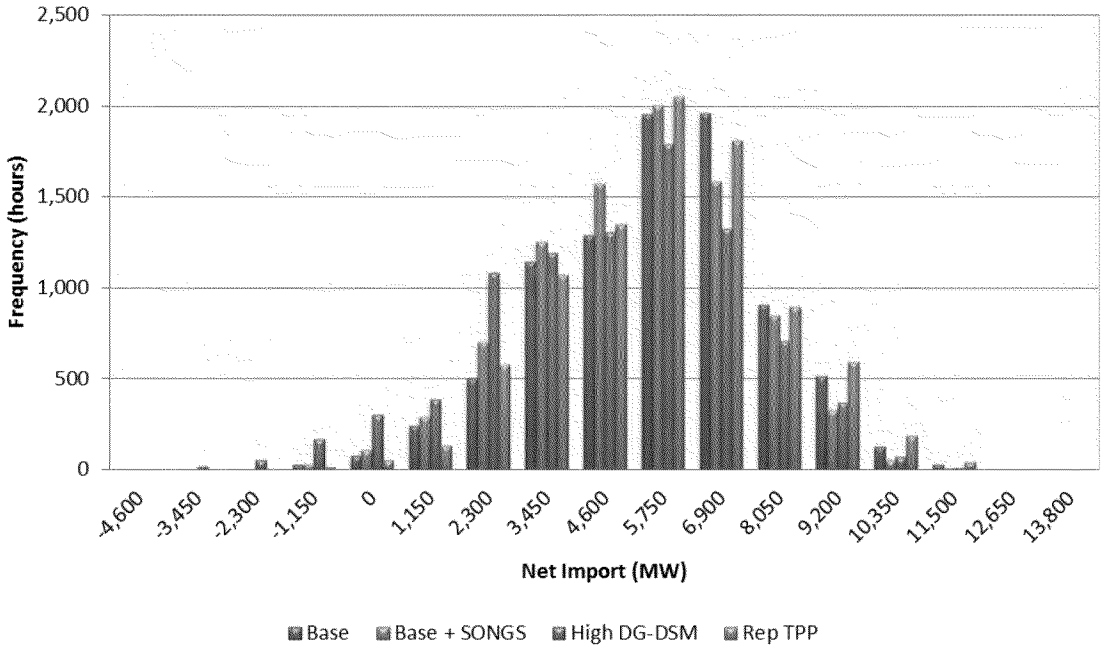


Figure 9 ISO Energy Balance on 03/26/2022 – High DG/DSM Scenario

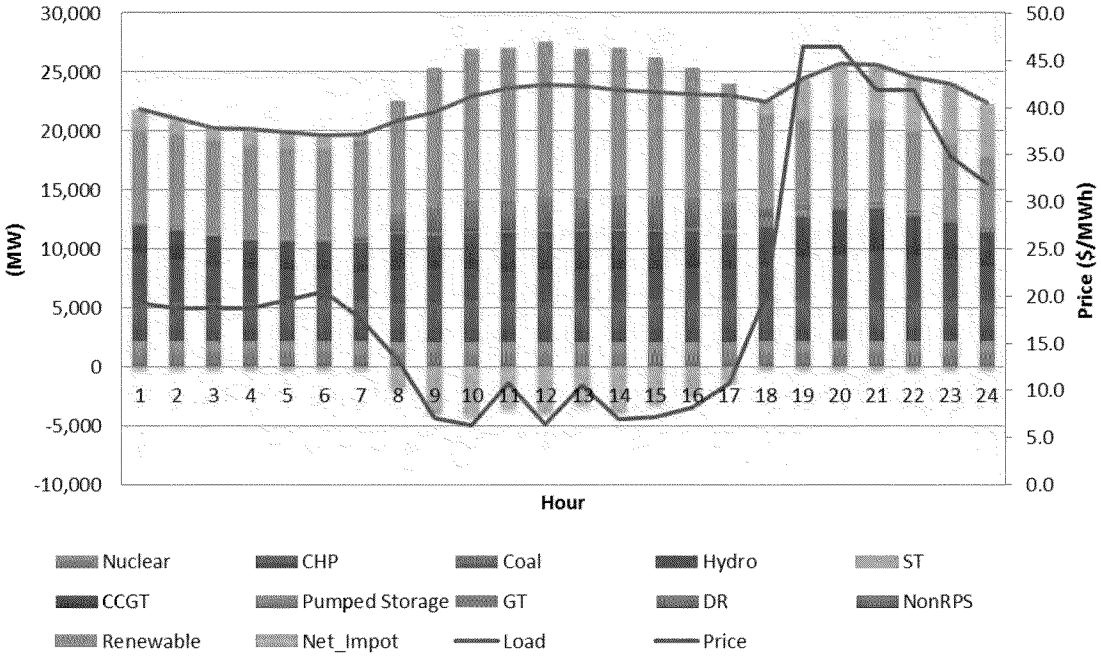


Figure 10 ISO Energy Balance on 07/22/2022 – Rep TPP Scenario

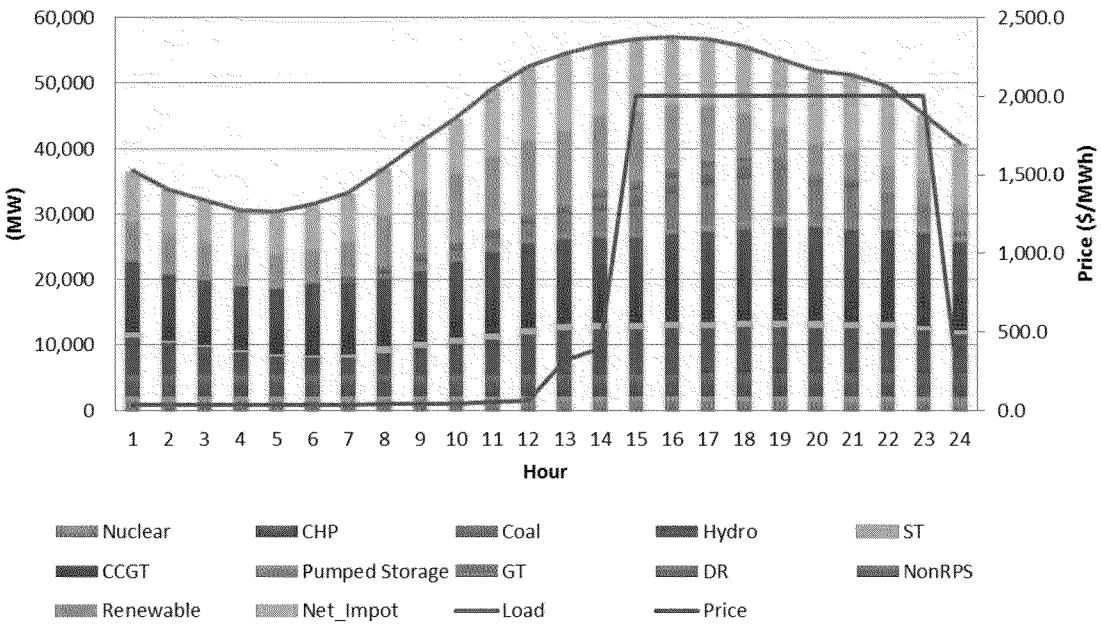
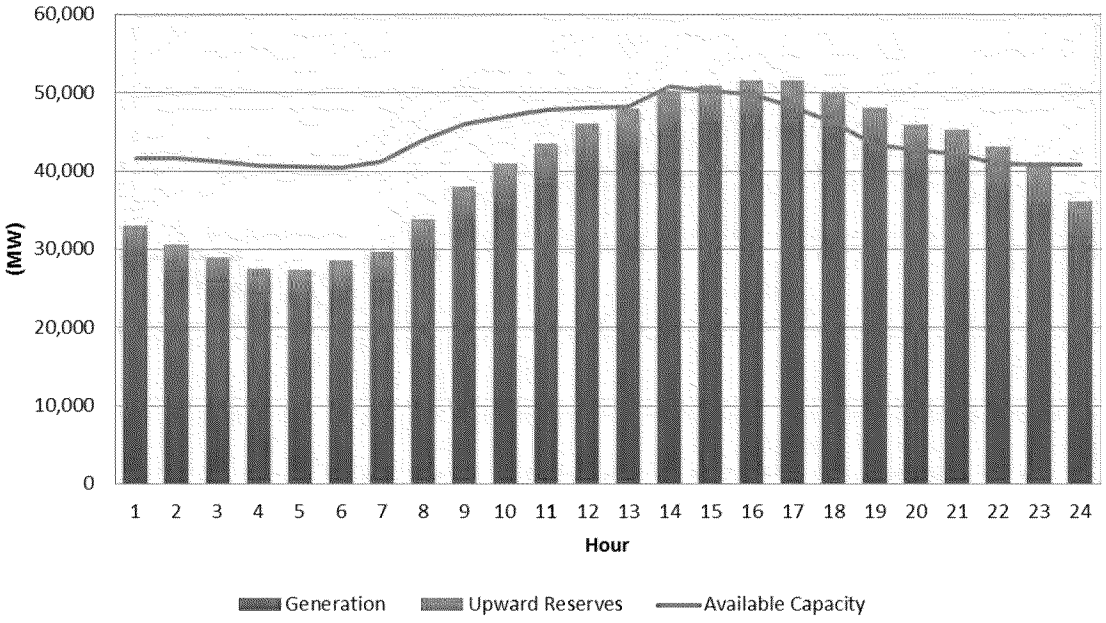


Figure 11 ISO Available Capacity Usage on 07/22/2022 – Rep TPP Scenario



VI. Future Improvements

In the upcoming CPUC 2014 LTPP study, some improvements to the models are envisioned. At this time the following four are put on the list. More may be added later as we start the discussion with stakeholders.

1) Economic curtailment of renewable generation

- The implementation of FERC 764 Order encourages renewable resources to submit economic bids in the ISO market. It will provide more flexibility for operation. It will also make economic curtailment of renewable generation more likely. The curtailment needs to be reflected in the production simulation model.
- The curtailment should be price triggered. By 2024 the ISO should have implemented a bid price floor of -$300/MWh. The triggering price of renewable generation curtailment should be in between -$250 and -$300/MWh.
- The economic curtailment should occur before load following-down or regulation-down requirement being cut short.
- The economic curtailment should be limited to only a portion of renewable generation by new projects of specific technologies (e.g., wind and solar).
- A follow-up analysis of the impacts of the curtailment needs to be conducted after simulations are done. The analysis needs to evaluate the policy impact of the curtailment, the cost to build more renewable resources to meet RPS goal, and alternative solutions to reduce or avoid curtailment.

2) Load shifts responding to time-of-use energy price

- Time-of-use (TOU) energy price is likely to be implemented before 2024. The model should reflect load shift responding to the TOU price.
- The shift should be optimized based on energy price.
- The load can be shifted only within the same day.
- Only a portion of load (e.g., a portion of residential) can be shifted.

3) Frequency response requirement

- Incorporate a frequency response requirement constraint to ensure that there is sufficient frequency responsive headroom to account for the ISO allocated portion of the interconnection frequency response requirements.

4) A stochastic production simulation model

- The stochastic production simulation model will be similar to the deterministic model, but has scope reduced to CA only.
- Focus will be on development of chronological stochastic variables.
- Simulations will be chronological in hourly and sub-hourly intervals.

Intermittent Resource Integration

Development and Application of a Stochastic Framework for System Planning Analysis

January 17, 2014

ABOUT THIS PAPER

This paper contains an overview of the analysis performed by Southern California Edison Company (SCE) for the renewable integration track (Track 2) of the 2012 Long Term Procurement Proceeding (LTPP) at the California Public Utilities Commission (CPUC). Following the cancellation of Track 2, SCE desires to continue discussion surrounding Track 2 for preparation of future work associated with the need to integrate increasing amounts of renewable generation into the California electric grid. This paper includes work that is part of SCE's ongoing analysis. Even through the Track 2 of the 2012 LTPP was cancelled, analysis associated with this paper, including assumptions, methodology and results, is subject to revision.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ANALYSIS TEAM AND CO NTACT INFORMATION

ANALYSIS TEAM

Martin Blagaich

Vidhi Chawla

Erin Childs

Tomislav Galjanic

Justin Kubassek

Megan Mao

Paul Nelson

Eric Wang

A special acknowledgement is due to Carl Silsbee who initially identified the need to undertake the work presented in this paper. Carl's guidance and advice to the team during the work effort has been critical to the development of SCE's methodology.

SPECIAL THANKS

The methodology presented in this paper would not be possible without the help of many people at SCE who helped manage and advance the project, including:

Kevin Duggan

Andrea Horwatt

Lujuanna Medina

Rebecca Meiers-De Pastino

Carol Schmid-Frazee

Alan Wong

In addition, SCE would like to thank the California Public Utilities Commission who helped set up the multiple workshops that allowed SCE to get valued feedback from parties throughout California. SCE would also like to thank the California Independent System Operator (CAISO) for their help and support in developing inputs and assumptions that were used in SCE's analysis.

CONTACT INFORMATION

For any questions or comments, please contact Megan Mao at Megan.Mao@sce.com or Martin Blagaich at Martin.Blagaich@sce.com

EXECUTIVE SUMMARY

This report outlines Southern California Edison's (SCE's) framework and stochastic methodology for the analysis it performed for the renewable integration track (Track 2) of the California Public Utilities Commission's ("Commission's) 2012 Long-Term Procurement Proceeding (LTPP).[1] On September 16, 2013, before testimony was submitted, the Commission cancelled Track 2 of the proceeding and deferred the Track 2 issues to the 2014 LTPP. In the meantime, to continue the dialogue on the important Track 2 issue of what additional procurement, if any, is needed to meet need associated with integration of increasing amounts of intermittent renewable generation resources, SCE is making its work public. SCE's work was enhanced by input it received at two CPUC sponsored workshops[2] to vet SCE's analysis and methodology with the parties to the LTPP, as well as through direct communications with several parties.

STUDY QUESTION

The investor owned utilities (IOUs) are on target to provide 33% of their energy from renewable energy sources by 2020, a large portion of which is expected to come from solar photovoltaic and wind energy. In the past, when the primary source of generation was conventional, the primary variability in planning the power system was the amount of load on the system at any particular time. Unlike conventional resources, however, these renewable sources of energy are inherently intermittent and uncertain.

Conventional generation is more predictable and controllable than solar and wind generation and can be dispatched to meet load variations in a relatively predictable way. The intermittent characteristics of certain types of renewable generation thus present new system planning challenges.

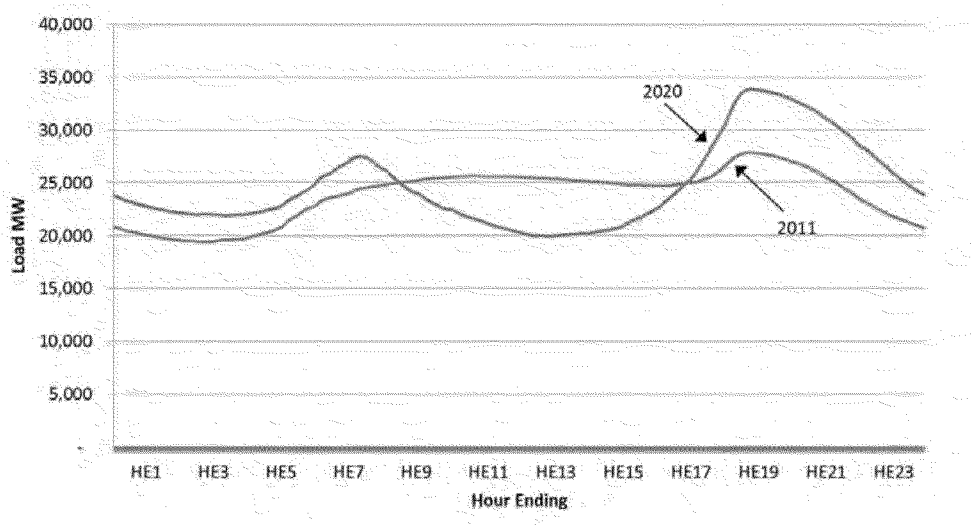
---

[1] Every two years, the CPUC holds a Long-Term Procurement Plan proceeding to review and adopt the IOUs' ten-year procurement plans. The LTPP proceeding evaluates the utilities' need for new fossil-fired resources and establishes rules for rate recovery of procurement transactions." (http://www.cpuc.ca.gov/PUC/energy/Procurement/LTPP/)

[2] Two workshops were held by the California Public Utilities Commission, during which SCE presented its methodology and results outlined in this paper. The workshop details can be found at: http://www.cpuc.ca.gov/PUC/energy/Procurement/LTPP/ltpp_history.htm

To analyze the impact of additional intermittent resources on the system, SCE created a daily net load by subtracting wind and solar generation from load on an hourly basis. This net illuminates the amount of load that must be satisfied by other types of generation. Because there is a daily pattern associated with intermittent resources, they have the potential to place new stresses on California's dispatchable generation, such as a steeper ramp in a shorter timeframe during the evening hours, a problem that is illustrated below in Figure 1.

Figure 1: CAISO Daily Average Net Load Shape, March 2011 and March 2020[3]



In this example, dispatchable resources could be required to increase production drastically in the evening hours, causing concern that the current dispatchable generation fleet may not be able to meet load demand. In addition to the issue illustrated above, the minute-by-minute variability of intermittent generation adds to the natural variability of load. As a result, potential swings in 5-minute net load may be too drastic for the existing generation fleet to meet. In both cases, the high penetration of intermittent generation places new pressures on California's dispatchable generation. These new pressures and the uncertain ability to meet future load demand with the new portfolio mix motivated SCE to conduct the renewable integration analysis presented in this report.

The question that SCE addresses in its analysis is "What additional resources and characteristic types are needed to satisfy system reliability standards in the face of increasing amounts of intermittent renewable generation?" SCE chose reliability

[3] 2011 Historical vs. the 2020 Environmental Case assumptions from the 2010 LTPP

standards as the metric for the analysis because they represent an acceptable economic tradeoff between construction of new generation and possible system outages. The specific reliability standards that SCE used in its study are discussed later in this paper.

STUDY OVERVIEW

SCE adopted annual loss of load expectation (LOLE) as the system reliability metric. SCE estimated future LOLE for the California electricity system and compared this to the "1 event in 10 year" standard. This standard means that generation resources should be built such that only one event of system outage is expected in a 10 year period. This standard is used throughout the energy industry, and represents the reasonable tradeoff between the costs of new generation and frequency of system outages.

SCE also used a stochastic study. This method is able to adequately address the system resource need when substantial amounts of intermittent resources are used to meet load. A stochastic study is one in which certain inputs are varied to create a pool of potential outcomes, each of which may have a different likelihood of occurring. The benefit of a stochastic study is that it allows many possible conditions to be analyzed. Each condition can then be taken into consideration to find the total likelihood of a system outage.

SCE founded its decision to perform analysis in a stochastic manner on two main principles. First, many of the inputs that are large drivers of potential system need are inherently uncertain. For example, load, wind generation, and solar generation are known to vary in ways that cannot always be predicted. Stochastic analysis allows the system planner to determine a level of need that reflects a broad range of possibilities for the stochastic input variables. Second, in order to determine the expected, or most likely, outcome, it is necessary to consider the potential outcomes of a number of possible scenarios. These scenarios can then be weighted based on their likelihood of occurring to give an overall picture of what is most likely to occur. So, without stochastic analysis, the study would not be able to accurately identify whether reliability standards would be met.

To perform the analysis, SCE created key stochastic inputs and ran them through a generator dispatch production simulation model, developed on a software tool called

PLEXOS[4]. Specifically, SCE used PLEXOS to simulate many potential days that might occur in 2022. SCE then used the production simulation results to calculate the system outages that occurred in each of the different days and seasons. SCE combined the results to give a total expected loss of load probability, which represents the likelihood of a system outage (or loss of load) occurring in 2022.

The conclusion of the SCE analysis was that the CAISO system had no need for additional resources in 2022 for the purposes of system reliability at this time. Specifically, SCE's analysis showed that if all generation resources expected to be available in 2022 are included in the analysis, less than one event was expected to occur in 10 years. This falls below the "1 in 10" standard described above. SCE concluded that no additional generation would be required for renewable integration at this time.

CONCLUSIONS AND NEXT STEPS

Although the analysis performed by SCE was not officially submitted as testimony in the LTPP process, it represents a significant advancement in system need analysis. SCE hopes that its framework and methodology will create a foundation for future stochastic analyses, which should become the standard for future system planning studies.

However, this work is only the beginning of an ongoing process to understand the potential system needs caused by increased intermittent generation. The analysis performed by SCE in anticipation of the Track 2 of the LTPP was focused on answering a very specific question. Because of that focus, the analysis left many other questions still unanswered. Among the unanswered questions are the potential issues associated with over-generation and the diversity of policy solutions that might be called upon to address over-generation.

With many of these types of questions in mind, SCE is continuing to study the potential impacts of intermittent generation. As a part of these efforts SCE proposes to collaborate with many other interested parties in California to develop solutions to these challenges. In the spirit of these collaborative efforts, SCE welcomes comments, questions or other feedback to any of the work in this paper, in the hopes of continuous improvement of SCE's study design and analysis.

[4] PLEXOS is produced by Energy Exemplar. Additional information can be found on their website (http://www.energyexemplar.com).

1.0 BACKGROUND

California's 33 % renewable portfolio standard (RPS) has created new challenges for resource planners and regulators charged with the responsibility for developing planning methodologies that more accurately reflect the impact of intermittent renewables on the power system.

One of the main concerns with increased renewable penetration is whether additional flexible resources will be needed to provide the level of "ramping" needed to reliably integrate renewables onto the grid. Specifically, increased levels of intermittent renewable generation resources in the California Independent System Operator (CAISO) and other parts of the Western Electricity Coordinating Council (WECC) interconnect have raised concerns regarding the adequacy of existing and planned generating resources to adjust their output level (i.e., "ramp") quickly enough to follow the variations in net load (customer load requirements less intermittent renewable resource production). Insufficient "ramping" capability can lead to imbalances in generation and net load, leading to firm load interruptions (outages) as a means of system control. Flexible generation, distinct from must-take or inflexible generation, is required in order to meet these ramping needs.

In the past, reliability studies in California have assessed the adequacy of generation resource fleet to meet peak load on an unexpectedly hot day when customer loads are high. These studies have not considered the flexibility needs of the system as the conventional generation resources that have traditionally provided most of the power source are typically flexible in operation and could be assumed to easily meet system flexibility needs. However, the presence of large amounts of solar generation is expected to lead to substantial increases in daily ramping requirements in late afternoon hours, as lighting load increases just as solar output diminishes, as seen in Figure 1. This phenomenon occurs principally during winter and springtime conditions, when overall loads are relatively low and conventional generation resources are often shut down, both due to lack of demand and the high amount of solar generation. In summertime conditions, solar generation is partly coincident with peak loads that are driven by air conditioning demand, and the influence of high penetrations of renewable power is not as dramatic. However, additional solar generation is shifting the net peak to a time later in the afternoon as solar output diminishes, potentially creating other operating problems. These changes in how existing flexible generators must be dispatched to meet net load raise concerns as to

whether additional flexible resources will be required to meet system flexible ramping needs.

2.0 FRAMEWORK

2.1 SCE'S ANALYSIS UTILIZED STOCHASTICS TO CAPTURE THE INHERENT UNCERTAINTIES OF KEY VARIABLES

In preparation for this analysis, SCE developed the stochastic-based methodology described in this paper. Stochastic studies are a useful way of incorporating and analyzing uncertainties of customer load, wind and solar generation, and fleet availability. As such, stochastic studies are particularly well suited to analyze and determine whether additional capacity and flexibility is needed in the CAISO system to meet the needs of a system with high levels of intermittent renewable generation. SCE's analysis (1) captures inherent uncertainties in certain variables; (2) gives probabilities for a range of future outcomes, and (3) ties system need to reliability standards, such as for loss of load.

2.2 SCE'S ANALYSIS USES THE CPUC'S 2012 LTPP BASE CASE SONGS OUT SCENARIO

On December 20, 2012, the Commission issued Decision (D.) 12-12-010, adopting final LTPP Track 2 assumptions and scenarios. D.12-12-010 invited the CAISO to utilize certain standardized planning assumptions and scenarios to conduct operational flexibility modeling.[5] As a result of SCE's June 7, 2013 announcement that it is permanently ceasing nuclear generation at the San Onofre Nuclear Generation Station ("SONGS"), SCE chose to study the Early SONGS Retirement Scenario. The Early SONGS Retirement scenario includes the original base case assumptions, except that SONGS is retired.

2.3 SCE'S ANALYSIS USES THE 1 EVENT IN 10 YEAR RELIABILITY STANDARD

[5] D.12-12-010 at OP 2. The purposes of the assumptions and scenarios are threefold. "First, the assumptions and scenarios are intended to inform the Commission of any procurement need to meet operating flexibility (also known as renewable integration). Second, the assumptions and scenarios analyze whether adequate resources exist to meet the planning reserve margin, after accounting for any local area and operating flexibility authorizations. Third, the assumptions and scenarios inform the three large IOU's bundled procurement plans of the assumptions utilized in assessing their bundled load for the rolling five plus years." D.12-12-010 at p. 4. As discussed, a stochastic analysis, as opposed to the deterministic analysis, best serves these purposes.

The industry standard for optimal electric system reliability is the "1 event in 10 year" standard.[6] SCE evaluated its results by this standard. SCE defined "event" as any day in which a Stage 3 System Emergency occurred for at least one five minute period. A Stage 3 System Emergency occurs when operating system reserve levels drop below 3% of energy requirements.[7] During a Stage 3 System Emergency, the CAISO is authorized to initiate "rolling blackouts" to ease pressure on the power grid. If SCE's analysis showed less than 1 event in 10 years, then no additional resources would be needed to maintain system reliability. The converse is also true. If the results showed more than 1 event in 10 years, additional resources would be needed to maintain system reliability.

2.4 SCE ANALYZE D THE KEY INPUTS STOCHASTICALL Y

Although all model inputs in the 2022 planning year are uncertain, limited computer capabilities make modeling all inputs stochastically impractical. Specifically, treating too many variables stochastically may impede the computing system's ability to perform a sufficient number of stochastic samples to properly test the range of potential outcomes on the system. Moreover, reliable results can be generated by treating the key determinants of flexibility need stochastically and other variables deterministically. SCE chose to analyze variables stochastically that satisfied two criteria: (1) drivers of system needs that have a large impact on reliability; and (2) variables that are sufficiently uncertain and uncontrollable. Any inputs that met both of these criteria were modeled stochastically. Any that did not were modeled deterministically.

---

[6] *See, e.g.,* D. 04-01-050 at p. 10-11, fn.9 (requiring the IOUs to maintain a 15% planning reserve margin and noting that the resources necessary to meet that demand, even under stressed conditions such as hot weather or unexpected plant outages, is "traditionally based on a "1-in-10" year hot weather scenario.")

[7] The 3% assumption is conservative because it has been shown that system reliability can be maintained even if firm load is not curtailed until reserve levels drop below 1.5% of load.

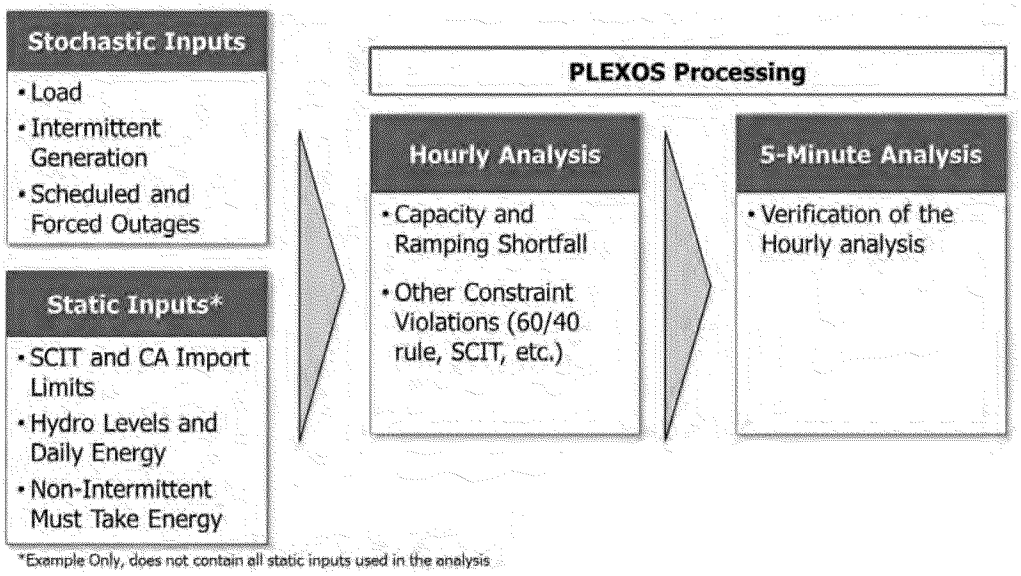
3.0 STOCHASTIC METHODOLO GY

3.1 SCE'S STOCHASTIC MET HODOLOGY OVERVIEW

SCE's analysis identifies the probability of a need for additional resources in 2022 by running a production simulation model using representative days in each season. SCE's analysis used the PLEXOS generation dispatch simulation software as it did in the 2010 LTPP. But for this analysis, SCE utilized a different methodology. First, as briefly discussed in Chapter 2, SCE chose to analyze certain inputs stochastically. Second, rather than sequentially modeling all of the days in a year, SCE's methodology samples and analyzes representative days within a season. The results determine the likelihood and the corresponding magnitude of any potential Stage 3 Emergency.

As demonstrated in Figure 2: Analysis Flow Chart, SCE's methodology is a three step process. First, all deterministic and stochastic inputs are created and fed into the model. Next, the model performs an hourly dispatch of generators to test if all system needs can be met. Finally, the model dispatches generators with 5-minute granularity to verify that the hourly results accurately capture intra hour ramping needs.

Figure 2: Analysis Flow Chart[8]



[8] At the time of this analysis, SCE's 60/40 rule and SDGE's 75/25 rule were still in effect. The elimination of these constraints will be updated in future analysis.

3.2 SELECTION OF STOCHAS TIC VARI ABLES

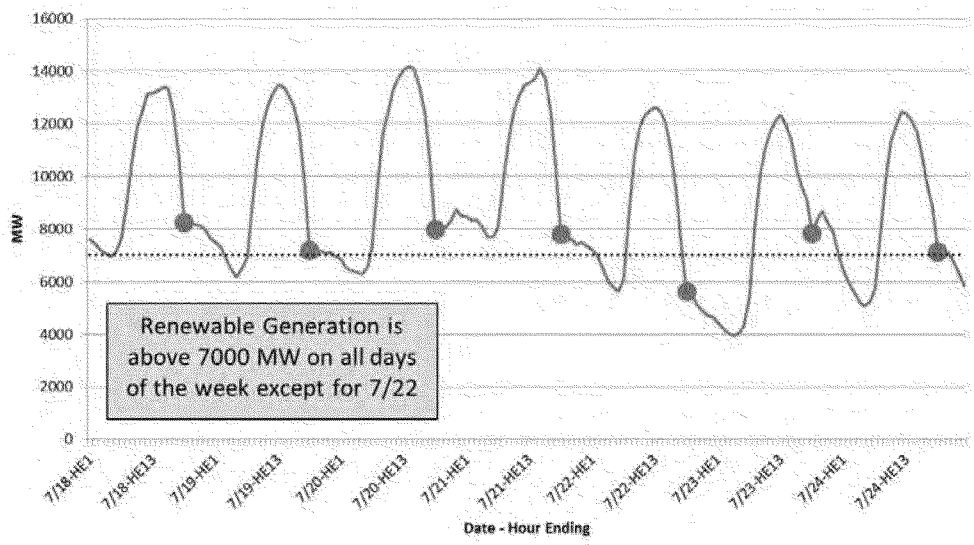
Load, intermittent generation, and generator outages meet the two criteria listed in Chapter 2.4 above of being significant drivers of system need that have a large impact on reliability, and are highly uncertain. Thus, for this analysis, all three were chosen to be stochastic inputs.

The July 15th, 2013 preliminary results of the CAISO's deterministic analysis as illustrated in Figure 3 show why the stochastic approach is necessary for studying these variables. According to the CAISO's deterministic results, July 22 was the peak day and the only day with a ramping shortfall of 2,600 MWs at 7:00 p.m. This shortfall occurred when there was ~6,786 MW of renewable generation and ~2,518 MW of generation on outage. Renewable generation for that day, when compared to other renewable generation days within the month, had an 87 % chance of being higher at 7pm, potentially reducing the amount of shortfall. Total system outages additionally had a 45 % chance of being lower at 7pm, again potentially reducing the amount of shortfall in the system. This analysis was based on CAISO's preliminary results, and SCE recognizes that CAISO continues to refine its results. This is just one example of why load and intermittent generation variability is critical to being analyzed stochastically[9].

[9] As noted, these section and figures are based on preliminary results that may have changed.

As described above, the key driver for this shortfall was a significant drop in renewable generation that day, along with high loads. The CAISO's deterministic model could not speak to the likelihood of what combination of events would result in a ~2,000 MW shortfall. Highlighted in Figure 3 is the level of renewable generation on a specific day at 7pm. All are higher than the day of shortfall, 7/22-HE19 on Figure 3: Example Renewable Generation During Peak Summer Days. When using a deterministic approach, there is too much emphasis on a day that in reality may have a low probability of occurrence. Stochastic analysis, in addition to capturing events like this, will also accurately capture the probability of such events resulting in system outages.

Figure 3: Example Renewable Generation During Peak Summer Days



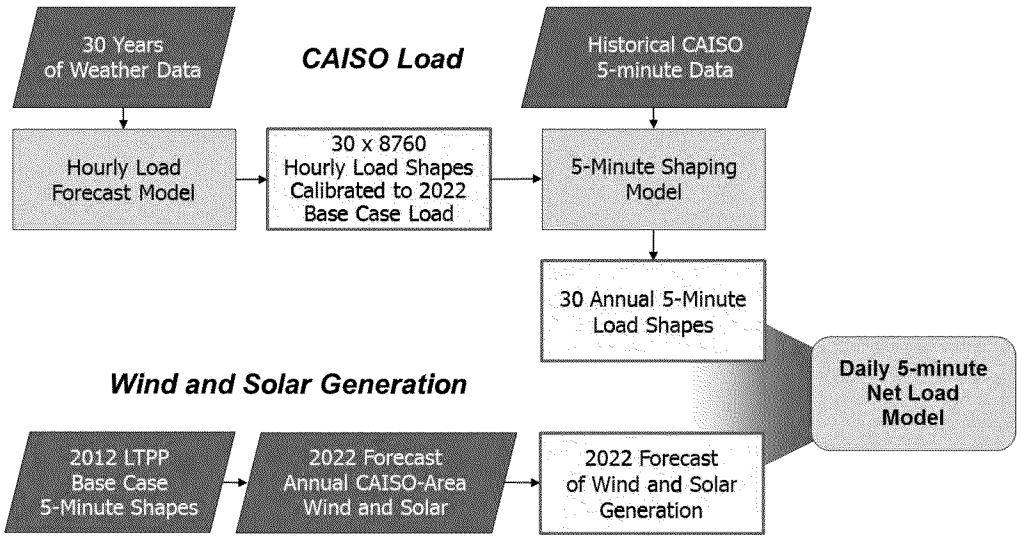
Having learned in the 2010 LTPP studies that the variability of load, intermittent generation, and generation outages can significantly impact results, SCE included them as stochastic variables in its analysis in this analysis. By contrast, SCE did not model hydro resources (both run-of-the-river and dispatchable hydro) stochastically. While it is possible for certain types of hydro to fit the two criteria used to choose stochastic variables, there was insufficient data available at the time of this analysis to create a stochastic distribution for hydro resources.

3.3 DETAILS OF THE NET LOAD STOCHASTIC MODELING

SCE'S FORECAST FOR LOAD, WIND, AND SOLAR

Prior to running PLEXOS, SCE developed a population of many possible future daily net loads. To develop the 5-minute load, SCE utilized weather data to develop expected energy usage patterns, which were then calibrated to the California Energy Commission's (CEC's) load forecast for 2022. For wind and solar energy production, SCE relied upon the CAISO's forecasts. Figure 4 summarizes how these inputs were combined to create the daily 5-minute net load. The individual elements of the net load creation process are discussed in greater detail below.

Figure 4: Net Load Creation Process Flow Chart



LOAD FORECAST

Weather has a significant impact on the variation of load that occurs during each month. To capture this variation, SCE followed a three-step process.[10] First, SCE developed a regression model to assess the impact of temperature variability on load by using 12 years – 2001 through 2012 – of recorded weather station data to correlate these weather conditions with the CAISO's hourly load. The hourly load model was

[10] Details of the forecast models are included in Appendix A.

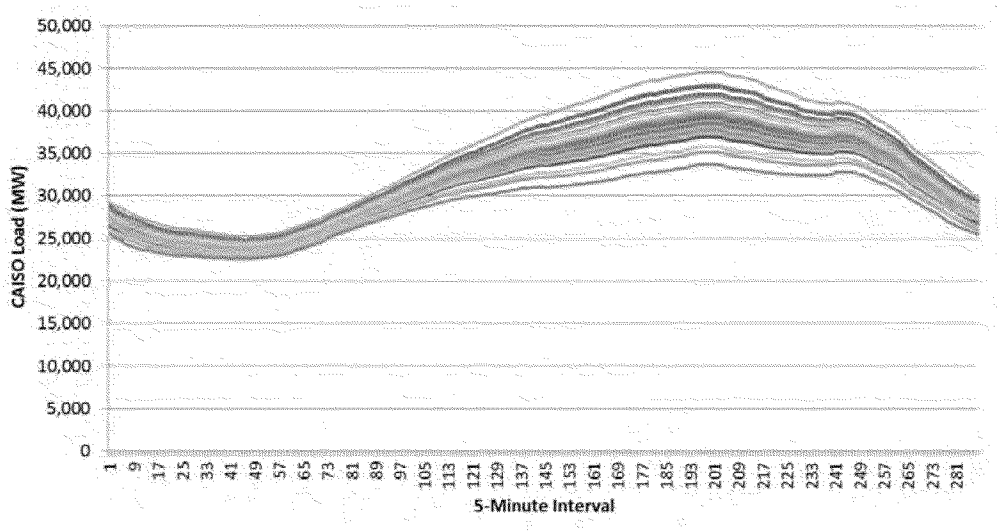
developed based on observed hourly load and daily minimum and maximum daily temperatures at seven representative weather stations across the CAISO territory.[11] Additional variables included in the model are trend (to capture load growth over time), hours of light, season, day of week, and holidays. These variables were identified as being the biggest drivers of customer behavior across the year. Second, SCE used the regression model estimates with 30 years of historical weather (1983-2012) data to create 30 years of hourly load profiles. SCE scaled the resulting 30 years of hourly forecasts so that their average peak and energy, collectively, were consistent with the CPUC's load forecast for 2022. Third, SCE converted the hourly load profiles to 5-minute granularity profiles. To accomplish this, SCE used a second regression model that established relationship between 5-minute and hourly loads based on the CAISO's 2010 5-minute and hourly load data ("the 5-minute shaping model"). Figure 5: 5-Min and Hourly Load Comparison shows the comparison of the hourly forecast and the five minute shaping for August 19, 2022 for one weather scenario.

Figure 5: 5-Min and Hourly Load Comparison

[11] Los Angeles (downtown), Riverside, Sacramento, San Jose, San Francisco, Fresno, San Diego.

Figure 6 below shows the thirty load shapes created in the above described process for August 19, 2022. The thirty load shapes show that there is greater variability in the afternoon than in the morning and evening. That result is reasonable given that air conditioning load is a primary driver of summer time customer demand. Thus, SCE's analysis is producing results that are consistent with expectations.

Figure 6: 30 Example Load Days for August 19, 2022



WIND AND SOLAR FORECAST

For wind and solar shapes, SCE relied upon the data provided by the CAISO.[12] Specifically, SCE used the CAISO's 1-minute forecast of California's total wind and solar generation in 2022 to create 5-minute wind and solar generation shapes. First, SCE took the statewide data and scaled it down to match the expected generation on the CAISO system. Second, SCE converted the 1-minute shape to 5-minute shapes by averaging the 1-minute generation over each 5-minute period in the annual shape. These 5-minute shapes were then used to generate SCE's net load scenarios.

NET LOAD CREATION

Net load is defined as load less wind and solar generation, which captures the impact of their combined variation. SCE combined the five-minute forecasts for load, wind, and solar to create a population of daily net load shapes representing the many

[12] Step 1 of the CAISO's deterministic analysis, which was conducted using the 2012 LTPP Base Case assumptions, developed the wind and solar data upon which SCE relied.

different net load shapes that could occur in 2022. While SCE created 30 years of possible load shapes, only one year of wind and solar shapes were available. To create a larger pool of daily net load shapes, SCE randomized the daily load, wind, and solar shapes within each month. Doing so for every possible combination of daily profiles by month dramatically increases the number of net load shapes to study. Using the month of January as an example, SCE created the following number of samples:

30 years of weather * 31 daily load * 31 solar * 31 wind = 893,730 net load shapes.

Repeating this process for each month of the year created just over 10.15 million net load shapes.

NET LOAD STRATIFICATION

Modeling all 10.15 million net load shapes within a reasonable timeframe is neither practical nor necessary. Sampling is a common technique for selecting a subset of observations within a larger population to estimate the characteristics of the whole population. To efficiently sample the net load population, SCE used a several step process to group or "stratify" similar net load shapes around similar characteristics of interest. This approach, called "stratified sampling," is often used to reduce the number of samples needed to accurately study a population. This method also allows for a more accurate estimation of groups that are more interesting in the analysis such as a day with both high net peak and high ramping needs compared to an average day. The stratification process is described below.

First, understanding that days with high net peak load and/or faster ramp requirements have a higher likelihood of insufficient generation to meet need, SCE calculated the daily peak net load and maximum three hour ramp for each observation in the population of net load shapes[13]. Doing so enabled SCE to identify the more extreme events in which there would be a higher likelihood of insufficient generation resources to meet need. As described below, SCE used peak load and ramp rate requirements to develop the groups or strata.

Second, SCE grouped the 10.15 million net load shapes by month. SCE further grouped those months into four seasons based on the observed relationship between peak net load and 3-hour ramp rate. Using these relationships, and recognition of

[13] The relationship between these factors and system emergencies are shown Chapter 6

seasonal issues, such as Spring hydro conditions in April and May, similar months were grouped into seasons. The months were grouped into the following Seasons:

Winter: November through March

Spring: April and May

Summer: June through August

Fall: September and October [14]

Third, SCE stratified the seasonal groups by percentiles of net peak load and three hour ramp for each season so that net load shapes having a greater expected likelihood of outage would be selected for analysis with higher frequency. SCE accomplished this by first grouping by net peak and then by three hour ramp. The complete process placed each net load group into one of 120 groups by four seasons, five net peak load, and six maximum 3-hour ramp groups. Table 1 illustrates the results of this process for Summer.

Table 1: Number of Daily Net Load Shapes by Summer Stratification Groups

		3 Hour Ramp Strata					
Net Peak Strata		> 99	95 - 99	90 - 95	50 - 90	25 - 50	< 25
	> 99	260	1,039	1,299	10,390	6,494	6,494
	95 – 99	1,039	4,156	5,195	41,559	25,975	25,975
	90 - 95	1,299	5,195	6,494	51,949	32,468	32,468
	50 – 90	10,390	41,559	51,949	415,594	259,746	259,746
	< 50	12,987	51,949	64,937	519,492	324,683	324,683

Finally, looking at Table 1, a random sample of 20 days was drawn from each cell. Taking the 99th percentile for net peak load and ramp requirements as an example, because 20 days represents a higher proportion of that group, the analysis is better able to estimate the impact of that group on the electric system reliability.

[14] SCE chose to include September as a Fall, as opposed to a Summer, month because SCE's data showed that the relationship between net load and ramp requirements in September's bore a closer relationship to that observed in October than that observed in the months classified as Summer.

CORRELATION BETWEEN LOAD, WIND, AND SOLAR

SCE was concerned that weather conditions impacting load could also impact the production of intermittent wind and solar generation. SCE therefore explored whether any correlation between load and intermittent, *i.e.*, solar/wind, generation should be included in the analysis. The impact of significant correlation would be the creation of a net load shape that did not exist in our data set, such as a peak load day with zero solar output. SCE performed a variety of analyses to test peak load-solar, peak load-wind, and maximum ramp-solar and maximum ramp-wind correlation. To determine if any correlation should be included in SCE's analysis, SCE examined data from each month of the year. For instance, for the month of September, SCE did not find that high load days were correlated with higher or lower wind production than low load days. This phenomenon can be explained by the scope of SCE's analysis, which is statewide and thus benefits from diversity. Overall, SCE concluded that its existing method sufficiently accounted for variations between wind and solar production and their relationship to load such that no further modification of the method was required. SCE therefore concluded that the method was a reasonable approach to introducing net load variability into its study. See Appendix B for more details on the correlation analysis.

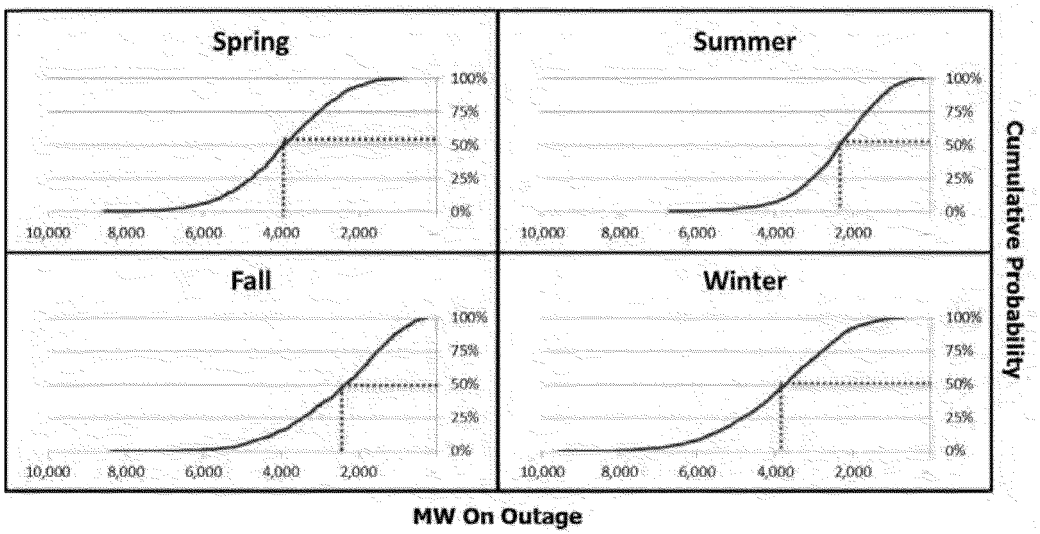
3.4 GENERATOR OUTAGE MOD ELING

OUTAGE SCHEDULE CREATION

SCE's analysis classifies generator outages into two types: (1) planned or scheduled maintenance, or (2) unplanned or forced outages. Even though maintenance is scheduled and thus predictable, both types of outages are treated stochastically because it is difficult to separate the reasons for generator availability. First, both types of generator outages impact the system irrespective of the reason for the outage. Regardless of reason, generation is not available to meet system needs. Second, scheduled outages do not have perfect foresight in practice and can vary throughout a season. It is therefore important to capture the wide range of scheduled outage possibilities available. Third, because the CAISO treats all outages that provide 72 hours or more of advanced notice as planned, there are outages that are identified as planned but that are uncontrollable like forced outages. For these reasons, forced and planned outages are both modeled stochastically.

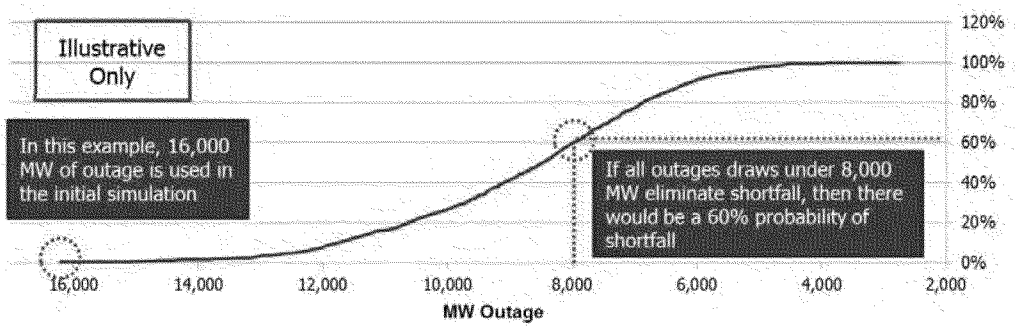
A three step process is used to incorporate outages in SCE's analysis. First, repeated random sampling of the generation fleet is performed using PLEXOS to create a distribution of possible outage events (both forced and planned) for each season. Both forced and scheduled outage rates used in the sampling are based on the CAISO deterministic database (including maintenance rates, forced outage rates, and mean time to repair). The outage cumulative distribution curves in Figure 7 represent the probability that a certain number of MWs or a greater number of MWs will be on outage.

Figure 7: Outage Curves for Each Season



Second, the highest MW quantity of generation outage in each season's outage curve is used for the PLEXOS run. Generally speaking, if the highest outage draw does not result in a shortfall, no other outage draw can. All other outage draws will have fewer MW on outage and thus more resources will be available to the system. If, however, a shortfall is observed with the highest outage draw, then the overall probability of having an outage level that results in shortfall is calculated. SCE calculates that probability, as show in Figure 8, by first determining which of the outage draws would have eliminated the shortfall. Then, SCE calculates the probability of those individual outage draws, which would have eliminated the shortfall, occurring.

Figure 8: Example of Outage Testing



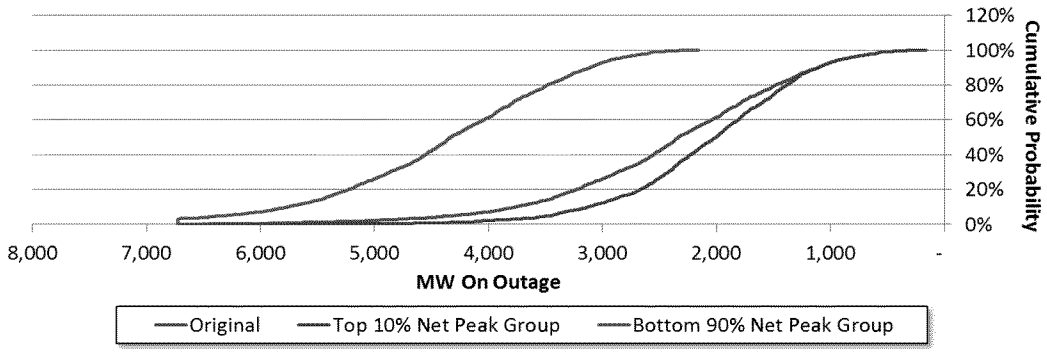
Third, in addition to analyzing the original outage distribution, the cumulative distribution process can be used to test the effect of shifting scheduled maintenance out of a high net peak day or into a different season. This allows the analysis to account for correlation between scheduled maintenance and high load days, as well as allow for the testing of maintenance shifting as a solution for system need.

SCHEDULED MAINTENANCE CONTROL

To conserve resources and to avoid reliability problems on peak load days, the CAISO, which must approve all generators' maintenance schedules, can sometimes decline to approve maintenance on days in which they are expecting particularly slim margins between net load and available capacity. SCE's analysis takes into account the CAISO's control over scheduled maintenance by including the option of retaining 1,000 MWs that would otherwise be unavailable due to scheduled maintenance on high net load days and deferring all other maintenance to low net load days. Accordingly, SCE's analysis allows more availability of MWs on peak load and potential outage days than the CAISO's deterministic analysis. By doing so, SCE can evaluate the impact of scheduled maintenance on potential Stage 3 Emergencies and can more accurately evaluate the actual number of MW that would be available on peak days.

Maintenance shifting tests are performed after the PLEXOS runs are completed and increase or decrease the amount of scheduled maintenance within each sample. In the example chart below , Figure 9, the group of days with the highest net load peak is modified such that, at maximum, only 1,000 MW of scheduled maintenance is allowed. The group of days with the lowest net load peak is modified so that at least 2,000 MW of scheduled maintenance is scheduled within those draws.

Figure 9: Summer Outage Curve with Scheduled Maintenance Shifting



Currently, there is not a known maximum scheduled maintenance cap that can be placed on high stress days. While scheduled maintenance is, by name, controllable, the control is not always absolute. For example, a specific maintenance may require such a large fixed cost for performance that it is impossible to move. In addition, any forced outage that gives sufficient notice before going on outage (typically 72 hours or greater), is usually classified as a scheduled outage even though it is not controllable.

Table 2 shows how the scheduled outage cap affects the loss of load expectation. While the changes can be significant in magnitude, the increased MW need does not result in a need for additional resources (discussed fully in Chapter 6).

Table 2: Outage Shifting Effects on MW Deficiencies

Scheduled Outage MW Cap	Expected Stage 3 Emergency Events	MW Deficiency[15]
2,000	1.70	700
1,500	1.49	500
1,000	1.24	300
500	0.97	0
0	0.71	0

[15] Approximated value. Deficiencies do not account for all the MWs that have been authorized for procurement (see Chapter 6)

3.5 SELECTION OF HYDRO P ROFILE

As discussed previously, both categories of hydro are modeled deterministically. Like the 2010 LTPP System Analysis, SCE's analysis uses 2005 as the source year for its hydro data for both run-of-the-river and dispatchable hydro generation. Run-of-the-river data is in the form of hourly generation shapes for every hour of the year. Dispatchable hydro data comes in the form of parameters for every week of the year.

Run-of-the-river generation uses a fixed generation shape for all draws in the season. The generation shape is created by first identifying the total daily energy for each day of the year. Then, for each season, the day with the lowest daily energy is selected[16]. The hourly generation shape for that day is used as the representative day for run-of-the-river hydro generation for the season.

With respect to dispatchable hydro, SCE's analysis uses parameters for dispatchable hydro that allow PLEXOS to simulate a dispatch that meets system needs, while ensuring that dispatchable hydro operates in a feasible manner. These parameters include (1) daily energy (2) maximum hourly output (3) minimum hourly output (4) maximum ramp up and (5) maximum ramp down. Daily energy for the representative day for the season is calculated by using the lowest weekly energy in a season divided by seven to reduce it to a daily energy target. The minimum hourly output is taken from the week that was used to calculate the daily energy. Maximum hourly output and upwards and downwards ramping limitation were the highest value in the season.

3.6 PLEXOS MODIFICATIONS AND MODELING

To create a reliable set of results, SCE had to run PLEXOS many times. To do this, SCE had to improve run time by identifying elements that determine system reliability and relaxing parameters that lack a significant relationship to reliability. For instance, SCE limited the economics associated with least cost dispatch of individual generating resources because finding the least cost dispatchable resource portfolio was not vital to determining system deficiency in the analysis.

A second major modification was to consolidate the region outside of California into a single region with aggregated generation. CAISO relies on imports from outside of California in order to meet its load throughout the year. Many factors can potentially

[16] Since not modeled stochastically, the lowest energy value within a season was used t o create a conservative assumption for studying capacity and upward flexibility deficiencies.

contribute to the CAISO's ability to import energy, including but not limited to transmission line capacities, available generation outside of CAISO, and simultaneous import limitations (also known as the California and Southern California Import Transmission (SCIT) Import Limits). In the 2010 LTPP, SCE observed that the California Import Limit was binding during peak hours, meaning that there was neither a dearth of available generation for import nor a lack of line capacity for imports.

4.0 MODELING ASSUMPTIONS

On December 20, 2012, the Commission issued Decision (D.) 12-12-010, adopting final LTPP Track II assumptions and scenarios and invited the CAISO to use them in its analysis. Most of the assumptions used in SCE's analysis are consistent with the decision; however, departures were necessary to perform a stochastic analysis.

4.1 LOAD

As described previously, SCE's analysis uses 30 years of weather data to produce 30 potential load forecasts for the CAISO area for 2022. SCE attempted to align its distribution of load forecasts with the Base Case SONGS Out load assumptions. To avoid distorting daily ramps when scaling annual load shapes, however, peak and annual energy do not exactly match the Base Case SONGS Out load assumptions. Table 3 below compares the mean and median peak and annual energy from SCE's analysis against the peak and annual energy of the Base Case SONGS Out load assumptions. Table 4 contains more detailed information about the percentile distribution of the load shapes used in SCE's analysis.

Table 3: CAISO Area Load Forecast

	Peak (MW)	Energy (GWh)
SCE Analysis Mean	51,656	245,816
SCE Analysis Median	51,453	245,736
Base Case SONGS Out Load Assumption	51,058	245,342
Difference from SCE Analysis Mean	1.20%	0.20%
Difference from SCE Analysis Median	0.80%	0.20%

Table 4: CAISO Area Load Forecast Percentile Distribution

	Peak (MW)	Energy (GWh)
Max	59,145	250,902
90%	54,586	248,909
75%	53,542	246,976
50%	51,453	245,736
25%	49,936	244,540
10%	47,282	243,489
Min	46,115	240,838

After SCE created the CAISO load forecasts described above, SCE split the load forecasts among the different regions based on their 2022 load share in the IEPR Forecasts[17]. The breakdown of load share by percentage is shown in Table 5 below.

Table 5: CAISO Area Load Forecast Regional Distribution

Region	Load Share
SCE	44.6%
PG&E VLY	25.4%
PG&E Bay	20.2%
SDGE	9.8%

4.2 RENEWABLE GENERATION BUILDOUT

SCE's analysis only models renewable generation within California. The renewable generation build out used in SCE's analysis is taken from the result of the CAISO's deterministic analysis of the Base Case SONGS Out renewable generation assumption. Table 6 below shows the breakdown of annual GWh of renewable generation expected in 2022.

Table 6: Renewable Generation Buildout by Region and Technology

Technology	PG&E Bay	PG&E VLY	SCE	SDGE
Biomass	85	4,561	1,557	127
Geothermal	418	11,054	2,966	
Small Hydro		3,699	1,576	
Wind	1,065	3,959	7,569	912
Biogas	32		764	
Solar Thermal			3,399	
Solar PV	256	5,520	9,012	2,077

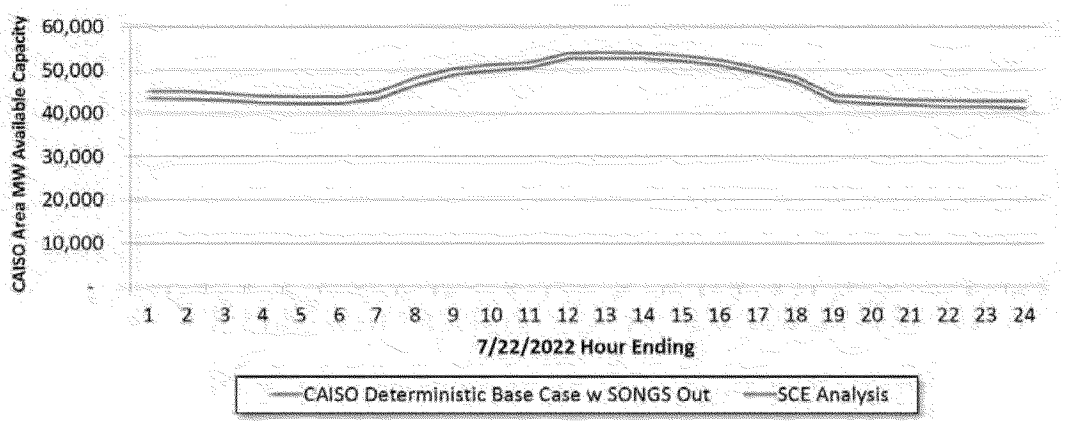
[17] 2012 IEPR Mid Forecasts - Mid_Case_LSE_and_Balancing_Authority_Forecast.xls – Form 1.5a

For the CAISO area, renewable generation falls into two categories: intermittent and non-intermittent. Intermittent generation, including wind, solar thermal, and solar photovoltaic, are part of the net load creation process described earlier. Non-intermittent resources, including biomass, geothermal, small hydro, and biogas, were analyzed at an hourly granularity to match the result of the CAISO's deterministic analysis of the D. 12-12-010 Base Case SONGS Out renewable generation assumption. Unlike wind and solar, a single daily generation shape was used for each season. To produce a generation shape, SCE chose the day with median daily energy in the season and used the hourly generation shape of that day as representative of all of the days in that season. SCE used the same procedure to analyze California Municipality renewable generation of all types.

4.3 GENERATION FLEET

SCE based its generation fleet assumptions on the inputs the CAISO used for its deterministic analysis of the Base Case SONGS Out assumptions, according to the results it published on July 13, 2013. To verify the fleet was similar, SCE performed a comparison of fleet capacity for July 22, 2022, the peak day. To perform a reasonable comparison, SCE modified its database so that: (1) all generation would be accounted for by removing outages; (2) renewable generation specific to July 22, 2022 in CAISO's analysis was replicated in SCE's analysis for capacity comparison purposes; and (3) to avoid the complications associated with choosing a CAISO import limit, the CAISO's import capability was not included in potential capacity. The results of the comparison can be seen in Figure 10:

Figure 10: CAISO and SCE Analysis Available Capacity Comparison

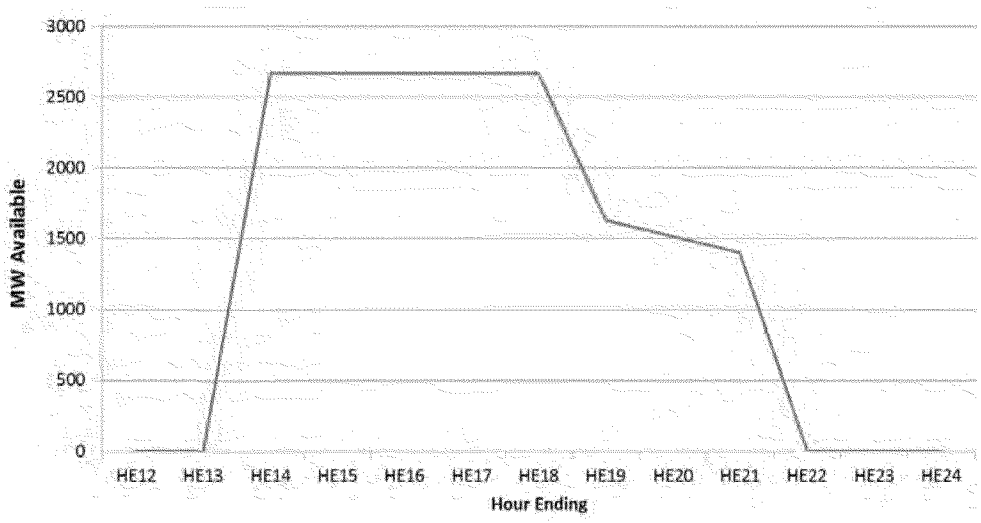


Overall, the fleet MW capacity used in SCE's analysis is 1% lower than that used by the CAISO for its deterministic Base Case SONGS Out analysis. This difference is caused by the difference in general capacity and rating between the approximately 350 units that make up the CAISO area in SCE's analysis. SCE also based other fleet characteristics, such as ramp rates and outage rates, on the CAISO's assumptions for the CAISO's Base Case SONGS Out deterministic analysis.

4.4 DEMAND RESPONSE

SCE, in collaboration with the CAISO, Pacific Gas and Electric Co. (PG&E), and San Diego Gas & Electric Company (SDG&E), developed a realistic demand response forecast for 2022 that takes into account the availability of demand response after 6:00 p.m. Figure 11 shows the demand response forecast for a summer day in 2022.

Figure 11: Demand Response Summer Available Capacity



SCE's analysis assumes that demand response capacity is only available between HE14 and HE22, which approximate peak demand periods. However, it should be noted that many demand response programs in California are not restricted to those hours and some programs can be called 24 hours a day. Therefore, SCE's assumed availability could be conservative.

4.5 RESERVE REQUIREMENTS

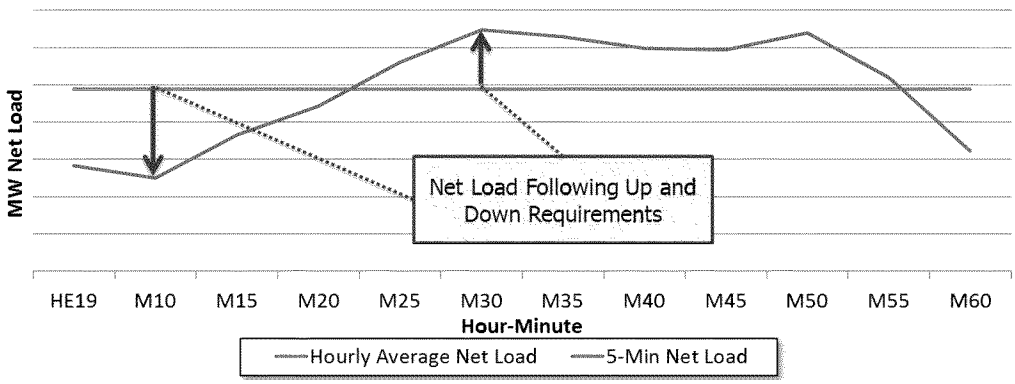
SCE used the following assumptions for ancillary services and reserve requirements within the analysis:

Regulation Up and Down: For both Regulation Up and Regulation Down, SCE set a regulation requirement for CAISO equal to 1.5% of the CAISO's load, as recommended by the CAISO.

Spinning and Non Spinning Reserves: Requirements are equal to 3% of CAISO load (each) and matches the assumptions used in the 2010 LTPP and CAISO's deterministic runs in the 2012 LTPP Track 2.

Net Load Following Up and Down: Figure 12 shows how Net Load Following Up and Down requirements are calculated in SCE's analysis to be, respectively, the maximum or minimum difference between hourly average load and 5-minute net load. To make sure that sufficient capacity is committed to meet net load changes within the hour, Net Load Following is incorporated in the hourly analysis.

Figure 12: Net Load Following Definition Example



4.6 TRANSMISSION BUILDOU T

Transmission paths and line ratings within California and connecting California to the rest of the Western Electricity Coordinating Council (WECC) that SCE used in its analysis generally match the assumptions used in the 2010 LTPP System Analysis. The only change is the inclusion of a new region to model the transmission constraints faced by SDG&E and the surrounding area. The new region is known by its two substation's names -- IV/ECO -- and is designed to capture the constraints on the Southwest Power Link (SWPL) and Sunrise transmission lines. Transmission paths and line ratings within the WECC, except for within California, were aggregated.

5.0 ASSESSMENT OF SCE'S METHODOLOGY

To test the accuracy of SCE's analysis, SCE conducted two checks: (1) a check that reliability results for 2012 were reasonable, and (2) a comparison between the results of SCE's and CAISO's analysis given similar inputs for 2022.

5.1 2012 RELIABILITY CHECK

The objective of the 2012 reliability check was to confirm that the reliability evaluation produced by SCE's analysis were reasonable given our understanding of generation sufficiency for that year. Specifically, SCE's analysis produced a 0.04 probability of a loss of load Stage 3 Emergency in 2012.

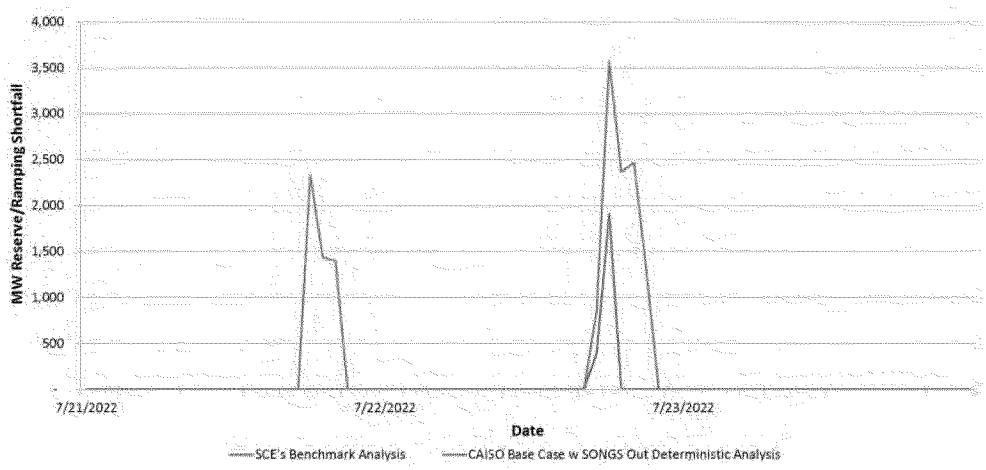
5.2 SCE'S ANALYSIS VALIDATION AGAINST THE CAISO'S DETERMINISTIC ANALYSIS FOR 2022

To confirm that SCE's modifications to PLEXOS did not skew results, SCE analyzed the key inputs the CAISO used for its deterministic analysis. If the reliability results of that analysis were the same as or similar to the CAISO's, it would confirm that SCE's modifications were reasonable. The results were fairly close. In fact, for the three days tested, SCE's analysis was more conservative in finding need because in each instance it found more need than the CAISO's analysis for that same input.

The following key inputs exactly matched those used in the CAISO's deterministic analysis of the Base Case SONGS Out for July 21 – 23, 2022: (1) the CAISO area load, (2) MWs of all types of renewable generation, and (3) ancillary service and ramping reserve requirements, including regulation up and down, load following up and down, spinning reserves, and non-spinning reserves.[18] SCE's analysis produced the following reserve shortfall using chronological hourly analysis:

[18] The about 2,500 MW of generation on outage within the CAISO region that SCE used approximately matched the CAISO's deterministic analysis of the Base Case SONGS Out for July 21 – 23, 2022.

Figure 13: CAISO Benchmark Results



As demonstrated in Figure 13 and as discussed above, SCE's analysis produced more conservative results. The extra reserve shortfall can be explained by (1) reduced available capacity due to generator capacity and rating differences; and (2) anomalies in the CAISO's import of reserves.[19]

[19] SCE understands that the CAISO has corrected the anomaly in its deterministic analysis.

6.0 RESULTS

6.1 SUMMARY OF RESULTS

SCE's analysis of the D.12-12-010 Base Case SONGS Out shows that there is no need for additional resources in the CAISO area in 2022 for the purpose of maintaining system reliability with integration of increasing intermittent renewable resources. This conclusion is based on the results of SCE's analysis, which shows that the CAISO area will have sufficient resources to meet the 1-in-10 year reliability standard. The resource assumptions employed in SCE's stochastic analysis are similarly based on the resource build out used by the CAISO in their SONGS Out Base Case deterministic analysis. These assumptions, however, fail to account for up to potentially 3,500 MW of existing or authorized resources in 2022.

6.2 SCE'S ANALYSIS FINDS NO NEED FOR ADDITIONAL RESOURCES IN 2022

The result of SCE's PLEXOS runs showed 1.3 expected Stage 3 Emergency events in ten years. Including up to 3,500 MWs of potential, existing, or authorized resources in 2022 that are not accounted for in the Base Case assumptions will reduce the 1.3 events to less than one, which satisfies the reliability standard. SCE therefore presently concludes that there is no need for additional system resources in the CAISO area in 2022.

6.3 RELIABILITY VIOLATIONS ARE MOST LIKELY TO OCCUR IN SUMMER AND FALL

SCE's analysis found that the highest potential for Stage 3 System Emergencies exists between June and October. As discussed more fully in Chapter IV of this testimony, this potential is more directly related to peak net load than to maximum three hour ramp.[20] Table 7, Table 8, and Table 9 illustrate this phenomenon. First, Table 7 sets forth the number of calculated potential Stage 3 System Emergency events by season.

[20] Maximum three hour ramp is defined for this analysis as the greatest change in net load between two time periods that are three hours apart.

Table 7: Expected Stage Emergencies by Season

Season	Expected Stage 3 Emergencies
Spring	0
Summer	0.89
Fall	0.35
Winter	0
Total	1.24

Second, Table 8 and Table 9 are heat maps that demonstrate the relationship or lack thereof between potential Stage 3 System Emergency events, peak net load, and three hour ramp requirements. The heat map shows that the majority of potential Stage 3 System Emergencies occurred in the samples with the highest net peak for both summer and fall. For fall, the greatest potential for Stage 3 System Emergencies did not occur in the samples with the highest three hour net ramp, but rather during the samples with mid-range ramping needs.

Table 8: Probability (%) of Stage 3 System Emergencies within Summer Net Load Groups

Summer		Net Load Peak Group		
		< 95%	95%-99%	99% +
3 Hour Net Load Ramp Group	< 25%	0%	0%	1%
	25% - 50%	0%	0%	6%
	50% - 90%	0%	0%	4%
	90%-95%	0%	0%	45%
	95%-99%	0%	0%	80%
	99% +	0%	0%	78%

Table 9: Probability (%) of Stage 3 System Emergencies within Fall Net Load Groups

Fall		Net Load Peak Group		
		< 95%	95%-99%	99% +
3 Hour Net Load Ramp Group	< 25%	0%	0%	0%
	25% - 50%	0%	0%	0%
	50% - 90%	0%	0%	11%
	90%-95%	0%	0%	17%
	95%-99%	0%	0%	2%
	99% +	0%	0%	0%

6.4 RESULTS CONFIDENCE INTERVALS ARE RELATIVELY NARROW

As discussed in Chapter IV, SCE analyzed 2,400 net load samples from a pool of 10 million. Although the 2,400 samples were carefully selected to represent the range of net load stress days that might reasonably be expected to occur, SCE conducted an additional analysis, known as bootstrapping, to estimate the confidence that could be placed on the estimated probabilities. One of the main problems with using a small sample from such a large pool is the potential for a few extreme samples to heavily skew the results. For example, if one sample found 10,000 MW of need, but all other samples found 0 MW of need, the one sample would skew average MW of need. To mitigate and evaluate the magnitude of this problem, SCE utilized the statistical technique known as bootstrapping.

Bootstrapping is a statistical technique commonly used for studies with large potential sample populations where only a small number of samples can actually be analyzed. In SCE's bootstrapping analysis, SCE studied a sample from the pool of 2,400 and then returned it to the pool before taking a new sample – a process known as sampling with replacement. SCE resampled the 2,400 samples 10,000 times. As a result, a distribution of outcomes was created that represents the uncertainty in the results.

The results show a small range in the potential number of Stage 3 System Emergencies. Table 10 below illustrates the confidence range associated with the particular 2,400 load sample that was drawn. The confidence intervals produced by the bootstrapping analysis show that even an extreme case (95%), the 500 MW need to maintain the reliability standard is less than the 3,500 MWs of resources that are expected to be available in 2022 but that are not accounted for in the Base Case SONGS Out assumptions.

Table 10: Confidence Intervals for SCE's Analysis

Category	5th Percentile	Mean	95th Percentile	Standard Deviation
Stage 3 Emergencies	1.00	1.24	1.49	.15
MW Deficiency	0	300	500	N/A

6.5 CAPACITY RESERVE MARGIN CHECK AFFIRMS NO NEED

Using the resource assumptions for the Base Case with SONGS Out, SCE calculated a reserve margin of approximately 120% for the entire CAISO's territory, which includes PG&E's, SCE's and SDG&E's service territories. Based on calculated reserve margin, SCE presently finds that there is no need for additional resource procurement in 2022 to maintain a 115% reserve margin.

6.6 THE BASE CASE SONGS OUT ASSUMPTIONS DO NOT ACCOUNT FOR UP TO 3,500 MW OF POTENTIAL, EXISTING, OR AUTHORIZED RESOURCES IN 2022

As discussed above, SCE chose to perform its stochastic analysis based on the D.12-12-010 SONGS Out Base Case assumptions. Because the CAISO also performed its analysis using, among other cases, the SONGS Out Base Case, SCE chose to align its resource assumptions with those of the CAISO so that the CAISO's deterministic analysis and SCE's analysis could be compared. There were, however, notable differences between the resources that D.12-12-010 assumed would be available in 2022 and those that SCE believes will be available. Specifically D.12-12-010's Base Case SONGS Out resource assumptions do not take into account approximately 3,500 MW of available resources.

First, the assumptions accounted for only 1,000 MWs of thermal generation procured to meet Local Capacity Requirements (LCR). There is currently approximately 1,500 MWs of thermal LCR authorized for procurement– up to 1,200 MW[21] of thermal resources in the Los Angeles Basin Region and approximately 300 MWs of thermal resources in the Big Creek / Ventura area[22]. The assumptions therefore omitted approximately 500 MWs of available thermal generation.

Second, because D.12-12-010 assumes that all thermal generation will retire after 40 years of generation, the Base Case SONGS Out assumptions did not account for the potential continuing availability of 1,700 MW of thermal generation resources past the presumed 40 year lifespan. Many thermal resource generators continue to operate for more than 40 years. For example, Table 11 below provides a list of generators that have been operational for more than 40 years.

Table 11: Generators with 40 or More Years of Operation

Generator	Years of Operation	Generator	Years of Operation
Morro Bay 3 Morro Bay CA	51	Mandalay 1	54
Morro Bay 4 Morro Bay CA	50	Mandalay 2	54
Moss Landing 6 Moss Landing CA	46	Mandalay 3	43
Moss Landing 7 Moss Landing CA	45	Ormond Beach Gen 1	42
AES Alamitos 1	57	Ormond Beach Gen 2	40
AES Alamitos 2	56	AES Redondo Beach 5	59
AES Alamitos 3	52	AES Redondo Beach 6	56
AES Alamitos 4	51	AES Redondo Beach 7	46
AES Alamitos 5	47	AES Redondo Beach 8	46
AES Alamitos 6	47	Miramar 1	41
AES Huntington Beach 1	55	Encina 1	59
AES Huntington Beach 2	55	Encina 2	57
El Segundo Power 4	48	Encina 3	55
South Bay GT1	57	Encina GT1	55

As set forth in Table 12, the 40 year retirement assumption results in the presumed retirement of over 1,700 MW (Nameplate) of thermal generation in the CAISO area.

---

[21] 1,000 MWs of thermal resources are authorized for Los Angeles Basin. Another 200 MW for Los Angeles Basin may be procured, but are not required to be thermal.

[22] D. 13-02-015 at OP 1.

Table 12: Resources Retired in 2022 Due to 40 Year Retirement Assumption

Generator	MW (Nameplate)	Generator	MW (Nameplate)
Oakland1	55	ELCAJNGT_1	16
Oakland2	55	Kearn2AB1	15
Oakland3	55	Kearn2AB2	15
Broadway 3 Pasadena	65	Kearn2CD1	15
Coolwtr1	63	Kearn2CD2	14
Coolwtr2	81.5	Kearn3AB1	16
CoolwtrS3	245.3	Kearn3AB2	15
CoolwtrS4	245.9	Kearn3CD1	15
Ellwood1	54	Kearn3CD2	15
Etiwand3	320	KearnGT1	16
Etiwand4	320	Miramar1	18
Pasadna1	22.3	Miramar2	18
Pasadna2	22.3	ELCAJNGT_1	16

In sum, had the assumptions accounted for the omitted MWs discussed above, SCE's analysis utilizing the assumptions would not have indicated a need for the procurement of new resources. Accordingly, SCE presently concludes that additional resources are not needed in 2022 to maintain system reliability.

7.0 CONCLUSIONS AND FUTURE WORK

This study develops a stochastic framework for the analysis of capacity and upward flexibility deficiencies within the CAISO area. The methodology developed in this paper, as well as the results produced for the 2022 Base Case with SONGS Out Scenario and the model validation tests, show the methodology is robust and reasonable for the purpose of studying capacity and upward flexibility needs.

Going forward, this analysis can be expanded to address new questions facing the system in future years. The questions can include, but are not limited to:

- The effects of over-generation and the economic trade-offs it creates with system reliability concerns and potential solutions (generation curtailment, energy exports, additional flexible resources, etc.).
- Defining the potential issues and additional resource need created by multiple types of forecast error.
- Developing an economic analysis to determine which future scenarios provide the best outcome for energy customers.
- Translating system need created by increased intermittent resources into drivers of need and potentially into an existing or new reliability metric.

In addition, aspects of the methodology created for the analysis of capacity and upward flexibility deficiencies can be improved for future analysis. These include improvements made to the forecast and control of maintenance schedules, the net load creation process, and hydro analysis, in addition to many other potential improvement areas.

CAISO Demand Volatility Modeling for the Stochastic System Need Analysis Project (SSNAP)

Statistical modeling and historical weather observations were used to create 30 scenarios of 5-minute CAISO load forecasts for year 2022. The effort was separated in two steps: (1) generate hourly CAISO load forecasts based on 30 years of actual weather data for the CAISO territory (1983-2012), and (2) apply expected 5-minute load shapes to each hourly load scenario created in Step 1. The forecasted 30 load scenarios were normalized to the Base Case CAISO load forecast scenario by using the metrics of total and peak annual load to ensure that the forecasted CAISO load distribution is centered on the Base Case scenario. The modeling details are described in the sections below:

Model Specification

Step 1: Hourly CAISO Load Volatility Scenarios

The weather scenarios were created using 30 years of actual weather data from 1983 to 2012[23]. The key weather variables needed for the analysis were Minimum and Maximum daily temperatures at 7 weather stations representing load in the CAISO territory:

- Los Angeles Downtown
- Riverside
- Sacramento
- San Jose
- San Francisco
- Fresno
- San Diego

The hourly regression framework was used to create load forecasts based on weather and other variables estimated to materially impact load. Specific

[23] It is assumed that the weather in the last 30 years would be representative of the possible weather patterns in 2022.

features and variables considered in the load forecasting model used are summarized below (with further model details provided in Appendix I):

- Temperatures from seven stations are considered;
- Temperature effects are considered seasonally;
- Economic situation is considered by use of a trend variable
- Calendar and holiday effects on load are considered

The model for each hourly period can be written as

$$\sum$$

Where:

= hourly load (MW) for date t at hour h,

= coefficients to be estimated,

= k regressor variables, and

= random error term.

Depended Variable: Hourly total ISO load (**KWH**)

Predictor variables:

- Average CAISO system temperature;
- Trend;
- Hours of light;
- Monthly and quarterly dummy variables;
- Day of week dummy variables;
- New Year's Day and surrounding days;
- Martin Luther King, Jr. Day;
- President's Day;
- Memorial Day;
- Daylight Savings;
- Easter;
- Independence Day;
- Labor Day;
- Thanksgiving Day;
- Veteran's Day;
- Christmas;
- Before and after holiday;
- Interaction variables.

The model was estimated using 12 years (2001-2012) of recorded hourly CAISO load and corresponding explanatory variable data. Afterwards, the raw model forecast results were fine-tuned by day of the week.

Step 2: 5-minute Shaping of Hourly Volatility Scenarios

The regression framework was also utilized to estimate expected 5-minute CAISO load shapes as a function of the average load during the corresponding hour and two adjacent hours. The 5-minute load data provided by CAISO for the year 2010 was selected for regression model specification. The 5-minute load model can be generally described as below:

$$\sum (\quad - \quad - \quad)$$

Where:

= hourly load (MW) for hour h (1 to 24) and 5-min period p (1 to 12),

= coefficients to be estimated (k=1 to 12),

$-$ = average hourly load (MW) for hour h (1 to 24),

= monthly indicator (dummy) variable (k=2 to 12)

= random error term.

TECHNICAL APPENDIX B – CORRELATION OF LOAD, WIND, AND SOLAR

Concerned that weather conditions impacting load could also impact the production of intermittent wind and solar generation, prior to developing net load samples, SCE explored whether correlation between load and intermittent solar / wind generation should be included in the analysis based on the matched 5-minute profiles obtained from the CAISO. This is an issue because it has been documented that wind output from certain zones declines when nearby load is also high due to the temperature impacts. However, SCE is using combined data for the CAISO footprint so diversity can occur which could reduce the correlation impact. SCE performed a variety of analyses to test peak load-solar, peak load-wind, and maximum ramp-solar and maximum ramp-wind correlation. As discussed below, although SCE found that wind and solar production varied significantly by month, SCE did not find sufficiently conclusive evidence supporting the need to include correlation between load and solar/wind generation within its analysis.

SCE therefore concluded that the method it used to create its net load population was a reasonable approach to introducing net load variability into its study.

Figure 14: Daily Wind and Solar Production by Month

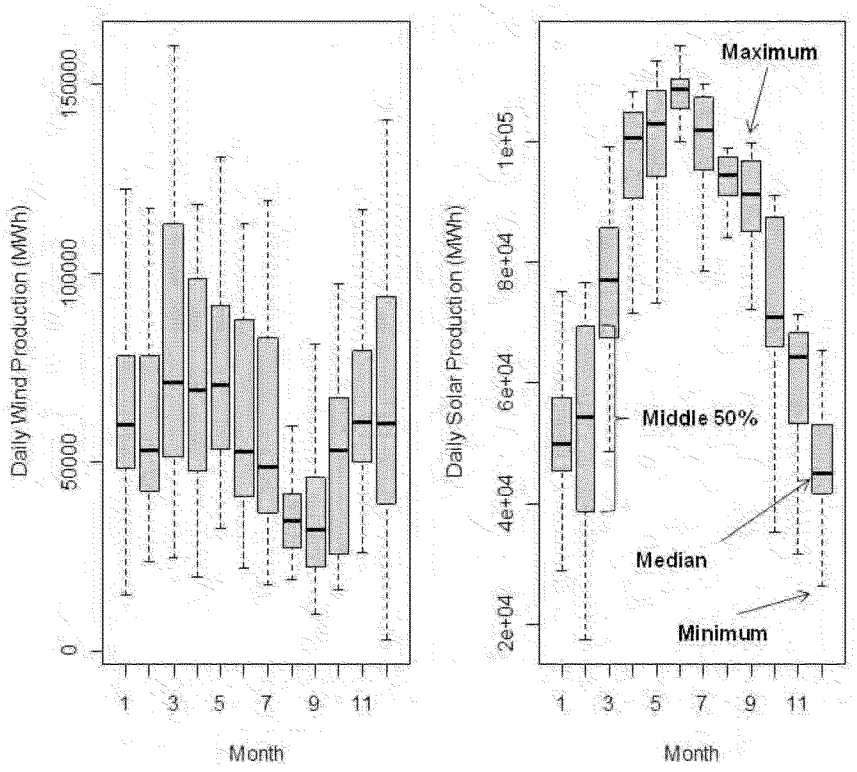


Figure 14 plots the distribution of daily wind and solar production by month using boxplots. The thick black line represents the median of the distribution. The grey box contains the middle fifty percent of all data points in the distribution. In this case, each data point is a single day's total solar or wind production. The dash lines contain all other data points. After examining these plots, it is clear that characteristics of wind and solar daily production changes throughout the year. This is most evident for solar, which peaks in production during June and rapidly diminishes during the winter months. Wind production shows a similar, though much less pronounced, pattern that peaks in the spring months. To ensure that the relationship between time of year and solar and wind production is preserved, SCE's net load process only matches load, solar, and wind profiles from the same month.

Figure 15: Solar Production at the Time of System Daily Peak versus System Daily Peak Load

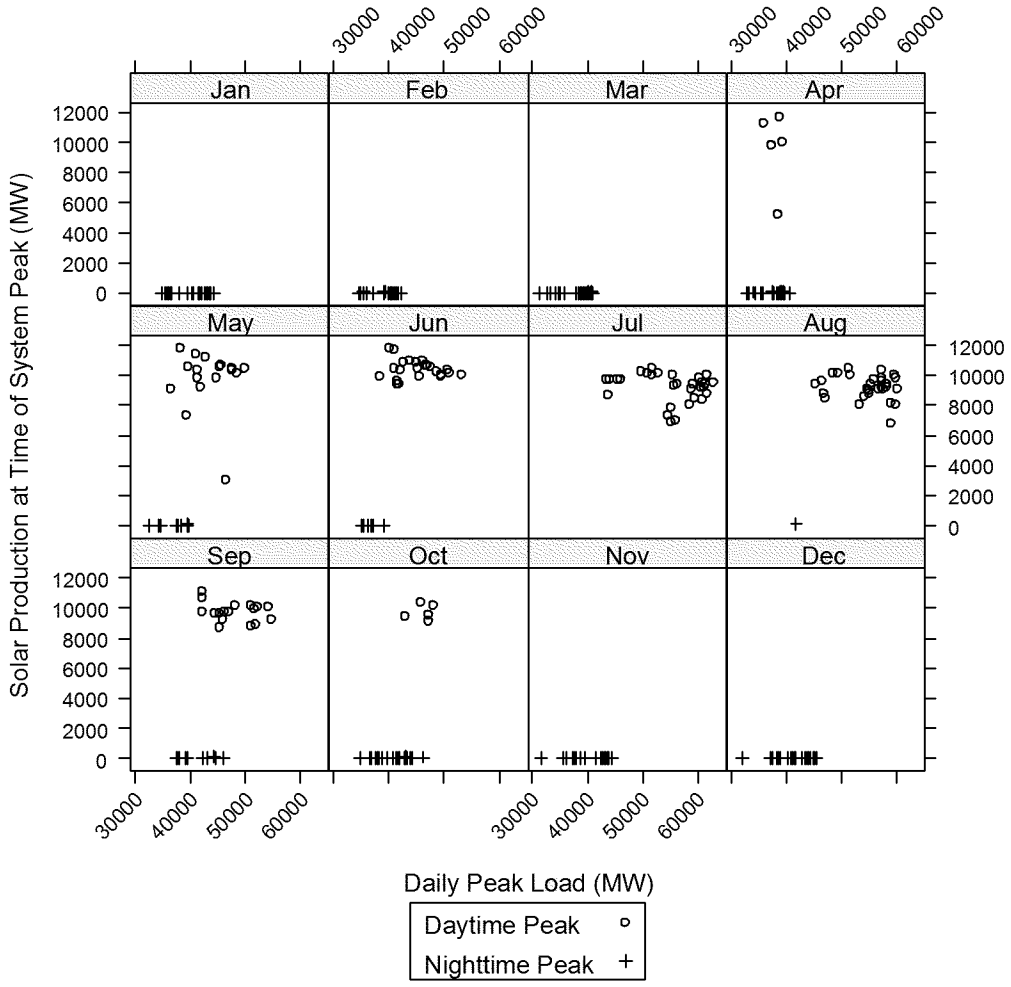


Figure 15 plots daily peak load against solar production in that same hour. Each point on the chart represents a single day from the dataset. The magnitude of peak load for each observation is recorded on the x-axis and the corresponding level of solar production is on the y-axis. SCE further grouped these data by time of day. Nighttime peaking days are denoted by a plus sign while daytime peaking days are denoted by a circle. This was done to isolate this variable's effect on solar production, which is important for understanding spring and fall months, which will have both daytime and nighttime peaking days. As expected, the time of day in which the peak load occurs has a significant impact on the amount of solar generation at that time. Specifically,

nighttime peaks have no corresponding solar production. This can be seen by the distinct clustering around the zero on the y-axis. SCE's daily profile sampling technique, which maintains patterns in production and load throughout the day, captures that important relationship. Perhaps more important, however, the data shows that there does not appear to be a strong downward or upward sloping relationship between daily peak load and solar production for daytime peaking days. That is, higher/lower levels on the x-axis are not associated with higher/lower levels on the y-axis and vice versa. Because temperature is a primary driver of peak load, we can conclude that no strong relationship exists between California average temperature and total California solar production. SCE suspects that this is due, in part, to the diversity of weather patterns across the State and the sheer number of different solar sites, which are represented by the aggregate shapes. Because SCE's Stochastic analysis is based on typical daily load shapes by season, the kinds of pattern variation shown in Figure IV-7 are already reflected in SCE's stochastic modeling. As a result of this assessment, SCE assumed for the purpose of stochastic modeling that load and solar production are uncorrelated within each monthly population.

Figure 16: Wind Production at the Time of System Daily Peak versus System Daily Peak Load

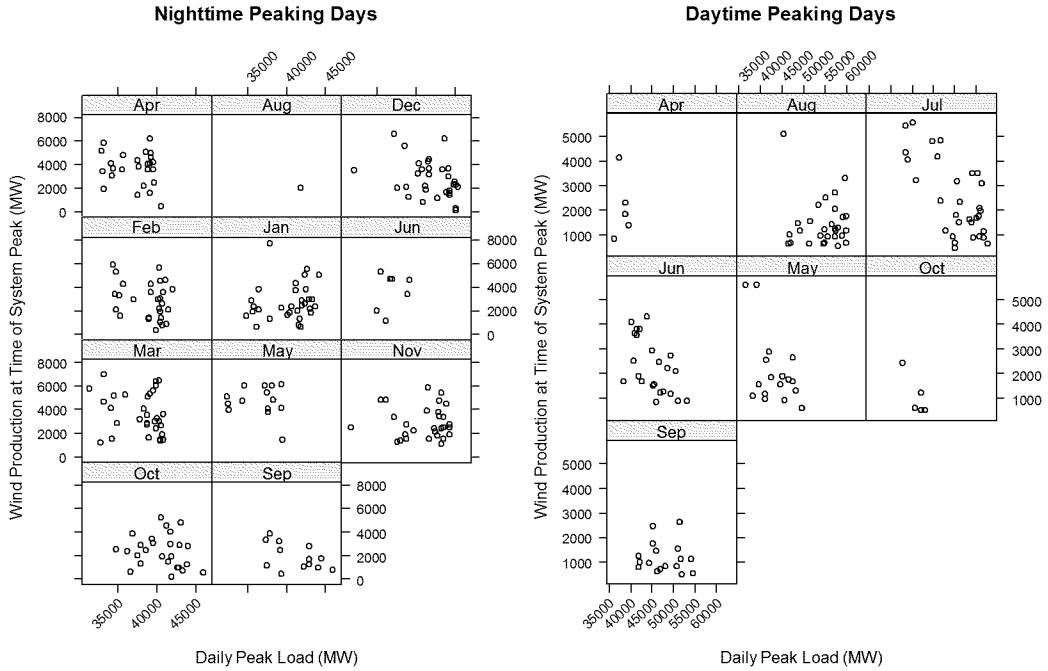


Figure 16 repeats the above analysis for wind in which each daily load peak is plotted against the observed wind production during the peak by month (ordered aiphabetically). Once again, daytime peaking days and nighttime peaking days are separated. Note that the nighttime peaking days are presented on the panels on the left and daytime peaking days are presented panels on the right. Some months have only nighttime peaking days or daytime peaking days. Those months are excluded from the charts in which they do not have any data. In both sets of data, peak load and corresponding wind production are scattered throughout each panel. That phenomena indicates that, as expected, peak wind production is substantially more variable than solar. For the majority of months, no downward or upward sloping relationship appears to exist between daily peak load and corresponding wind production. While some relationship could exist for June and July, SCE opted not to modify its sampling framework given the substantial variation in wind production within these months and an already limited sample sizes.

Energy+Environmental Economics
101 Montgomery Street | Suite 1600 | San Francisco, CA 94104 | 415.391.5100 | www.ethree.com

Renewable Energy Flexibility (REFLEX) Model

MODEL DESCRIPTION FOR PG&E/EPRI FLEXIBILITY MODELING FORUM

April 15, 2014

1. Introduction

The Renewable Energy Flexibility (REFLEX) model is a tool to calculate the need for power system flexibility under high renewable penetration and to evaluate alternative strategies for meeting power system flexibility needs. While it has long been known that high renewable penetration can cause operating challenges on certain days, no methodology has existed to provide the context for those days; that is, to determine the frequency with which they occur and the full set of conditions that can cause challenges. REFLEX performs stochastic production simulation that captures a broad range of system operating conditions. This enables calculation of the likelihood, magnitude, duration and cost of reliability violations to characterize flexibility constraints and inform potential solutions. REFLEX provides an economic framework for determining cost-effective flexible capacity investments by trading off the cost of investments in new flexible resources against the value of avoided flexibility violations.

The REFLEX approach combines the results of:

(1) A conventional reliability model (E3's Renewable Energy Capacity Planning (RECAP) Model) that calculates Loss of Load Probability (LOLP), Loss of Load Expectation (LOLE) and Expected Unserved Energy (EUE) resulting from a shortage of capacity without considering operational issues that might limit access to capacity resources, and

(2) A flexibility model (REFLEX) that calculates EUE and Expected Overgeneration (EOG), both hourly and within-hour (denoted as EUE_{WH} and EOG_{WH}), resulting from an insufficiency in flexibility that prevents the system from meeting all upward and downward ramping requirements caused by the variability, uncertainty and diurnal patterns of variable resource output.

REFLEX calculates these metrics using a stochastic unit commitment and dispatch model that simulates performance across a broad range of load, wind, solar, and hydro behavior, and resource outage conditions, through Monte Carlo draws of operating days. Costs of flexibility violations including EUE and EOG on the hour and within hour are included in calculating the lowest total system operating cost, allowing assessment of the relative value of resource portfolios or flexibility solutions in meeting future capacity and flexibility needs.

2. Uses of Approach or Model

The REFLEX approach can be used to perform the following analyses:

* Calculate "pure capacity" reliability performance metrics including LOLP, LOLE, EUE, Planning Reserve Margin (PRM), Effective Load-Carrying Capability (ELCC), and Resource Adequacy (RA) requirements (using the RECAP Model).
* Calculate expected flexibility violations for a given portfolio of renewable and conventional resources:
    - Hourly Expected Unserved Energy (EUE);
    - Within-hour Expected Unserved Energy (EUE_{WH});
    - Hourly Expected Overgeneration (EOG); and
    - Within-hour Expected Overgeneration (EOG_{WH}).
* Calculate the cost of expected flexibility violations given exogenously-specified values VUE, VUE_{WH}, VOG, and VOG_{WH} that are applied to the flexibility violation metrics EUE, EUE_{WH}, EOG, and EOG_{WH}.
* Identify the importance of each of the many potential flexibility constraints that could cause flexibility violations under some conditions. These include:

- Upward ramping capability on multiple time scales: 5 minutes, 20 minutes, 1 hour, 3 hours, 5 hours;
- Downward ramping capability on multiple time scales: 5 minutes, 20 minutes, 1 hour, 3 hours, 5 hours;
- Minimum generation levels;
- Start time;
- Shut-down time;
- Minimum run times;
- Minimum down times.

* Calculate the expected quantity of renewable output curtailment given exogenously-specified values VUE, VUE_{WH}, VOG, and VOG_{WH}.
* Calculate the value of potential flexibility solutions at avoiding flexibility violations. Potential solutions include:
  - Conventional resources (various technologies);
  - Fast-ramping resources (various technologies);
  - Conventional (downward) demand response;
  - Advanced (upward and downward) demand response;
  - Energy storage resources (various technologies);
  - Improved forecasting;
  - New transmission to external markets;
  - Market design changes such as shortened scheduling and commitment windows; and
  - Many others.

3. Description of REFLEX Model Approach

3.1. ESTIMATING NEEDS AND REQUIREMENTS

"Pure capacity" needs and requirements are calculated using RECAP, a conventional LOLP model. Pure capacity is added until the system performance meets or exceeds a user-specified benchmark such as one loss of load event in ten years.

Rather than estimating exogenous *requirements* or needs for the load following and regulation ancillary services [1], REFLEX calculates flexibility *violations* endogenously as part of the methodology. Violations occur when the system is unable to meet upward or downward ramping demands, on either the hourly or within-hour level, due to insufficient flexibility. Outputs from multiple REFLEX runs can then be used to develop guidelines or requirements for the composition of the conventional resource portfolio, similar to how LOLP studies today are used to inform the determination of an appropriate PRM. Since renewable curtailment is a critical strategy for maintaining reliable operations at high penetration, solutions can be assessed for their value in avoiding renewable curtailment.

3.2. REPRESENTATION OF UNCERTAINTIES

REFLEX considers a range of stochastic variables including load, wind production, solar production, hydro availability, and resource outages. This is done through Monte Carlo draws of operating days. To enforce relevant correlations between the stochastic variables, days are binned into low load, medium load and high load day types for both weekdays and weekends for each month. Loads are drawn first, and wind and solar shapes are then drawn from the same day type bin. Hydro is assumed to be independent of load, wind and solar.

Multiple years of data are included in the library of operating conditions. Neural network regression is used to develop a long time series (30+ years) for load based on recent

[1] Regulation is generally defined as generation responding to energy imbalance on a scale of 4 seconds - 5 minutes. Load following resources are those that meet deviations between the hourly generation set points and the regulating resources.

conditions. For wind and solar, the maximum available data is used. For energy limited resources such as demand response and hydro, the model enforces daily energy budgets, Pmin and Pmax constraints, and maximum upward and downward ramp rates. Daily hydro energy use over the past 30 years populates draws for the daily hydro budget.

3.3. COMMITMENT AND DISPATCH DECISIONS

Once an operating day is drawn, REFLEX utilizes a mixed-integer program, optimal unit commitment and economic dispatch algorithm (based on PLEXOS or ProMax) to minimize operating costs throughout the day. An additional day is drawn both before and after the measured day in order to enforce reasonable start and end conditions. Forecast error and variability are incorporated at the day-ahead and hour-ahead timesteps to inform unit commitment decisions through the use of surfaces that estimate EUE and EOG as a function of the MW and MW-min. committed. The surfaces are incorporated into the optimization by adding two new terms to the cost function in the model objective: VUE_{WH} * EUE_{WH} and VOG_{WH} * $EOG*_{WH}$.

3.4. TRANSMISSION

To date, REFLEX has been run for single zones with no internal transmission constraints. External interties are assigned Pmin, Pmax, and multi-period maximum upward and downward ramp rates based on historical data. Internal transmission constraints can also be enforced if they are important determinants of flexibility constraints, and if run times allow for the inclusion of these constraints while ensuring statistically robust sampling of operating days.

4. Inputs and Sources of Data

REFLEX requires the same types of inputs required by other production simulation models , with some additional data to ensure appropriate representation of "tail events" for stochastic variables:

* Conventional resource operating parameters (heat rates, ramp rates, start times, minimum run times, minimum down times, fuel costs, start-up costs).
* Demand-side resource characterizations (notice time, number of hours per call, number of calls per month/year).
* Historical hydro resource availability data.
* Historical weather data.
* Hourly load data from the Balancing Authorities.
* Minutely load, wind and solar data (to construct the within-hour flexibility surfaces) from the Balancing Authorities.
* Hourly wind and solar profiles.
* Societal value parameters for Value of Unserved Energy (VUE) and Value of Overgeneration (VOG), both hourly and within-hour.
* Capital costs and O&M costs for new resources.

5. Model Outputs

Core RECAP model outputs are:

* Loss of Load Probability (LOLP);
* Loss of Load Expectation (LOLE);
* Loss of Load Frequency (LOLF);
* Expected Unserved Energy (EUE) resulting from an insufficiency in "pure capacity"; and
* New resources needed (in MW) to achieve benchmark reliability performance.

Core REFLEX model outputs from a given run are:

* Expected Unserved Energy (EUE) resulting from an insufficiency in flexibility;
* Within-hour Expected Unserved Energy (EUE_{WH});

* Expected Overgeneration (EOG);
* Within-hour Expected Overgeneration (EOG_{WH}); and
* Total value of flexibility violations in each of the above categories.

Many additional, diagnostic outputs can be generated:

* Maximum upward and downward ramping demands on multiple time scales (5 minutes to 5 hours);
* Maximum unserved energy and overgeneration by hour;
* Total available upward and downward ramping capability for each hour, on multiple time scales (5 minutes to 5 hours);
* Load following reserve procurement during each hour;
* Day-ahead operating reserve committed for each day;
* Flexibility statistics by season, month or day type.

Cost metrics for candidate solutions can be investigated through sequential REFLEX model runs:

* Total reduction in each type of flexibility violation, in MWh;
* Total reduction in flexibility costs for each type (in $);
* Total reduction in fuel, emissions, and O&M costs (in $);
* Benefit-cost ratio, calculated as the reduction in flexibility and other system costs divided by the annualized capital and variable costs of the solution.

6. Sample Results

The following charts and tables show results from the E3 study "Inv estigating a Higher Renewable Portfolio Standard for California".

Table 1: Reliability statistics related to capacity need for 2030 RPS scenarios

	LOLF (events/year)	LOLE (hours/year)	EUE (MWh/Year)

33% RPS	0.15	0.29	371
40% RPS	0.09	0.16	245
50% RPS Large solar	0.07	0.12	193
50% RPS Diverse	0.03	0.04	50

Table 2: Resource need/surplus to achieve 1in-10 LOLF in 2030 (MW)

Scenario	Resource need/(surplus) (MW)
33% RPS	615
40% RPS	(150)
50% RPS Large Solar	(762)
50% RPS Diverse	(2764)

Table 3: Incremental 2030 renewable resource additions and contribution to resource adequacy for 40% RPS and 50% RPS scenarios

	RPS Installed Nameplate Capacity (MW)	RPS Resource Adequacy Contribution (MW)	ELCC of Incremental wind and solar PV
33% RPS	28,544	11,292	40%
From 33% to 40% RPS	8,332	765	9%
From 40% to 50% Large Solar	11,904	612	5%
From 40% to 50% Diverse	8,194	2,614	32%

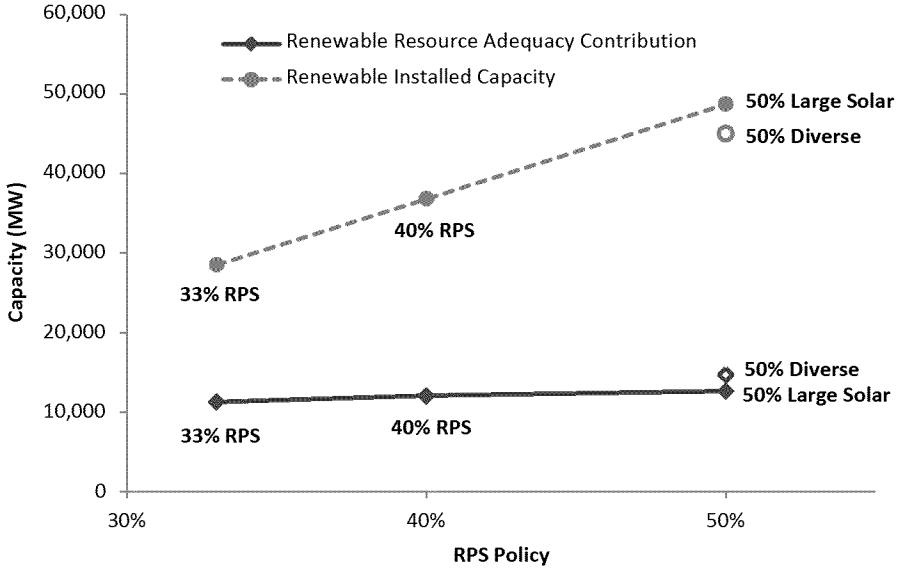


Figure 1: Renewable nameplate capacity and resource adequacy contribution for the 33% RPS, 40% RPS, 50% RPS Large Solar 50% Diverse Scenarios

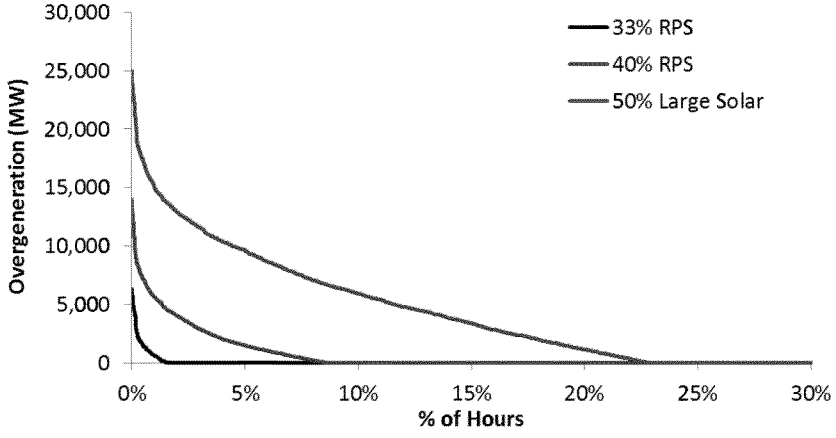


Figure 2: Duration curves of overgeneration events in 2030 RPS scenarios

Table 4: Overgeneration statistics for the 33% RPS, 40% RPS and 50% RPS Large Solar Scenarios

Overgeneration Statistics	33% RPS	40% RPS	50% RPS Large Solar
Total Overgeneration			
GWh/yr.	190	2,000	12,000
% of available RPS energy	0.2%	1.8%	8.9%
Overgeneration frequency			
Hours/yr.	140	750	2,000
Percent of hours	1.6%	8.6%	23%
Extreme Overgeneration Events			
99th Percentile (MW)	610	5,600	15,000
Maximum Observed (MW)	6,300	14,000	25,000

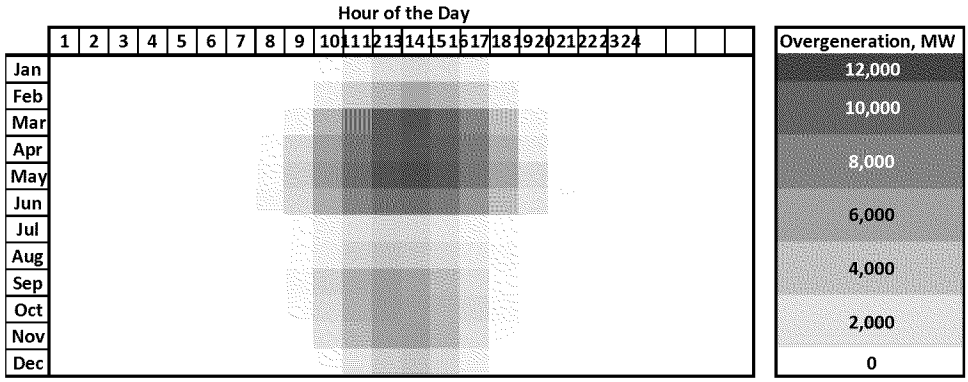


Figure 3: Average hourly overgeneration in the Study Area that must be mitigated by month-hour in the 2030 50% RPS Large Solar Scenario

Table 5: Marginal overgeneration (% of incremental MWh resulting in overgeneration) by technology for various 2030 RPS scenarios

Technology	33% RPS	40% RPS	50% RPS Large Solar	50% RPS Diverse
Biomass	2%	9%	23%	15%
Geothermal	2%	9%	23%	15%
Hydro	2%	10%	25%	16%
Solar PV - Large	5%	26%	65%	42%
Wind	2%	10%	22%	15%

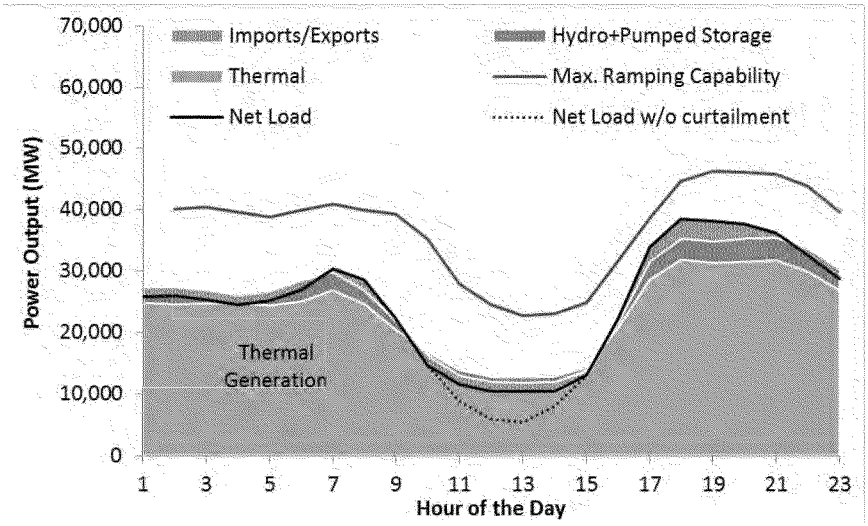


Figure 4: Conventional fleet performance and flexibility on the representative day with the largest net load ramp

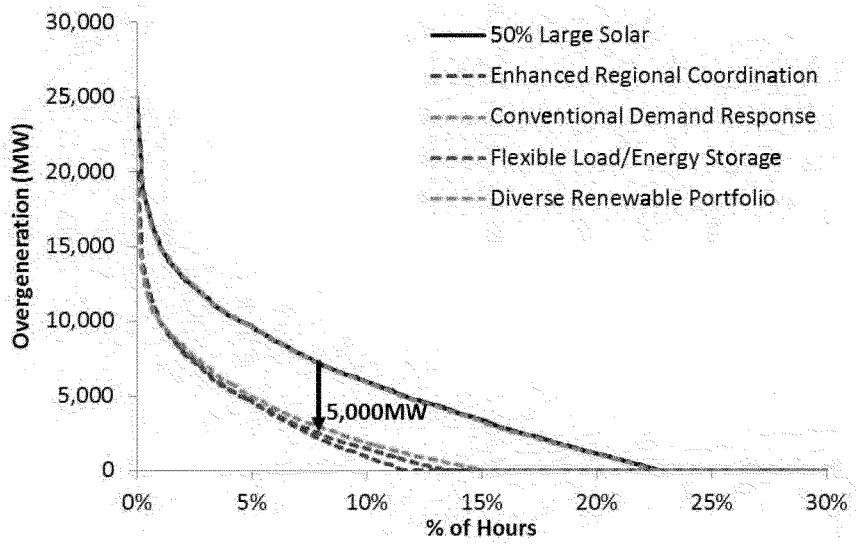


Figure 5: The effect of each solution category on overgeneration for the 50% RPS Large Solar Scenario

7. Known Limitations

7.1. PROCESSOR SPEED

REFLEX utilizes a mixed-integer programming (MIP) approach to calculating optimal unit commitment and economic dispatch. This is necessary in order to accurately characterize the limitations on the ability of large generators to start, stop and ramp. MIP is processor-intensive, which means that REFLEX runs may require lengthy computation times. Ensuring that a sufficiently broad distribution of system conditions is generated may require eliminating calculations and constraints that have little impact on the final solution. For this reason, REFLEX runs conducted to date for California have not considered internal transmission constraints. Other simplifications might include simplification of heat rate curves or aggregation of multiple generating units into a single range that can be represented linearly.

The simplifications that are required to ensure acceptable run times will be different for each system. Smaller systems with fewer resources will likely require fewer simplifications due to the reduced number of calculations required.

7.2. DATA AVAILABILITY

REFLEX's conclusions depend on accurately characterizing the full distribution of operating conditions the system is likely to face. This includes both low and high load conditions, low- and high hydro conditions, low and high wind and solar generation. Wind data availability has been relatively limited thus far; the National Renewable Energy Laboratory published three years (2004-2006) of estimated wind production data as part of its Western Wind and Solar Integration Study, however, this dataset has known limitations such as its characterization of coastal wind regimes. Solar output profiles are available that span a longer historical record (1998-2009 through Solar Prospector®). NREL is in the process of preparing additional wind datasets for release in the near future.

7.3. AVAILABILITY OF RAMPING CAPABILITY OVER EXTERNAL INTERTIES

Limits on the availability of ramping capability over external interties stem from a variety of factors including flexibility constraints in neighboring regions, transmission constraints, and the manner in which wholesale power is transacted in the Western Interconnection (e.g., bilateral trading of heavy-load hour (6 AM – 10 PM, Mon.-Sat.) and light-load hour (all other hours) blocks of power). REFLEX utilizes historical information to develop Pmin, Pmax and maximum upward and downward ramp rates over periods from 1 to 9 hours in length for external interties; however, future intertie use may be different from the past.

7.4. AVAILABILITY OF RAMPING CAPABILITY FROM HYDRO GENERATION

Hydro generation is very difficult to model accurately given operational constraints related to river flows and reservoir elevations that are specific to each project, as well as the fact that multiple projects are frequently arranged in a sequential, cascading fashion. REFLEX utilizes historical information to develop daily energy budgets, Pmin and Pmax values, and maximum upward and downward ramp rates over periods from 1 to 9 hours in length for hydro facilities; however, future hydro performance may be different from the past.

8. Planned Improvements

E3's plans for technical model improvements largely are centered around increasing the value of the analysis that can be done subject to computing power limitations. These include, specifically:

* Investigate appropriate convergence criteria;
* Develop smart sampling techniques to increase the sampling power.

We are also interested in conducting additional studies to understand how sensitive our early results are to a number of potentially important parameters. These include:

* Testing alternative RPS levels and RPS portfolios;
* Testing alternative configurations of the fleet of thermal resources (i.e., changes in once-through cooled plants, alternative CCGT, CT and IC technologies);

* Develop zonal models to test whether flexibility issues could be more significant in local areas, particularly SP26 which is expected to have a higher renewable penetration than NP26;
* Test the effectiveness and processor requirements for conducting 5-minute dispatch in place of the within-hour flexibility surfaces (surfaces would still be used for commitment decisions, but EUE and EOG would be calculated from a 5-minute dispatch rather than directly from the surfaces);
* Develop more robust estimates of the potential for California to export power during overgeneration conditions;
* Test changes to intertie ramping constraints;
* Develop techniques to simulate the effect of the California ISO – PacifiCorp EIM;
* Develop detailed studies of the effect of dispatchable wind, solar and geothermal resources – how fast could the resources ramp and what would be the implications for regulation and load following requirements; and
* Review data regarding conventional resource operating characteristics to ensure proper and consistent treatment of start times, minimum run times and minimum down times, startup costs, ramp rates, etc.

Finally, we are interested in conducting additional studies to increase understanding of the solutions and their interactions with each other, such as:

* Increased regional coordination
* Renewable resource diversity
* Flexible loads
* Flexible generation
* Energy storage

I. SERVM Introduction

Many electric simulation tools fall into one of two categories:

- **Resource Adequacy Models -** Traditional resource adequacy models are designed around an architecture that allows for fast simulation of thousands of iterations of unit performance, weather conditions, and other stochastic variables. Resource adequacy models are ideally suited for identifying the frequency and magnitude of firm load shed events or other events that occur during resource constrained periods. Emergency operating procedures and energy limited resources are modeled to very fine levels of granularity to capture their impact on resource adequacy. In order to perform fast simulations, these tools frequently make simplifications on the economic considerations used for commitment and dispatch or they ignore economic considerations altogether. In summary, resource adequacy tools are ideal for understanding the frequency and magnitude of high-impact, low-probability events such as firm load shed.
- **Production Cost Models -** Production cost models are designed around providing more detailed assessment of the economics of electric generation. These models typically take into consideration more detailed unit variables including operational constraints and cost components. Unit commitment algorithms that take into account minimum uptimes, minimum downtimes, startup times, ramp rates, complex heat rate curves, ancillary service capabilities, and other variables require significant computation time.

While there are many other features of the two classes of models, the above items are of significant interest for assessing the impact of flexibility constraints on a system. The fact that firm load shed is, or should be, a very infrequent occurrence requires that the system be tested under a wide range of conditions to assess its capability to meet net load shapes while respecting all unit constraints. Resource adequacy models provide the capability to assess a wide range of conditions, but may not be able to adequately consider unit constraints. Production cost models are limited by processing time requirements when attempting to consider a wide range of conditions, but provide robust capabilities for considering unit constraints.

SERVM was initially designed as a hybrid resource adequacy and production cost model by the Southern Company in the mid 1980's. It has been in continual enhancement since that time to provide the full range of capabilities of both classes of models. The designers of SERVM recognized that resource adequacy can be greatly influenced by economics, and that the economics of the system can be greatly influenced by the high-impact/low-probability type of events considered by resource adequacy models. SERVM performs a full economic commitment on a weekly basis taking into account relevant unit variables as well as short-term load and resource forecast error. As system conditions materialize in the simulation, SERVM performs updates to the commitment on various time frames. If shortages or unit outages occur, SERVM has access to resources consistent to the opportunities that a dispatcher would

have in similar conditions. Price, energy constraints, reliability requirements, and ancillary service requirements can all be considered to perform remedial actions to maximize reliability and minimize cost.

A typical implementation of SERVM includes performing hourly chronological simulations for the full 8760 hours in a year for the following combination of discrete variables:

- **30 Distinct Load Shapes -** Each load shape is derived from the application of a neural network model (containing the weather/load relationship of a given system) to the actual weather conditions in a historical year. The load shape has 8760 consecutive hourly load points for each region being modeled. Some years will have more extreme weather conditions than other years, resulting in more extreme load conditions. Some years will reflect more or less diversity amongst load shapes in neighboring regions.
- **6 Load Forecast Error Points -** Load can grow faster or slower than expected during the long-term procurement planning process. SERVM uses input probabilities of load forecasting error to represent this uncertainty. The hourly load for each weather shape is multiplied by 6 distinct load forecast error points to create 180 distinct cases. Each of these cases will be simulated independently.
- **100 Unit Performance Draws -** Forced outages, partial outages, common mode outages, and start-up failures can occur randomly. SERVM simulates these events stochastically. Fifty full 8760 simulations are performed for each of the above mentioned cases to capture the variation in unit performance that can occur over a year. In addition to unit performance variation, other variables can be treated stochastically, including renewable output, transmission availability, and short term load forecast error.

Economic Commitment and Dispatch in SERVM

SERVM performs a weekly commitment for each region being studied using a proprietary dynamic programming technique. The large scale optimization problem to commit adequate generation to meet the full load and ancillary service requirements for every hour is broken into a number of sub-problems. The first sub-problem includes meeting load for every hour up to the minimum load of the week. For this problem, minimum up-time and down-time and start-up time constraints can be ignored or relaxed. The next sub-problems are then set up to meet remaining unserved load. For each subsequent sub-problem up to the final sub-problem which fully meets load plus operating reserve requirements, the unit constraints become more critical and all relaxations are progressively dismissed. The selection of resources to optimally meet the need in each sub-problem is performed using a proprietary indexing technique.

Results are saved from each weekly commitment for use in an evolutionary algorithm to adjust the commitment for subsequent iterations. Optimality is tested and adjustments are made to the commitment through the use of phantom load and generation variables to further refine the commitment in each subsequent iteration. This evolutionary algorithmic approach also allows for optimal commitment among zones that satisfies import/export constraints.

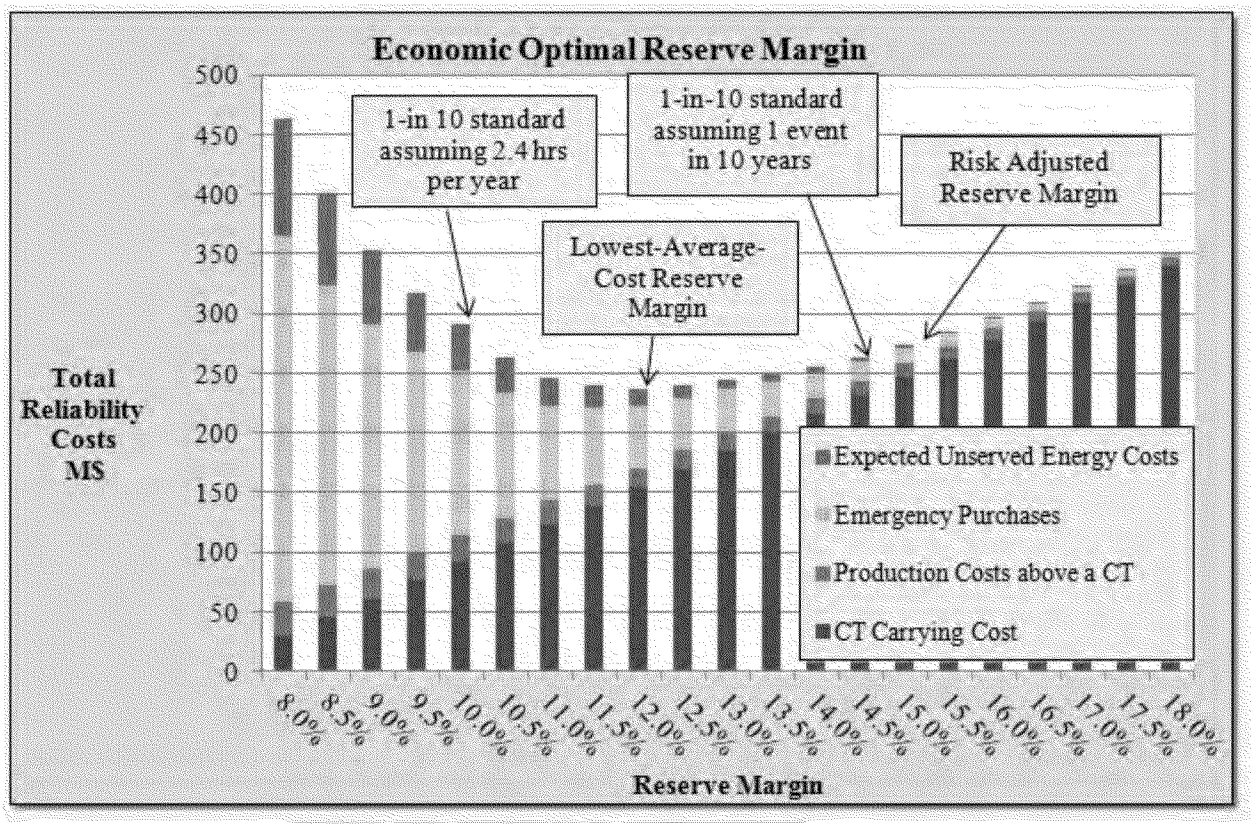
Each hour in the simulation, peaking resources are used to modify the commitment in the event of unexpected unit outages or load forecast error. Also in each hour, SERVM looks four hours ahead to identify needed changes to baseload or intermediate resource commitments.

Since the commitment algorithm will result in different magnitudes of operating reserves each hour, an economic dispatch routine is used every hour (and intra-hour if necessary) to identify the exact operating point, ancillary service contribution, and ramping capability of each unit.

II. Overview of Applications of the SERVM Modeling Approach

A. Identification of Planning and Resource Adequacy Requirements

SERVM has been adopted by a number of utilities and regulators for identifying appropriate reserve margin targets based on both economic and reliability criteria. From a reliability standpoint, SERVM can identify the reserve margin that provides a targeted level of reliability such as the 1-day-in-10 LOLE. From an economic standpoint, simulations at various reserve margins can identify the trade-off between paying for more capacity and the economic benefit provided by that capacity in terms of reduced production costs, market purchases, and unserved energy.



*Data in the figure above is illustrative only.

B. Estimate the Contribution of Resources to Reliability and Flexibility Requirements

SERVM has been used to identify the reliability and economic contribution of a wide range of technologies including wind, solar, hydro, pumped storage, energy storage, demand response, and conventional thermal resources with various operating constraints. Typical operating constraints evaluated include storage capacity, contractual or environmental constraints (emissions limits or max calls per day, month, year, etc.), ramp rates, temperature restrictions, minimum uptimes, minimum downtimes, startup times, fuel availability, outage rates, and intermittency of wind/solar resources.

The approach used to evaluate the effect of these constraints includes the following steps:

1. Model a base case.
2. Model change cases with incremental unconstrained capacity.
3. Model change case with the same incremental nameplate capacity of a resource having the particular constraints or operating characteristics being evaluated.
4. Calculate the ratio in reliability benefit (as measured in unserved energy or loss of load expectation) provided by the two incremental resources. This ratio is used to define the reliability contribution of the resource being evaluated.

C. Estimate Amount and Operating Characteristics of New Resources to Meet Identified Need

Astrape believes that the correct approach to any least-cost/best-fit portfolio construction is to perform the needs assessment and the resource selection in a single step. Many modeling approaches first identify the magnitude of the need in an initial set of simulations, and then separately assess the economic efficiency with which different types of resources meet that need. However, we have found that the magnitude of the need is often a function of the economics of the type of resource being used to serve the need.

For example, the optimal reserve margin studies mentioned in Section A can be performed with multiple types of resources; different resource types may identify a different economic optimum reserve target. If the capacity cost of a particular resource class is very inexpensive, the optimal resource plan may mean selecting more resources than necessary to achieve a previously identified reserve margin target. SERVM can be used to perform this integrated assessment not only for various conventional resources, but also for transmission improvements and fuel supply strategies.

A similar philosophy can apply to multiple types of studies. For demand response procurement, instead of identifying a target block of generic capacity and seeking to fill it with generic demand response resources, optimization of a range of demand response programs and constraints may lead to a varying range of demand response options (each with different cost) depending on the characteristics of the demand response resources. Similarly, flexibility needs can be met in a variety of ways:

- Selection of more resources
- Selection of different types of resources

- Changes to operating procedures
    - Overcommit conventional generation and curtail additional renewable generation
    - Improve scheduling flexibility
    - Develop reserve sharing agreements

The economic and reliability-related viability of pursuing solutions across the range of possible opportunities could be assessed through a series of co-optimizations that include various combinations of the options above. Since needs will vary based on the type of solution being tested, SERVM simply uses a reliability target or an economic metric as an objective function and allows the results of the simulations to determine not only the type, but also the magnitude of the ideal solution. Separate needs assessments and economic assessments are not necessarily needed.

III. Detail of SERVM Modeling Approach

A. Model Inputs

Most of the model inputs and sources are described in detail in the 'Probabilistic Reliability Modeling Inputs and Assumptions - Part One' document listed on the following website, developed as part of a probabilistic reliability modeling initiative at the California Public Utilities Commission:
http://www.cpuc.ca.gov/PUC/energy/Procurement/RA/Probabilistic+Modeling.htm

However, for this description, we will summarize some of the salient inputs that have significant impact on flexibility modeling.

1. **Load –** 30+ load shapes are developed using historical weather and current or projected weather/load relationships. This effort will identify the magnitude and variability of loads (average peak load is typically scaled to equal the forecasted peak load). In addition, diversity of loads between regions will also be quantified in this effort.

2. **Demand-side Resources –** Program definitions with capacities, contractual constraints, and dispatch rules are defined in the model. Typical constraints considered include hours per day, hours per month, hours per season, hours per year, days per week, and call duration. Dispatch rules include order in emergency operating procedures or dispatch shadow price. Demand side response magnitude can also be modeled as a function of price or load, or can be modeled as a distinct profile.

3. **Supply-side Resources –** Supply side resources are broken into the following categories:

    a. Fossil – This category contains conventional resources with start times greater than one hour.
    b. Nuclear – These resources are modeled as must-run in all periods.
    c. Turbines – This category includes resources with start times of one hour or less.

d. Renewable – Renewable generators whose output is dependent on weather patterns; these resources are non-dispatchable and are not economically triggered.
e. Pumped Storage – This category contains all energy storage resources.
f. Hydro - Hydro facilities that are not pumped storage; they are modeled as one of three subtypes – emergency, scheduled, or run of river.

4. **Capital and Operating Costs –** Capital costs are considered outside of the SERVM model. Enhancements are planned to incorporate capital cost considerations into the SERVM framework. Variable operating costs are fully modeled in SERVM including fuel costs, start-up costs, emission costs, and O&M costs.

5. **Hydro Availability –** Hydro resources can be modeled in one of four available categories, allowing for significant flexibility in capturing environmental, economic, weather, ancillary service contribution, and load considerations in dispatching hydro. Hydro energy and dispatch data is input, corresponding to the 30+ weather years and reflects variation in historical rainfall patterns. Input variables include capacity and energy by month as well as minimum and maximum flow information on hourly, daily, weekly, or monthly bases. Also, for emergency hydro resources, emergency operating procedure dispatch rules can be input. Hydro resources are typically aggregated based on zones or river systems. Inputs can be calibrated to reflect actual historical aggregate output or ancillary service contributions.

6. **Transmission modeling –** SERVM utilizes a pipe-and-bubble representation of the transmission system. Currently up to 300 zones can be defined, but the number of zones defined directly affects processing speed; as a result, it is typically best to model up to approximately twenty zones. Imports and export constraints can be defined between any zones, and the architecture allows for nesting of internal zones.

   SERVM can also produce input files for power flow models such as PSS/E to test that the transfers forecast by SERVM are reasonable.

7. **Ancillary service requirements –** All of the following inputs can be defined as a fixed amount by hour of day, month, or year, or they can be defined as a function of load (or as a combination of both). Most of these components can be set as firm requirements (shed firm load to maintain), or they can be enforced with varying degrees of firmness.
   a. *Regulation up requirements* - SERVM can either shed firm load to maintain regulation or dip into regulation requirements if necessary to avoid shedding firm load.
   b. *Regulation down requirements* - The model uses the regulation down requirements and an input cost of curtailment as part of the commitment algorithm to minimize total production costs. To the extent forecast uncertainty and volatility affects the actual ability to meet regulation down requirements, remedial actions can be performed to restore regulation down. One of the remedial actions is curtailment of generation.
   c. *Spinning reserve requirements* – The model can either shed firm load to maintain spinning reserve requirements or dip into spinning reserve requirements if necessary to

avoid shedding firm load. Spinning reserves can always be depleted partially for the 90 minutes after a contingency.

d. *Spinning reserve target* - This category represents load following reserves. These reserves will be procured subject to incremental resource dispatch cost being less than the coincident market price.
e. *Non-spinning reserve requirements* - Spinning reserves or quick start units can serve this category. If this category is not set to impose firm load shedding, it is used simply for reporting reserve shortages.

B. Simulation Process

1. Case Setup - Each combination of weather year and economic load growth uncertainty defines a unique case. In addition to selecting all the global data that applies in every case (conventional resource definitions, transmission constraints, operating procedures), SERVM selects all information that corresponds to the selected weather year. This will include for every region the following:
    a. Load shapes
    b. Wind/Solar/Conventional output profiles
    c. Hydro capacities and energies
2. Commit run of river and scheduled hydro blocks - Scheduled hydro shaves peak loads subject to input constraints and run of river reduces load at equal level all hours of the day.
3. Schedule planned outages during low load periods
4. Perform simulation for each iteration
    a. Schedule fixed profile and must run resources for entire year. Also, to the extent fixed profile and must-run resources do not cover the minimum load for the year, commit additional base load resources for the entire year. This process will include forced outages and derates for all resources modeled with outage data.
    b. Initialize unit operational status. Full availability and partial outage availability inputs will be used stochastically to determine whether each unit begins the iteration available or unavailable. Depending on the starting status, a time to fail or time to repair value will be selected
    c. Perform weekly commitment -
        i. First, Monte Carlo techniques are used to draw uncertainty values from uncertainty distributions for load, solar, and wind resources. The uncertainty values drawn represent day ahead uncertainty from historical forecast error. Each category of uncertainty values uses specific parameters to ensure appropriate uncertainty is selected.
            1. For load uncertainty, the actual daily peak load as a percentage of normal annual peak load is a parameter for selection. Higher load periods have different load forecast uncertainty characteristics than low load periods.

2. Wind uncertainty utilizes the actual wind output at every hour as a parameter for selecting an appropriate uncertainty value. If the actual wind is at full output, only under forecast of wind is possible. If actual wind is at zero output, then only over forecast is possible. A complete range of historical actual wind and forecast wind are used to populate the input distribution.
3. Solar uncertainty utilizes the actual solar output compared to the maximum achievable output given perfect solar conditions at every hour for selecting appropriate uncertainty. A "blue-sky day" profile as well as forecast uncertainty distributions based on percentage of blue-sky day output must be input.

ii. Adjust the remaining expected load by uncertainty values. This will be the net load used by the commitment algorithm. Note that during commitment, SERVM assumes units that are on outage will remain on outage and units that are available will remain available. As units fail or return to service, the commitment will change.

iii. A dynamic programming technique is used to develop a commitment that fully meets the remaining expected net load and ancillary service requirements while respecting all unit constraints. The algorithm is described in more detail in the SERVM introduction. The commitment developed in this section will be enforced for resources with start-up times of greater than one hour. However, while resources with startup times of one hour or less are used to bound the commitment for the longer-lead resources, their commitment from the week ahead commitment is not enforced. They will actually be committed in a separate commitment algorithm that is performed each hour.

d. Simulate each hour

i. Adjust commitment based on updated load and resource forecasts. Adjustment is based on identified need over the subsequent 4 hour period, but can affect commitment over balance of week if incremental resources provide value.

ii. Change unit statuses

1. If a unit will fail (or go to partial outage or maintenance outage or planned outage) this hour, identify intra-hour timing of failure. Also, adjust commitment for each failure.
2. Start units scheduled to start
3. Begin shutdown of units scheduled to shut down

iii. Use resources committed plus available short-lead resources to calculate energy and ancillary service prices for each region

iv. Subject to import/export constraints and hurdle rates, transfer power between regions to balance market prices.

v. Commit short-lead resources. This will include CTs, demand response resources, and any emergency operating procedures.

vi. Simulate each intra-hour period

1. Implement unit failures as identified at the beginning of the hour for this period.
2. Dispatch all resources to economically optimal levels subject to ancillary service requirements, ramp rates and other constraints.
3. If shortages or overgeneration are present, perform remedial actions to restore. Remedial actions include re-committing short-lead resources, shedding firm load, or curtailing generation.

vii. Consolidate metrics and decrement counters

e. Report iteration-specific metrics

5. Report consolidated metrics

C. Simulation Performance

SERVM typically performs a single 8760-hour simulation on an hourly interval for a 15-region system with 2000 generators in approximately 1 minute. Performing the simulations on a 5 minute interval increases processing time to approximately 3 minutes. Cases can be spread across multiple cores on multiple machines to assist in processing times for large studies. The speed at which SERVM can complete an 8,760 simulation and provide accurate commitment and dispatch allows for thousands of iterations to be examined in a short period of time.

IV. Typical Model Outputs

Most metrics can be reported hourly, monthly, annual, or iteration-specific, and either by region, by unit, or system-wide.

a. Peaking capability deficiency
    i. MWh of EUE, LOLH, LOLE
b. Upward ramping capability deficiency
    i. MWh of Expected Unserved Ramping Energy, Loss of Ramping Hours, Loss of Ramping Days or Events
c. Downward ramping capability deficiency
    i. Curtailment MWh, Curtailment Hours, Curtailment Days
d. Other shortage metrics
    i. Frequency of calling demand side resources
    ii. Frequency of implementing various emergency operating procedures
    iii. Reserve shortage energy
e. Production cost
    i. Fuel burn (MMBtu & cost), O&M costs - reported by region and by unit
f. Market purchase costs
    i. Energy and prices
g. Societal costs
    i. Reserve shortage costs: Apply a cost to all MWh where the simulation is below target operating reserve levels
    ii. Unserved energy costs
h. Market price forecasts

i. Optimal mix of new resources to procure
    i. Not a direct output of the model. Production costs and reliability metrics can be compared across multiple sets of simulations with various solutions.
j. Requirements for different ancillary service categories
    i. Not a direct output of the model. Simulations can be performed with various requirements pre-defined and the economic/reliability trade-off can be compared to identify optimal ancillary service requirements.

V. Sample Results

See link to full report: http://www.astrape.com/?ddownload=934

VI. Known Limitations

- Pipe and Bubble versus full AC power flow
- Modeling concentrating solar power facilities requires fixed dispatch profiles

VII. Planned Improvements

- Inclusion of capital and FO&M costs directly in inputs and outputs of model.
- Automated expansion planning techniques
- Explore the possibility of implementing a DC transportation model; currently working with EPRI to allow SERVM and full power flow model to communicate to understand the impact of not modeling all internal transmission constraints not captured by zone topology.

Introduction

Meeting California's 33 percent renewable energy goal will introduce substantial wind and solar capacity to the system. System operators will need to manage the system under considerable day-ahead uncertainty about generation and considerable variability during the operating day. It is expected that these operating problems can be mitigated, and operating costs can be minimized through probabilistic forecasts of renewable generation and unit commitment algorithms that take into account the uncertainty in renewable generation.

The LLNL Grid Simulation Model explicitly models the day-ahead uncertainty in renewable generation using a physics based atmospheric model. It then executes a day-ahead unit commitment using a stochastic optimization algorithm that takes into account the uncertainties in the day ahead forecast. The real time dispatch is modeled on a five-minute time step using the actual realized renewable generation over the operating day. By accounting for the day ahead uncertainty and the unit commitment given this uncertainty, it is expected that this model will provide amore realistic assessment of the operation of the system, the operating costs, and the prices for energy and ancillary services during the operating day. It has been applied to estimate the value of demand response for regulation and the value of energy storage for both regulation and price arbitrage.

Uses of Model

Automated demand response and energy storage resources are inserted into the model and utilized in concert with other resources in the system to estimate the value they could provide in a system context. This value is estimated by identifying the avoided costs of the conventional hydro and fossil resources that they displace for providing regulation, load following, and energy arbitrage functions.

Approach

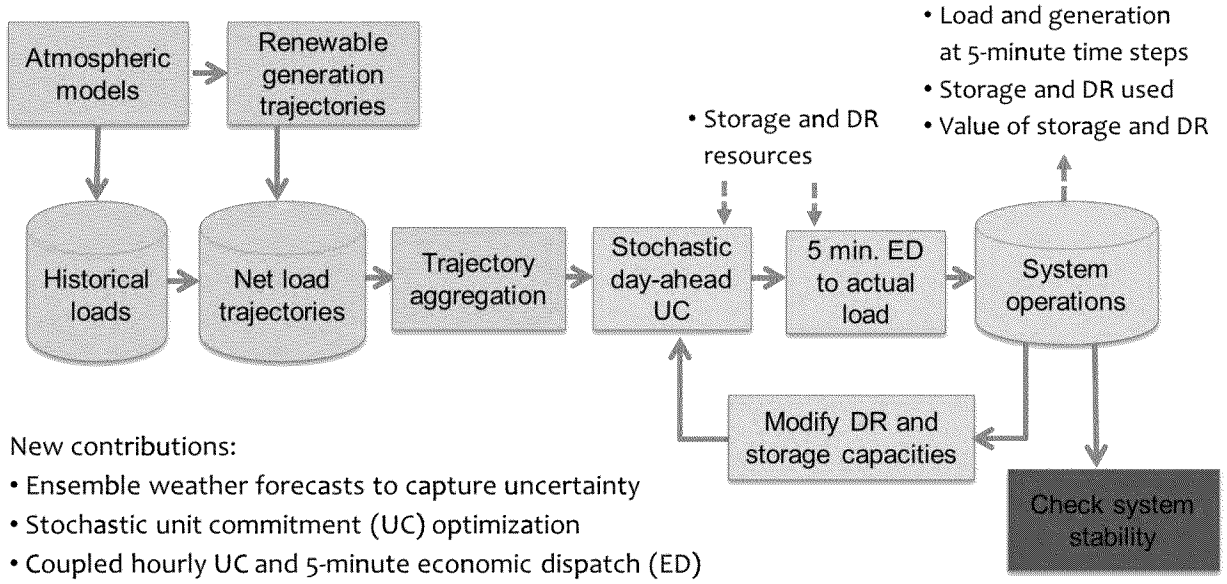
To conduct an analysis to estimate the value, a number of different models of system components were developed and coupled. These models and the overall analysis process are depicted in **Figure LLNL-1**. Three basic types of models developed and used for the analysis, indicated by the colors of the processes in the figure, are:

- **Weather Forecast and Renewable Generation Models (Blue):** Models of the weather and renewable generators that produce trajectories of renewable generation that are subtracted from gross load to get the net load that must be met by dispatchable resources.
- **Stochastic Production Simulation Model (Yellow):** Probability-based production simulation model that produces usage, revenues, costs, and prices for all resources and services. This includes a process loop that allows for repeated production simulation

runs with different levels of storage and demand response capacity. This loop is used to estimate the marginal value of additional increments of capacity.

- **Electromechanical Simulation Model (Green):** Model that simulates regulation resources and checks the controllability of the system to ensure that it meets electrical engineering standards.

Figure LLNL-1: Renewable generation, production simulation, and resource evaluation process



As indicated in the figure, the analysis approach incorporates three new capabilities:

- **Ensemble weather forecasts with uncertainty** – Scenario specific uncertainty in the weather forecast is based upon variation in atmospheric physics sub-models applied to weather conditions for that particular day. This approach is in contrast with current practice of using a single forecast for each day and one uncertainty value for each season that is an average from historical data.
- **Stochastic unit commitment optimization** – The production simulation model minimizes expected cost taking into account the entire ensemble of possible renewable generation trajectories produced by the weather model. Current practice is to minimize cost for a single trajectory, and to add safety factors at additional costs.
- **Coupled hourly and 5-minute timescales** – The production simulation model utilizes two different timescales for the unit commitment and economic dispatch to perform the optimization.

As indicated in the upper left portion of the figure, at the beginning of each day in the year first principles physics-based atmospheric models are used to develop a set of renewable generation trajectories and to perturb historical loads. The atmospheric model incorporates the initial state of the atmosphere at the beginning of the day, and models the physics of the atmosphere throughout the day. The model produces wind velocity and solar insolation data that are

passed to models of renewable resources. A key uncertainty in the forecasts is the nature of the detailed physics behaviors that will dominate during the day. Accordingly, the model provides an ensemble of 30 possible renewable generation trajectories each day by incorporating 30 different configurations of physics sub-models in the atmospheric simulation model. These sub-models represent the propagation of solar and terrestrial radiation throughout the atmosphere, the microphysics of cloud formation and evolution, turbulent mixing, and land surface processes. The Weather Research and Forecasting (WRF) code, an open source code maintained by the National Center for Atmospheric Research, was used to develop the model.

As indicated in the figure, the atmospheric forecast models also influence the forecasted load. To provide consistent pairs of renewable generation and corresponding load trajectories, historical loads must be adjusted for the difference in temperatures between the realized and hypothetical weather trajectories. After this adjustment is made to the loads, renewable generation is subtracted from load for each trajectory to obtain an ensemble of 30 net load trajectories indicated in the figure. This ensemble of 30 trajectories captures the day-ahead uncertainty in the weather and renewable generation, including trajectories with high net load ramp rates that tend to stress the system. Finally, a trajectory aggregation process has been developed and used to ease the computational burden associated with the stochastic day-ahead unit commitment optimization algorithm.

After the trajectory aggregation step, a stochastic production simulation model is formulated using standard modeling software and data sets developed by California Independent System Operator. Probability-weighted load and renewable generation trajectories are used as input to the production simulation model. A stochastic optimization algorithm finds the unit commitment schedule that minimizes expected cost for the net load trajectories. As indicated in the figure, the code then performs economic dispatch at five-minute time steps over the operating day using the trajectory of weather and loads that were actually realized over the day. This process includes commitment and dispatch of demand response and storage resources, as indicated in the figure. The PLEXOS production simulation model developed by Energy Exemplar, LLC was used to conduct the analysis.

The solution generated by this process provides detailed information on the operation, costs, and revenues of demand response, storage, and all other resources in the system. It also provides estimates of the prices of regulation and load following services at each five-minute interval during the year.

Finally, regulation requirements and system stability studies were conducted. The software used was based on the KERMIT package, developed by DNV-Kema Corp. Some of the functionality in the Matlab-based code KERMIT were re-implemented in C++ to increase throughput to levels needed for this project, in coordination with DNV-Kema.

Over 3,000 days were simulated using this process under various sets of assumptions. Running the WRF, PLEXOS, and KERMIT models on high performance computing systems with thousands of cores allowed us to complete this three million core hour analysis campaign in a reasonable amount of time.

Inputs and sources

The Electric Power Research Institute, the California Energy Storage Alliance, and the Demand Response Research Center provided the data and assumptions describing energy storage and demand response resources that are used in the models. The California Independent System Operator provided the production simulation model and other supporting data. CAISO's High Load PLEXOS model was used for the study.

Load following requirements were derived from a combination of the expected hourly ramp and the variation in renewable generation and load forecasts, with further adjustments and minimum levels for periods of the day with high intra-hour variation, namely the morning and evening peaks. Regulation and reserve requirements were set using the values in the CAISO high load scenario.

Model outputs

The key outputs of the PLEXOS production simulation model of interest for this study were the values of Lagrange multipliers, or shadow prices, associated with constraints on energy balance, load following, and regulation. These shadow prices provided marginal values of energy, load following, and regulation resources, respectively. Using repeated runs of the simulation model with storage and demand response capacity additions, the marginal value capacity increments for these resources was established. The model also outputs the dispatch for all generators in a system and the transmission line flows between regions. The regulation simulation generates a detailed record of the control actions of a given generator in response to a second by second load pattern, which allows an assessment of the effectiveness of system control for various scenarios.

Sample results for 2022 LTPP Base Scenario

The model was developed and exercised for the planning year 2020. Loads and resources would have to be modified to address 2022 planning issues. For example, the model was developed and results were generated before the decision to retire the San Onofre Nuclear Generating Station. This resource would have to be removed from the model.

Known limitations

We believe that this model provides an accurate representation of uncertainty and variability caused by intermittent renewable generation in the day ahead markets and how the operator could manage it. However, the unit commitment and economic dispatch procedures did not model the intra-hour uncertainty caused by variations in renewable generation. It assumes perfect information for simulation of five minute economic dispatch operations. As such, it complements other modeling approaches that use statistical models to capture hourly and sub-hourly variations in loads and renewable generation, but do not incorporate numerical weather simulation features. The regulation modeling performed with the KERMIT software and its derivatives did take into account uncertainty at one minute timescales.

Planned improvements

There is currently no funding in place for modification and subsequent use of the model. Opportunities are being explored.

Appendix B: Glossary

Variability, Operating Uncertainty, and Statistical Variance

Throughout this document, references are made to both variability and uncertainty. These are important terms when considering operating flexibility, and their use can differ throughout the industry—they are often used interchangeably. Also, the modeling approaches reviewed in this report account for variability and uncertainty in different ways.

This document refers to variability as changes in the magnitude of a variable over time (e.g., load, and wind and solar generation) for which the changes are known, or assumed to be known to the model when making unit commitment or dispatch decisions. For example, variability is reflected in the changes of an hourly or minutely profiles used as inputs in a production simulation model.

Operating Uncertainty is used here to refer to changes in magnitude of input variables which are unknown when the model makes decisions. Examples of uncertainty are unknown weather, which is represented by distributions of historical weather in planning models, or probability distribution of forecast errors of load, wind and solar generation that a planning model could include in unit commitment and dispatch decisions.

Models may also account for the long-term statistical variance of conditions that the system may be exposed to, for example, a year with 50 percent of normal hydro availability and 20th percentile peak temperatures. This type of long-term uncertainty is referred to here as statistical variance, and is typically accounted for by drawing from a distribution of possible conditions over many iterations.

Operating Flexibility

The ability for the system operator to adjust supply from resources (generation or demand-side) in order to respond to changes on the system that are predicted in advance (variability) as well as changes that are not predicted in advance (uncertainty).

Local Reliability/Local Capacity Requirements

In planning "Local" refers to issues that exist within small, transmission-constrained regions within the larger system. Local capacity requirements are driven by the need to provide reliability within these "Local Capacity Requirement" areas. The models under consideration do not currently have the resolution to provide this level of analysis, rather they consider the whole CAISO system as one region with no internal transmission constraints, or they break the CAISO system into several large "Zones" or "Bubbles" that generally correspond to the California utility territories with constraints on energy transport between the zones.

Expected Unserved Energy (EUE)

Expected Unserved Energy is a probabilistic metric measuring generation adequacy, typically stated as expected MWh/year. EUE is most often calculated by completing a large number of annual simulations with randomly drawn unit outages, load levels and sometimes renewable generation. Over this large number of iterations it is possible to calculate the expected annual number of MWh of firm load that is not met. Some models also estimate MWh of unmet reserves as well.

Typically, EUE analyses assume that any resource not on outage is available to meet load and reserve requirements and do not account for operational constraints such as startup time or ramp rate; however some models, such as REFLEX, include unserved energy driven by these operational flexibility constraints.

Stage 3 Emergency

A Stage 3 Emergency is called when on-line Reserves fall below the minimum requirements (amount can vary- usually around 3 percent of the minimum Operating Reserve total). Under a Stage 3 Emergency, the Independent System Operator may call on the utilities to reduce "firm load" by implementing rotating outages. This is a last resort, used only when a climbing demand for energy is close to surpassing the available supply. (Source: www.caiso.com/Pages/AboutTodaysOutlook.aspx.)

Loss of Load Expectation (LOLE), Loss of Load Frequency (LOLF), and Loss of Load Hours (LOLH)

Loss of Load Expectation (LOLE) is a probabilistic metric measuring generation adequacy typically stated as expected "days per year." LOLE is most often calculated by completing a large number of annual simulations with randomly drawn unit outages, load levels and sometimes may include renewable generation uncertainty. Over this large number of iterations it is possible to calculate the expected number of loss of load events per year. Typically, LOLE analyses assume that any resource not on outage is available to meet load and reserve requirements and do not account for operational constraints such as startup time or ramp rate. The threshold for loss of load may include some amount of reserve above firm load. This metric can also be referred to as Loss of Load Frequency (LOLF) and stated as "events per year."

Loss of Load Hours (LOLH) is a similar probabilistic metrics measuring generation adequacy stated as expected "hours per year."

Some models, such SERVM, calculate probabilistic metrics such as LOLE that include outage events driven by operational flexibility constraints such as startup time or ramp rate as well as events driven by economic commitment and dispatch decisions. This is different from the traditional metric, which simply assumes any resource not on outage is available without limitation. We have attempted to differentiate such metrics from traditional metrics in this report.

Expected Over-Generation (EOG)

Expected Over-Generation (EOG) is a probabilistic metric measuring downward flexibility, typically stated as expected MWh/year. Unlike the traditional interpretation of EUE, this metric is only driven by operational flexibility constraints such as minimum generation levels, startup time and ramp rate. EOG is calculated by models such as REFLEX or SERVM by completing a large number of simulations with randomly drawn unit outages, load levels and renewable generation. Over this large number of iterations it is possible to calculate the expected annual number of MWh of must-take generation that is in excess of firm demand—Expected Over-Generation.

Over-Generation

CAISO defines over-generation as a condition where supply exceeds demand on the CAISO system. Such conditions can arise in a planning model when the sum of must-take generation (e.g., renewables) and must-run generation (e.g., units needed for reliability purposes) exceeds load, or when generating units are unable to balance the system during steep downward ramps.

Unit Commitment and Dispatch Decisions

At any given point in time, the system operator must decide for a future point in time, which units should be generating (unit commitment) and what level (dispatch). These decisions may be made in a planning model with perfect foresight (e.g., the model knows exactly what the load, wind and solar output will be in the future) or with some forecast uncertainty, similar to the way such decisions must actually be made.

Recourse

The ability for a model to account for uncertainty by adjusting a prior decision in response to change. Models can make an initial decision that minimizes the expected cost of that decision's consequences—as REFLEX does. Models can also characterize the uncertainty at different points prior to the actual realization of load, wind and solar, by including forecast errors.

Deterministic Versus Stochastic Modeling

A parameter is treated as deterministic or certain when it is given a single value or fixed profile in a model (i.e., the 2022 CAISO load in the CAISO/PLEXOS model is given a fixed profile).

A parameter is treated as stochastic or uncertain when it is given a range of possible values in a model—for example, most models discussed here treat generator outages as a stochastic parameter.

Reserves (Regulation, Contingency, and Load Following)

Reserves are unloaded capacity (i.e., MW of generation that are online or can quickly come online). In planning models, reserves are required to manage generation/load variability and uncertainty that occur on very short timescales (regulation), on slightly longer timescales (load following), and to manage unexpected and large-scale outages (contingency reserves, both spinning and non-spinning).