Pacific Gas and Electric Company

EPIC Final Report

Program	***Electric Program Investment Charge (EPIC)***
Project	***EPIC 2.07 – Real Time Loading Data for Distribution Operations and Planning***
Reference Name	***EPIC 2.07 – Real Time Load Forecast***
Department	*Electric Asset Management: Grid Modernization*
Project Sponsor	Ferhaan Jawed
Project Business Lead	Anne-Lise Laurain
Contact Info	EPIC_Info@pge.com
Date	November 28, 2018
Version Type	Final
Version	

Table of Contents

List of Tables

List of Figures

Table of Acronyms

AMI	Advanced Metering Infrastructure
Amps	Amperes
AOR	Area of Responsibility
APE	Absolute Percent Error
CEC	California Energy Commission
CPUC	California Public Utilities Commission
DER	Distributed Energy Resources
DMS	Distributed Management System
EPIC	Electric Program Investment Charge
GHG	Greenhouse Gas
GIS	Geographic Information System
IGP	Integrated Grid Platform
IOU	Investor-owned utility
kV	Kilovolt
kVa	Kilovolt-amperes
kVAR	Kilovolt-amperes reactive
kW	Kilowatt
kWh	Kilowatt Hour
MAPE	Mean Absolute Percent Error
MdAPE	Median Absolute Percent Error
MW	Megawatt
PG&E	Pacific Gas & Electric Company
PV	PhotoVoltaic

RMS	Root Mean Square
SCADA	Supervisory Control and Data Acquisition
SDP	Service Delivery Points
sAPE	symmetric absolute percent error
TD&D	Technology Demonstration and Deployment
V	Volt

1 Executive Summary

This report summarizes the project objectives, technical results and lessons learned for Electric Program Investment Charge (EPIC) Project 2.07 *Real Time Loading Data for Distribution Operations and Planning*, as listed in the EPIC Annual Report.

1.1 Project Objectives

The purpose of this project was to determine the best way to use SmartMeter™, Supervisory Control and Data Acquisition (SCADA), photovoltaic (PV) system generation, Geographic Information System (GIS), weather, and other relevant data within an analytical framework to deliver better real-time and forecasted loading information for Distribution Operators, Distribution Engineers, and Planners (collectively, the "End Users"), and assist them in making more informed decisions in their day-to-day activities.

Currently, the End Users only have access to real-time and historical amperes (amps), megawatt (MW) and mega volt amp (MVAR) data when a device is SCADA-equipped. With limited deployment of SCADA sensors on the distribution grid and limited real-time loading data, it is difficult for End Users to make detailed circuit-level decisions with confidence. Today, for non-SCADA devices, end users primarily access either peak load data or aggregated raw Advanced Metering Infrastructure (AMI) historical loading converted to amps. The peak load data is calculated using load flow models based upon customer kilowatt-hours (kWH) consumption as well as customer kilowatt (kW) (where available) and represents the seasonal peak. The aggregated raw AMI data typically uses the past 30 days of data with a lag of 2 days. This project will provide real-time and forecasted loading visibility to the End Users, allowing them to take more informed decisions for planned and unplanned events.

1.2 Project Overview

Pacific Gas and Electric Company (PG&E) contracted with an external vendor that had previously worked on load forecasting, and had developed proprietary models for both bottom-up and top-down forecasting. Due to their proprietary nature, only limited descriptions of these models are provided in this report. A new reconciled model was developed through this project, which aimed to determine which of the two previously-developed models provided the best forecast, based on the data available.

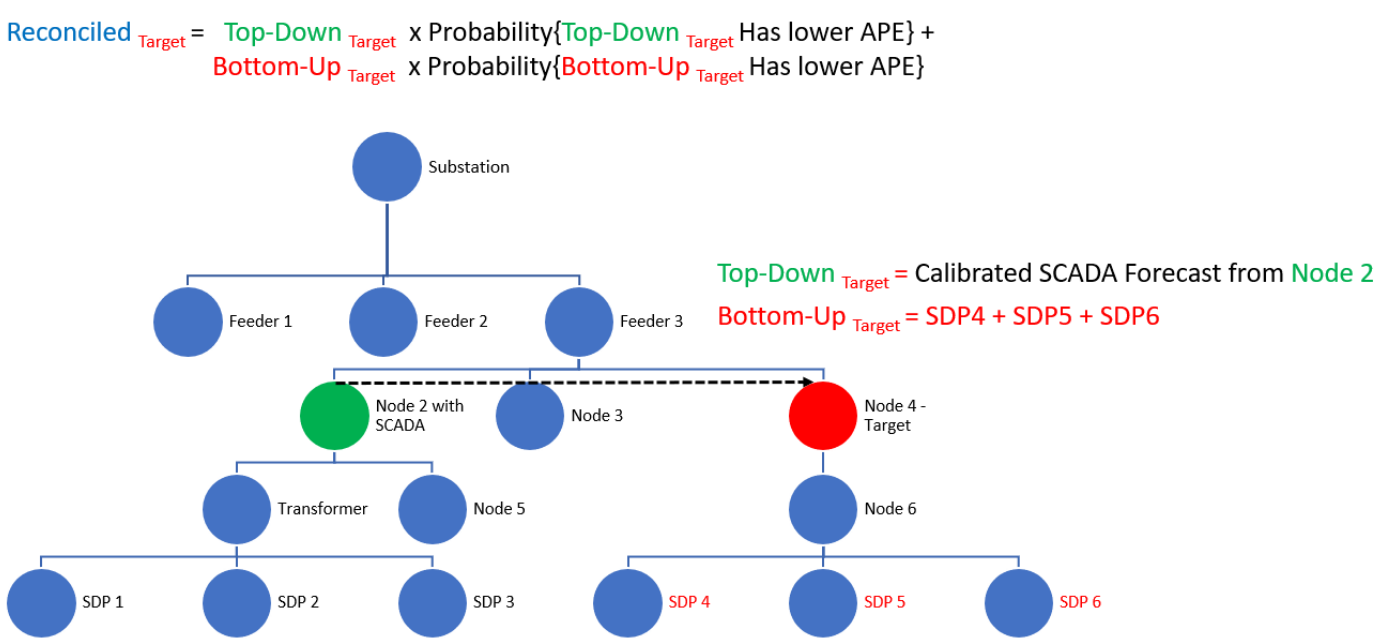
There were two sequential tasks in this project. First, a retrospective forecast demonstration (task 1) of the achievable precision for short-term forecasts at distribution nodes was conducted. It used a sample of 38 feeders from PG&E's service territory to validate that real-time forecast could improve the day-to-day operation of PG&E's system. Next, a broader near real-time implementation (task 2) of the analytics used in task 1 was conducted, with the key differences in scope and scale presented in Table 1. For task 2, a real-time environment was set up, using roughly 20% of PG&E's service territory. As SmartMeter™ data was reported with a 24 hour delay at midnight every day into the database used for this project, the forecast provided data for the day before and the current day to fill this gap.

Table 1: EPIC 2.07 Scope Comparison – Task 1 and Task 2

Scope Component	Task 1	Task 2
Overall Focus	Achievable Forecasting Precision of Analytics	Integration for Test Deployment
Demonstration Area	38 Feeder Sample	Two complete Area of Responsibility (AOR) regions – approx. 676 feeders from 188 substations, and 1.6 million Service Delivery Points (SDP)
Topological Constraints	As-Built Topology	As-Switched Prevailing Topology
Integration with Enterprise applications	No System Integration	Near Real-Time integration with other Enterprise applications
User Interaction	UI Demo Only	PG&E Login Access to Deployed Software Platform

For both tasks 1&2, a SDP-Level Forecast Model using SmartMeter™ data and using a proprietary peer model for customers without SmartMeter™ was built. Similarly, a SCADA forecast model was built using mainly data from SCADA, where SCADA data was available. Then, a bottom-up forecast model was developed using data from the SmartMeter™, GIS, weather data, and power factor estimates and a top-down forecast model was produced for non-SCADA distribution nodes calibrated by the SCADA model forecasts. Finally, a reconciled model was created to determine which of the two previous models provided the best forecast, based on the data available at the time of the analysis. Figure 1 summarizes this description.

Figure 1: Simplified Depiction of Partial Topology for a Substation with Three Feeders



Additional details of the scope and approach for the two tasks can be found in Sections 3 and 4.

1.3 Key Accomplishments

The following were the project's key accomplishments:

- Successfully built a platform to ingest and process AMI, SCADA, weather, PV generation, and topological data in real time, at a scale never before implemented at PG&E
- Produced hourly load forecasts from 2 days in the past to cover latency in receiving SmartMeter™ data, to seven days in the future for all distribution device classes of interest and individual customer meters in two of the eight AOR regions within PG&E's service territory in under 4 hours
- Developed a reconciled forecasting method that leveraged both bottom-up kWh converted to amps per phase and top-down forecasts using SCADA data
- Integrated as-switched topology into forecasts
- Produced stability and confidence models for distribution node forecasts

1.4 Key Takeaways

Top-Down/Bottom-Up Reconciled Forecast Accuracy

The project team compared the forecasted mean amps per phase to actual amps per phase from SCADA meters to assess the accuracy of the forecasts deployed in the project. The team used 317 SCADA nodes for validation, which were screened to ensure that input data of sufficient quality was available to generate the forecasts. The reconciled forecasts, which can be produced at any energized distribution node, had a median absolute percent error (MdAPE) of 32% across all hours in the validation sample[1] in task 2. This means that half of the forecasts across all validation nodes and all hours of the day in the seven day evaluation period were within 32% of the target. As for within-node accuracy, half of the validation nodes had a 28% MdAPE or less, and half had higher than 28% MdAPE.

SCADA Model Forecast Accuracy

In practice, devices with SCADA instrumentation will have the data history from the corresponding meters available to better train the forecast models. These SCADA forecasts had an overall MAPE of 6.7%, varying according to the hour of the day and device type being forecasted.

Outsourcing Model Development

It should not be assumed that external vendors with experience in data science will be able to seamlessly process and apply operational PG&E data in complex analytical projects such as this without significant support from PG&E subject matter experts. Extensive collaboration was required to assist the vendor in processing the various input data needed for the load forecasting models. Also, working vendors that employ proprietary forecasting models can limit visibility and in turn make it more difficult to assist them in troubleshooting issues and improving on their methods.

[1] The validation sample was a set of SCADA nodes whose readings passed a data quality screening, to alleviate concern of the percent errors being influenced by data quality issues present in the SCADA amps readings.

Forecast Stability and Confidence Analysis

Neither the SCADA nor reconciled forecasts were biased significantly from the actual amps per phase. Overall, 22% of the reconciled estimates were 50% or more below the actual amps per phase, and 14% were 50% or more above the actual amps per phase. Only 0.2% of SCADA forecasts were underestimates by 50% or more and 4.4% were overestimates by 50% or more. The confidence models correctly flagged these large forecast errors 80% of the time for SCADA models and 85% of the time for the reconciled models. These flags could alert End Users to make more conservative assumptions of the loading than what the forecasts show in those instances.

Analysis of Systematic Error in Load Forecasts

The analysis of deviations between bottom-up aggregations of historical actual AMI meter kWh converted to amps per phase with measured amps per phase at SCADA nodes showed that there was a very similar error distribution in the amps per phase using SDP-level kWh actuals as with SDP-level kWh forecasts, used in the aggregations. This finding suggests that if a more accurate method was used for converting kWh to amps per phase at a target node, the forecast error could potentially be reduced significantly.

1.5 Challenges and Resolutions

The main challenges in developing this project's load forecasting capabilities were to make up for the limitations in data available at all the distribution nodes, and to ingest and utilize input data for the forecasting analytics. These challenges and resolutions are summarized in Table 2 and Table 3.

Table 2: Forecast Process Methodological Challenges and Resolutions

Forecast Process	Challenge	Project Team Resolution
Coverage of nodes in the distribution grid without metering infrastructure	Partial coverage of distribution nodes with SCADA meters and line sensors	Aggregated forecasted SmartMeter™ data corresponding with the target distribution node.
Coverage of nodes in the distribution grid without metering infrastructure	About 4% of SDPs without SmartMeters™	Peer models were assigned and persisted for all SDPs for use in forecasting, when an AMI based model could not be produced for various reasons.
Utilization of most recent model input data available	Two-day delay in availability of SmartMeter™ data	Calibrated SmartMeter™ based forecasts with low-latency SCADA data from metered assets on the same feeder as the target distribution node.
Unit Conversion	SDP level data in kWh and target node forecasts needed in amps	Converted aggregated kWh to amps per phase, accounting for the target node phase configuration and operating voltage (see 4.1.1.1.).
Accounting for switching	The SDPs contributing to the loading at a given distribution node changes according to switching by distribution operators	Integrated Distribution Management System switching state into the SDP aggregation methodology. Historic switch positions stored for all switches.
Topology and network changes	New assets (meters / grid devices) are installed, removed / exchanged.	Updates to device master data and topology regularly captured and integrated with the existing data

Table 3: Data challenges and Resolutions

Data Class	Data Challenge	Project Team Resolution
Topology	No timestamp for abnormal/ normal switch transitions	Transition timestamps were assigned from metadata from first subsequent abnormal state file without record for the switchable device.
SDP-level time series data	Incomplete monthly kWh data available for non-AMI SDPs	Peer models required the average daily kWh for calibration. For non-AMI SDPs the team received a limited transfer of this data, so mean average daily kWh by customer class was used as a proxy for this peer model input.
SDP-level time series data	PV forecast and actual data refresh data inconsistent	Forecasted net demand (kWh delivered minus kWh received), implicitly accounting for PV generation instead of explicitly with forecasted and actual PV.
Loading data at asset nodes in distribution topology	Changing tag name conventions	The project team rewrote the SCADA/Topology mapping code as needed.
Loading data at asset nodes in distribution topology	Irregular transfers for ingestion	The vendor team worked with PG&E Information Technology (IT) to reinstate transfer schedule or restart the data ingester when irregular file formats are encountered.
Loading data at asset nodes in distribution topology	Data quality	The project team applied a trigger for distribution node forecasts excluding SCADA data of insufficient quality for use in this project, when common issues impacting forecast quality were identified.
Multiple	Changing data exchange protocols – resulting from lessons learned through course of work	The project team rewrote data ingest pipelines as needed.
Data corrections	Correcting erroneous time series data (meter reads) in No SQL database	Delete and re-insert approach was used instead of updating the wrong records. This turned out to be a more accurate option, though time consuming.
Data volume growth	Exponential increase in data volume resulted in data handling issues	Pre-aggregated data was stored only for frequently accessed data points and only for a specified time period. Forecasts only at SDP level were persisted. Most of the aggregations were calculated real time
Large volume of PV forecast data	Large volume of PV forecast data consumed a major portion of compute and storage resources	Less granular PV data (daily forecasts) were used instead of hour a head, six hour a head forecasts

1.6 Recommendations

The following recommendations should be considered by any organization that endeavors to implement real-time load forecasting capabilities similar to those developed within this project:

- Test forecasting methods on smaller sets of data before deploying capabilities at scale
- Ensure that the data platform has sufficient processing power to support computationally-intensive forecasting methods, and ensure that the ability to scale up is addressed in platform development
- Deploy data completeness and quality flag tracking analytics in parallel to forecasts to provide deeper insight to potential issues with forecast accuracy and available data to produce the forecasts
- Focus on the quality of the underlying models & assumptions for the aggregation of actual kWh converted to mean amps per phase. The accuracy of the aggregation is impacted by assumptions for power factor and system losses, as well as the load flow model. Distribution engineers and data scientists would need to work together to address these issues and ensure strong forecasting models are employed.
- Perform robust validation to compare performance across the different forecasting methods
- Follow standard data science techniques to ensure that the classification tree model of the reconciled forecast method is robust to prevent overfitting

1.7 Conclusions

While challenges remain for improving the analytics deployed in this project for use by End Users, this project demonstrated that on-demand level forecasts for loading at distribution nodes are feasible. Load forecasting is not an easy task, but having multiple models to use across various levels of the grid helps to compensate for the potential gaps of each model. Overall, the results produced by the load forecasting models were inconclusive. Also, the use of median to assess forecast accuracy can hide the impact of outliers.

On average, the forecasts were on target, as indicted by the minimal bias observed. One of the key takeaways was that systematic errors – i.e. unrelated to the stochastic models driving the forecasts – can be significant drivers of forecast errors, and represent a logical next step for anyone attempting to drive down forecast errors. All the models built into this project assumed a set power factor per customer class, no system losses, no power flow analysis, and did not take into consideration any reactive power supplied by capacitor banks. These limitations of this approach would require expertise from both distribution engineers and data scientists to improve upon. Additionally, data quality flags for SCADA data measurement readings and bolstering the tracking of inbound model input data would further improve the forecast accuracy.

A challenge with working with an external vendor is that it can limit the visibility into their methods, especially when their methods are proprietary. This can make it difficult to support them to improve their models. Moreover, without a full understanding of the models, building in additional functionality is challenging, and this makes a utility dependent upon the external vendor for future enhancements.

2 Introduction

This report documents the EPIC 2.07 – Real Time Load Forecast project achievements, highlights key learnings from the project that have industry-wide value, and identifies future opportunities for PG&E to leverage this project.

The California Public Utilities Commission (CPUC) passed two decisions that established the basis for this demonstration program. The CPUC initially issued Decision (D.) 11-12-035, *Decision Establishing Interim Research, Development and Demonstrations and Renewables Program Funding Level*[2], which established the EPIC on December 15, 2011. Subsequently, on May 24, 2012, the CPUC issued D. 12-05-037, *Phase 2 Decision Establishing Purposes and Governance for Electric Program Investment Charge and Establishing Funding Collections for 2013-2020*[3], which authorized funding in the areas of applied research and development, technology demonstration and deployment (TD&D), and market facilitation. In this later decision, CPUC defined TD&D as "the installation and operation of pre-commercial technologies or strategies at a scale sufficiently large and in conditions sufficiently reflective of anticipated actual operating environments to enable appraisal of the operational and performance characteristics and the financial risks associated with a given technology."[4]

The decision also required the EPIC Program Administrators[5] to submit Triennial Investment Plans to cover three-year funding cycles for 2012-2014, 2015-2017, and 2018-2020. On November 1, 2012, in A.12-11-003, PG&E filed its first triennial EPIC Application with the CPUC, requesting $49,328,000 including funding for 26 Technology Demonstration and Deployment Projects. On November 14, 2013, in D.13-11-025, the CPUC approved PG&E's EPIC plan, including $49,328,000 for this program category. On May 1, 2014, PG&E filed its second triennial investment plan for the period of 2015-2017 in the EPIC 2 Application (Application (A.) 14-05-003). CPUC approved this plan in D.15-04-020 on April 15, 2015, including $51,080,200 for 31 TD&D projects.[6]

Pursuant to PG&E's approved 2015-2017 EPIC triennial plan, PG&E initiated, planned and implemented the following project: EPIC 2.07 – Real Time Load Forecast. Through the annual reporting process, PG&E kept CPUC staff and stakeholders informed on the progress of the project. The following is PG&E's final report on this project.

---

2 http://docs.cpuc.ca.gov/PublishedDocs/WORD_PDF/FINAL_DECISION/156050.PDF.
3 http://docs.cpuc.ca.gov/PublishedDocs/WORD_PDF/FINAL_DECISION/167664.PDF.
4 D.12-05-037 pg. 37.
5 PG&E, San Diego Gas & Electric Company (SDG&E), Southern California Edison Company (SCE), and the California Energy Commission (CEC).
6 In the EPIC 2 Plan Application (A.14-05-003), PG&E originally proposed 30 projects. Per CPUC D.15-04-020 to include an assessment of the use and impact of electric vehicle energy flow capabilities, Project 2.03 was split into two projects, resulting in a total of 31 projects.

3 Project Summary

This section summarizes the industry gap that the project addresses, as well as the project's objectives, the scope of work, and major tasks, milestones, and their corresponding deliverables.

3.1 Issues Addressed

Currently, the loading information available to distribution operators, distribution engineers, and planners (collectively referred to as "End Users") is either seasonal peak device level amp data or historical AMI raw aggregations converted to device and line level amp values in non-SCADA areas. With limited deployment of SCADA sensors on the distribution grid and limited real-time loading data, it is difficult for End Users to make detailed circuit-level decisions with confidence. This challenge is compounded with the need to preemptively plan outages, requiring a forecasted view of future loading characteristics that even a robust deployment of SCADA sensors cannot provide. Improving load forecasting will improve the following use cases:

- **Scheduled Outages**: Distribution operators (DO) and distribution operations engineers (DOE) must develop switch plans to re-route power serving end customers that is currently flowing through assets which need to be de-energized temporarily for maintenance or repairs. In some circumstances the switch plans can become complex if there is not enough capacity through direct tie-points on the circuit, thereby requiring cascading load transfers.
- **Emergency Load Transfer**: Unscheduled outages from damaged distribution assets result in loss of power to end customers whose load is served by power flowing through the damaged asset. DOs are responsible for safely restoring power as quickly as possible by re-routing load through nodes with capacity to accommodate the incremental load from the customers experiencing the outage.
- **Capacity Planning**: System planning engineers estimate loading throughout the distribution grid based on a single peak loading value recorded at the bank level, often by an onsite reading of an electro-mechanical meter, when no SCADA exists, and determine which downstream assets are most at risk of failure due to overloading, requiring equipment repairs or replacement.

3.2 Project Objectives

The objective of this project was to configure analytics using Smart Meter, SCADA, PV, GIS, Weather, and other data to deliver better real-time and forecasted loading information for End Users and assist them in making more informed decisions in their day-to-day activities.

3.3 Scope of Work and Project Tasks

3.3.1 Task 1: Retrospective Forecast Demonstration

Objectives: Demonstrate the achievable precision for short-term (real-time to 7 days ahead) forecasts at blind spots (nodes in the distribution topology without SCADA or line sensors) in PG&E's distribution system at topological nodes without sensors collecting loading time series data.

Outputs: The following project task outputs are documented in the Task 1 results of Section 4 of this report.

3.3.1.1 Bottom-Up Forecast Accuracy Validation

Bottom-up load forecasts were applied using models of end customer SmartMeter™ data, weather data, GIS data, and power factor estimates. Volt-amps were converted to average amps per phase by dividing the volt-amps by the high-side operating voltage at the target node. Since validation was against instantaneous root mean square (RMS) reads at the top of each hour, bottom-up estimates consisted of averages of the kWh-derived estimates from that matching timestamp and the following hour, since kWh timestamps are on the hour end.

3.3.1.2 SCADA Model Forecast Accuracy Validation

SCADA models of phase-level amps were configured for use in reconciled top-down/bottom-up forecasts.

3.3.1.3 Reconciled Top-Down/Bottom-Up Forecast Accuracy Validation

The reconciliation process was designed to be a systematic means of determining conditions where the bottom-up forecasting approach should be used over the top-down, or vice versa, or whether a combination of the two would be expected to outperform either one on its own. A machine learning classification tree model was used for estimating which of the models would give the lowest forecast error, given several candidate input variables:

- Count of end customer Service Point IDs comprising the load at the target node
- Number of SCADA nodes on the feeder corresponding with the target node
- Historical forecast precision at nodes on the feeder
- Coefficient of variation of the top-down estimates

3.3.1.4 Sensitivity Analysis

The forecasts produced in this task used actual weather data, rather than forecasts. The sensitivity analysis quantified the expected change to forecast accuracy statistics with respect to weather forecast error variance.

3.3.1.5 Stability and Confidence Analysis

The project team quantified the frequency of forecasts with error of 50% or more. This was termed a “stability analysis”. Subsequently, the project team applied a confidence model where it modeled the accuracy as a function of forecast conditions, in an effort to provide the End User a flag that would indicate an elevated risk for the forecast error being above some threshold (50% was used in this task).

3.3.2 Task 2: Limited Scale Live Prospective Forecast Demonstration

Objectives: Demonstrate the feasibility of deploying the Task 1 analytics in an on-demand tool for PG&E End Users, covering two AORs, comprising approximately 30% of PG&E’s electricity customer population.

Outputs: The following project task outputs are documented in the Task 2 results of Section 4 of this report.

3.3.2.1 Bottom-Up Actual kWh Converted to Mean Amps Per Phase Accuracy Validation

To validate the process of aggregating SDP level net demand to distribution node levels, this analysis was conducted with actual AMI meter readings, rather than forecast models at the customer level. This provides insight into the maximum achievable precision for a bottom-up forecast, since there is no demand or weather forecast error, only unaccounted non-AMI SDPs.

3.3.2.2 Top-Down/Bottom-Up Reconciled Forecast Accuracy Validation

This validation was conducted to compare distribution node forecast accuracy, without the benefit of SCADA metrology, from a live setting in Task 2 to the retrospective controlled setting in Task 1.

3.3.2.3 SCADA Model Forecast Accuracy Validation

A validation of SCADA models, using historical SCADA meter readings for training, was conducted to provide insight into the expected forecast precision by hour and lead time (up to 7 days into the future) at SCADA nodes, which are active on many important switchable or otherwise critical assets to monitor throughout the distribution grid.

3.3.2.4 SDP Model Forecast Accuracy Validation

While not expected to be of primary interest to DOs, the accuracy of SDP forecasts was provided for stakeholders at PG&E who have a need to forecast SDP level net demand.

3.3.2.5 Stability and Confidence Analysis

As with Task 1, the stability and confidence models were summarized to inform End Users on the risks of forecast errors above 50%.

4 Project Activities, Results, and Findings

This project consisted of two sequential tasks demonstrating forecasting analytics within distribution topologies. The following table summarizes the differences in scope of the two tasks.

Table 4: EPIC 2.07 Scope Comparison – Task 1 and Task 2

Scope Component	Task 1	Task 2
Overall Focus	Achievable Forecasting Precision of Analytics	Integration for Test Deployment
Demonstration Area	38 Feeder Sample	Two complete AOR regions – approx. 676 feeders from 188 substations and 1.6 million SDPs (about 20% of PG&E's territory)
Topological Constraints	As-Built Topology	As-Switched Prevailing Topology
Integration with Enterprise applications	No System Integration	Near Real-Time integration with other Enterprise applications
User Interaction	UI Demo Only	PG&E Login Access to Deployed Software Platform

- **Overall Focus**:
  - Task 1 demonstrated the achievable precision for short-term (real-time to 7 days ahead) forecasts at blind spots in PG&E's distribution system at topological nodes without sensors collecting loading time series data.
  - Task 2 demonstrated the feasibility of applying the Task 1 analytics as an on-demand tool for PG&E staff.
- **Demonstration Area**:
  - Task 1 analytics were applied to 38 feeders selected for the project.
  - Task 2 expanded to cover the distribution system in two AORs.
- **Topological Constraints**:
  - Task 1 analytics were configured to the as-built topology for the 38 feeders. The as-built topology accounts for the possible paths of energized assets in the distribution grid. Actual energized paths are specified from the as-built topology by applying the positions of switches from open to closed. Therefore, the as-built topology includes redundant options for routing power when certain assets must be de-energized for performing maintenance, or in the case of an outage.
  - Task 2 analytics were applied to the prevailing as-switched topology, utilizing an ongoing stream of data identifying switches currently in a temporary or normal configuration, denoting the switches being open or closed at a given time.
- **Integration**:
  - Task 1 analytics were applied to a batch of data sets in one-time delivery. All forecasts were produced retrospectively, using actual weather and distributed PV generation data.
  - Task 2 analytics were applied to PG&E data dynamically, supporting user-specified on-demand forecasts for both past and future date spans, using forecasted weather data.

- **User Interface (UI) Integration**
  - Task 1 analytics precision results were summarized in report and slide deck format only
  - Task 2 analytics results are available to users both in the UI visualization and for download to a comma separated value (CSV) file, with corresponding precision summarized in this report.

4.1 Task 1: Retrospective Forecast Demonstration

4.1.1 Technical Development and Methods

4.1.1.1 Unit Conversion

The unit of measurement for the forecasts at nodes upstream of SDPs was amps per phase. SCADA meters installed on assets in nodes in the distribution system measured phase-specific amps and volts. Amps were of primary interest to End Users since it is used for load transfer calculation. Forecasting loading by phase would have been preferable to End Users from feedback provided to the project team, however data for mapping SDP loading to each phase was not available, so mean amps across the phases was the target metric for the forecasts in the project.

Since SDP level consumption was measured in kWh, conversion was first necessary to go from kWh to kVA, and then from kVA to amps per phase, accounting for the operating voltage and phase configuration at the target node. Power factors were assumed to be static at 0.95 for residential customers and 0.85 for non-residential customers. The following formula was used for converting bottom-up kWh to amps per phase at the target node in Task 1:

$$Amps\ per\ phase = \frac{\sum kVA_{Target}}{\sqrt{3} \times kV_{OperatingVoltage\_Target}}$$

$\sum kVA_{Target}$ was the bottom-up kWh converted to kVA by dividing the kWh by the power factor by customer type.

$kV_{OperatingVoltage\_Target}$ was the operating voltage at the target distribution node where the forecast was made.

4.1.1.2 Forecasting Methodology

The forecasting approach consisted of five components:

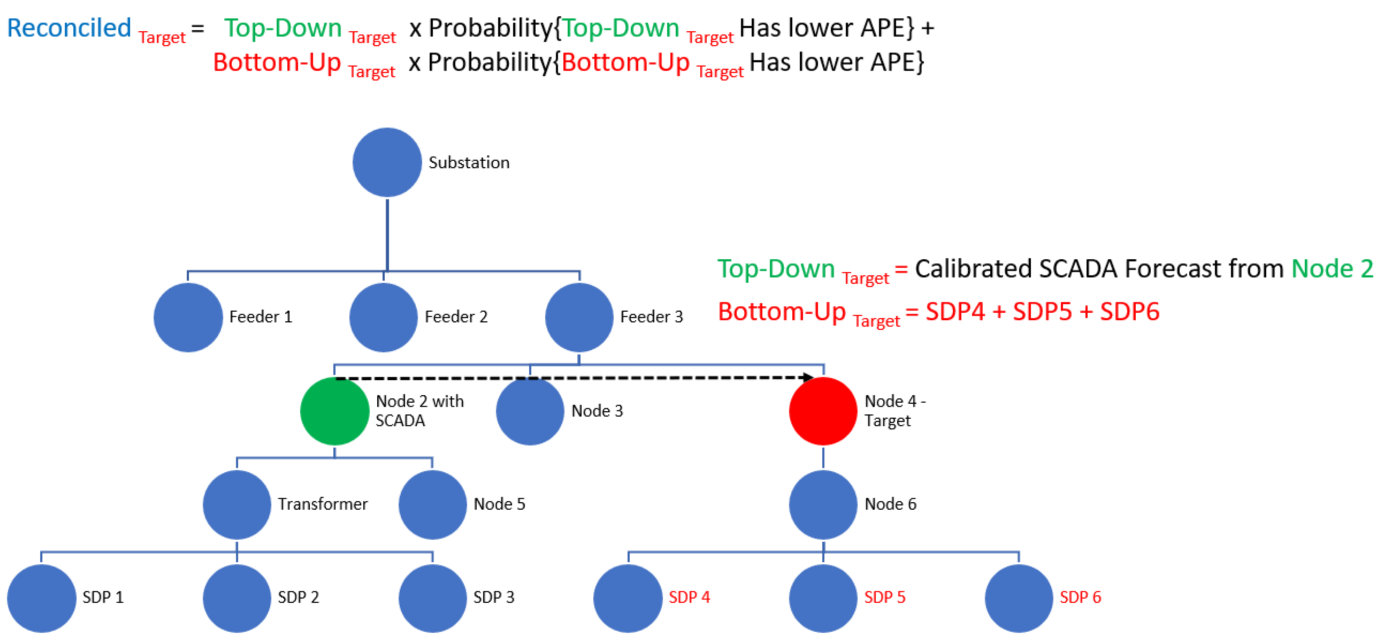
1. **SDP-Level Forecast Models**. The project team fit models for net delivered hourly kWh for every SDP with SmartMeter™ data. This included SDPs with 15-minute readings, which were aggregated to hourly for consistency. For non-SmartMeter™ SDPs, peer models were estimated to account for their contribution to loading. The SDP models accounted for calendar effects and weather sensitivity.
2. **SCADA Load Models**. Models fit directly from historical amps per phase data were fit for each SCADA node, to forecast future loading at the SCADA nodes, and to be provided as input to the top-down forecasts described below. The SCADA models also accounted for weather sensitivity and calendar effects.
3. **Bottom-Up Aggregations of Forecasted kWh Converted to Amps per Phase**. Bottom-up distribution node level forecasts consisted of aggregations of the SDP-level forecasts downstream of the target node of interest, using the topology data provided. The bottom-up aggregations are dependent on SDP-level forecasts, the results of which are detailed in aggregate in Section 4.1.3.1 and at the granular SDP level in Section 4.2.3.4.
4. **Top-Down Forecasts**. Top-Down Forecasts for the target distribution node were produced by calibrating SCADA model forecasts from other nodes within the feeder with SCADA instrumentation or line sensors. For some feeders in the 38 validation feeder set, SCADA data was not available or was not of sufficient quality for

producing a top-down forecast. Top-Down forecasts utilize SCADA model forecasts, the results of which are detailed in Section 4.1.3.2 and Section 4.2.3.3.

5. **Reconciled Top-Down/Bottom-Up Forecasts**. For distribution nodes where top-down and bottom-up forecasts could both be produced, reconciled models predicted the probability of the top-down or bottom-up forecasts having the lower APE, given various tracking conditions, trained on forecasts for SCADA nodes with historical top-down and bottom-up forecasts. Reconciled forecasts used this model to weight the top-down and bottom-up forecasts according to those probabilities.

Figure 2 shows a simplified topology with node-level expansions down to the SDP-level for one of the three feeders for a substation, with one of the feeders containing a target node for forecasting (in red). Forecasts were generated using a bottom-up approach aggregating the forecasted load for the downstream SDPs and a top-down approach using the readings from another node on the feeder with SCADA instrumentation (in green). The reconciled forecast is a weighted average of the top-down and bottom-up forecasts using the probability outputs of the model.

Figure 2: Simplified Depiction of Partial Topology for a Substation with Three Feeders



The weights in the weighted average come from a classification tree model configured for determining the circumstances when the bottom-up or top-down model performed best, applied to the target node. This reconciled model was then applied to point to the final forecast estimate for a given node. The accuracy of the forecasts was evaluated on SCADA nodes where actual amperage was recorded, and sensitivity of that accuracy with respect to temperature forecast inaccuracy was quantified in a sensitivity analysis.

Bottom-up load forecasts were applied using models of end customer SmartMeter™ data, weather data, topology data, and power factor estimates. Volt-amps were converted to average amps per phase by dividing the volt-amps by the high-side operating voltage at the target node. Since validation was against instantaneous RMS reads at the top of each hour, bottom-up estimates consisted of averages of the kWh-derived estimates from that matching timestamp and the following hour, since kWh timestamps are on the hour end.

Top-down forecasts were applied using SCADA measurements of current by phase, weather data, SmartMeter™ data, and topology data. While expected to be useful more narrowly than bottom-up forecasts because of the scarcity of SCADA nodes on the test set of 38 feeders, top-down models were tested for whether it would outperform bottom-up forecasts in certain conditions due to the one-hour latency of SCADA data compared with 48-hour latency for SmartMeter™ data.

The reconciliation process was designed to be a systematic means of determining conditions where the bottom-up forecasting approach should be used over the top-down, or vice versa, or whether a combination of the two would be expected to outperform either one on its own.

A machine learning classification tree model for estimating which of the three models would give the lowest forecast error was fit, given several candidate input variables:

- Count of end customer SPIDs comprising the load at the target node
- Number of SCADA nodes on the feeder corresponding with the target node
- Historical forecast precision at nodes on the feeder
- Coefficient of variation of the top-down estimates

To validate the forecasts, the average amps per phase from hold-out SCADA nodes as “truth” were used for validation. As noted in section 2.3.2, this limited the number of top-down forecasts which could be validated.

The validation process considered three criteria:

1. **Accuracy**. How large are the forecast errors, on a percentage basis?
2. **Stability**. How often are forecast percent errors very large?
3. **Confidence**. Can large forecast percent errors be predicted? How often do large errors not get flagged?

Accuracy. The accuracy metrics were mean absolute percent error (MAPE), which is commonly used in forecasting applications, as well as median absolute percent error (MdAPE), which is useful for showing a measure of center of APE in the presence of large outliers which may drive up the mean. For each hourly interval, the deviation between the forecasts and the actual load given in the target SCADA validation nodes was computed. The absolute percent error (APE) was then computed by taking the absolute value of the deviation and dividing by the actual value. The MAPE and/or MdAPE was then computed by averaging over the APE values. These error metrics can be computed for a vast number of subsets of the conditions for which forecasts are generated, e.g., for a specific hour of the day or for temperatures within a specified range, so long there were observations supporting it.

Stability. Stability of the forecasts was analyzed by computing the proportion of forecasts outside of a threshold of absolute percent error. Through stakeholder interviews it was made clear that operator trust in the forecasts could be compromised by large forecast errors, even when relatively infrequent compared to more moderate errors, so the stability assessment quantified the likelihood of those errors occurring.

Confidence. With the understanding that large, infrequent forecast errors for loading may have a disproportionately large impact on operators in their switch plans, a model was developed for predicting the error of the forecasts based on tracking metrics on the inputs to those forecasts. The output of this model could be a companion confidence-indicator flag for the load forecasts themselves, which could inform operators to use the load forecast with caution or defer to more conservative backup options such as the seasonal peak load estimates.

4.1.1.3 Data Summary

The Table 5 provides a summary of the various data elements used in the project. Overall, there were adequate data provided to implement this project's forecasting approach.

Table 5: Summary of Key Metrics for Critical Data Elements to Task 1

Data Element	Measure	Value	Project Team Comments
Topology	Sample set of feeders for the EPIC 2.07 Project	38	Breakdown by area: Diablo (2), Fresno (25), San Jose (9), Sonoma (2)
SmartMeter™ readings	Meters on the 38 EPIC 2.07 feeders with interval data	79,496	Missing data (2% of the meters) did not hold us back on the project, but complete data would increase forecast accuracy in future phases
SCADA readings extracted from PI	Nodes with amps for all phases on set of 38 feeders used in EPIC 2.07 project	52	18 of the 38 EPIC 2.07 feeders had one or more of these nodes with SCADA instrumentation. Only seven of the nodes provided data of sufficient quality for this project's forecasting, and were thus used in validation.
Bank-Level MW and Power Factor data	Feeders with Bank Level Power Factor Data	11	Did not use in this project since too high-level for aggregations to distribution nodes. Instead used assumed power factor by customer class.
Line Sensor Data	Distribution nodes accounted for in CSV files of interval line sensor data	15	Line sensor data contains phase-level amps, useful in top-down models and validation. Not all files were sufficiently populated for use in the project.
Weather Data (Hourly weather actuals)	Weather stations with sufficient historical data for retrospective load forecasts	25	Nearly complete data for all weather stations across the service territory was available
PV Generation (Meter Level, Hourly Intervals)	Meters with PV Generation interval data matched to the demonstration feeder set topology	6,652	83 Meters for which PV_Gen data are available have not been linked to a transformer in the topology.

4.1.2 Challenges

4.1.2.1 Data Challenges

The following data challenges listed in Table 6 were encountered over the course of the project.

Table 6: Data Challenges Encountered on Task 1

Data Issue	Impacted Project Tasks	Resolution Status	Notes
Irregular SCADA load patterns	Top-Down Load Forecasts, Validation	No	Several SCADA nodes had loading time series with apparent unaccounted switching or linear patterns which were unsuitable for extrapolation to other nodes. Such nodes were insufficient for use in this project and excluded from the validation task.
Zero-valued SmartMeter™ kWh readings	Bottom-up load forecasts	Yes	Zeros were excluded from the kWh models. It was determined they occur when there is a meter swap, outage, and when net generation occurs from PV system.
Missing Map of Non-SmartMeter™ SPIDs to Topology	Bottom-up load forecasts	No	Count of non-SmartMeters™ by district suggests likely very low percentage shares of non-SmartMeters™, so this had minimal impact on this project, but in future work will need this mapping to facilitate peer interval load modeling.

4.1.2.2 Technical Implementation Issues

Table 7 summarizes the technical issues encountered in the project.

Table 7: Task 1 Technical Implementation Issues for Task 1

Technical Implementation Issue	Impacted Project Tasks	Notes
Conversion of kWh to Amps	Bottom-Up Load Forecasts	For validation nodes, converted by dividing aggregate kWh by high-side operating voltage, then dividing by square root of 3. There may be special wiring configurations that would require different formula in the future.
Static Topology	All Forecasting Tasks	For operational deployment, will need real-time topology to ensure accurate forecasts can be made. Refreshed SCADA models may be required for changes in topology, and bottom-up aggregations will depend on accurate configuration.
Forecasting Instantaneous Readings	All Forecasting Tasks	Bottom-up forecasts will be “smoother” than instantaneous readings because of the demand aggregation to hourly energy. Higher frequency SCADA readings than one per hour may improve precision.
Load Imbalance Across Phases	All Forecasting Tasks	Average amps per phase can mask severe loading issues for imbalanced lines. In future work, should seek to account for imbalances in models, using available data and engineering principles.

4.1.3 Results and Observations

This section contains the results of the analyses described for Task 1 in Section 3.

4.1.3.1 Bottom-Up Forecast Accuracy Validation

Bottom-up load forecasts were applied using models of end customer SmartMeter™ data, weather data, topology data, and power factor estimates. Volt-amps were converted to average amps per phase by dividing the volt-amps by the high-side operating voltage at the target node. Since validation was against instantaneous root mean square (RMS) reads at the top of each hour, the bottom-up estimates consisted of averages of the kWh-derived estimates from that matching timestamp and the following hour, since kWh timestamps are on the hour end.

The bottom-up load estimates for target nodes were first generated in terms of kWh, and were then converted to average amps per phase. This section presents the precision at distribution nodes for both the kWh forecasts and those forecasts converted to amps and compared to actual spot amp readings at the validation SCADA nodes. The two sets of estimates are presented separately because the kWh converted amps will have incremental error due to

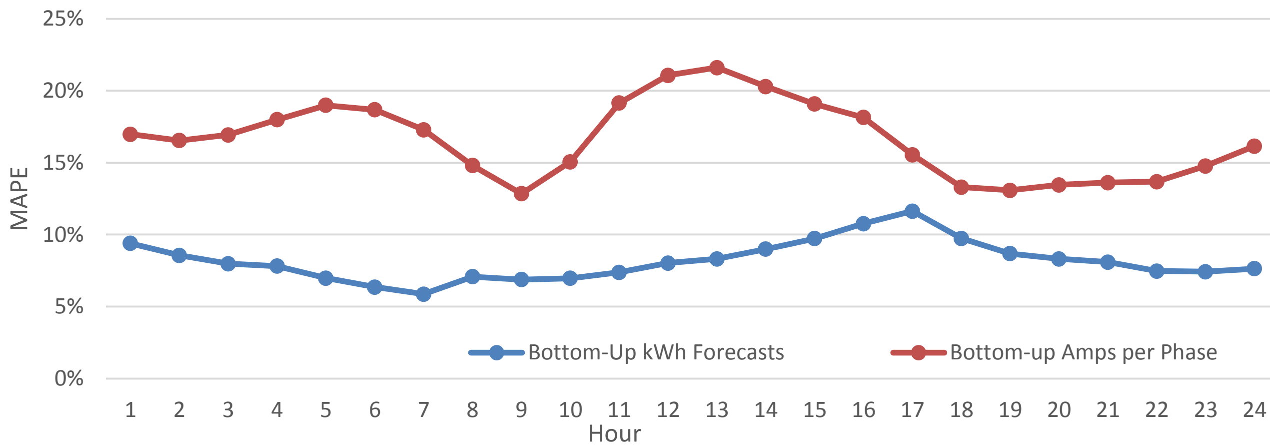
- Technical losses not being accounted for currently
- Instantaneous SCADA readings not directly corresponding with hourly energy
- Possible differences in the conversion formula due to special cases of line wiring configurations

The figure below shows the MAPE by hour for kWh forecasts across the validation nodes, in blue, and the corresponding MAPE for those bottom-up forecasts converted to average amps per phase (in red).

- The Bottom-Up Amps Per Phase Forecast MAPE in red measure the precision of the aggregated forecasted kWh from SDPs at the validation SCADA nodes, compared to the SCADA amps per phase from the SCADA meters at those nodes.
- The bottom-Up kWh Forecast MAPE is a measure of the precision of forecasted aggregated kWh compared to actual aggregated kWh, where both the forecast and actuals are aggregated up from SDP SmartMeters™.

As shown on Figure 3, the kWh MAPE ranges from 6% at hour ending 7 to approximately 12% in hour ending 17. Across all hours, the average kWh MAPE for the kWh forecasts was 8%. For the reasons listed above, the bottom-up MAPE for amps per phase is higher, ranging from 13% to 23% across the hours of the day.

Figure 3: Bottom-Up MAPE for kWh and Amps Per Phase Forecasts at Validation Nodes, by Hour



4.1.3.2 SCADA Model Forecast Accuracy Validation

SCADA models of phase-level amps were configured for use in reconciled top-down/bottom-up forecasts.

Across all SCADA nodes on the 38 feeders, the median MAPE is summarized in Table 8 by forecast lead time across the columns, and measure type down the rows. The Ex Ante lead time column refers to any larger lead time than seven days. Measure types RMS_CURRENT_1, RMS_CURRENT_2, and RMS_CURRENT_3 are associated with line sensor data, which only contains current by phase. SCADA meters measured both current and voltage per phase. Voltage forecast precision is also provided. Voltage is far more stable over time than current, and accordingly the MAPE is significantly less. The row with measure type labeled “Amps A+B+C” is the mean amps across the phases, which was ultimately the benchmark for forecast validation.

Table 8: Median MAPE Across All SCADA Nodes in the 38 Feeder Set by Measure Type

Measure Type	Hour Ahead (%)	1-Day Ahead (%)	2-Day Ahead (%)	3-Day Ahead (%)	7-Day Ahead (%)	Ex Ante (%)
AMPS_A	5.2	11.5	12.5	13.1	14.7	17.8
AMPS_B	5.8	13.4	15.4	16.1	16.4	20.3
AMPS_C	5.2	11.5	12.1	12.4	13.4	16.9
RMS_CURRENT_1	3.9	8.8	9.0	9.0	9.2	9.3
RMS_CURRENT_2	3.9	8.9	9.0	9.1	9.4	9.7
RMS_CURRENT_3	3.7	9.1	9.2	9.3	9.5	9.8
Amps A+B+C	**4.1**	**10.5**	**11.1**	**11.6**	**12.9**	**15.0**
VOLTS_A	0.2	0.4	0.4	0.4	0.4	0.4
VOLTS_B	0.2	0.4	0.4	0.4	0.4	0.4
VOLTS_C	0.2	0.4	0.4	0.4	0.4	0.4

After conducting a data quality review of the SCADA node interval data series, several SCADA nodes were flagged as not being appropriate for use in this project, due to having a significant share of potentially anomalous or unexplained readings or having dramatic structural shifts in the time series typical of switching activity, but not accounted for in the abnormal states switch records database. Notably, several of these series had very strong precision because the anomalous load pattern was highly predictable by the regression models. The Table 9 summarizes the SCADA forecast precision for the seven nodes that were used in the forecast validation task, after removing the remaining nodes with unexplained anomalous load patterns. Other than being 1% higher for hour ahead forecasts, the MAPE for the average amps across the phases (AMPS A+B+C) was unchanged or slightly lower in the final set of validation nodes for the different lead times.

Table 9: Median MAPE Across Validation SCADA Nodes in the 38 Feeder Set by Measure Type

Measure Type	Hour Ahead (%)	1-Day Ahead (%)	2-Day Ahead (%)	3-Day Ahead (%)	7-Day Ahead (%)	Ex Ante (%)
AMPS_A	5.4	8.0	9.0	9.5	10.2	12.9
AMPS_B	7.9	9.8	10.6	11.0	11.7	13.7
AMPS_C	5.8	8.4	9.1	9.4	10.0	12.6
RMS_CURRENT_1	8.0	16.2	16.4	16.5	16.9	18.4
RMS_CURRENT_2	10.1	20.5	20.6	20.8	21.1	22.5
RMS_CURRENT_3	7.6	15.2	15.3	15.4	15.8	17.1
Amps A+B+C	**5.2**	**10.6**	**11.2**	**11.6**	**12.3**	**14.8**
VOLTS_A	0.2	0.4	0.4	0.4	0.4	0.5
VOLTS_B	0.2	0.4	0.4	0.4	0.4	0.5
VOLTS_C	0.2	0.4	0.4	0.4	0.5	0.5

4.1.3.3 Reconciled Top-Down/Bottom-Up Forecast Accuracy Validation

The reconciliation process was designed to be a systematic means of determining conditions where the bottom-up forecasting approach should be used over the top-down, or vice versa, or whether a combination of the two would be expected to outperform either one on its own. A machine learning classification tree model was used for estimating which of the models (either top-down or bottom-up) would give the lowest forecast error, given several candidate input variables:

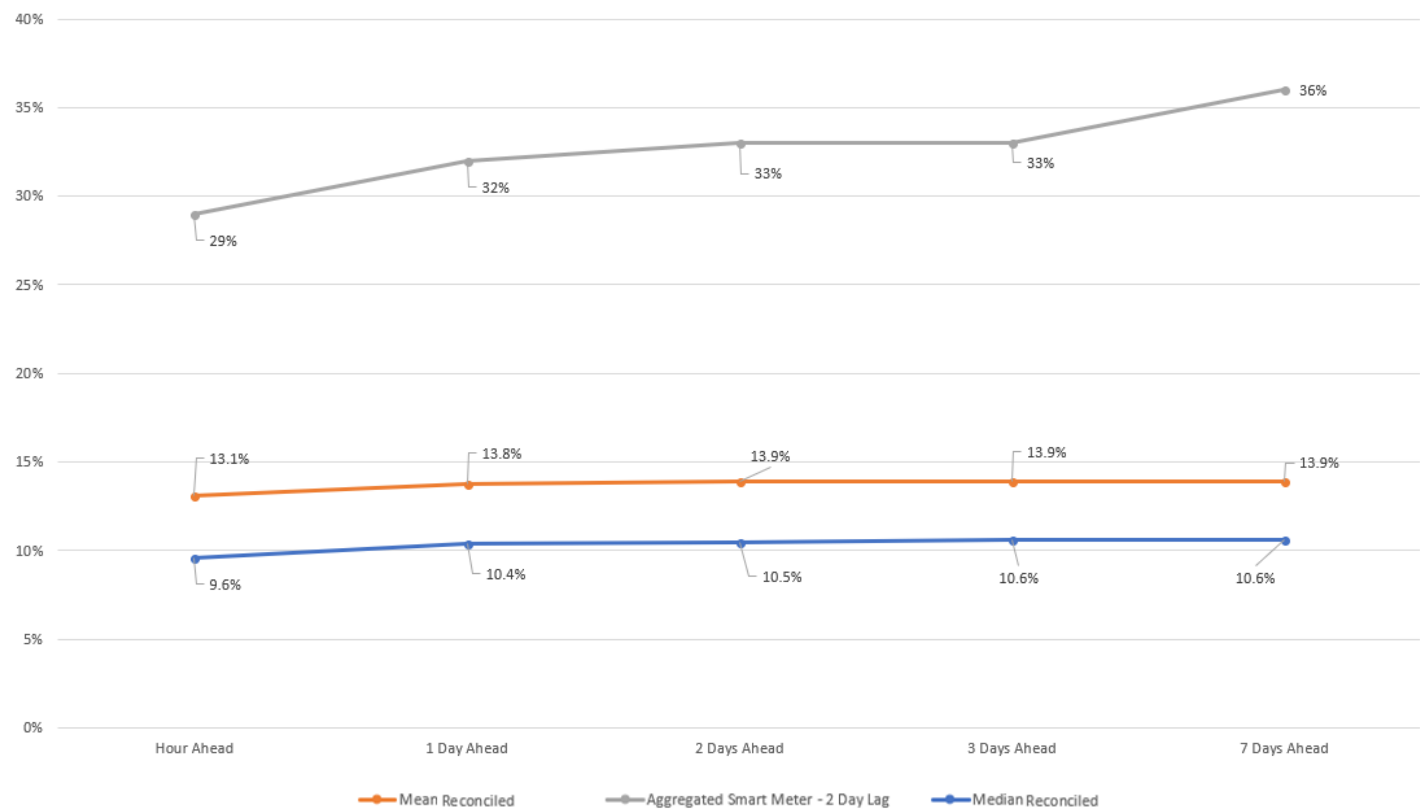
- Count of end customer SPDs comprising the load at the target node
- Number of SCADA nodes on the feeder corresponding with the target node
- Historical forecast precision at nodes on the feeder
- Coefficient of variation of the top-down estimates

Accuracy

The top-down/bottom-up reconciled forecast median absolute percent error (MdAPE) ranged from 9.6% for hour-ahead forecasts to 10.6% for 7-day ahead forecasts, as shown in the blue curve in the Figure 4. The corresponding range of the mean was 13.1 – 13.9%, corresponding with the orange series in the figure.

The project team compared the MAPE results to two-day lagged bottom-up kWH aggregations converted to mean amps per phase from SmartMeters™ downstream of the validation nodes, to evaluate the value of the forecasts compared with non-forecasted 2-day old converted kWh aggregations, which is the current latency of the data provided to End Users. Two days is the typical latency of SmartMeter™ data transmission to the central storage repository at the utility. By comparison, the MAPE for the two-day lag aggregated SmartMeter™ based estimate for loading (without a forecast model) ranges from 29 – 36%, shown in the grey curve below.

Figure 4: Comparison of the Current Model with Reconciled Forecasts by Lead Time – Validation Nodes



Both the median and mean APE are given in the figure above to illustrate that the error distribution is skewed, with a small share of very high errors causing the mean to be higher than the median. In this distribution for hour-ahead forecast APE, for example 90% of the forecasts had an APE less than 28%, 95% had an APE less than 35%, and 99% had an APE less than 59%.

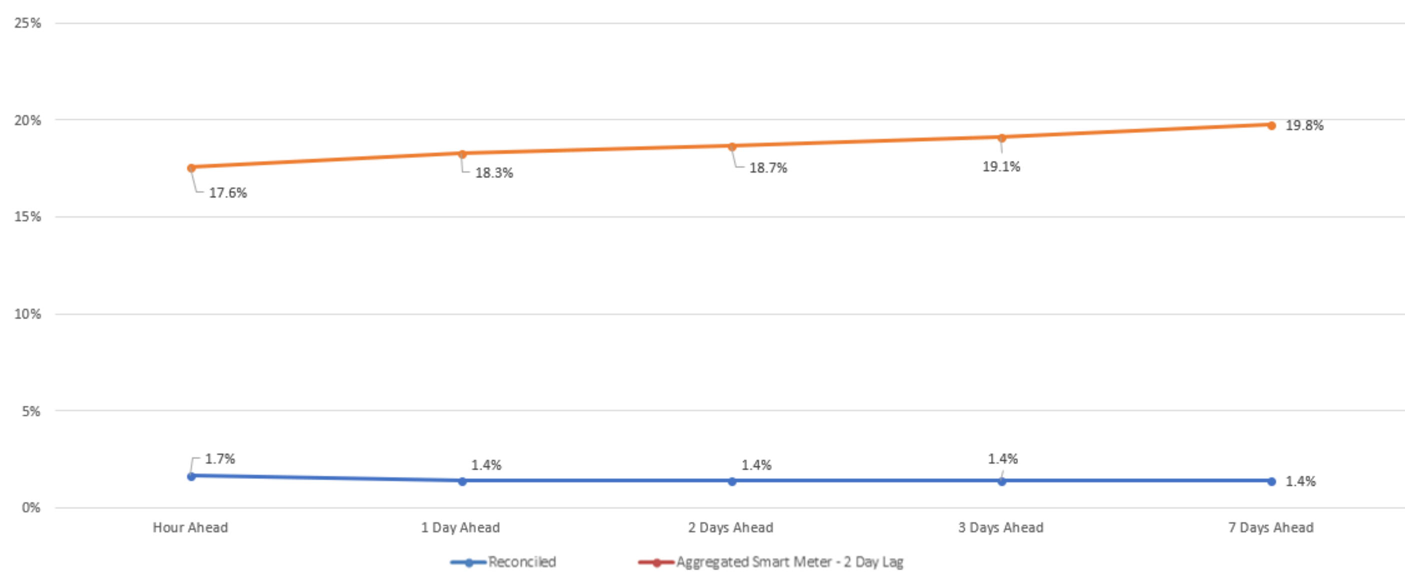
For the full set of validation nodes, which includes two nodes where only bottom-up forecasts could be made due to an insufficient number of SCADA nodes on the feeder, the MAPE was 16% for hour-ahead and day-ahead, and was steady at 17% for lead times of two days through seven days.

4.1.3.4 Stability and Confidence Analysis

In response to an inquiry by a PG&E subject matter expert, the project team quantified the frequency of forecasts with error of 50% or more. This was termed a “stability analysis”. Subsequently, the project team applied a confidence model where it modeled the accuracy as a function of forecast conditions, in an effort to provide the End User a flag that would indicate an elevated risk for the forecast error being above some threshold (50% was used in this task).

To measure the forecast stability, the percent of the forecasts with absolute percent error over 50% was computed, and compared it to the corresponding percentage for the backward-looking aggregated SmartMeter™ estimate. While the threshold of 50% was chosen arbitrarily, it gives an indication of the likelihood of a forecast being significantly off target, which in some use cases could have disproportionately severe consequences. At all lead times, fewer than 2% of the reconciled forecasts exceeded 50% error. By comparison, 18-20% of the backward-looking aggregated SmartMeter™ estimates exceeded 50% error. The stability estimates are summarized in the Figure 5.

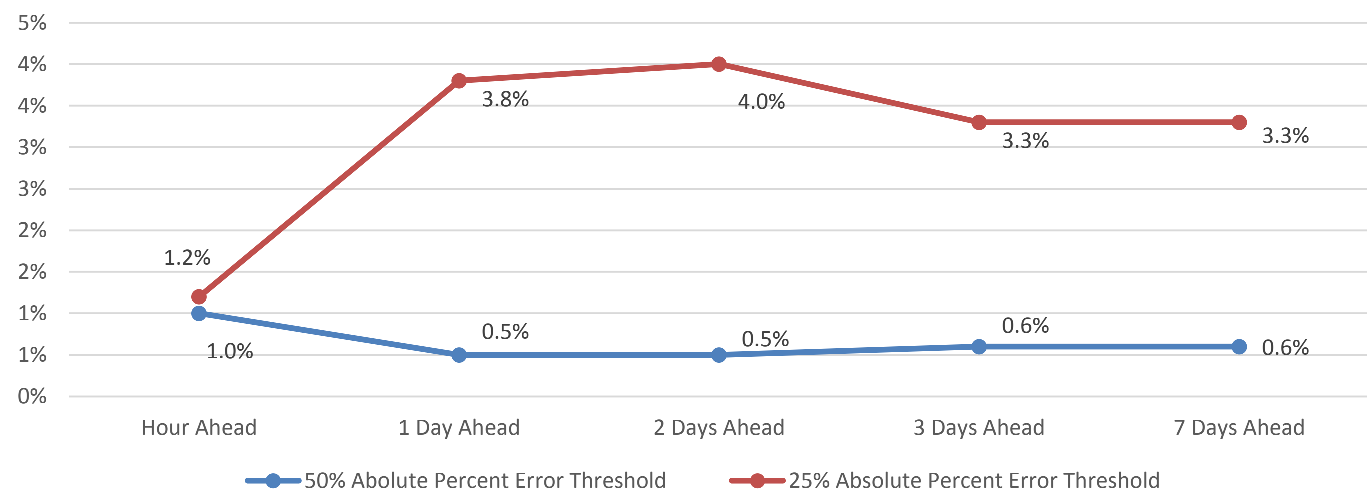
Figure 5: Stability Comparison: Reconciled Forecasts vs. Aggregated SmartMeter™ with 2-Day Lag: Percent of Forecasts with APE at Least 50%



To minimize occurrences of DOs or DOEs unknowingly using load forecasts with large errors in their switch plans, the project team developed a model for flagging forecasts which have inputs associated with larger load forecast errors. The confidence metric used in this evaluation is the percent of forecasts with absolute percent error at least 50% which the model failed to flag, and the percent of overall forecasts which were flagged.

The Figure 6 gives the frequency of a forecast having at least a 25% or 50% error and not being flagged by the EPIC 2.07 confidence model. Increasing the percent of overall forecasts which are flagged will result in fewer un-flagged large forecasts.

Figure 6: Confidence Illustration for Reconciled Forecast Model: Percent of Reconciled Forecasts with APE Above the Threshold, and Not Flagged



The similarity in the percentages for lead times of one day or more is due to those forecasts being driven mostly by the bottom-up load forecasts, whereas the top-down models influence the reconciled forecast more at the hour-ahead lead time. The value of the top-down model degrades with respect to the bottom-up model as the lead time increases.

4.1.3.5 Sensitivity Analysis

The forecasts produced in this task used actual weather data, rather than forecasts. The sensitivity analysis quantified the expected change to forecast accuracy statistics with respect to weather forecast error variance.

The table below summarizes the expected precision degradation for a typical distribution node due to temperature forecast error for lead times up to seven days, with the forecast error impact given in the right-most column of the table. Impacts were computed using historical mean absolute error (MAE) for temperature forecasts provided by PG&E's meteorology department (second column of the table), which were converted to standard deviation (third column of the Table 10) assuming forecast errors were equally likely to be positive or negative for each of these lead times, and follow a normal distribution.

Table 10: Expected Load Forecast Degradation Due to Temperature Forecast Error for a Typical Distribution Node

Forecast Lead Time	Temperature Forecast Error MAE	Temperature Forecast Error Standard Deviation	Expected Load Forecast MAPE Increase Due to Weather Forecast Error
Same Day	1.6	2.01	1.7%
1 Day Ahead	2.0	2.51	2.5%
2 Days Ahead	2.3	2.88	3.1%
3 Days Ahead	2.6	3.26	3.8%
7 Days Ahead	3.9	4.89	7.0%

This MAPE impact relationship was developed by comparing bottom-up forecasts for kWh to actual bottom-up kWh for the validation nodes. Bottom-up estimates were used because they are applicable for all distribution nodes, whereas top-down estimates are only used when there is sufficient SCADA instrumentation on the feeder containing the target node. Additionally, this offered a more conservative estimate of the potential accuracy degradation because top-down estimates are less weather sensitive because they incorporate recent trends in loading, whereas the bottom-up estimates did not. The sensitivity was computed on the aggregate kWh forecasts rather than load per phase, since the loading conversion was deterministic – any additional error resulting after the conversion was considered systematic (and correctable, with additional engineering data), rather than stochastic.

The combined error distribution for bottom up forecasts is given below

$$E_i \sim N(0, \sum_j \sigma_j^2 + \gamma^2 (\sum_j \beta_{ji})^2)$$

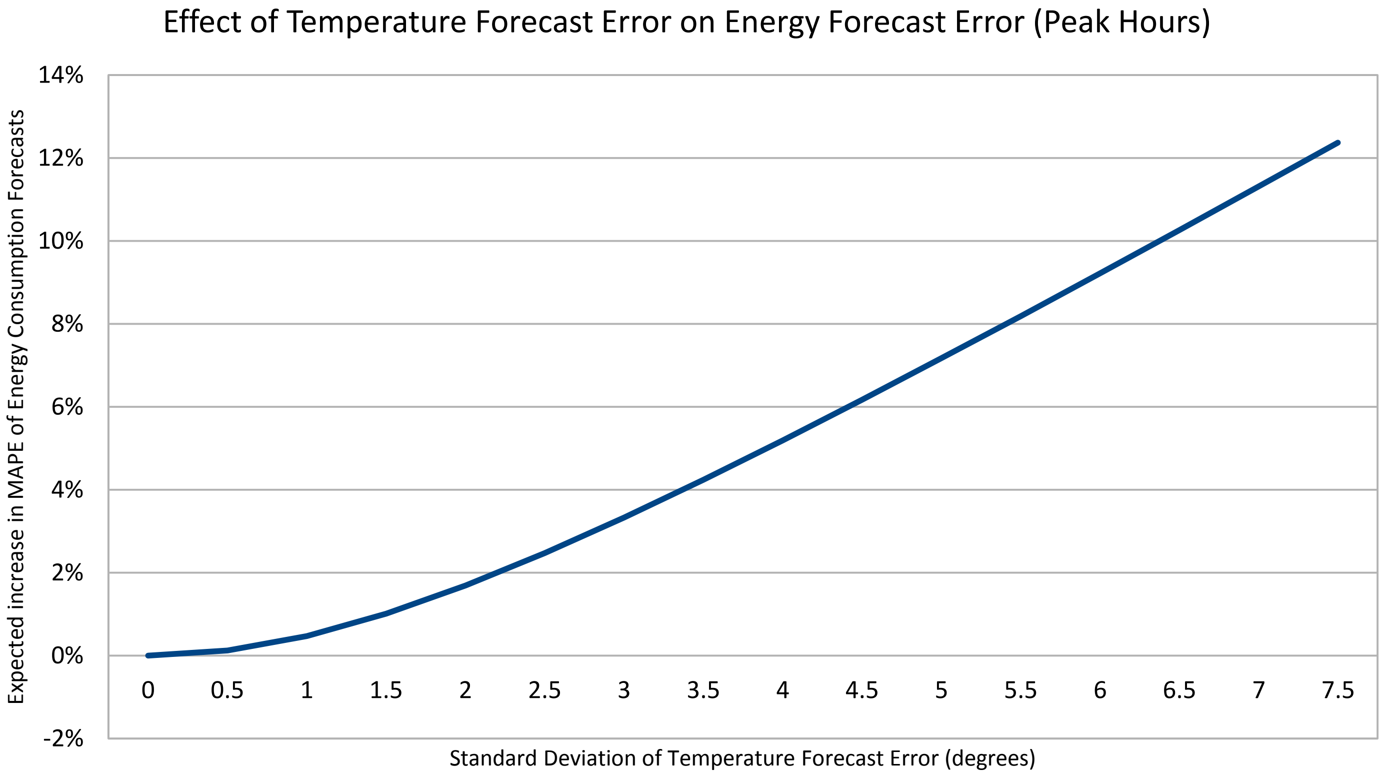
E_i is the error for distribution node i
N(a,b) is the normal distribution function with mean a and variance b. Here the mean is zero and the variance is $\sum_j \sigma_j^2 + \gamma^2 (\sum_j \beta_{ji})^2$
σ_j^2 is the variance for the load forecast error for SDP_j contributing to the loading on node i
γ^2 is the variance for the temperature forecast error, which is largely driven by the lead time for the forecast
$\beta_{ji}$is the regression coefficient for temperature associated with SDP j for node i

Figure 7 contains the increase in the MAPE for the load forecasts at the validation nodes with respect to the standard deviation of the temperature forecast error. For a standard deviation of 5, hourly temperature forecasts would be within 5 degrees of the actual temperatures (in either direction) about 68% of the time, and within 10 degrees about 95% of the time. To estimate the increase in MAPE for a forecast lead time, PG&E should estimate the corresponding standard deviation of the hourly temperature forecast errors (as was done in the table above), and locate the point on the curve with that value along the x-axis.

To convert MAE to standard deviation under the 2-sided error representation, the following formula is applied, assuming absolute error follows a half-normal distribution.

$$\sigma = \frac{\sqrt{\pi}(MAE)}{\sqrt{2}}$$

Figure 7: Effect of the Temperature on the Forecast Error



The expected month-specific MAPE degradation estimates are given in Table 11 below. The MAE, and corresponding MAPE impact, of weather forecasts is seasonal with a peak in June and trough between November and January, depending on the lead time. The shading in Table 11 corresponds with the MAPE impact, with yellow being the lowest and red being the highest.

Table 11: Monthly MAPE degradation estimates due to temperature forecast error (%)

Month	Day-Of	1-Day Ahead	2-Days Ahead	3-Days Ahead	7-Days Ahead
Jan	1.3	2.1	2.3	2.9	4.0
Feb	1.7	2.3	2.9	3.3	4.3
Mar	1.7	2.5	2.9	3.3	5.7
Apr	1.7	2.3	2.9	4.0	8.7
May	1.7	2.9	3.6	4.3	8.2
Jun	2.3	3.1	4.7	5.9	11.1
Jul	1.9	2.7	3.6	4.3	7.7
Aug	1.7	2.3	2.9	3.8	6.7
Sep	2.1	2.9	4.0	5.5	10.0
Oct	1.7	2.5	2.7	3.1	5.5
Nov	1.3	1.7	2.3	2.5	5.0
Dec	1.5	2.3	2.1	2.5	5.5

4.2 Task 2: Limited Scale Live Prospective Forecast Demonstration

4.2.1 Technical Development and Methods

This section describes the technical procedures used to produce and validate the distribution node forecasts specific to Task 2. The general methodology is described in Section 4.1.1, Technical Developments and Methods.

4.2.1.1 Data Ingestion and Computational Architecture

The computational platform was deployed to three PG&E servers composed of 32 CPU cores and 64GB of Random Access Memory each, and used in Task 2 to maintain synchronization with multiple operational PG&E data systems, as well as scale and run the analytics developed and described in Task 1. Two of the three servers were used for processing analytic scripts and one of the three servers was used for data ingestion and general platform orchestration.

A number of analytic scripts were developed or configured for use in Task 2. Several large scale analytic scripts utilized an API to the Apache Spark framework to accomplish parallel execution of the analytic across each worker node in the vendor platform cluster. Table 12 details the analytic scripts deployed to the vendor platform in fulfillment of system specifications.

Table 12: Analytic Process Execution Frequency

Analytic Type	Description	Execution Frequency
SDP Training Script	Training of hourly load forecasting model for each individual SDP.	Weekly
SDP Forecasting Script	Hourly forecasts across a 7-day forecast horizon, along with a 2-day “back-cast” for each individual SDP.	Nightly
SCADA Forecasting Script	Training and forecasting of hourly amps for any SCADA device at or downstream of a provided device node. This is an on-demand API because of the impact of dynamic switching events and the value/cost ratio of storing these forecasts for every device versus those of interest to operational decisions.	On-demand
Reconciliation Script	Combines bottom-up SDP and top-down SCADA forecasts to produce a single forecast for a selected device node and its downstream devices. This is an on-demand API because of the impact of dynamic switching events and the value/cost ratio of storing these forecasts for every device versus those of interest to operational decisions.	On-demand
Peer Model Assignment Script	Assignment of each SDP to a peer class used as a proxy model when an SDP is not AMI-enabled or there was insufficient data to compute a forecast.	Monthly

4.2.1.2 Unit Conversion

As described in Section 4.1.1.1, this project involved conversion from kWh to kVA and ultimately to amps per phase at the target distribution nodes. With Task 2 involving significantly more feeders than in Task 1, the project team received guidance, summarized in Table 13, from PG&E distribution engineers for the correct conversion formula under various phase designation scenarios.

Table 13: Mean Amps Per Phase Formula by Phase Designation at Target Node

Phase Designation	Mean Amps per Phase Formula
abc	kVAh/(kV Operating voltage*√3)
ab	kVAh/ kV Operating voltage
bc	kVAh/ kV Operating voltage
ac	kVAh/ kV Operating voltage
a	kVAh/(kV Operating voltage /√3)
b	kVAh/(kV Operating voltage /√3)
c	kVAh/(kV Operating voltage /√3)

The Phase Designation column in Table 13 provides the different phase representations found in the distribution system nodes. Designation “abc” is for a 3-phase wiring configuration, with the phases labeled “a”, “b”, and “c”. Two-phase (e.g., “bc”) and single phase (e.g., “a”) wiring was also found. Operating voltage refers to the normal voltage at the target distribution node for the forecast.

The conversion from kWH to Amps assumes 1) set power factor per customer class 2) no system losses 3) no power flow analysis which then does not take into consideration any reactive power supplied by capacitor banks. These are the limitations of this strategy.

4.2.1.3 Forecasting Under As-Switched Topology

To account for the as-switched topology in the forecasts, switch transition data was used along with the as-built Distribution Management System (DMS) topology tables to determine which SDPs were energized through the target node being forecasted at a given time. The key data sources for this were:

- **DMS LINE Table**. Provides as-built parent/child relationships for each distribution node. Since this is as-built, it contains redundant relationships that would be filtered out depending on the switch states at a given time. For example, an ID may have two upstream parent IDs, but only one of them would represent an energized line segment at a given time.
- **Temporary State Files**. Each node ID that is in a temporary (abnormal) state is included in a separate record in files provided every five minutes. The current status is always opposite to the normal status field in the record in this file. The current status field provides the phase-specific switch position value, with a “0” always indicating open, and “111” indicating the switch is closed for each of the three phases, for example. Since phase-level connections from node to node were not provided, parent-child segment between nodes were considered energized if the current status included a “1”. For example, “010” would be considered a closed switch.

A timestamp field identifies when the switch position transitioned from its normal to temporary state. There was no explicit timestamp for an ID transitioning back from the temporary to normal state, so the timestamp embedded in the 5-minute frequency file that first did not include an ID that had been in a temporary state was used as a proxy for the temporary-to-normal state transition timestamp.

- **Normal Status Bulk File**. This file provided the normal switch state position for all switches, for reference when a switch did not go into an abnormal state at any time during the data history.

4.2.1.4 Data Challenges

Linking SCADA Data to GIS Topology Data. To properly associate SCADA data with its corresponding node in the distribution topology, a mapping convention was needed. Three different mapping conventions were needed to correctly process SCADA data covering multiple device types. Through the course of the project, SCADA tag name conventions changed a few times, which required reconfiguring the mapping functions within the vendor platform data pipeline services.

Maintaining Temporal Switch States Across Time. Due to forecasting analytics training on historical data across significant historical periods of time (greater than 1 month), historical switch states had to be maintained in addition to the current switch state. The temporary switch sates file from DMS (every 15 minutes) includes switches in an abnormal state position. The normal state of all the switches is received from GIS. Both the files need to be merged to get the point in time position of all the switches. The timestamp of a return to normalcy can be inferred as at some point between two file deliveries.

4.2.1.5 Forecast Validation

As in Task 1, the forecast validation process considered three criteria: Accuracy, Stability, and Confidence. Details on these validation metrics can be found in Section 4.1.1.1, with details specific to Task 2 below.

SDP Accuracy. The accuracy metric for SDPs in Task 2 was symmetric absolute percent error (sAPE), which is defined as follows:

$$sAPE = \frac{|F_t - A_t|}{\frac{|A_t| + |F_t|}{2}}$$

Where F_t and A_t $are\ the\ forecast\ and\ actual\ values\ in\ time\ t$

Whereas the calculation of absolute percent error (APE) uses only the *actual* quantity in its denominator, the calculation of sAPE is different in that it leverages both the *actual* quantity and the *forecasted* quantity. This is beneficial when true values are very small and error measures such as MAPE can be dramatically inflated. This often lends itself to a misleading sense of error magnitude. Using mean or median measures of sAPE, however, facilitate a nearly identical interpretation to those which are based on APE, while partially addressing the issue of the measure being undefined when the actual reading is close to zero, as can happen with SDP kWh readings. Using this notation, the mean sAPE would be sMAPE and median sAPE would be sMdAPE[7].

[7] sMAPE and sMdAPE will be undefined when both the forecast and actual are zero.

Stability. Stability was only quantified for distribution node forecasts and not SDP forecasts.

Confidence. Same process used for Task 1, described in Section 4.1.1.1, was used for Task 2.

4.2.1.6 Data Summary

The Table 14 provides a summary of the various data elements used in Task 2 of the project. The platform data ingest processors were adapted to PG&E file formats for each of the following data sources described in Table 14. Each data source was delivered in a delimited file format, parsed, and stored in the vendor platform for subsequent analytics. The platform bulk loaded 2 years of historical data for the 2-AOR scope of Task 2.

Table 14: Summary of Critical Data Elements to the EPIC 2.07 Task 2 Project

Data Class	Data Elements	Coverage	Refresh Frequency for Ingest to Platform
Topology	• DMS tables: LINE, NODE, LOAD, CAPACITOR, DEVICE • SDP to transformer maps. • Abnormal switch state indicators • System-wide normal state switch positions	Full PG&E Distribution System, except normal state switch positions and SDP to transformer maps which are limited to the 2 AORs (5 & 6) 676 feeders from 188 substations in AOR 5 & 6	• Abnormal Switch state indicators updated every 5 minutes • SDP to transformer maps refreshed monthly
SDP-level time series data	• 15 and 60 minute kWh delivered and received from AMI meters • SDP-level PV generation forecasts and actuals	• 1,539,194 Unique SDPs • 65,896 SDPs Without SmartMeter™ Data	Daily, at an average lag of 2 days
Loading data at asset nodes in distribution topology	• SCADA meter - current by phase • Line sensor data – RMS current	2,344 SCADA/Line Sensor SCADA meter count varies by feeder (1 to 10+)	Daily, at an average lag of 6 hours
Weather Data (Hourly weather actuals)	Weather stations with sufficient historical data for retrospective load forecasts	Hourly actual and forecast temperatures from 160 PG&E stations throughout the system	Daily

4.2.2 Challenges

4.2.2.1 Data Challenges

Considerable efforts were made to scale up the small-scale backcasting analytical project in Task 1 to support forecasts for 2 AORs at an on-demand basis in Task 2. Configuration of analytics from a static environment to a dynamic data setting was an intensive process, as was the back end and user interface engineering configuration of the data feeds and storage to support the scheduled and on-demand analytics. Various issues forced unexpected delays to get the processes online, hence the limited seasonal validation in this report.

The following data challenges listed in the table below were encountered and addressed to enable the forecast demonstration in Task 2.

Table 15: Data Challenges Encountered on Task 2 of the EPIC 2.07 Project

Data Class	Data Issue	Project Team Resolution
Topology	No timestamp for abnormal-to-normal switch transitions	Transition timestamps assigned from metadata from first subsequent abnormal state file without record for the switchable device.
SDP-level time series data	65,896 SDPs without AMI meter data	Peer models assigned for all SDPs as backup for use in forecasting
SDP-level time series data	Incomplete monthly kWh data available for non-AMI SDPs	Peer models required the average daily kWh for calibration. For non-AMI SDPs the team received a limited transfer of this data, so mean average daily kWh by customer class was used as a proxy for this peer model input.
SDP-level time series data	PV forecast and actual data refresh inconsistent	Forecasted net demand (kWh delivered minus kWh received), implicitly accounting for PV generation instead of explicitly with forecasted and actual PV.
Loading data at asset nodes in distribution topology	Changing tag name conventions	Rewrote SCADA/Topology mapping code
Loading data at asset nodes in distribution topology	Irregular transfers for ingest	Worked with PG&E IT to reinstate transfer schedule or restart data ingester when irregular file formats are encountered.
Loading data at asset nodes in distribution topology	Data quality	Apply trigger for reconciled forecast to exclude SCADA nodes that did not have sufficient quality for use in this project, when nodes had missing or zero-valued readings
Multiple	Changing data exchange protocols – resulting from lessons learned through course of work	Rewrote data ingest pipelines multiple times

4.2.2.2 Technical Implementation Challenges

This project demonstrated the feasibility of distribution node level forecasts generated on-demand using as-switched topology data and ongoing feeds of loading data from SDPs and distribution SCADA and line sensor data, and weather data. Through this project, the following technical challenges were noted:

- **Forecast Data tracking**. The solution demonstrated in this project requires a number of time series data feeds for the forecast model training and estimation processes. The complexity and volume of the key data elements was a continual challenge to the project team. A robust tracking system for data would enable proactive mitigation of issues which could affect the forecasting process, and maximize the performance of the analytics.
- **IT Resourcing**. Several ongoing analytics jobs in this project were extremely taxing on the system provisioned, and often competed with each other for resources, requiring intensive monitoring and schedule coordination by the project team and remedial actions when components of the system went down, or costing in terms of turnaround time. Appropriate resource sizing in support of similar future deployments which plan for several simultaneous intensive analytics would facilitate consistency in the execution of key analytical processes.

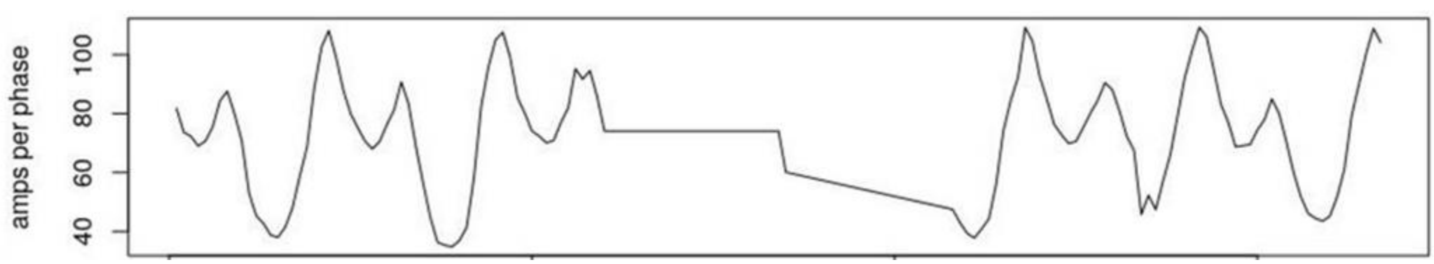
4.2.3 Results and Observations

The findings documented in this section provide insight to the question of whether the precision measured on a handful of validation nodes in Task 1 are maintained in a broader implementation across the 2 AORs in Task 2, accounting for the as-switched topology.

To validate forecasts, mean amps per phase at SCADA nodes were treated as "truth." The project team selected validation SCADA nodes which did not have missing data, constant-valued readings, or linear interpolations between readings, as they were assumed to be linked to the meter reading algorithm and storage of the readings, rather than the actual load being measured.

In Figure 8, the time-series plot of actual mean amps per phase illustrates linear interpolations used by PG&E's SCADA data management application. Because of this interpolation, this particular SCADA node was not used for validation.

Figure 8: Example SCADA node with linear interpolations for actual amps in two spans within the displayed range



To narrow the scope of evaluation, the project team selected 300 SCADA nodes which had fewer than ten hourly intervals flagged for being part of one of an apparent straight-line linear interpolation (as shown in the figure above) over a seven-day forecasting evaluation period starting June 4, 2018. This sample provided a sufficient number of SCADA nodes to validate accuracy with suitable precision.[8] The on-demand forecasting platform was not online early enough to cover multiple forecasting seasons. A seven-day evaluation period was chosen so that the forecast lead time could be evaluated (currently individual forecasts can go up to a seven-day lead time).

---

8 See figure on p. 15 showing relative precision of samples at various confidence levels for an example load research example. https://aeic.org/wp-content/uploads/2013/07/AMI_MDMWhitePaperFinal2.pdf.

4.2.3.1 Bottom-Up Actual kWh Converted to Mean Amps per Phase Accuracy Validation

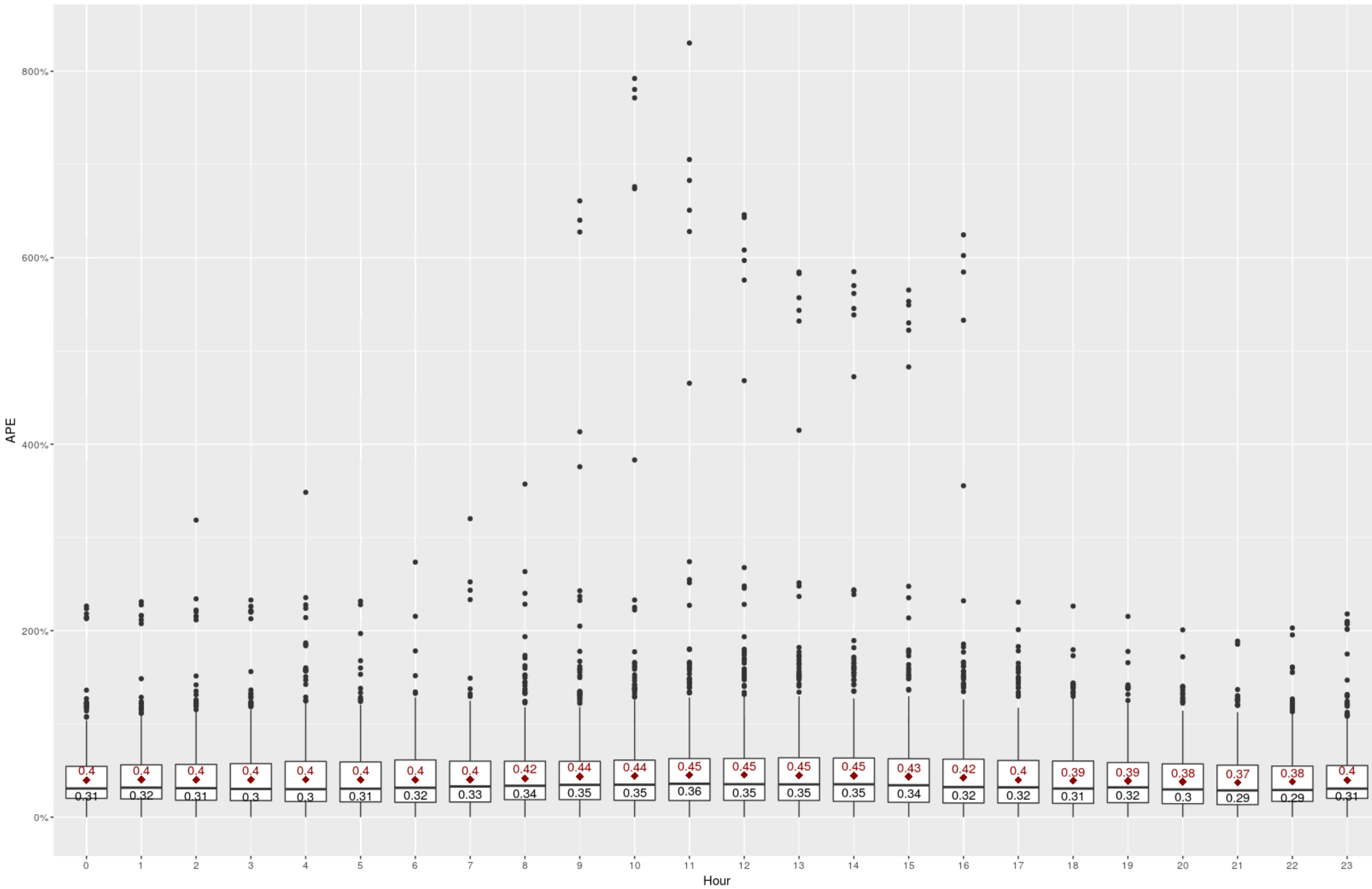
Prior to validating the reconciled forecast, the project team evaluated the accuracy of the bottom-up conversion process of kWh to mean amps per phase, using actual kWh from SmartMeters™. The reconciled forecasting accuracy is dependent on several factors apart from the forecasts themselves, including:

- Accuracy of distribution topology linkages from the target node and its corresponding downstream SDPs
- Accuracy of assumed power factors for converting kWh at the SDP level to kVA for aggregation
- Accuracy of the operating voltage and phase designation at the target node, and conversion formula from kVA to mean amps per phase
- Alignment between loading for timestamps from SCADA measurement extracts from data management system and hour-ending kWh (i.e., which timestamp(s) in the SCADA data extracts would reflect loading from kWh in AMI extracts?)
- Accuracy and completeness of switching records

While it was beyond the scope of this project to validate these systematic dependencies, the project team did compare aggregations of the bottom-up actual kWh to actual amps per phase from SCADA meters at validation nodes. The discrepancies suggest that an engineering review of the assumptions listed above, and other factors may be beneficial, and bolster the achievable accuracy of the forecasting analytics deployed.

Figure 9 represents the distribution of the APEs for the bottom-up aggregation of actual kWh by hour for all the validation nodes for the seven-day span starting June 4, 2018.

Figure 9: Boxplots of Bottom-Up APEs (Where Actual Historical kWh Are Converted to Amps), by Hour, for All Validation Nodes and for the 7 Day Span Starting June 4, 2018



This distribution shows the median (in black) and the mean (in red). The magnitude observed suggests that the achievable accuracy for forecasts derived from bottom-up aggregations of kWh is systematically limited, but may be improved after resolving potential issues noted in the bullets above.

4.2.3.2 Top-Down/Bottom-Up Reconciled Forecast Accuracy Validation

The reconciled forecasts were produced using the method described in Task 1 of this report, and configured for the as-switched live forecasting context. For the reconciled forecast validation, the SCADA data at the target nodes were not used in producing the forecasts, rather they were withheld to serve as an evaluation benchmark for "ground truth." For the reconciliation forecast validation, the project team started with the 317 validation nodes, then selected those on feeders with two or more validation nodes, since the validation process required one SCADA node to be held out, and the top-down forecast feeding into the reconciliation model requires at least one SCADA node forecast. For the resulting set of 289 validation nodes, the feeder list was divided in two groups – one for reconciliation model training (132 validation nodes) and one for the holdout group (157 validation nodes) for validation.

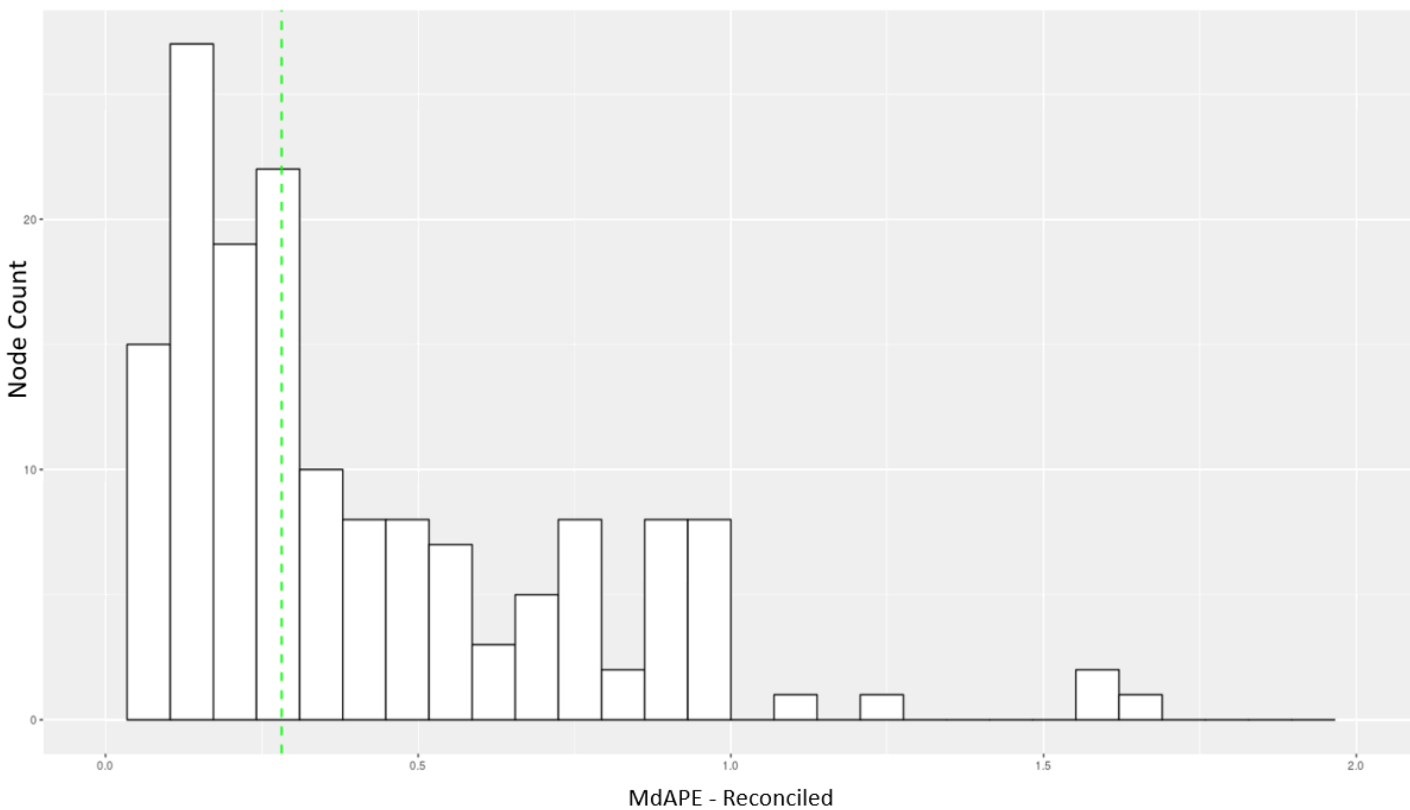
Reconciled forecasts were produced for the seven-day forecast starting June 4, 2018 on 289 SCADA nodes, which were located on feeders which had at least two SCADA meters. Top-down and bottom-up forecasts were fit for 132 of the nodes, with the remaining 157 serving as a holdout validation test sample for the reconciliation model.

As with Task 1, a machine learning classification tree model for estimating which of the three models would give the lowest forecast error was fit, given several candidate input variables, with the following terms used in the model:

- Count of end customer SDPs comprising the load at the target node
- Number of SCADA nodes on the feeder corresponding with the target node
- Count of end customer SDPs comprising the load at the SCADA nodes on the same feeder as the target node
- Difference between the estimated power factor for the target node and the power factors from the SDPs downstream from the SCADA nodes
- Proportion of total estimated kWh for the target node from SmartMeter™ based models

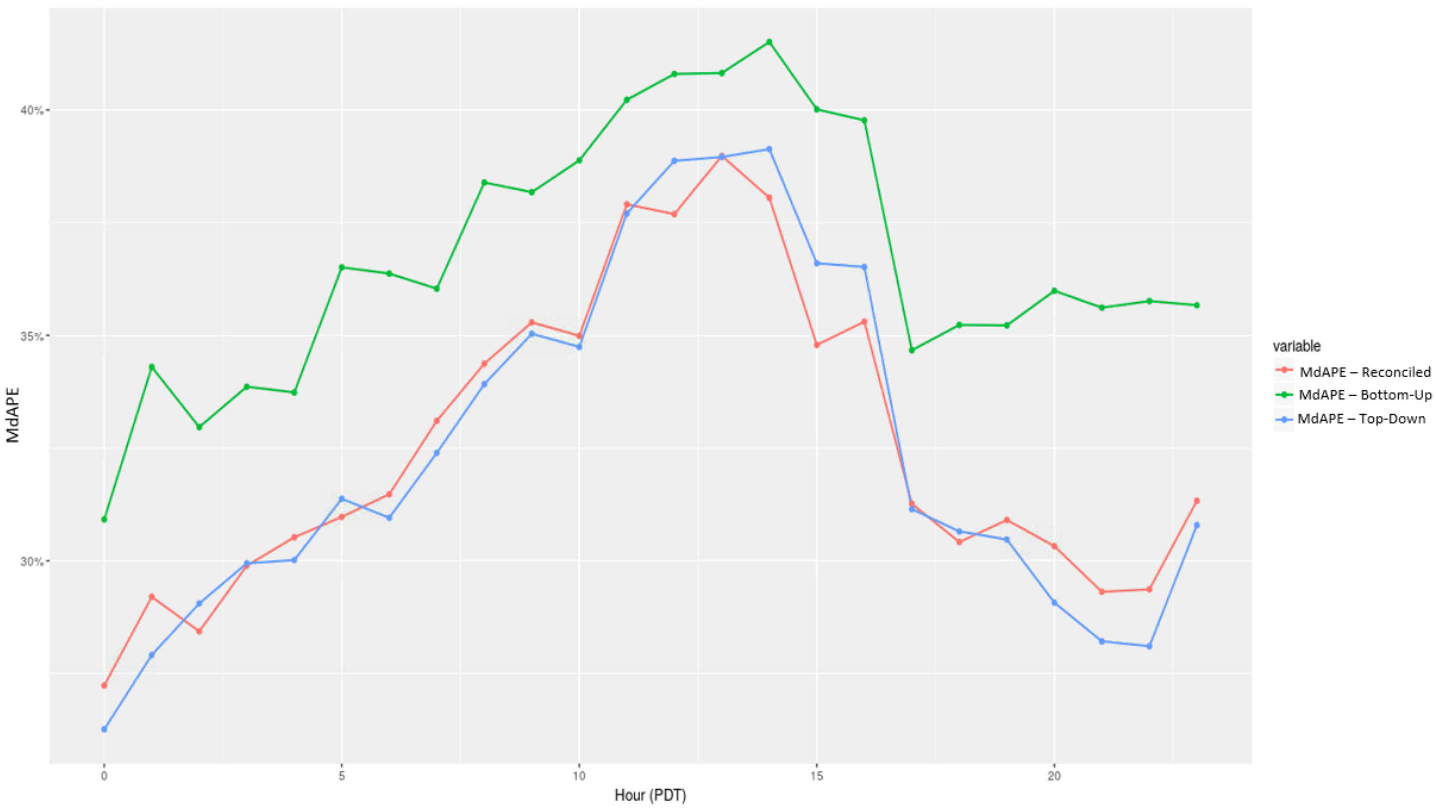
The distribution across the validation nodes is shown in the following figure. The median test node in the validation sample had MdAPE of 28%, shown in the dashed green line in the histogram.

Figure 10: Distribution of Reconciled Forecast Median Absolute Percent Error – by Validation Node



In Figure 11 below, the MdAPE for the reconciled forecast was 32% across all hours in the validation sample. By hour, the MdAPE peaks in midday, coincident with solar irradiance and temperature in early afternoon. The reconciled forecast and the top-down forecast were nearly identical across the hours of the day, with the bottom-up MdAPE higher, at 37% overall.

Figure 11: Reconciled Forecast Median Absolute Percent Error by Hour and Forecast Type on Validation Nodes

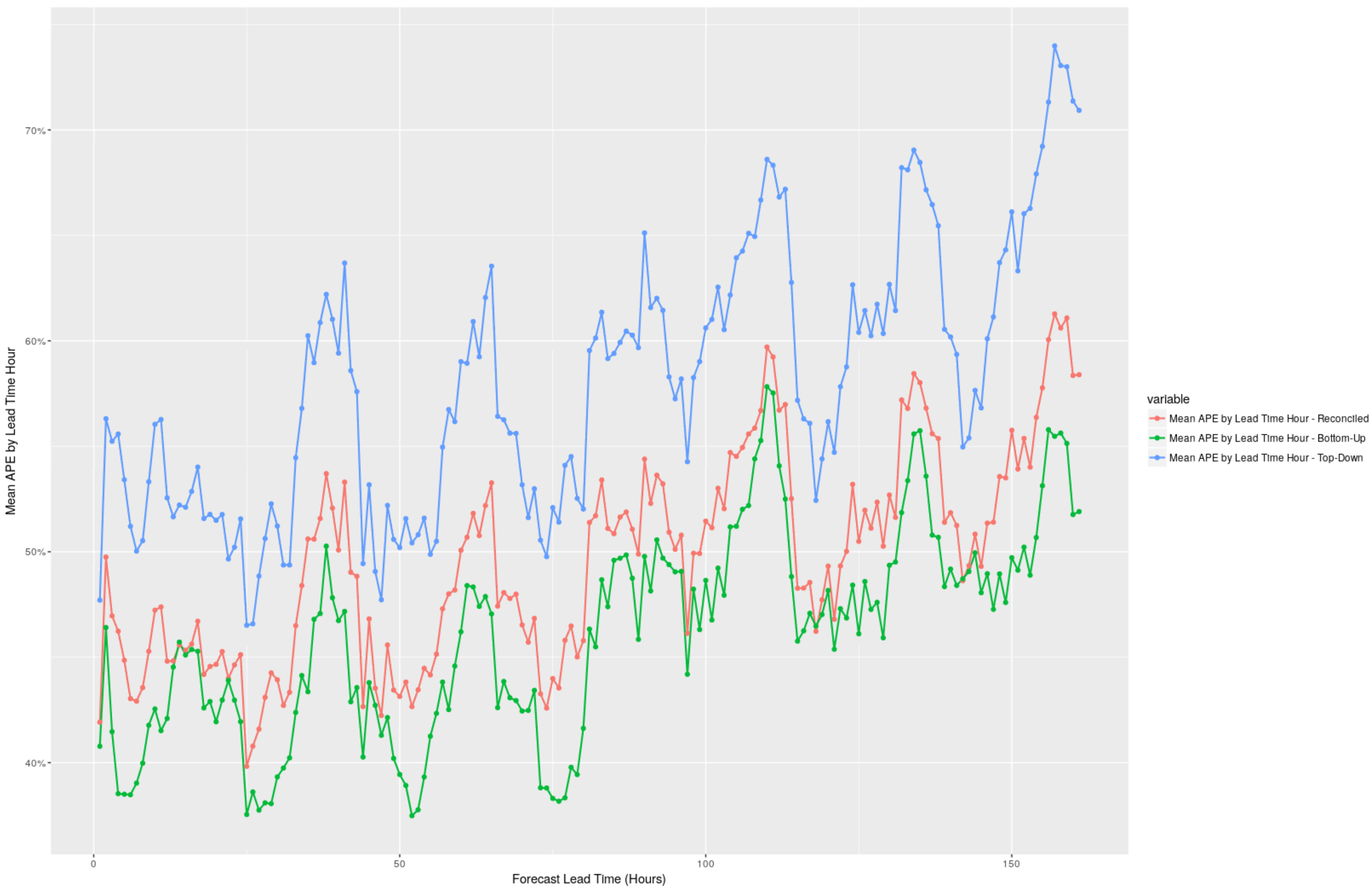


However, in Figure 12 below, the same data is used as in Figure 11 but the mean is displayed instead of the median. Now, the top-down forecast appears to perform worse than the reconciled forecast and that of the bottom-up forecast. This shows that there is still work required to obtain a robust solution. It does reveal that the bottom-up model is more robust than the top-down model, probably due to the impact of SCADA data limitations on top-down model performance.

Figure 12: Reconciled Forecast Mean of Absolute Percent Error by Hour and Forecast Type on Validation Nodes

The Figure 13 shows the upward trend in mean of the APEs by lead time hour (this time normalized by starting hour through the seven day forecast horizon, termed the ‘lead time hour’), while maintaining the inter-day seasonality shown in Figure 12.

Figure 13: Forecast Mean Absolute Percent Error by Lead Hours and Forecast Type



4.2.3.3 SCADA Model Forecast Accuracy Validation

Supervisory Control and Data Acquisition (SCADA) meters are most commonly installed at line recloser or circuit breaker nodes, and provide near real-time loading information for DOs and other End Users managing the distribution system. While useful, the SCADA meter data collected does not include forecasts for future loading. The forecasting analytics from this project can provide End Users insight into the loading on the devices with SCADA meters up to seven days in the future.

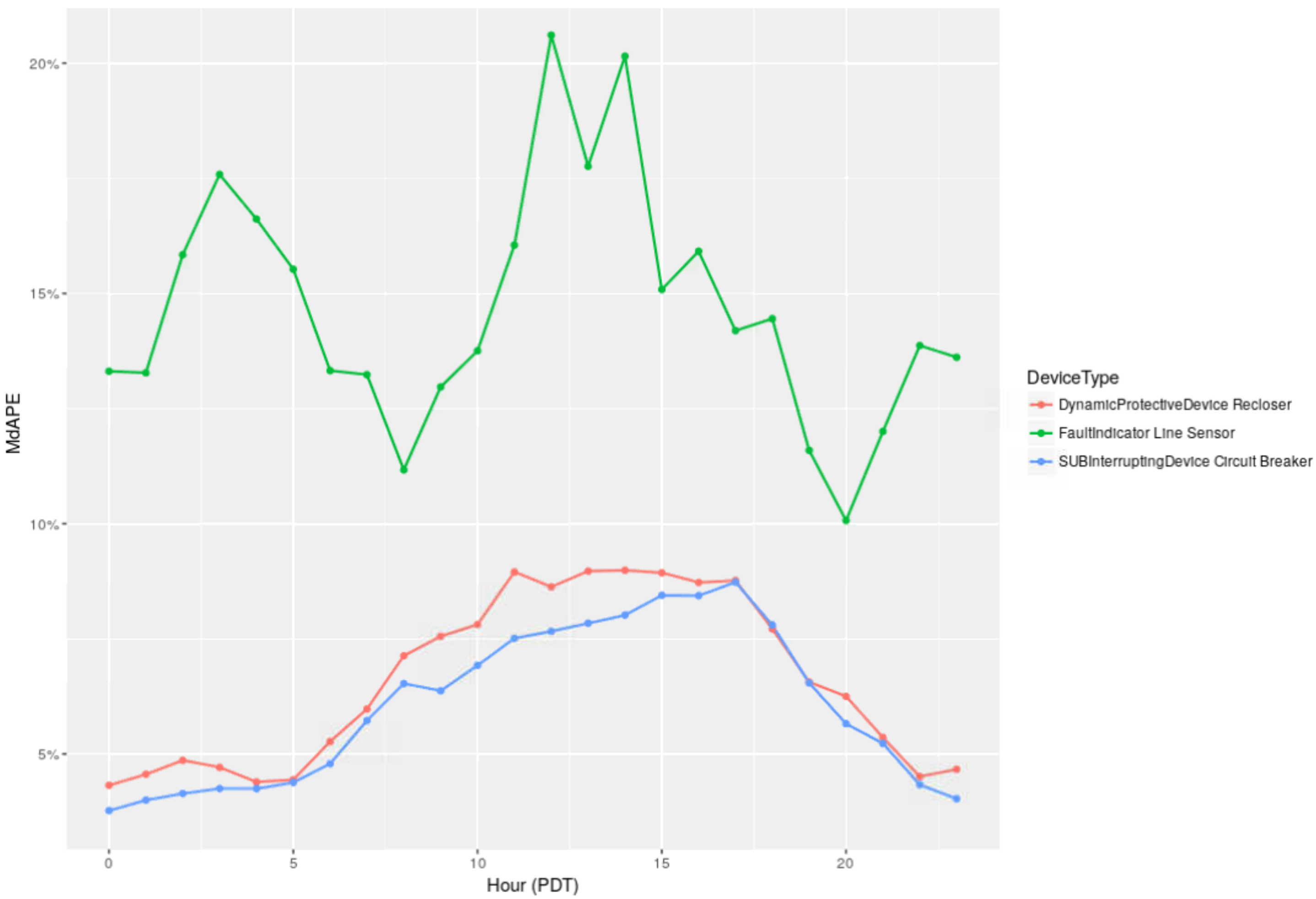
While the reconciled forecast methodology of which results are presented in the previous section can be produced for any distribution node that the topology connects with downstream SDPs (including SCADA nodes), The SCADA model forecasts presented in this section are trained on the actual historical data collected directly from the SCADA meters, rather than a bottom-up approach. Because of the direct connection to historical load data to train on, the SCADA models are typically more accurate than their bottom-up or reconciled forecast counterparts, but are applicable only to the specific node the model is trained on.

In Task 2 the project team deployed SCADA model analytics to nodes in the two AORs. To evaluate the accuracy of the SCADA data models, the project team analyzed the MdAPE for the validation nodes. A preliminary analysis showed that there was not a significant difference in the hourly MdAPE profile between forecasts generated on the two dates, so the

7 day forecasts starting June 4 forecasts are included in the summary plots below. Across all nodes and device types, the MdAPE was 6.7%.

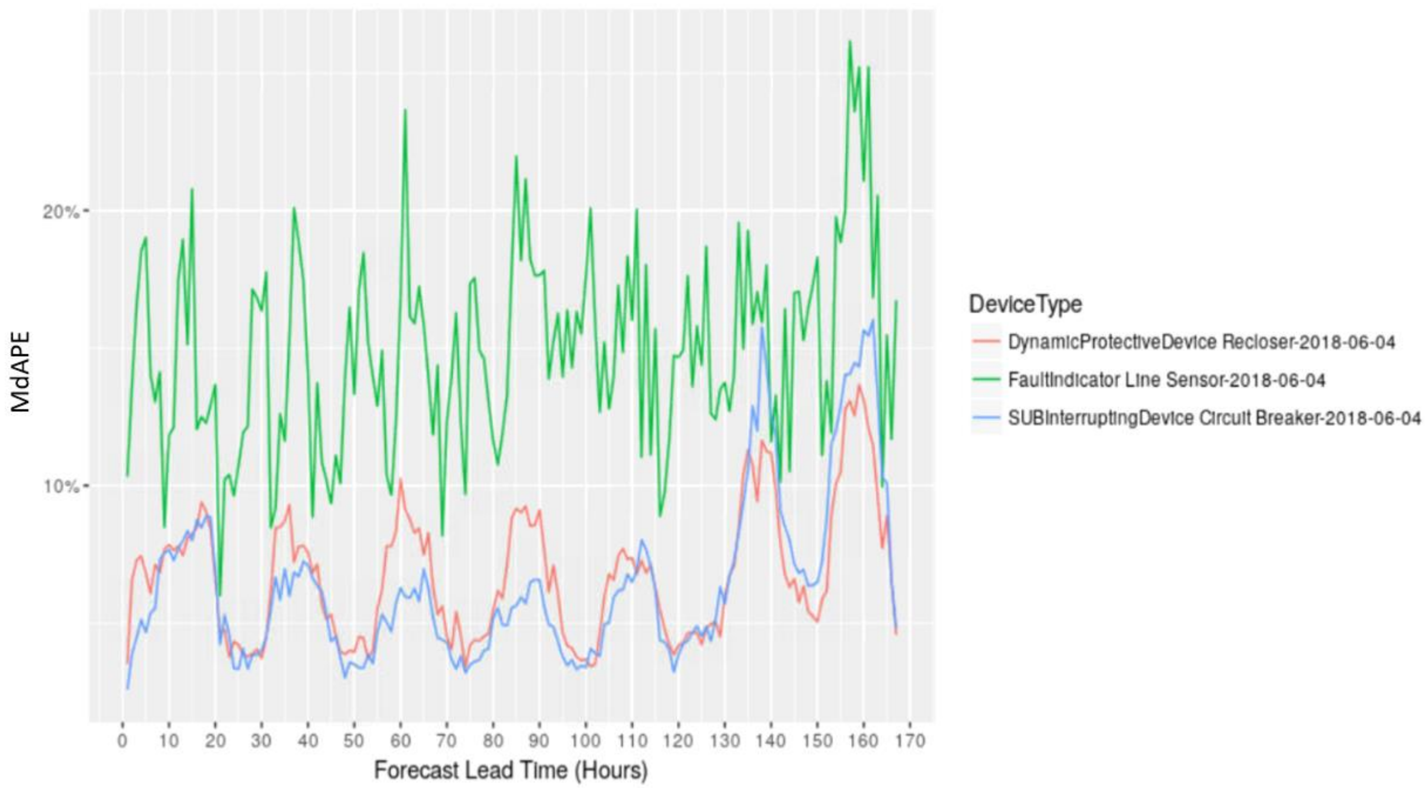
The Figure 14 below shows the MdAPE by hour and device type where the SCADA meter is installed. The MdAPE curve is much smoother and lower magnitude for Dynamic Protective Device Reclosers (reclosers) and SUB Interrupting Device Circuit Breakers (circuit breakers) likely because there were relatively fewer Fault Indicator Line Sensors (line sensors). Additionally, the line sensor readings may have had unidentified quality issues that impacted model performance and could have been installed on nodes with more volatile load. The MdAPE peaks in the middle of the day, coinciding with the period of peak solar irradiance and temperature, likely because solar irradiance was not used in the SCADA models deployed.

Figure 14: SCADA Model Forecast Median Absolute Percent Error by Hour and Device Type



The Figure 15 shows the MdAPE by lead time, to illustrate the level of degradation in the forecasts further in the future. There is an increase in the MdAPE for the circuit breaker and reclosers after lead hour 120, with no apparent trend in the first 120 hours, or in the Line Sensor time series throughout.

Figure 15: SCADA Model Forecast MdAPE by Lead Time and Device Type



4.2.3.4 SDP Model Forecast Accuracy Validation

This section describes the predictive performance of the SDP-level load forecasting models[9] driven by SmartMeter™ data flowing to the project team in 1 hour or 15 minute intervals. The SmartMeter™ time series data was labeled according to its channel, with separate channels for delivered kWh and received kWh. Received kWh represented net energy flowing from the customer to the distribution grid, due to PV systems or another distributed energy resource (DER). The SDP modeling objective was to predict the net delivered energy at an hourly level (15 minute data was aggregated to hourly for consistency across the SDPs), with net delivered energy equal to delivered minus received kWh. Net hourly delivered kWh precision estimates were captured for a randomly selected sample of 50,000 customers across both May and June of 2018. Details of the distribution of this sample are as follows:

- 42,973 Residential, 6,993 Non-Residential
- 3,576 SDPs with net meters for PV installations

The **Figure 16** and Figure 17 depict predictive performance in terms of symmetric MAPE by hour and by customer type (residential and commercial/industrial) for the months of May and June, respectively. The customer type field[10] maps to each SDP, with values “Res” (residential), “Com/Ind” (commercial or industrial), and very infrequently compared to these, a street lighting indicator, which was not included in the analysis. Both figures show that commercial/industrial customers had a lower sMAPE than residential customers overall, and particularly so between midday business hours.

[9] A SDP in this analysis corresponds with a unique service point identifier (SP_ID) value.
[10] The customer type field was sourced from the monthly periodic master meter data refresh to the project team.

Figure 16: SDP Level Symmetric MAPE by Hour and Sector – May

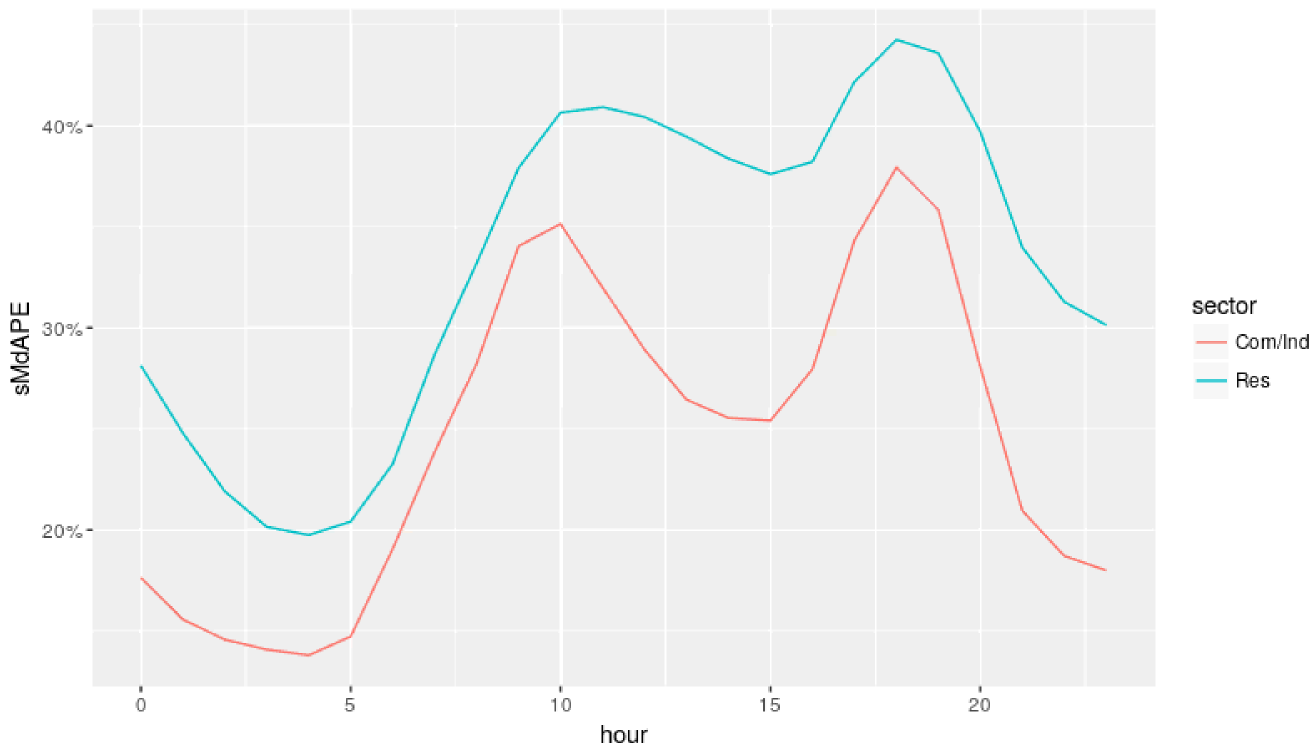


Figure 17: SDP level Symmetric MAPE by Hour and Sector - June

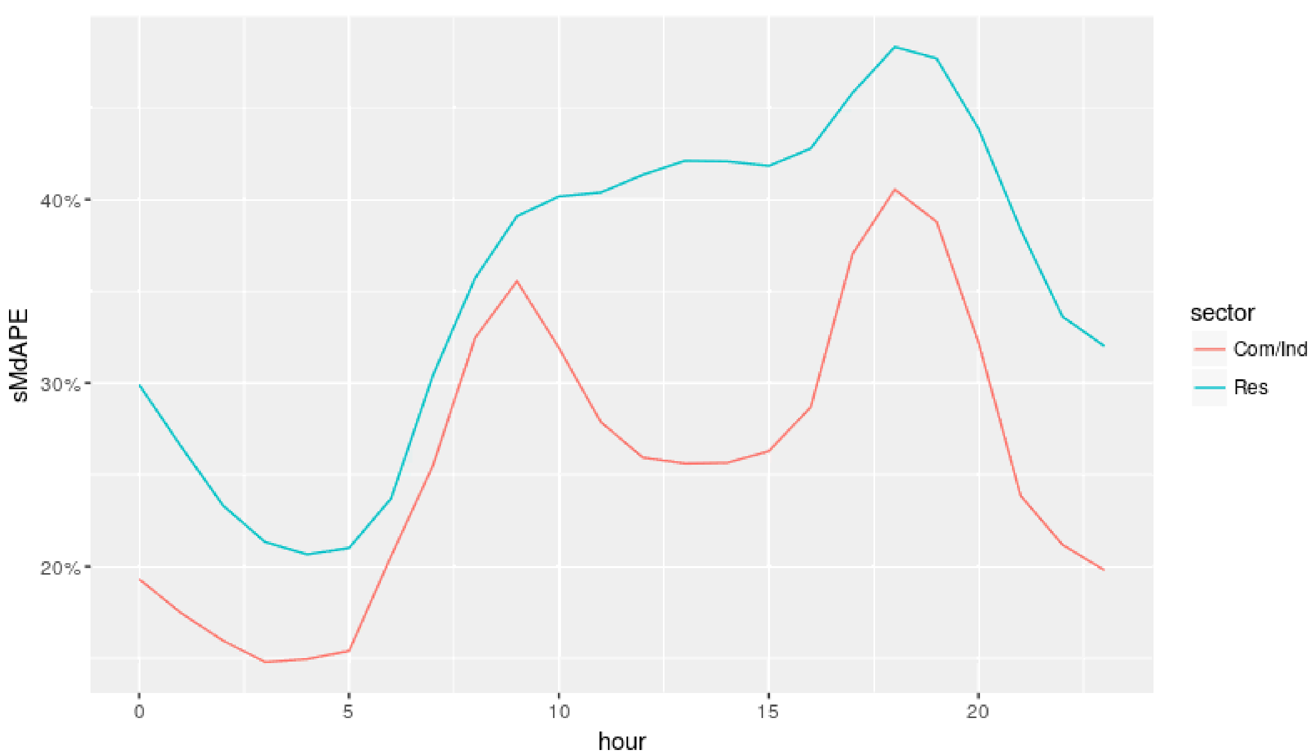


Figure 18 and Figure 19 depict sMAPE by hour and by whether or not the SDP has a PV installation, as indicated from the presence of a net energy meter. There is a late morning and an evening spike in sMAPE for PV SDPs coincident with hours most likely to have net demand from the grid close to zero. With both actual and estimated kWh close to zero, sMAPE is inflated because the denominator approaches zero.

Figure 18: SDP Level Symmetric MAPE by Hour and PV Indicator - May

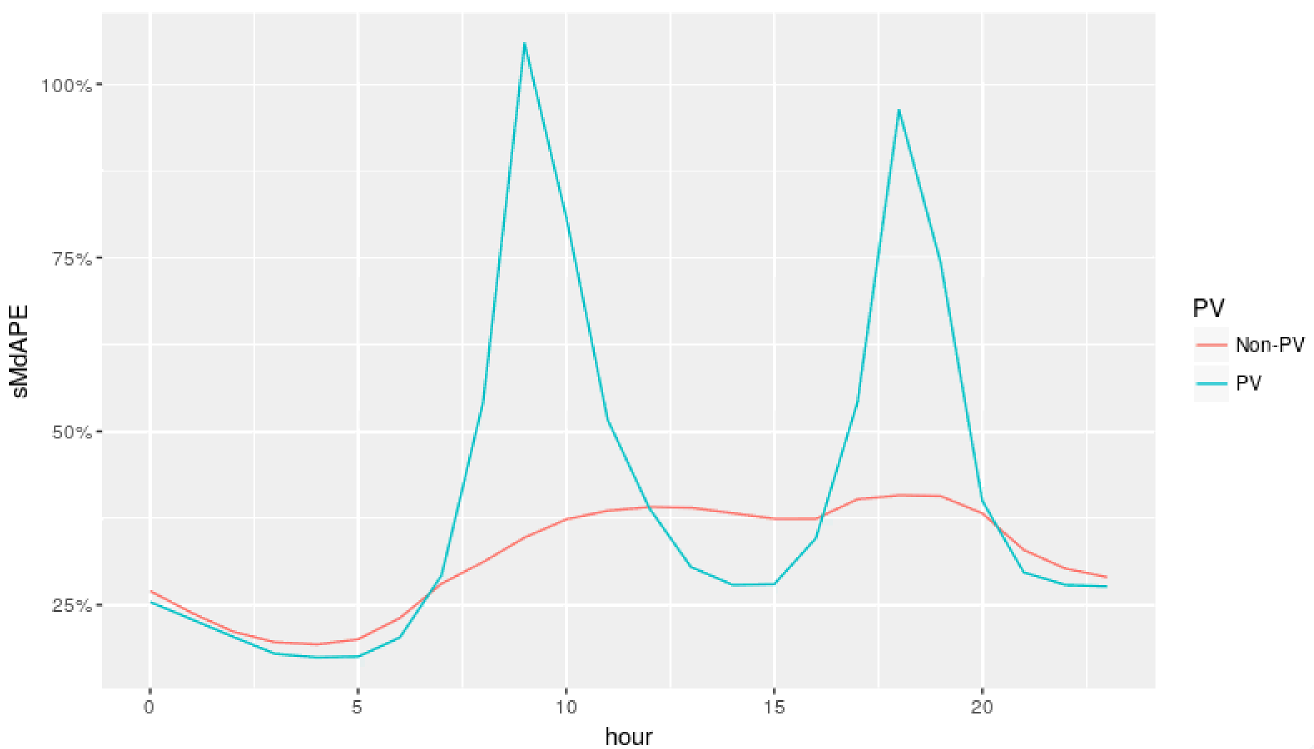


Figure 19: SDP Level Symmetric MAPE by Hour and PV Indicator - June

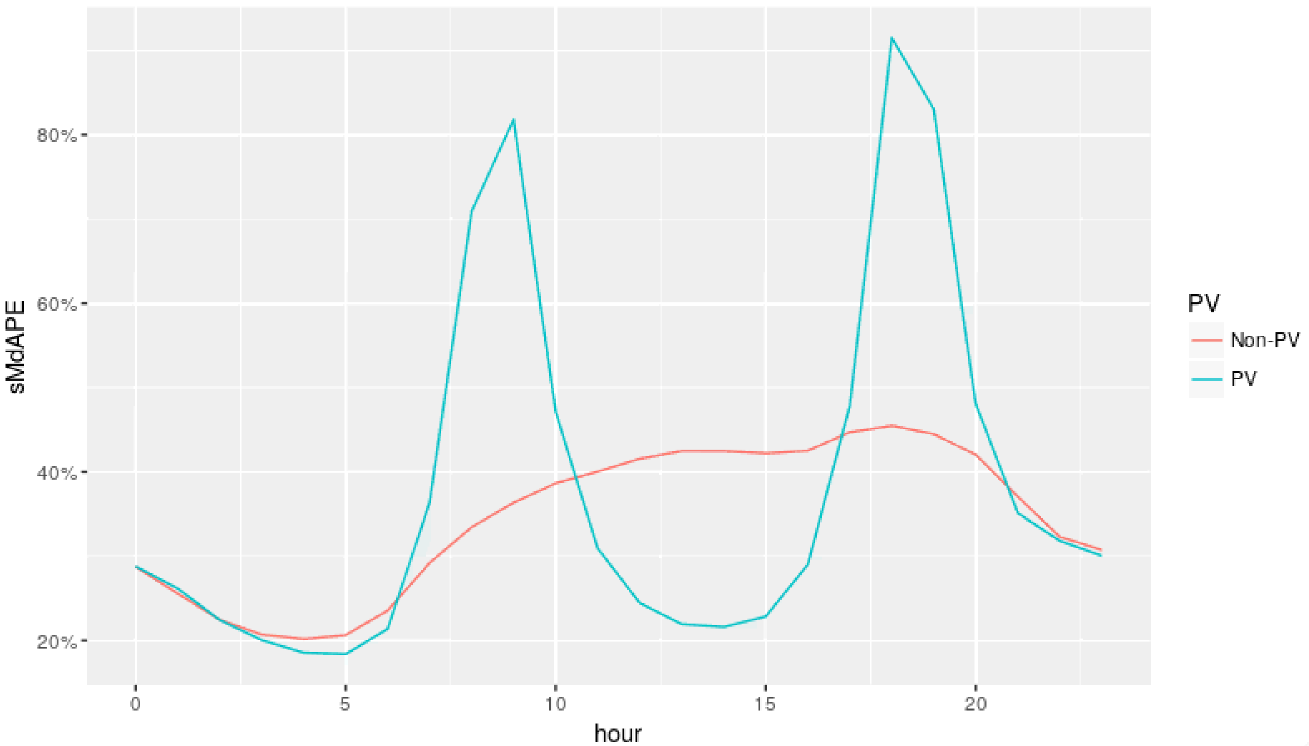
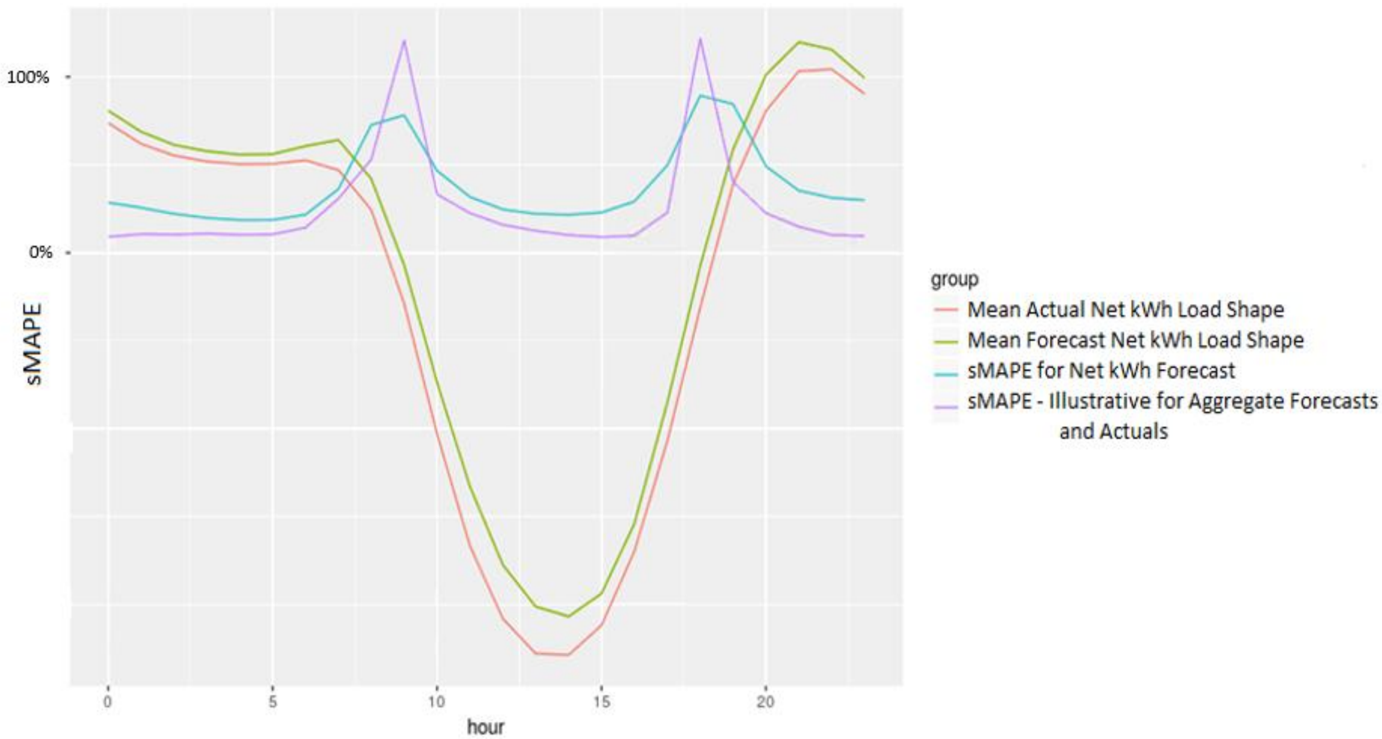


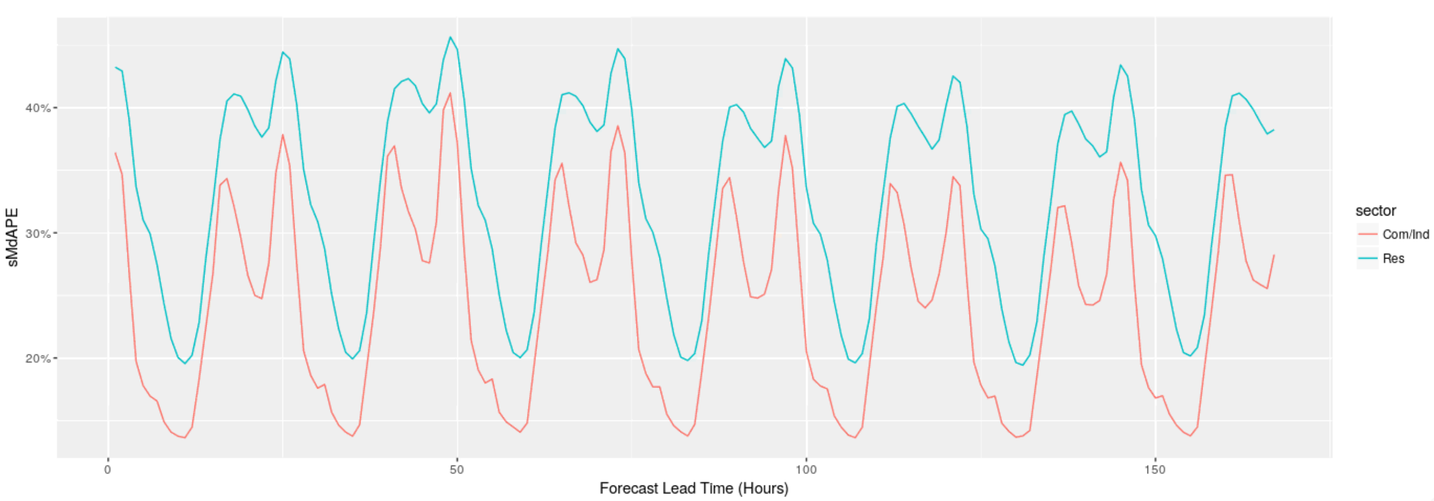
Figure 20 provides additional context for the dual spikes in MAPE for the PV customers, with the actual and forecasted load overlaying on the sMAPE plot (the axis scale corresponds with sMAPE rather than kWh). The purple sMAPE curve corresponds with the aggregate actual and forecast curves in green and red, for illustrative purposes. As noted above, the peak in the sMAPE occurs when the mean net kWh approaches zero.

Figure 20: SDP-Level Symmetric MAPE by Hour with Load Shape for Forecast and Actual for Context



To measure the impact in the sMAPE of longer lead time temperature forecasts (which tend to have more error than shorter lead time forecasts), the project team examined sMAPE by lead time for the 168 hours following the forecast generation on June 4, 2018. Figure 21 shows no apparent upward trend in MAPE across the lead hours.

Figure 21: SDP Level Symmetric MAPE as a Function of Forecast Lead Time by Sector



4.2.3.5 Stability and Confidence Analysis

By PG&E subject matter expert request, the project team computed the overall proportion of forecasts with error 50% or higher produced by bottom-up, reconciliation and SCADA methods at SCADA nodes where the validation was conducted. By PG&E subject matter expert request, the project team computed the overall proportion of forecasts with error 50% or higher produced by bottom-up, reconciliation and SCADA forecasts at SCADA nodes where the validation was conducted. SCADA models require time series at the node being estimated, whereas bottom-up and reconciliation models can be applied at any distribution node. SCADA model requires time series at the node being estimated, whereas bottom-up and reconciliation models can be applied at any distribution node. Table 16 summarizes the mean ratio of forecast to actual load for each model type, as well as the probability (Pr) of the ratio being less than 0.5 (i.e., forecast lower than the actual by 50%) and greater than 1.5 (i.e., forecast higher than the actual by 50%).

Table 16: Stability of Forecasts Using SCADA Models vs. Reconciliation Models

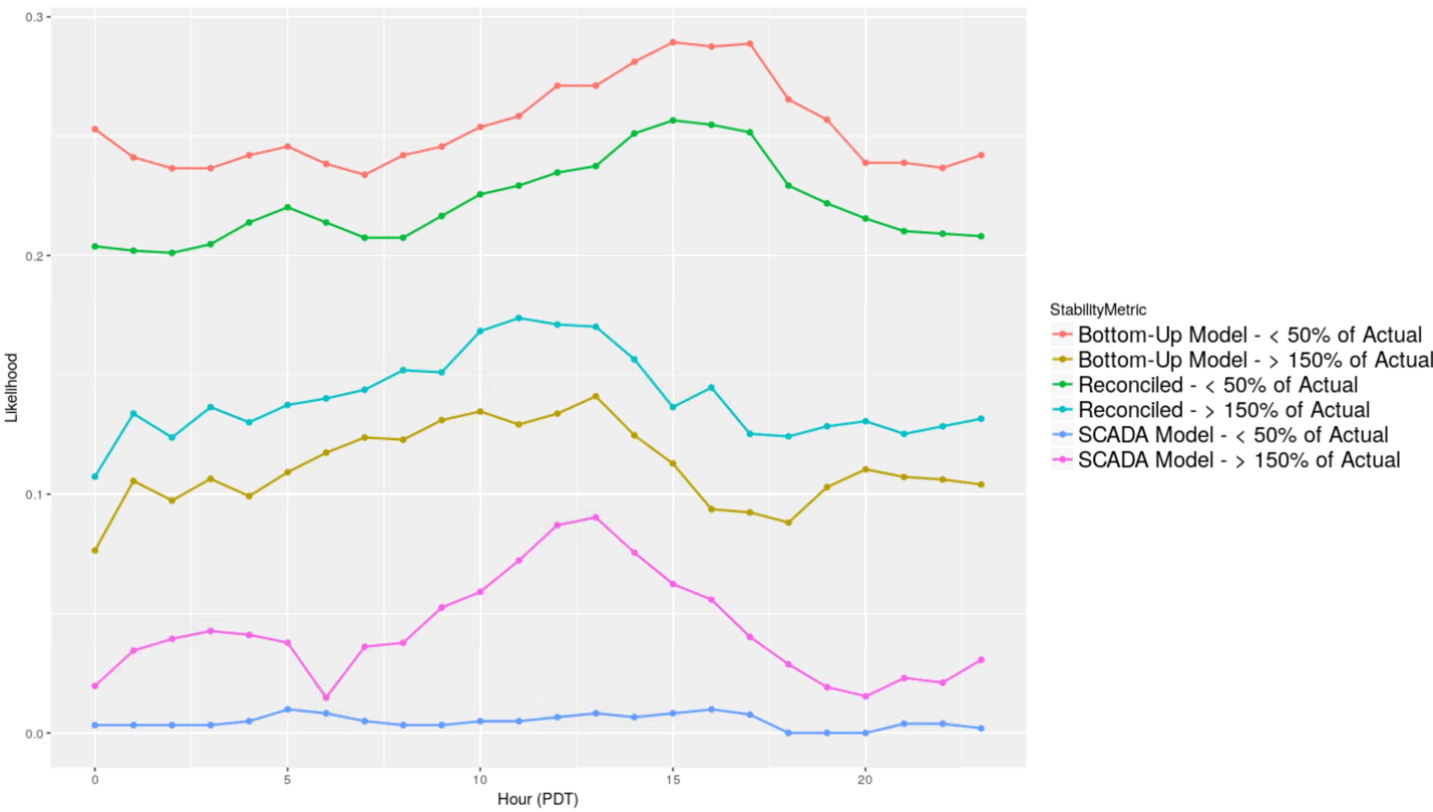
Model Type	$\text{Median}\left\{\frac{Forecast}{Actual}\right\}$	$\Pr\left\{\frac{Forecast}{Actual}\right\} < 0.5$	$\Pr\left\{\frac{Forecast}{Actual}\right\} > 1.5$
Bottom-Up	0.975	.254	.112
Reconciled	1.009	.222	.141
SCADA	1.028	.005	.044

The median ratio of bottom-up model forecasts to actual amps per phase was 0.975, indicating a slight underestimate. The median ratio for reconciled estimates was 1.009, less than one percent above the actual on the median. The SCADA model ratio was 1.028, 2.8% above the actual on the median, across the nodes and hours in the validation group.

Approximately 25% of the bottom-up forecasts were less than 50% of the actual amps per phase, and 11% were 50% or more above the actual amps per phase value. A slightly lower percentage of the reconciled forecasts (22%) were 50% or below the target, and a slightly higher percentage (14%) were 50% or more above target, as compared with the bottom-up forecasts. Only 0.5% of the SCADA forecasts were less than 50% of the actuals, and 4.4% were at least 150% of the actual amps per phase. Only 0.5% of the SCADA forecasts were less than 50% of the actuals, and 4.4% were at least 150% of the actual amps per phase.

Figure 22: Likelihood of at Least a 50% Error by Forecast Type shows the stability by model type across the hours of the day. The bottom-up and reconciled models have a similar hourly stability profile. The SCADA model stability is flat for the 50% or less series and has a pronounced peak for the 150% or more series near hour-ending 12-13.

Figure 22: Likelihood of at Least a 50% Error by Forecast Type



The confidence model, as with Task 1, was a classification tree model for flagging intervals with absolute percent error of 50% or more based on tracking metrics associated with the node being forecasted.

Tested out of sample, the confidence model correctly flagged reconciled model based forecasts 85% of the time. The confidence model correctly flagged reconciled model based forecasts 85% of the time. For SCADA models, the confidence model correctly identified 80% of the relatively low number of errors greater than 50% of the actual amps per phase.

5 Value Proposition

The purpose of EPIC funding is to support investments in technology demonstration and deployment projects that benefit the electricity customers of PG&E, SDG&E, and SCE. EPIC 2.07 has demonstrated that a near real-time load forecast was possible.

The load forecasting analytics developed and configured in this project are for distribution operators, operations engineers, capacity planners, and others who are responsible for managing the loading on the PG&E distribution grid on a real-time and forward looking basis.

Unlike existing analytical tools which provide loading estimates from backward-looking analytics based on seasonal peaks or lagged energy consumption, these predictive analytics are calibrated to the expected conditions for the time the forecast is to be made, using a multitude of attribute and interval data sources.

Near real-time load forecasts can provide information which will help distribution operators restore power outages faster and more safely due to improved understanding of available loading capacity across all nodes in the distribution grid, at a lower cost than a significant increase in SCADA installations where blind spots exist currently. Even with full SCADA instrumentation, forecast models like those developed in this project would need to be implemented to provide visibility beyond real-time to the upcoming seven to ten days.

5.1 Primary Principles

The primary principles of EPIC are to invest in technologies and approaches that provide benefits to electric ratepayers by promoting greater reliability, lower costs, and increased safety. This EPIC project contributes to these primary principles in the following ways:

- **Greater reliability**: Providing forecasted load visibility to distribution engineers and operators to allow them to manage distribution switching for both planned and unplanned events more quickly and with less switching steps.
- **Lower costs**: Having accurate predictive load forecasts may reduce the number of switching steps required to perform maintenance and restoration and therefore reduce operational costs.
- **Increased safety and/or enhanced environmental sustainability**: Having look ahead visibility into dynamic grid load conditions will allow operators to better plan for maintenance and enhance operational decision making awareness.

5.2 Secondary Principles

This EPIC project contributes to the following three secondary principles: societal benefits, GHG emissions reduction, economic development; and efficient use of ratepayer funds.

- **Societal benefits:** Having a forecasted visibility of the grid may support the operator in restoring power faster as well as allow planners to find the best time and conditions for maintenance, thereby impacting fewer customers.
- **Efficient use of ratepayer funds:** With load forecasts, distribution engineers may optimize the system on a shorter time horizon given dynamic loading conditions. Their goal is to maintain reliable power as safe and efficiently as possible.

5.3 Accomplishments and Recommendations

5.3.1 Key Accomplishments

- Successfully built a platform to ingest and process AMI, SCADA, weather, PV generation, and topological data in real time, at a scale never before implemented at PG&E
- Produced hourly load forecasts from 2 days in the past to cover latency in receiving SmartMeter™ data, to seven days in the future for all distribution device classes of interest and individual customer meters in two of the eight AOR regions within PG&E's service territory in under 4 hours
- Developed a reconciled forecasting method that leveraged both bottom-up kWh converted to amps per phase and top-down forecasts using SCADA data
- Integrated as-switched topology into forecasts
- Produced stability and confidence models for distribution node forecasts

5.3.2 Key Recommendations

The following recommendations should be considered by any organization that endeavors to implement real-time load forecasting capabilities similar to those developed within this project:

- Test forecasting methods on smaller sets of data before deploying capabilities at scale
- Ensure that the data platform has sufficient processing power to support computationally-intensive forecasting methods, and ensure that the ability to scale up is addressed in platform development
- Leverage data completeness and quality flag tracking analytics to screen and prepare data prior to conducting any forecasting
- Deploy data completeness and quality flag tracking analytics in parallel to forecasts to provide insight to potential issues with forecast accuracy and available data to produce the forecasts
- Focus on the quality of the underlying models & assumptions for the aggregation of actual kWh converted to mean amps per phase. The accuracy of the aggregation is impacted by assumptions for power factor and system losses, as well as the load flow model. Distribution engineers and data scientists would need to work together to address these issues and ensure strong forecasting models are employed.
- Perform robust validation to compare performance across the different forecasting methods
- Follow standard data science techniques to ensure that the classification tree model of the reconciled forecast method is robust to overfitting:
    - Select optimal hyperparameters on training set (needs to include all parts of the model – feature selection, dimensionality reduction, the learning model itself) using grid search nested inside a cross-validation to find the best hyperparameters for each fold. The best hyperparameters minimize model error.
        - Examine the optimal hyperparameters for each fold to ensure model stability
        - Once stability has been established, then record the optimal hyperparameters
    - Use the optimal hyperparameters to train the model on the training set
        - Use cross-fold validation to calculate a robust in-sample error
    - Iterate through the above two steps with different classification algorithms (linear classifiers, decision tree, random forest, extreme gradient boosting, etc.) to identify the best performing algorithm
    - Run the final model on the test data and examine final model error

5.4 Technology Transfer Plan

5.4.1 Path to Production

Near real-time load forecasting capabilities should be part of PG&E's future Integrated Grid Platform (IGP), where all of the input data required for the forecasts is planned to be integrated. The learnings from this project will be used to inform the design of the forecasting capabilities in IGP. Improved meter phasing data is planned to be made available through deployment of algorithms develop in EPIC 2.14 – *Automatically Map Phasing Information*, which will improve load flow modeling and the aggregation of kWh converted to mean amps per phase.

5.4.2 Investor-Owned Utilities' Technology Transfer Plans

A primary benefit of the EPIC program is the technology and knowledge sharing that occurs both internally within PG&E, and across the other Investor-Owned Utilities (IOU), the CEC and the industry. In order to facilitate this knowledge sharing, PG&E will share the results of this project in industry workshops and through public reports published on the PG&E website. Specifically, below are the information sharing forums where the results and lessons learned from this EPIC project were presented or plan to be presented:

Information Sharing Forums Held:

2018 DistribuTECH
San Antonio, Texas | Jan 2018

2017 Utility Analytics Week
San Antonio, Texas | Nov 2017

5.4.3 Adaptability to Other Utilities and Industry

The following findings of this project are relevant and adaptable to other utilities and the industry:

Near Real-Time Environment with Ongoing Data Feed Integration

This project used numerous data sources and ingested a large amount of data in near real-time, including the as-switched model. T. For optimizing the performance of the application, a combination of relational database for complex queries, No SQL database to store time series data and in memory data store for performance improvement was used. The architecture is horizontally scalable and can be easily extended to other AORs by provisioning additional servers. The learning of this project can be applied to other upcoming projects such as for the implementation of an IGPor Distributed Energy Resource Management System (DERMS).

Real-time Load Forecast

Based on the data available at other Utilities, some of the models used in this project could be reproduced to build their own load forecasting engine. Depending on the data available at each Utility, various options are possible. For example, a SCADA forecast with a top-down model may be the only possible option if the utility does not have SmartMeter™ installed. This type of work requires both a good understanding of the grid, as well as strong data scientists to create models that would be performant.

5.5 Data Access

Upon request, PG&E will provide access to data collected that is consistent with the CPUC's data access requirements for EPIC data and results.

6 Metrics

The following metrics were identified for this project and included in PG&E's EPIC Annual Report as potential metrics to measure project benefits at full scale.6 Given the proof of concept nature of this EPIC project, these metrics are forward looking.

D.13-11-025, Attachment 4. List of Proposed Metrics and Potential Areas of Measurement (as applicable to a specific project or investment area)	Reference
1. Potential energy and cost savings	
b. Total electricity deliveries from grid-connected distributed generation facilities More accurate predictive forecast will allow PG&E to operate the grid more reliably and have a higher penetration of DER.	5.2
3. Economic benefits	
a. Maintain / Reduce operations and maintenance costs Due to the increase visibility, less switching steps would be required, lowering the operation costs.	5.1
5. Safety, Power Quality, and Reliability (Equipment, Electricity System)	
a. Outage number, frequency and duration reductions Real-time load forecast allows operating engineers to have higher visibility and act faster when outages occur.	3.1
c. Forecast accuracy improvement Forecast distribution loads on each phase in short-term scenarios allow flexible switching that can lead to increased safety, power quality and reliability.	3.1
i. Increase in the number of nodes in the power system at monitoring points If implemented, the top-down forecast will provide a proxy estimation of the load at each distribution nodes.	3.1
7. Identification of barriers or issues resolved that prevented widespread deployment of technology or strategy	
a. Description of the issues, project(s), and the results or outcomes Combining all the data sources near real-time to get a better understanding of the grid	1.5
b. Increased use of cost-effective digital information and control technology to improve reliability, security, and efficiency of the electric grid (PU Code § 8360) To get the equivalence of visibility of the distribution grid with other technics, it will take more time and cost more.	5.1
c. Dynamic optimization of grid operations and resources, including appropriate consideration for asset management and utilization of related grid operations and resources, with cost-effective full cyber security (PU Code § 8360) This project defined the requirement for future IGP applications	5.2
8. Effectiveness of information dissemination	
d. Number of information sharing forums held Twice	5.4.1

7 Conclusions

While challenges remain for improving the analytics deployed in this project for use by PG&E End Users, this project demonstrated that on-demand level forecasts for loading at distribution nodes are feasible. Load forecasting is not an easy task, but having multiple models to use across various levels of the grid helps to compensate for the potential gaps of each model. Overall, the results produced by the load forecasting models were inconclusive. Also, the use of median to assess forecast accuracy can hide the impact of outliers.

On average, the forecasts were on target, as indicted by the minimal bias observed. One of the key takeaways was that systematic errors – i.e. unrelated to the stochastic models driving the forecasts – can be significant drivers of forecast errors, and represent a logical next step for anyone attempting to drive down forecast errors. All the models built into this project assumed a set power factor per customer class, no system losses, no power flow analysis, and did not take into consideration any reactive power supplied by capacitor banks. These limitations of this approach would require expertise from both distribution engineers and data scientists to improve upon. Additionally, data quality flags for SCADA data measurement readings and bolstering the tracking of inbound model input data would further improve the forecast accuracy.

A challenge with working with an external vendor is that it can limit the visibility into their methods, especially when their methods are proprietary. This can make it difficult to support them to improve their models. Moreover, without a full understanding of the models, building in additional functionality is challenging, and this makes a utility dependent upon the external vendor for future enhancements.