Toward Scalable Stochastic Unit Commitment - Part 1: Load Scenario Generation

Yonghan Feng, Ignacio Rios, Sarah M. Ryan, *Member, IEEE,* Kai Spürkel, Jean-Paul Watson, *Member, IEEE,* Roger J-B Wets, and David L. Woodruff, *Member, IEEE*

***Abstract*—Unit commitment decisions made in the day-ahead market and during resource adequacy assessment are critically based on forecasts of load. Traditional, deterministic unit commitment is based on point or expectation-based load forecasts. In contrast, stochastic unit commitment relies on multiple load scenarios, with associated probabilities, that in aggregate capture the range of likely load time-series. The shift from point-based to scenario-based forecasting necessitates a shift in forecasting technologies, to provide accurate inputs to stochastic unit commitment processes. In this paper, we discuss a novel scenario generation methodology for load forecasting in stochastic unit commitment, with application to real data associated with ISO-NE. The accuracy of our methodology is consistent with that of point forecasting methods. The resulting sets of realistic scenarios serve as input to rigorously test the scalability of stochastic unit commitment solvers, as described in a companion paper.**

***Index Terms*—Demand forecasting, load modeling, stochastic processes, scenario generation, stochastic unit commitment.**

I. Introduction

CONSTRAINTS on thermal generation unit operation requires them to be committed well in advance of when they may be needed to provide power. Typically, scheduling decisions for a day d are made on day $d-1$ based on forecasts of uncertain quantities such as hourly load and renewables output; such quantities are generally aggregated across the buses in a load zone. In the context of traditional deterministic unit commitment procedures, such forecasts take the form of point or expected-value quantities – representing a single time series for each forecasted quantity. Uncertainty associated with such forecasts is addressed by maintaining a non-trivial level of generation reserves, which compensate for deviations from the predicted quantities as day d operations proceed.

In contrast, stochastic unit commitment procedures [1], [2] assume the availability of a number of forecast scenarios, each representing a distinct time-series of the forecasted quantities. Throughout, we use the term *scenario* in a narrow sense, representing a full specification of all random data required to specify a unit commitment problem, with associated probability of occurrence. In aggregate, the set of scenarios should represent the range of possible behaviors on day d. By explicitly representing forecast uncertainty through sets of scenarios, it should be possible to significantly decrease the generation reserve margins and consequently reduce overall system operation costs [3]. However, the need for multiple scenarios imposes fundamentally novel requirements on forecasting technologies, which have yet to be adequately addressed.

Our goal in this paper is to present approaches and data sources for generating quantifiably accurate and realistic load scenarios for use in stochastic unit commitment. We focus on load, as opposed to renewables production, to tractably scope our study. The procedures described below extend to wind and solar plant production, which we will address in a future contribution. We demonstrate the ability of RTO and other operator entities to generate accurate load scenarios for use with stochastic unit commitment procedures, using data that is readily available. Our experiments proceed in the context of publicly available data from ISO-NE. The resulting scenarios are then used in the companion to the present paper [4] in order to rigorously test the scalability of a stochastic unit commitment solver.

The remainder of this paper is organized as follows. Our methodology for data transformation and fitting of load regression models is detailed in Section II. We then describe our procedures for generating load scenarios in Section III. Comprehensive experimental results are then presented for data associated with ISO-NE in Section IV. We conclude with a summary of our results in Section V.

II. A Novel Load Forecasting Methodology

When forecasting load, the information available to operators on day $d-1$ includes weather forecasts for day d, historical records of previous weather forecasts, and historical actual system loads. Historical system load data exhibit temporal patterns that vary according to season of the year, day of the week, and hour of the day. While some temporal load patterns are predictable based on knowledge of business hours and diurnal light patterns, the portion of load derived from heating and cooling (both industrial and residential) depends strongly on weather. And while numerical weather prediction models have become increasingly accurate over the past few decades, there remains significant uncertainty associated with day-ahead weather forecasts. The challenge for system operators is to form an accurate and comprehensive picture of the day-ahead load, which not only includes point forecasts of the load in each hour, but also acknowledges the precision (or lack thereof) associated with those forecasts. To address this challenge, we introduce a novel optimization-based method to

Yonghan Feng and Sarah Ryan are with Iowa State University, in Ames, Iowa. Jean-Paul Watson is with Sandia National Laboratories in Albuquerque, New Mexico. Roger Wets and David Woodruff are with the University of California Davis, in Davis, California. Kai Spükel is with the University of Duisburg-Essen, in Duisburg, Germany. Ignacio Rios is with the University of Chile, in Santiago, Chile.

Manuscript received November TBD, 2013; revised XXX.

develop a stochastic model for the load on day d based on the weather forecast available on day $d-1$.

A. Background

Common methods for short-term load forecasting can be categorized as either based on artificial intelligence or statistical techniques [5]. Methods from artificial intelligence, such as neural networks, are widely used but do not provide probabilistic information that could be used to generate multiple probability-weighted scenarios. Among statistical approaches, which can provide the required probabilistic information, the most prevalent methods are time series and regression models. Due to limited space, we do not provide a complete review but refer the reader to recent surveys such as [6]. Instead, we now highlight samples of statistical approaches for load forecasting from the recent literature, focusing on achieved accuracy and limitations for purposes of scenario generation.

The weather variable most commonly used to predict load is temperature, due to its influence on heating and cooling requirements. Other variables considered include humidity and cloud cover, although their impacts on load are much smaller than that of temperature. Humidity increases load in the summer, again due to cooling requirements. In contrast, cloud cover increases load in the winter (due to increased lighting requirements) and reduces load in the summer.

Liu et al. [7] analyze the nonlinear relationship between temperature and load, using estimates derived from a nonparametric regression method. They fit a time series model to the residuals of the load-temperature regression and considered lags of 1, 24, and 168 hours in their day-ahead forecasting model. Using actual historical temperature and load data obtained from a US utility, they demonstrate an out-of-sample mean absolute percent error (MAPE) of 1.2% for their 24-hour-ahead forecasts. Hong et al. [5] develop a multiple linear regression model of load that considers temperature, hour, type of day, and month as independent variables; the model additionally contains a linear term trend, and terms to capture interactions among the independent variables. Using actual weather data to predict hourly loads for a US utility over a one-year time period, they obtain an out-of-sample MAPE of 4.6%. Black [8] also uses a multiple linear regression model to examine the influence of weather on load, but instead focuses on summer weekdays in the region served by ISO-NE. Time-of-day effects are captured through a separate regression model for each hour of the day, each considering temperature, humidity, solar and radiation as independent variables. The out-of-sample MAPEs yielded by these models averaged 2-3% for the whole New England region and 3-4% for individual subregions such as Connecticut and Southeast Massachusetts.

While "hind-casting" studies of the type above (which use *actual* historical weather data as input) are useful for identifying factors that influence hourly loads, they do not possess the accuracy or precision of load forecasts available in practice – which necessarily rely on day-ahead weather forecasts, as opposed to actual quantities. In terms of scenario generation, a drawback of time series-based methods is that when used to forecast more than one step ahead, uncertainty propagates through the lagged terms. This resorts in significant distortions in forecast variability for future time periods, rendering them unsuitable for building stochastic process models for day-ahead load forecasts.

An alternative approach to short-term load forecasting is to identify similar days within a historical database, where the similarity is based on weather, day of the week and time of year. For example, ISO-New England identifies up to five similar days drawn from the same season with the same day-type according to similarity of their actual temperatures to the forecast temperature of the given day as well as similarity of forecast loads in the last hour of the previous day [5]. Our method has some commonality with this approach, in that we create segments of days that are similar in some sense. Then, within each segment we employ a functional regression method to approximate the probability distribution of load in each hour of the day ahead.

B. Methodology Overview

We use a multi-step procedure to control for season and type of day, and then approximate the relationships between weather forecast data and the distribution of hourly load sequences within segments of similar days. Starting from a historical database of day-ahead hourly weather forecasts and corresponding actual loads, our load forecasting methodology proceeds as follows:

1) Identify *date ranges*, or "seasons," in which the relationship between weather and load – disregarding day-of-week effects – is likely to be similar. This ad hoc characterization qualitatively accounts for diurnal light patterns, heating vs. air conditioning impacts, and sociological factors such as whether school is in session.
2) Within each date range, transform the data to account for day-of-the-week and zonal differences within the system. Then, segment the data into bands based on forecast temperature. Data segmentation can in principle proceed using multiple forecast quantities (e.g., humidity). However, our experiments indicates these additional factors do not significantly improve forecast accuracy.
3) Within each segment, approximate the relationship between weather and load via a regression function. Additionally, approximate the distribution of residuals associated with the resulting regression model.

Following completion of the segmentation and approximation steps, the procedure for generating load scenarios for a given day d (with associated weather forecast generated on day $d-1$) is given as follows:

1) Identify the date range DR to which day d belongs. Within DR, identify the temperature segment to which the weather forecast for day d belongs.
2) Apply the regression function associated with the identified segment to the weather forecast for day d.
3) Generate forecasted load scenarios for day d using distributions of the forecast errors.
4) As necessary, perform inverse transformations of the load sequences to match the day of the week and the zone.

The segmentation and approximation steps are fully described in the remainder of this section. Details of the load scenario generation process are described subsequently in Section III.

C. Estimating regression curves

The main idea to build the regression curve is to consider the weather forecast from day $d-1$ to day d and with this information build a regression function. For this we use 2nd order epi-splines [9] that minimize the deviations from the observed load at the day d. For each day d in a given segment it's assumed that the following information is available: the hourly load, l_h^d, and the weather prediction for day d made on day $d-1$. We use the temperature for all months and in summer, we also add dew point.

We split the 24 hours of day d in N_R sub-intervals $(h_{k-1}, h_k]$ of length $\delta = 24/N_R$, and this determines the total number of coefficients that need to be estimated, $2 \cdot (N_R + 2)$. In the summer, our regression curves will be built by relying on two epi-splines of order 2, one associated with temperature, and a second one associated with dew point. Implicitly, our construction also assumes that the curves against which we are fitting are twice differentiable (not necessary C^2). In addition to the parameter N_R there is a curvature parameter, κ. The impact of the paramters is explored in Section IV-B.

Further, we assume that the load can be represented as the sum of two components:

- a non-weather component, ie, a component which doesn't depend on the weather forecast and which is related to the normal behaviour of the people at each segment,
- a weather component which depends on the weather forecast.

It's natural to think that the non-weather component depends on the segment considered: independently of the weather, people in winter use a different amount of energy than in the summer. So, for each segment we estimate the baseline load, which is the average load for each hour in the segment.

Full details of regression using epi-splines for predicting demand based are provided in [10]. We represent the regression function from the weather for day d to $\mathbb{R}^{24}$, which was fit using data from an some set of days $\hat{\mathcal{D}}$ as

$$r(d; \hat{\mathcal{D}}).$$

We use d when is the weather forecast and d^o when the observed weather is used to estimate transformation functions. A more formal statement would inclue the regression control parameters, but we omit these in the interest of clarity.

D. Data Transformation

Within a date range, we use transformations to combine data from disparate day types and zones, prior to constructing regression and error distribution models. Without such transformations, there is typically insufficient historical data to yield accurate forecasts. Inverse transformations are performed to create load forecasts for particular day types and zones.

Suppose we are given observed load profiles $l^d = (l_1^d, ..., l_{24}^d) \in \mathbb{R}^{24}$ for a range of dates $d \in \mathcal{D}$. A portion of the the load is dependent upon weather factors, but load profiles also depend on the *type* of the day, e.g., load patterns differ between weekends and weekdays. In our analysis, we consider six day types: one for each weekday and one representing weekend days and holidays. We denote the set of day types as $\mathcal{T}$. The set of all dates belonging to a day type $T \in \mathcal{T}$ is denoted by $\mathcal{D}_T$. Clearly, each date maps to a unique day type, such that $\bigcup_{T \in \mathcal{T}} \mathcal{D}_T = \mathcal{D}$ and $\mathcal{D}_T \cap \mathcal{D}_{T'} = \emptyset$ if $T \neq T'$.

While a regression model could be developed for each day type, this would decrease the amount of data available for fitting significantly. Instead, we compute a transformation for each day type to a standard reference day, which we arbitrarily select as Wednesday. The transformation is easily inverted, so that observed loads can be transformed to "Wednesday" and forecast loads can be transformed back to the original day type. In our analysis, we consider linear transformations to allow for straightforward forward and inverse computation. We use observed weather to find a linear transformation from each day type to Wednesday, but the transformation itself is not based on observed weather so load forecasts and scenario generation can be based entirely on data that is available in advance.

For each $d \in \mathcal{D}_T$, let $w^d = (w_1^d, ..., w_{24}^d)$ denote the expected loads for our reference day type (Wednesday) corresponding to the native day type loads $l^d = (l_1^d, ..., l_{24}^d)$. We assume that for $\forall T \in \mathcal{T}$, $\mathcal{D}_T \neq \emptyset$. For each $d \in \mathcal{D}_T$ of day type T, w^d is computed as a regression $r^o(d^o; \mathcal{D}_{wed})$, where d^o denotes the observed weather for day d.

Our goal is then to find a 24×24 matrix A_T such that

$$A_T \, l^d \;\approx\; w^d \quad \forall d \in \mathcal{D}_T. \tag{1}$$

We formally characterize Equation 1 as an optimization problem in which the coefficients of A_T appear as variables:

$$\min_{A_T} \sum_{d \in \mathcal{D}_T} || \, A_T \, l^d - w^d ||. \tag{2}$$

Note that this (very small) optimization problem might be linear or non-linear, depending on the choice of norm.

Depending on the relative size of the available dataset $\{l^d \,|\, d \in \mathcal{D}_T\}$ and the number of coefficients in A_T, it may be necessary to introduce additional constraints and regularization terms to formulation (2). Another potential with issue is that the resulting A_T may be ill conditioned, causing difficulties in the calculation A_T^{-1}. Finally, nonsingularity needs to be enforced in (2), which can be easily achieved by requiring all coefficients above and below the diagonal of A_T equal zero; we use this simple approach in the experiments reported below.

In practice, a balancing region is typically divided into zones, for which loads are forecast and reported. In order to increase the data available for regression and error distribution estimation, we additionally combine data from disparate zones in a fashion analogous to that described above, for converting data associated with different day types to a reference day type.

E. Segmentation

For each date range, we partition the weather data for the composite dates into distinct segments. The idea is to

limit regression and error distribution inaccuracy by only considering data with similar response characteristics. We segment based on the forecast temperature, as inclusion of additional weather variables (e.g., via k-means clustering) failed to improve prediction accuracy in our experiments. For each date range, we form a temperature distribution over which we introduce cutting points that define the segments.

Let t_*^d be a scalar representation of the hourly forecast temperatures for day $d \in \mathcal{D}$, where $\mathcal{D}$ is the set of days in the date range. In our experiments, we define $t_*^d = t_{12}^d$, although alternative metrics such as average hourly temperature can be substituted. We require scalar representations of hourly temperature vectors to prevent data for a given day to be mapped to multiple error categories (see Section III).

We estimate the probability density function $f_{t_*}(\cdot)$ of the temperature scalar t_*^d by fitting an exponential epi-spline [11], [12]. We denote the corresponding cumulative distribution function by $\mathbb{F}_{t_*}(\cdot)$. To obtain N_S segments of equal size, we introduce the break points $\{b_1, \ldots, b_{N_S+1}\} = \{0, 1/N_S, \ldots, (N_S-1)/N_S, 1\}$ for $\mathbb{F}_{t_h}(\cdot)$ and then calculate the limit temperatures for each segment $\mathcal{S}_i$ as

$$(\underline{t}_i, \bar{t}_i) = (\mathbb{F}_{t_*}^{-1}(b_i), \mathbb{F}_{t_*}^{-1}(b_{i+1})),\ i = 1, \ldots, N_S.$$

Finally, considering the limit temperatures of each segment $\mathcal{S}_i, i = 1, \ldots, N_S$, we group the days in the date range according to the rule:

$$d \in \mathcal{S}_i \Leftrightarrow t_*^d \in [\underline{t}_i, \bar{t}_i).$$

III. Scenario Generation

To capture the notion that both the mean load response and error distribution vary during the day, we split the day into parts and then categorize each portion according to the relative error. A regression model is constructed for each of the resulting day parts and associated error categories. Scenarios are then constructed by sampling from paths constructed by selecting a specific error category for each day part. This process is a specific instantiation of the general scenario generation methodology detailed in [10].

Let H be the set of hours that define a partition of the hours in a day, specified as follows:

$$H = \{H_i\}_{i=1}^{|H|}, H_1 = 1, H_{|H|} = 24, H_i < H_{i+1}.$$

The elements H_i represent the partition end-points, e.g., the i-th part of the day is given by the set of hours $\{H_i, \cdots, H_{i+1}\}$. For each partition boundary H_i, we compute the error observed regression error ϵ_i^d for each day $d \in \mathcal{D}$ as:

$$\epsilon_i^d = l_{H_i}^d - r_{H_i}(d; \mathcal{D}).$$

We estimate the distribution of these errors by fitting an exponential epi-spline [11], [12]. This process is graphically illustrated in Figure 1.

For each partition hour i, we denote the corresponding error probability density function by $f_{\epsilon_i}(\cdot)$. Categories within $f_{\epsilon_i}(\cdot)$ can then be defined through identification of break points of the associated cumulative distribution function $F_{\epsilon_i}(\cdot)$, as was performed for temperature segmentation. Specifically, to generate N_K equally sized categories, we select the break points

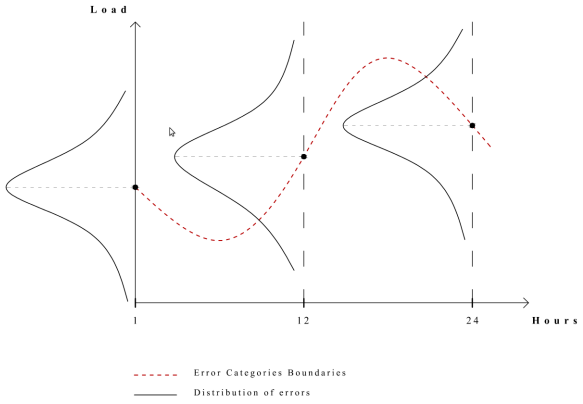

Fig. 1. Illustrative Load Regression Model and Error Distributions with Hour Partition $H = \{1, 12, 24\}$.

$\{k_1, \ldots, k_{N_K+1}\} = \{0, 1/N_K, 2/N_K, \ldots, (N_K-1)/N_K, 1\}$ in order to obtain equally weighted categories for each partition of the day. The resulting categories $\mathcal{C}_i^k$ are then defined as $\mathcal{C}_i^k = [\mathbb{F}_{\epsilon_i}^{-1}(k_i), \mathbb{F}_{\epsilon_i}^{-1}(k_{i+1}))$, where $i \in \{1, \ldots, |H|-1\}$ denotes the day partition and $k \in \{1, \ldots, N_K\}$ denotes the category.

Given an hour partition H and associated error categories $\mathcal{C}_i^k$, we sub-segment the days $d \in \mathcal{D}$ according to the observed regression model error at the corresponding partition i. Specifically, let $\mathcal{D}_i^k$ denote the set of days in the segment for hour i and error category k, defined as follows:

$$d \in \mathcal{D}_i^k \Leftrightarrow \epsilon_i^d \in \mathcal{C}_i^k.$$

For each sub-segment $\mathcal{D}_i^k$, we fit a regression model $r(d; \mathcal{D}_i^k)$, from which a vector of predicted hourly loads $\hat{r}^{d,k}$ are extracted for each day $d \in \mathcal{D}_i^k$. Each sub-segment has a different regression and hence, for all middle day part boundaries, there could be two points. We avoid discontinuities deriving a single regression curve per category, $\hat{r}^{d,k}$, by merging the regressions obtained at the limit hours. In particular, for $h \in [H_i, H_{i+1}]$,

$$\hat{r}_h^{d,k} = \left(\frac{H_{i+1}-h}{H_{i+1}-H_i}\right) \cdot \hat{r}_{h,i}^{d,k} + \left(\frac{h-H_i}{H_{i+1}-H_i}\right) \cdot \hat{r}_{h,i+1}^{d,k}.$$

This process is illustrated in Figure 2.

Given regression models for each category $\mathcal{C}_i^k$, the observed load forecast errors are computed as

$$\varepsilon_{h,d}^k = l_h^d - \hat{r}_h^{d,k}$$

. We obtain a corresponding probability density function $f_{\varepsilon_i^k}(\cdot)$ by fitting an exponential epi-spline. An illustrative example of this step is shown in Figure 3, where $N_K = 2$.

The error densities $f_{\varepsilon_i^k}(\cdot)$ serve as the primary input to the scenario generation process. The first step in scenario generation involves the identification of a set of distribution cut points $C = \{c_z\}_{z=1}^{|C|}$, subject to $c_1 = 0.0$ and $c_{|C|} = 1.0$. For each partition i and category k, we then calculate the conditional expected value of the error in each interval defined

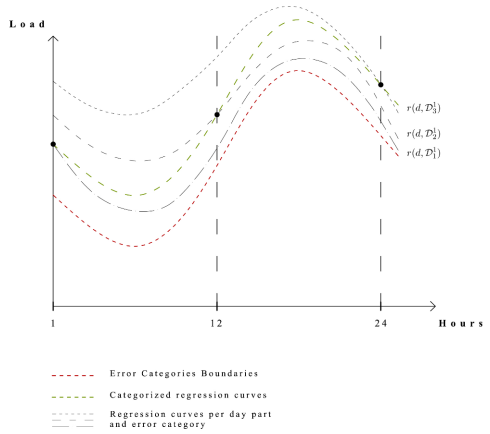

Fig. 2. Regression of the Error Category Construction.

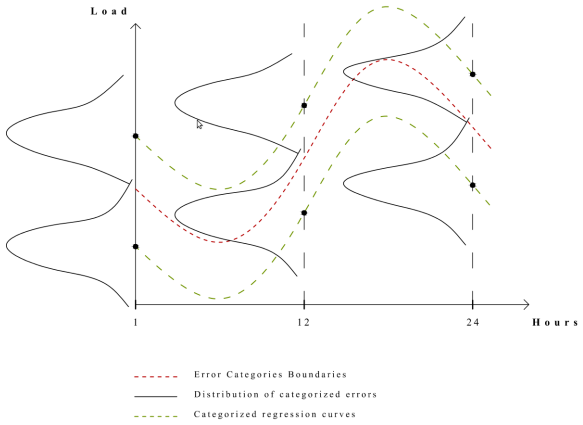

Fig. 3. Illustrative Error Category Regressions and Associated Error Distributions with Hour Partition $H = \{1, 12, 24\}$ and $N_K = 2$ Categories.

by a pair of adjacent cutting points:

$$\mathbb{E}\left[\varepsilon_i^k | \varepsilon_i^k \in [c_z, c_{z+1}]\right] = \frac{\int_{c_z}^{c_{z+1}} x \cdot f_{\varepsilon_i^k}(x) dx}{\int_{c_z}^{c_{z+1}} f_{\varepsilon_i^k}(x) dx} = \xi_i^{k,z}$$

where $\xi_i^{k,s}$ denotes the expected error in category k at cutting point z for hour i. The number of cutting points can vary per hour, as shown in Figure 4. In this illustrative example, $\mathcal{C} = \{0.0, 1.0\}$ for $i = 1$, but $\mathcal{C} = \{0.0, 0.5, 1.0\}$ for $i = 12, 24$.

Given regression models $(\hat{r}_{h,i}^{k,d})$ for each hour partition boundary i and category k, we compute loads at the partition boundaries via:

$$l_{H_i}^{d,k,z} = \hat{r}_{H_i}^{d,k} + \xi_i^{k,z}.$$

For each hour H_i and each category C_i^k, this step yields $|C|$ forecast load samples. The final step in our scenario generation process is to connect these samples in order to construct a set of paths that approximates the stochastic process representing load for the full day. This is simply done by calculating the scenario loads at time $h \in [H_i, H_{i+1})$ by assuming that the

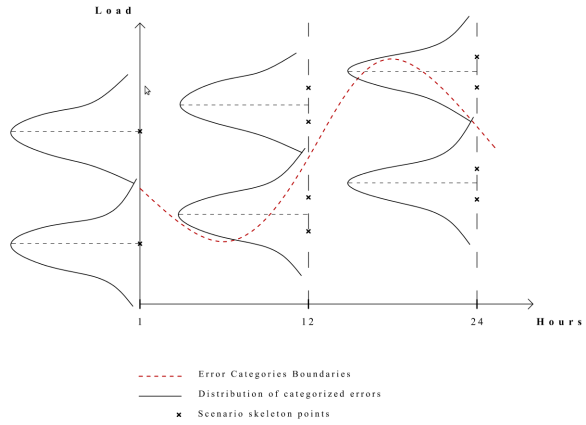

Fig. 4. Distribution Cutting Points for Scenario Generation when$H = \{1, 12, 24\}$ and $N_K = 2$. For $i = 1$, $\mathcal{C} = \{0.0, 1.0\}$. For $i \in \{12, 24\}$, $\mathcal{C} = \{0.0, 0.5, 1.0\}$.

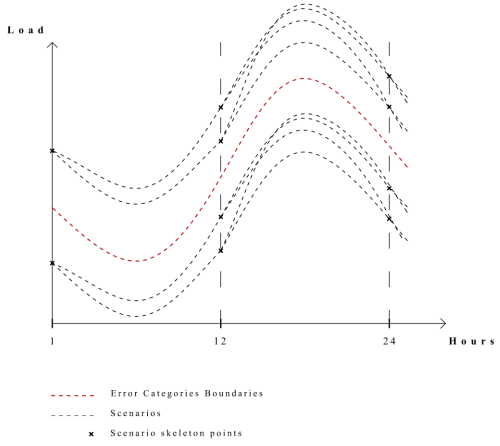

Fig. 5. Illustrative Scenario Paths Corresponding to the Cutting Points Shown in Figure 4.

deviation from the forecast varies between the deviation at hour H_i and hour H_{i+1}. This process is illustrated in Figure 5. Under this methodology, the number of paths (i.e., scenarios) generated is equal to

$$N_K \cdot (|C| - 1)^{|H|-1},$$

such that the number of scenarios is dictated by the values of the referenced parameters. Further, the generation process is deterministic, given a fixed set of historical input data.

IV. Experimental Results

Although substantially better MAPEs can be achieved using standard leave-one-out validation, we have generated our forecasts and scenarios by simulating a rolling horizon as would be seen by a real-world system operator. In our experiments, we consider data for ISO-NE. We begin by fitting the models using only data from 2009 and 2010, and consider operations during 2011. As we simulate the progression through the year, data from 2011 is added to the fit process as it becomes historical. Specifically, we execute the complete methodology

Load Zone	Weather Stations	Weights
ME	PWM	1.0
NH	CON	1.0
VT	BTV	1.0
CT	BDL, BDR	0.13, 0.87
RI	PVD	1.0
SEMASS	PVD	1.0
WCMASS	BDL, ORH	0.5, 0.5
NEMASSBOST	BOS	1.0

TABLE I
LOAD ZONES FOR ISO-NE, WITH WEATHER STATIONS AND CORRESPONDING WEIGHTS

described in Section II-B for each simulated day; the entire procedure takes minutes of wall clock time to complete, and is therefore feasible in practice. Because our interest is in demonstrating scenario generation methods, and not actually providing ISO-NE with load forecasts, we begin our simulation on January 2, 2011 and end on November 20, 2011. Excellent methods exist for dealing with the holiday season in the US [13], but their use is beyond the present scope. Further, we ignore August 28-30 of 2011, due to a hurricane event in the region. We partition the days of the year into the following date ranges: Winter (January 2 – March 31), Spring (April 1 – May 14), Summer (May 15 – September 14), and Fall (September 15 – November 20).

As reported in Table I, ISO-NE is divided into 8 load zones. Weather data for each zone is taken from one or two stations. For zones with two stations, an aggregate weather forecast is computed by weighting the composite station data appropriately. Historical load data was obtained by ISO-NE through their web site (http://www.iso-ne.com/markets/hstdata/znl_info/hourly/index.html). Hourly day-ahead temperature forecasts were provided directly by ISO-NE.

As indicated in Section II-D, we aggregate data from disparate zones to make more data available to the fit process. Zone aggregation proceeds as follows, by employing two reference zones: CT and NEMASSBOST. The RI, SEMASS, and WCMASS zones are aggregated with the CT zone, while the ME, NH, and VT zones are aggregated with the NEMASSBOST zone. This aggregation corresponds to a partition of ISO-NE in approximately northern and southern regions, which in turn share similar load characteristics. Within each partition, we select the zone with the greatest demand as the reference zone, minimizing the total load transformed.

A. Forecast MAPEs

We quantify load forecast accuracy as the Mean Average Percent Error (MAPE), denoted by MAPE(N_R, κ), as

$$\frac{\sum_{z\in\mathcal{Z}}\sum_{d\in\mathcal{D}}\sum_{h\in\mathcal{H}}\left(l_h^{z,d} - \mathbb{E}(l_h^{z,d})(N_R,\kappa)\right)/l_h^{z,d}}{|\mathcal{Z}|\cdot|\mathcal{D}|\cdot|\mathcal{H}|}\cdot 100 \quad (3)$$

where $\mathcal{Z}$, $\mathcal{D}$, and $\mathcal{H}$ respectively denote the sets of load zones, dates, and hours under consideration. N_R and κ denote regression fit parameters. The aggregated MAPEs obtained for each date range in 2011 are reported in Table II, considering $N_R = 24$, $\kappa = 500$, and a variable number of temperature segments N_S. The corresponding disaggregated (by zone) MAPEs are reported in Table III. We observe that segmentation of the data by temperature does improve load forecast accuracy, although the benefits either stagnate or decrease once $N_S \geq 7$. Further, the data in Table II exhibit load forecast accuracies that are consistent with those obtained by ISO-NE in practice, e.g,. see http://www.iso-ne.com/support/training/courses/wem101/10_forecast_scheduling_callan.pdf. However, our method additionally provides estimates of forecast variability, represented through a collection of load scenarios.

	Segments (N_S)			
Season	1	3	5	7
Fall	5.45	4.66	4.2	3.99
Spring	3.1	2.88	2.67	2.73
Summer	10.25	4.82	4.14	4.19
Winter	5.25	3.32	3.29	3.47

TABLE II
AGGREGATED ISO-NE MAPES FOR 2011

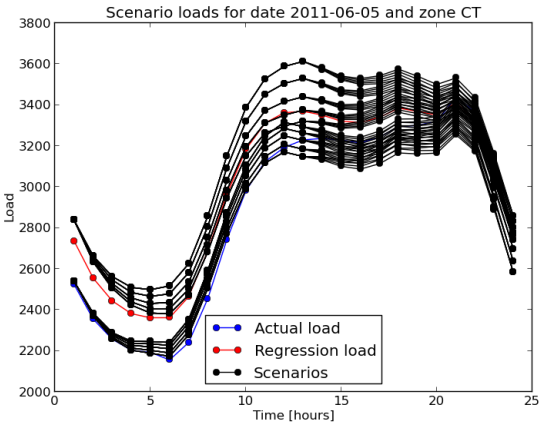

Fig. 6. Scenarios for Low Variance Load Forecast Day in 2011 for Zone CT

We now briefly consider inclusion of dew point in addition to temperature in the segmentation process, which intuitively can influence load by impacting cooling requirements. In this case, we segment by forming scalar weather quantity based on a linear combination of temperature and dew point, again using observations at $h = 12$. The resulting aggregate MAPEs are shown in Table IV, and indicate that inclusion of specific quantities can marginally improve load forecast accuracy in specific contexts.

Finally, we provide exemplars of forecast load scenarios for the CT zone in ISO-NE, for early summer days in 2011. We have selected low and high variance load 50-scenario cases, respectively shown in Figures reffig:lowvariancescenarios and 7. These two particular cases serve as test cases for assessing our stochastic unit commitment solver, described in the companion paper. All scenario sets generated for ISO-NE, either at the zonal or aggregate level, are available by contacting the authors.

		Segments (N_S)			
Season	Zone	1	3	5	7
Fall	NH	4.44	4.22	4.41	4.14
	VT	3.52	2.91	2.95	3.04
	ME	4.13	4.14	4.02	4.15
	CT	8.31	7.12	6.77	6.51
	RI	5.93	5.3	4.49	4.0
	SEMASS	5.75	4.99	4.4	3.84
	WCMASS	7.2	6.6	6.34	6.66
	NEMASSBOST	4.82	4.63	4.19	4.07
Spring	NH	3.14	3.46	3.69	3.93
	VT	3.23	3.18	3.04	3.22
	ME	4.25	3.9	4.07	4.0
	CT	3.46	3.64	3.42	3.89
	RI	3.24	2.97	3.27	3.32
	SEMASS	3.05	2.93	3.17	3.24
	WCMASS	3.74	3.5	3.8	3.98
	NEMASSBOST	3.24	3.41	3.35	3.48
Summer	NH	9.29	5.65	4.95	5.05
	VT	5.64	3.75	3.41	3.4
	ME	7.54	4.65	4.83	4.54
	CT	11.22	6.42	5.98	5.84
	RI	12.9	7.15	5.69	5.86
	SEMASS	12.72	6.76	5.6	5.74
	WCMASS	9.47	5.51	4.57	4.63
	NEMASSBOST	11.34	6.17	5.78	5.32
Winter	NH	4.99	3.88	3.78	3.82
	VT	4.28	3.77	4.02	4.17
	ME	4.1	3.87	3.65	4.05
	CT	6.17	4.35	4.21	4.32
	RI	5.23	3.81	3.56	3.84
	SEMASS	5.34	4.05	3.76	4.11
	WCMASS	5.46	3.69	3.84	3.95
	NEMASSBOST	5.42	3.54	3.62	3.67

TABLE III
ZONAL ISO-NE MAPES FOR 2011

	Segments (N_S)			
Season	1	3	5	7
Summer	10.55	4.73	4.06	4.06

TABLE IV
AGGREGATED ISO-NE MAPE FOR SUMMER 2011, USING SEGMENTATION BASED ON TEMPERATURE AND DEW POINT

B. Parameter Sensitivity for Model Fitting via Epi-Splines

To construct our regression models via dpi-splines, it is necessary to specify values for the following key parameters:

- N_R: The number of intervals into which the hours of a day are sub-divided.
- κ: The maximum curvature of the regression curve.

While it is necessary to specify a specific norm (e.g., L^1 or L^2) when solving the embedded optimization problems, in practice the choice of norm has almost no impact on the resulting load forecast MAPEs. In contrast, accuracy is more sensitive to the choice of N_R and κ, as we now demonstrate.

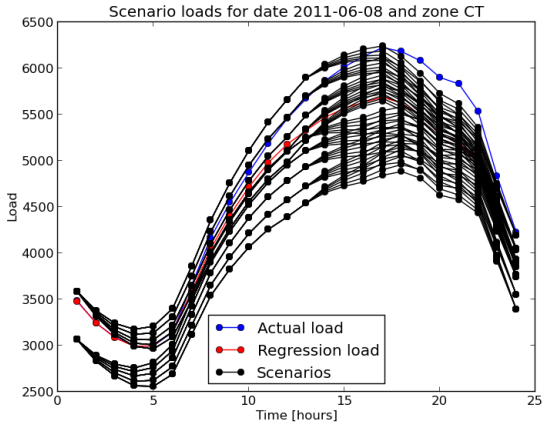

Fig. 7. Scenarios for High Variance Load Forecast Day in 2011 for Zone CT

First, we consider the impact of N_R on load forecast accuracy. We compute the MAPEs for $N_R \in \{6, 12, 18, 24, 32, 48\}$, fixing $N_S = 5$ and $\kappa = 500$. The results that the MAPE changes from about 4.7% to 4.5% over the range, with values of 24 and above yielding nearly equal MAPEs.

Next, we consider the impact of κ on load forecast accuracy. Fixing $N_R = 24$ and $N_S = 5$, we compute MAPEs for $\kappa \in \{20, 40, 60, 80, 100, 150, 200, 500\}$. Extremely small values of κ do adversely impact the MAPEs by severely restricting the curvature; however, over the range of κ we tested, there is essentially no sensitivity.

V. CONCLUSION

In this paper, we consider novel methods for obtaining distributions of forecast load in each hour of day d, based on weather forecasts available on day $d-1$. Our goal was to use the estimated trends and error distributions to generate probabilistic scenarios for the day-ahead load, for use as input for assessing stochastic unit commitment solution procedures. Our experiments indicate that our models (1) can be generated using data readily available by system operators, (2) can be produced in minutes of run time for multiple years of input data, (3) produce errors that are competitive in the aggregate with load forecasting procedures used by industry, and (4) obtain MAPE values for forecast load that are similar to those found in hind-casting studies that eliminate weather forecast uncertainty by focusing on strictly on actual or observed weather quantities. We continue to refine our methods, specifically by focusing on reducing prediction errors associated with peak load periods and adapting the approach to times of the year where temperature is not as strong a predictor of the load.

ACKNOWLEDGMENTS

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94-AL85000. This work was

funded by the Department of Energy's Advanced Research Projects Agency - Energy, under the Green Energy Network Integration (GENI) project portfolio. The authors would like to thank Dr. Eugene Litvinov and his research group at ISO-NE (in particular, Bill Callan) for their assistance with ISO-NE system and data sources.

An earlier version of some of the work presented here appeared in [14].

REFERENCES

[1] S. Takriti, J. Birge, and E. Long, "A stochastic model for the unit commitment problem," *IEEE Transactions on Power Systems*, vol. 11, no. 3, pp. 1497–1508, 1996.

[2] A. Papavasiliou and S. Oren, "A stochastic unit commitment model for integrating renewable supply and demand response," in *Proceedings of the 2012 IEEE Power and Energy Society Meeting*, 2012.

[3] P. Ruiz, R. Philbrick, E. Zack, K. Cheung, and P. Sauer, "Uncertainty management in the unit commitment problem," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 642–651, 2009.

[4] K. Cheung, D. Gade, C. S. Monroy, S. M. Ryan, J.-P. Watson, R. J.-B. Wets, and D. L. Woodruff, "Scalable stochastic unit commitment, part 2: Solver performance," *IEEE Transactions on Power Systems*, Under Review.

[5] T. Hong, M. Gui, M. Baran, and H. Willis, "Modeling and forecasting hourly electric load by multiple linear regression with interactions," in *Proceedings of the 2010 IEEE Power and Energy Society Meeting*, 2010.

[6] E. Feinberg and D. Genethliou, "Load forecasting," in *Applied Mathematics for Restructured Electric Power Systems*, 2005, pp. 269–285.

[7] J. Liu, R. Chen, L. Liu, and J. Harris, "A semi-parametric time series approach in modeling hourly electricity loads," *Journal of Forecasting*, vol. 25, pp. 537–559, 2006.

[8] J. Black, "Load hindcasting: A retrospective regional load prediction method using reanalysis weather data," Ph.D. dissertation, Department of Mechanical and Industrial Engineering, University of Massachusetts Amherst, 2011.

[9] R. J.-B. Wets and S. W. Bianchi, "Term and volatility structures," *Handbook of Asset and Liability Management: Theory and Methodology*, vol. 1, p. 25, 2006.

[10] R. J.-B. Wets, I. Rios, and D. Woodruff, "Multi-period forecasting and scenario generation with limited data," University of California Davis, Davis CA 95616, Tech. Rep., 2013.

[11] J. Royset and R. Wets, "Epi-splines and exponential epi-splines: Pliable approximation tools," University of California Davis, Tech. Rep., 2012.

[12] ——, "Fusion of hard and soft information in non-parametric density estimation," University of California Davis, Tech. Rep., 2012.

[13] T. Hong, "Electric load forecasting with holiday effect," https://sites.google.com/site/hongtao01/courses/holiday, 2012, presented at AEIC Load Research Workshop 2012, Orlando, FL.

[14] Y. Feng, D. Gade, S. M. Ryan, J.-P. Watson, R. J.-B. Wets, and D. L. Woodruff, "A new approximation method for generating day-ahead load scenarios," in *Proceedings of the 2013 IEEE Power and Energy Society General Meeting*, 2013.

Yonghan Feng received his B.S. and M.S. degree from Xi'an Jiaotong University, Xi'an, China. He is a Ph.D candidate in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University. His research applies optimization to power systems.

Ignacio Rios received his B.E. degree in Industrial Engineering from the University of Chile, Santiago, in 2011. He is currently working toward his M.S. degree in Operations Research at the University of Chile, and is working as a research assistant at the University of California, Davis, under the supervision of Professors Roger J-B Wets and David L. Woodruff.

Sarah M. Ryan (M09) received her Ph.D. degree from the University of Michigan, Ann Arbor. She is currently Professor in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University. Her research applies stochastic modeling and optimization to the planning and operation of energy and manufacturing systems.

Kai Spurkel received his B.Sc. from the University Duisburg-Essen, Germany, in 2012. His research interests include combinatorial and stochastic optimization.

Jean-Paul Watson (M'10) received his Ph.D. in Computer Science from Colorado State University. He is currently a Principal Member of Technical Staff in the Discrete Math and Complex Systems Department at Sandia National Laboratories, in Albuquerque, New Mexico. He leads projects involving optimization under uncertainty and general analytics for US government agencies, including the Department of Energy and Defense.

Roger J.-B. Wets is a Distinguished Research Professor of Mathematics at the University of California, Davis. His research interests include stochastic optimization, variational analysis, equilibrium problems in stochastic environments, and the fusion of hard and soft information in statistical estimation. His awards include Guggenheim and Erskine Fellowships, the SIAM-MPS Dantzig Prize, and the INFORMS Lanchester prize.

David L. Woodruff is Professor in the Graduate School of Management at the University of California at Davis. He received his Ph.D. in Industrial Engineering from Northwestern University. His research concerns computational aspects of multi-stage optimization under uncertainty. He has worked on applications in a variety of areas and has been involved recently in a number of applications in power systems.