Analysis and Synthesis of Load Forecasting Data for Renewable Integration Studies

Preprint

N. Steckler, A. Florita, J. Zhang, and
B.-M. Hodge
National Renewable Energy Laboratory

Presented at the 12th International Workshop on Large-Scale Integration of Wind Power Into Power Systems
London, England
October 22–24, 2013

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Conference Paper
NREL/CP-5D00-60270
November 2013

Contract No. DE-AC36-08GO28308

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at http://www.osti.gov/bridge

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone: 865.576.8401
fax: 865.576.5728
email: mailto:reports@adonis.osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
phone: 800.553.6847
fax: 703.605.6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/help/ordermethods.aspx

Cover Photos: (left to right) photo by Pat Corkery, NREL 16416, photo from SunEdison, NREL 17423, photo by Pat Corkery, NREL 16560, photo by Dennis Schroeder, NREL 17613, photo by Dean Armstrong, NREL 17436, photo by Pat Corkery, NREL 17721.

Printed on paper containing at least 50% wastepaper, including 10% post consumer waste.

Analysis and Synthesis of Load Forecasting Data for Renewable Integration Studies

Nicholas Steckler, Anthony Florita, Jie Zhang, Bri-Mathias Hodge
Transmission and Grid Integration Group
National Renewable Energy Laboratory
Golden, Colorado, USA

Abstract—**As renewable energy constitutes a greater portion of the generation fleet, so does the importance of modeling uncertainty as part of integration studies. In pursuit of optimal system operations, it is important to capture not only the definitive behavior of power plants but also the risks associated with systemwide interactions. Load forecasting is an area of renewable energy integration studies that is often neglected, chiefly because of a lack of available data. In this research, the dependence of load forecast errors on external predictor variables such as temperature, day type, and time of day was examined. The analysis was utilized to create statistically relevant instances of sequential load forecasts with only a time series of historic, measured load available. The creation of such load forecasts relies on Bayesian techniques for informing and updating the model, thus providing a basis for networked and adaptive load forecast models in future operational applications.**

Keywords-load forecasting; power demand; renewable integration; Bayesian probability

I. INTRODUCTION

Increasing amounts of variable and uncertain renewable generation are currently being introduced into the electric grid, resulting in more uncertainty in system operations with larger penetrations. Renewable integration studies examine the availability of electric power from scenario-based perspectives so that supply is best utilized to meet demand. Ensuring these studies rely on realistic and statistically relevant data and assumptions is of upmost concern and fundamental to the results gleaned from outcomes. As such, uncertainty must be incorporated into modeling tasks and is the first step toward dynamic models for informing system operations at high levels of renewable penetration.

Load forecasting plays an important role in integration studies [1, 2]. It consists of making predictions about future demand for electricity to optimize generation schedules, and with the acquisition of additional data these beliefs can be updated. Ongoing studies related to the Western Wind and Solar Integration Study [2], which simulated various scenarios of the unit commitment and dispatch problem, utilize load forecasts to better understand sensitivities and interactions between load and renewable generation. The commitment (scheduling) process involves determining which generating units will be turned on during future time periods and adjusting their output levels closer to the operating hour during dispatch. Because many thermal units require long start-up and shutdown times, planning is often performed on a day-ahead basis, whereas fluctuations in solar photovoltaic power plants can require planning updates on the order of seconds or less. With forecasting performed on a range of different timescales, providing different levels of insight about future load through uncertainty modeling can enhance existing operations.

The most commonly used forecasting methods include Similar Day, Time Series, Regression, Fuzzy Logic methods, Expert Systems, and Support Vector Machines [3]. Bayesian probability as applied to power systems has been researched to a lesser extent, with the most frequent approaches dominating the literature. Douglas et al. [4] proposed using Bayesian estimation in conjunction with a dynamic linear model to predict peak forecasts by using average temperature as a predictor variable. As expected, because of air-conditioning and gas-combustion heating trade-offs, it was shown that the load was most sensitive to temperature during the summer and least sensitive to it during the winter.

As a starting point for detailed uncertainty modeling, a question of interest in this research is whether external variables, such as temperature, influence the accuracy of load forecasting. Therefore, a characterization of error between day-ahead forecast and actual load data from the New York Independent Service Operator (NYISO) was established—specifically, its dependence on external predictor variables such as temperature, day type, and time of day. Examining such load forecast error distributions aids the understanding of overall performance obtained from state-of-the-art load forecast systems without the need to consider their mathematical details. Resulting observations can be utilized to improve existing forecast systems or train new approaches. To the best of the authors' knowledge, the proposed approach of directly characterizing load forecast errors is unique to the literature.

The necessary framework to characterize forecast errors and provide model insight for creating load forecast errors is provided by Bayesian probability, thereby allowing historically realistic load forecasts to be produced. Bayesian inference estimates the "after data" posterior probability distribution based on the "before data" prior probability distribution combined with likelihoods associated with the selected predictor variables [5]. Beliefs can be updated as new observations are made and the process recursively moves through time. By building a bottom-up model and conditionally assessing situations or hypotheses of interest, the accuracy of load forecasts can be significantly improved. This paper describes the methods and data used in the analysis, presents the results of analyzing the distributions of forecast errors, and suggests a modeling technique for creating historically accurate load forecasts. Finally, conclusions are drawn and implications for future studies and for power systems operations are outlined.

II. Data and Methods

A. Data Utilized

The data utilized was in the form of load forecasts, load actuals, and daily maximum temperatures for the NYISO operating region for the years 2009 through 2011. The data was obtained from the NYISO Market & Operations Data website [6]. The actual load data was provided on five-minute intervals, and the forecast load data was on hourly intervals. The five-minute data was averaged over each hour to create hourly data that aligned with the forecasts. All forecasts were performed by the operator on a day-ahead timescale. The load forecast error was equal to the hourly forecast minus the actual load data. For NYISO, the data was provided for all 11 operating regions individually. The regional loads were aggregated so that the electric demand of the entire NYISO could be analyzed. For the NYISO temperature data, the historical percentage of total load contributed by each region was taken as a weight value and was used to create a temperature data set representative of the entire NYISO operating region.

B. Methodology

1. Characterization of Forecast Errors

One of the objectives of this research was to evaluate how variables such as the type of day, time of day, and outside temperature correspond to the scale and frequency of load forecast errors in power systems. As shown in Fig. 1, three variables were selected as predictors for load forecast errors: business or nonbusiness day, time of day, and daily maximum temperature, which were represented by X_1, X_2, and X_3, respectively. All random continuous variables were discretized through binning in the development of Bayesian uncertainty modeling. The variables were classified into bins corresponding to ranges of values that the variables may assume during any hour. Forecast errors as a percentage of actual loads were classified into 0.05% increment bins across the entire range for any given year. The day type class referred to any hour falling on a weekday or non-holiday as a business day. A nonbusiness day is one that falls on a weekend or holiday. The time bins shown in Fig. 2 were chosen in blocks that are commonly used by independent system operators [7]. Last, the daily max temperature variable was binned such that each bin represented a different section of the total range of the temperature variable. The number of bins for the temperature was chosen to be five to ensure that an adequate and practical number of data points were classified into each bin. The bounds for each bin were optimized so that significance between error distributions for differing bins was maximized, as described in Section III-A.

A number of statistical tools were used for determining when forecast error distributions differed. All characterizations were done using the R statistical computing environment [8]. To compare distributions, the Kolmogorov-Smirnov (KS) test was used. For comparing against normal and hyperbolic distributions, the Shapiro-Wilk and *HyperCvMTest* tests were used, respectively [9]. Metrics such as mean absolute error were used for analyzing the differences between distributions of load forecast errors. Multiple visual aids were also utilized, such as histograms, kernel density plots, and time-series plots. The autocorrelation of the forecast errors was analyzed by using the *acf* function in R [8].

With the methods described above, the predictive ability of the three variables can be inferred. The way in which the forecast errors are distributed with varying combinations of predictor bins reveals correlation between the predictors and the predicted random variable.

2. Development of a Bayesian Probabilitic Model

The characterization results lay the foundation for the construction of a Bayesian probabilistic model. The relationships among the model's variables can be formally expressed through Bayes's Theorem of Eq. (1). This representation is a joint probability distribution of a set of n variables, $\{X_1, \dots, X_n\}$, as a directed acyclic graph and a set of conditional probability distributions (CPDs). Each node corresponds to a variable with an associated CPD that gives the probability of each state of the variable given every possible combination of states of its parents [10].

A special case of the Bayesian network is one in which each X_i has C as the sole parent and C has no parents. This special case is called a naïve Bayesian model, whose network can be analyzed using a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions [10]. For a known set of n predictors, the most probable value of the parent variable C is found to be the value that maximizes the conditional probability shown in Eqs. (1)-(4) [11, 12]. Similarly, a posterior probability distribution can be calculated by finding the probability of all possible values of C, in which the most likely value in that distribution is found by maximizing Eq. (4). The distribution is found by evaluating the equation for all values of C, instead of maximizing. The function *naiveBayes* from the R package "e1071" was used to create the necessary naïve Bayes object [13].

$$\mathrm{P(c|x)} = \frac{\mathrm{P(x|c) * P(c)}}{\mathrm{P(x)}} \tag{1}$$

$$\mathrm{c_{NB} = argmax\,P(x_1, x_2, x_3|c)P(c)} \tag{2}$$

$$\mathrm{P(x_1, x_2, x_3|\,c) = P(x_1|c) * P(x_2|c) * P(x_3|c)} \tag{3}$$

$$\mathrm{c_{NB} = argmax\,P(c) * P(x_1|y) * P(x_2|c) * P(x_3|c)} \tag{4}$$

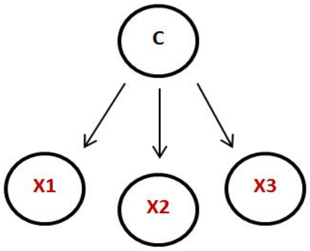

Figure 1. A naïve Bayes model.

Another important characteristic of forecast errors is their autocorrelation. The naïve Bayesian model produces a posterior distribution that produces samples with no sequential dependence. Although it includes a time-of-day predictor, the model returns a distribution of error values and expresses no temporal relationships between error values from different hours. As shown in Fig. 3, the actual forecast errors contained a large level of autocorrelation. To build this correlation into the synthesis of load forecast errors, the posterior distribution was modified based on the synthesized values from past hours.

The final probability distribution used for random sampling was calculated as shown below in Eq. (5). It is a multiplicative combination of the posterior distribution from the naïve Bayes model and an inverse squared function of *d*. The parameter *d* is the distance between every possible class value in the sorted posterior and the synthesized class from a specified time lag of *i*. A vector results from the inverse function, and by multiplying it into the original posterior, probability is increased for the class value of the specified lag. In this research, the inverse squared modifier was included only for *i*=1, that is the class from the past hour. More lag values can also be introduced, but that would make the fitting procedure more time consuming because every lag introduced would produce its own inverse squared modifier, including a new set α and β to direct the influence of *d.* An optimal lag value could be determined for more accurate representation of the autocorrelation trends found in the actual errors. The values α and β were determined by a fitting procedure that is based on mean absolute error and the KS test.

$$\mathrm{P}(\mathrm{c}|\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3,\mathrm{c}_{\mathrm{t-i}}) = \mathrm{P}(\mathrm{c}|\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3) * \frac{1}{(1+\beta * \mathrm{abs(d)})^{\alpha}} \tag{5}$$

$$\mathrm{d} = \mathrm{index(c)} - \mathrm{index}(\mathrm{c}_{\mathrm{t-i}}) \tag{6}$$

In the work by Lowd and Domingos [10], the predictive capability of naïve Bayes models was compared with that of Bayesian networks. Experiments on a large number of data sets showed that the two models take similar time to learn and are similarly accurate, but naive Bayes inference is orders of magnitude faster. Additionally, the conditional probabilities within the naïve Bayes model are easily updated as new observations are made.

III. Results

A. Characterization of Forecast Error Distributions

Characterization of the load forecast error distributions first required the determination of external variables correlated with the errors. From this analysis, three external variables were identified as predictors for hourly load forecast error: business day or nonbusiness day, time of day, and daily maximum temperature. The selected bin sizes are shown in Table I. The bins were chosen as those that minimized the KS comparison between errors in the specified bin range and the rest of the data set. A comparison between resulting binned forecast error probability distributions is shown in Figs. 2-4 in the form of kernel density estimates. The load forecast error was equal to the forecast (L_f) minus the actual (L_a) load. The normalized load forecast error e_L is defined as

$$e_L = (L_f - L_a) \big/ L_a \tag{7}$$

TABLE I. Selected Bins for Predictor Variables

(X1) Time Bin	(X2) Day Type Bin	(X3) Temperature Bin
(1) 03:00–06:00	(0) FALSE	(1) 14ºF to 41ºF
(2) 07:00–10:00	(1) TRUE	(2) 41ºF to 53ºF
(3) 11:00–14:00		(3) 53ºF to 64ºF
(4) 15:00–18:00		(4) 64ºF to 81ºF
(5) 19:00–22:00		(5) 81ºF to 96ºF
(6) 23:00–02:00		

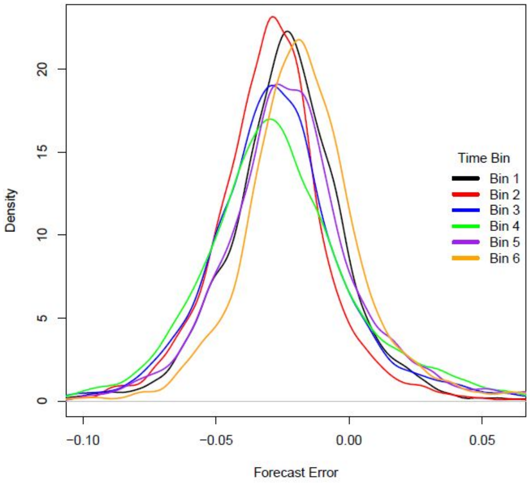

Figure 2. Forecast error distributions by time bin.

Temperature bins were especially correlated with differing error distributions, as shown in Fig. 4.

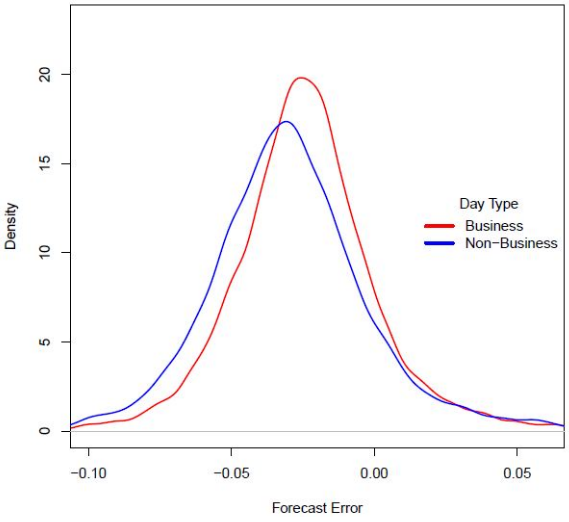

Figure 3. Forecast error distributions by day type bin.

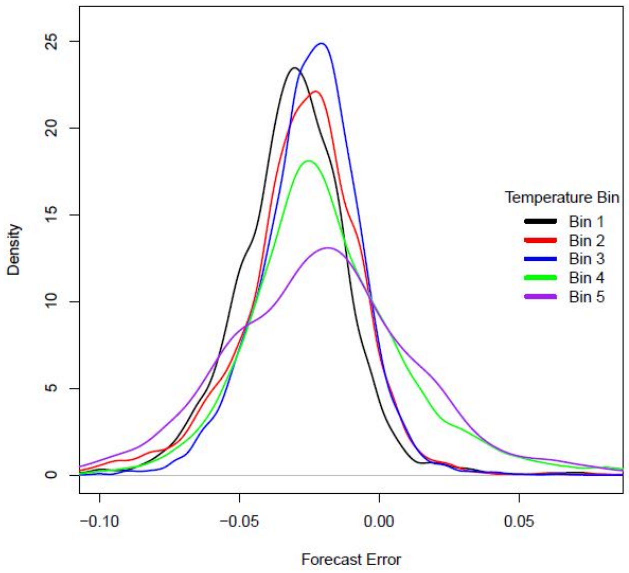

Figure 4. Forecast error distributions by temperature bin.

With six time bins, five temperature bins, and two day type bins, a total of 60 data subsets were generated. Some notable trends were observed from the distributions of forecast errors in these subsets:

- All subsets produced an underforecast mean (negative forecast error), indicating that the current load forecast methodology adopted at NYISO tended to underforecast.
- Almost all 60 data subsets presented a hyperbolic distribution, and 40 out of 60 satisfied (P-value > 0.025) a 0.975 confidence interval.
- The temperature bins led to the greatest variance between the error distributions.

The autocorrelation coefficients for the three-year data set of forecast errors are displayed in Fig. 5 and Table II.

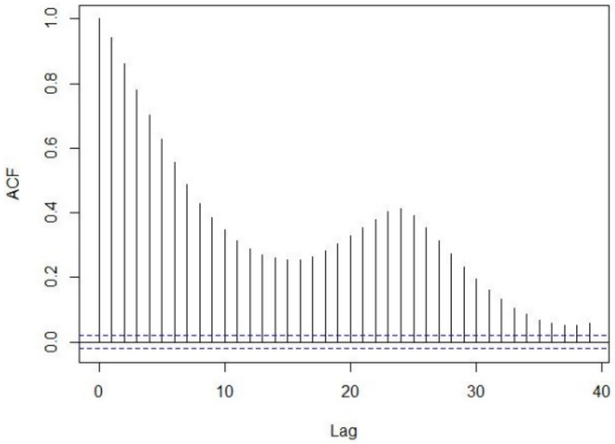

Figure 5. Autocorrelation plot for actual errors from 2011.

TABLE II. CORRELATION COEFFICIENTS

Lag (hours)	Coefficient
1	0.941
2	0.861
3	0.779
4	0.702
5	0.629
6	0.557
12	0.289
24	0.413
36	0.058

B. *Synthesis of Historic Load Forecasts*

The model described in Section II-B-2 was trained on the year 2009 and optimized on 2010. The coefficients α and β, which increase the likelihood of the occurrence of the error value from the past hour, were selected through a fitting procedure. This fitting process minimized the mean absolute error and the D-value from the function *ks.test*, which performed a KS test between actual and synthesized data for the year 2010. The fitting procedure determined values of α and β to be 1.7 and 1.25, respectively.

Once fully defined, the model was trained on the years 2009 and 2010 and was then used to synthesize load forecast errors for the 2011 data set. The synthesis was performed for 100 trials so as to yield a better representation of the variation in synthesized values. For every trial, an array of metrics (e.g., mean, median, standard deviation, autocorrelation coefficients with different hour lags, kurtosis, and D-value from KS test) were calculated; the average values of those metrics are displayed in Table III. The standard deviations associated with each metric are listed in Table IV to show the variation in these metrics throughout the 100 trials.

Both Table III and Table IV include a row labeled "Original posterior," which contains the metrics for another 100 trials in which forecast errors were synthesized using the same predictors but without including the inverse squared function. This was done to investigate the effects of the inverse squared function on the distribution and time series of forecast errors. In the case of synthesis using only the original posterior distribution, the lower D-value and lower autocorrelation coefficients represented a better synthesized forecast error distribution but a decrease in accuracy of the point forecasts (time series).

The second rows of Table III and Table IV describe the metrics for the case in which the inverse squared function was applied during synthesis. As a result, the distribution became less similar to that of the actual errors, although the autocorrelation values increased toward those of the actual errors. This was observed in the increase in D-value and increase in autocorrelation coefficients. It should also be noted that the standard deviation and the kurtosis of the synthesized errors decreased when the inverse squared function was applied. The phenomena described above is shown in Figs. 6 and 7, in which histograms of the actual and synthesized errors are displayed for a single representative trial. The histogram for synthesized errors visually reflects the metrics listed above, because the distribution had a narrower spread and was more concentrated around the mean. The tails of Fig. 6 also extended much farther than those of Fig. 7, which, along with the lower average kurtosis value, shows that the synthesized distribution was more platykurtic than the actual error distribution.

TABLE III. AVERAGE VALUES OF METRICS THROUGHOUT 100 TRIALS FOR METRICS DESCRIBING TWO SYNTHESIZED DATA SETS

	Mean	Median	Std	AC Lag 1	AC Lag 2	AC Lag 24	Kurtosis	D-value (*ks.test*)[a]
Original posterior	-0.0296	-0.03	0.022	0.047	0.044	0.043	2.110	0.071
Inverse squared	-0.0295	-0.03	0.015	0.742	0.558	0.052	1.783	0.136
Actual 2011 Error distribution	-0.0262	-0.03	0.028	0.941	0.861	0.413	2.760	

[a] D-value compares synthesized set to actual errors

TABLE IV. STANDARD DEVIATIONS OF METRICS THROUGHOUT 100 TRIALS FOR METRICS DESCRIBING TWO SYNTHESIZED DATA SETS AND COMPARISONS WITH THE ACTUAL DATA

	Mean	Median	Std	AC Lag 1	AC Lag 2	AC Lag 24	Kurtosis	D-value (*ks.test*)
Original posterior	0.000215	0.00	0.00017	0.0129	0.0124	0.0104	0.305	0.005
Inverse squared	0.000462	0.00	0.00050	0.0104	0.0155	0.0215	0.747	0.010

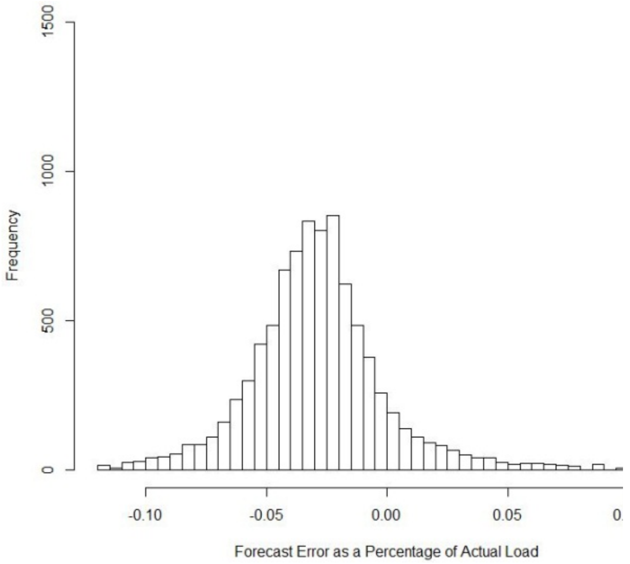

Figure 6. Actual forecast errors, 2011.

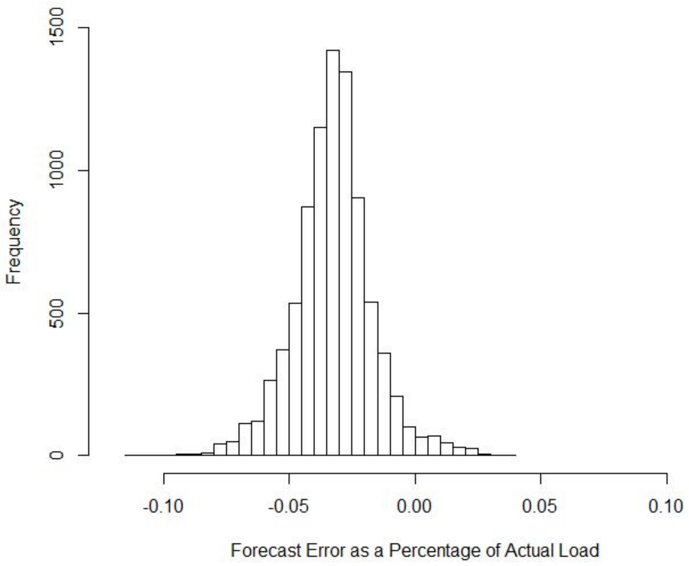

Figure 7. Single trial 2011, synthesized forecast errors.

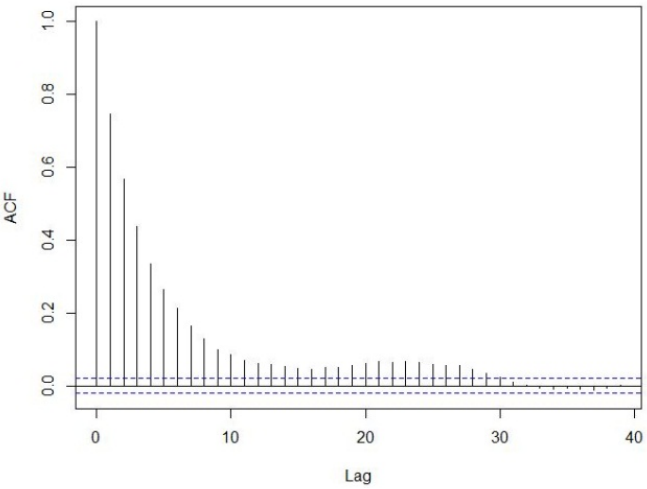

Figure 8. Single trial 2011, autocorrelation of synthesized forecast errors.

Experimenting with the modified naïve Bayesian model's performance in predicting load forecast errors has produced certain results. First, temporal applications such as load forecasting require modification of the model because it can provide only a distribution associated with the random variable. This means that sampling directly from the naïve Bayes posterior distribution can yield a very accurate distribution, but rapidly oscillating point forecasts that may not contain the temporal trends found in the target time series. Second, point forecasts can be improved through modification of the original posterior in the form of a multiplicative inverse squared function. In this research, the inverse squared method brought the synthesized values closer to the autocorrelative and temporal traits of the actual forecast errors.

The nature of the inverse squared modifier allows for change in coefficients involved along with the way that the modifier is multiplied into the original posterior. In combination with the result that the original posterior was able to successfully predict the overall distribution of errors, it is quite promising that this model may achieve greater predictive capability through the introduction of more layers of modifiers such as the inverse squared modifier. The performance of the model might be improved by optimally selecting the time lags and predictor variables, which will be investigated in future work.

IV. Conclusions

In this research, load forecast errors from the New York Independent Service Operator were characterized by investigating how the error distributions vary with exogenous predictor variables such as day type, time of day, and temperature. This analysis determined that these predictor variables are able to partition the forecast errors into conditional cases with significantly different distributions, which are very likely to follow the hyperbolic distribution.

Furthermore, the characterization of error distributions and identification of predictor variables was a means for developing a naïve Bayesian model for prediction of load forecast errors. The accuracy of this model was highly dependent on the method chosen to introduce autocorrelation into the synthesized data. By using the inverse squared multiplier, the synthesized data set approached autocorrelation values similar to those seen in the actual errors, although there was sensitivity in how much autocorrelation could be obtained without changing the original posterior distribution drastically.

Future development will involve increasing the complexity of the Bayesian network and tracking increases in accuracy. New predictors may be characterized and introduced alongside the ones used in this study. More levels of dependence may be introduced into the method for building autocorrelation into the data set. Additionally, the authors plan to return to the original posterior distribution and study its predictive capability in finer, conditional detail, as was done in the characterization of actual load forecast errors.

Acknowledgment

This research was supported by the U.S. Department of Energy under Contract No. DE-AC36-08-GO28308 with the National Renewable Energy Laboratory and by the Department of Energy Office of Science.

References

[1] EnerNex, Eastern Wind Integration and Transmission Study. Golden, CO: National Renewable Energy Laboratory, 2010.

[2] GE, Western Wind and Solar Integration Study. NREL/SR-550-47434. Golden, CO: National Renewable Energy Laboratory, 2010.

[3] B. U. Islam, "Comparison of conventional and modern load forecasting techniques based on artificial intelligence and expert systems," IJCSI Int. J. of Comput. Sci. Issues, vol. 8, 2011.

[4] A. Douglas, A. Breipohl, F. Lee, and R. Adapa, "The impacts of temperature forecast uncertainty on bayesian load forecasting," IEEE Trans. Power Syst., vol. 13, November 1998.

[5] J. K. Kruschke, Doing Bayesian Data Analysis A Tutorial with R and BUGS. Oxford: Elsevier, 2011.

[6] NYISO. Market & Operation Data, 2011. Available: www.nyiso.com/public/markets_operations/market_data/power_grid_data/index.jsp

[7] ERCOT, "Ercot methodologies for determining ancillary service requirements," ed., 2010, p. 9.

[8] R: A Language and Environment for Statistical Computing, ed. Vienna, Austria: R Foundation for Statistical Computing, 2010.

[9] D. Scott, HyperbolicDist: The Hyperbolic Distribution, ed., 2009.

[10] D. Lowd and P. Domingos, "Naive bayes models for probability estimation," presented at the International Conference on Machine Learning, 2005.

[11] D. Jurafsky, "Formalizing the naive bayes classifier," ed. Open Course Online, 2012.

[12] K. M. Leung. Naive Bayesian Classifier, 2007, 7/22. Available: http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf

[13] D. Meyer, "e1071," 1.6-1 ed., 2012.