A Naïve Multiple Linear Regression Benchmark for Short Term Load Forecasting

Tao Hong, Pu Wang, and H. Lee Willis, *Fellow, IEEE*

***Abstract*—Benchmarking issue in short term load forecasting has not received as much attention as it deserves. Although dozens of techniques have been reported to be applied to short term load forecasting, most of them are still on the theoretical level with insignificant practical value. None of them has been established to produce benchmarking models for comparative assessment. This paper proposes a naïve multiple linear regression benchmark for short term load forecasting, which is from the experience of helping a US utility develop the first in-house short term load forecasts. The proposed model has been served as a benchmark for this utility since 2009, and was in production use for a year with satisfying performance before a major upgrade. It has also been used for a Canadian utility for load forecasting purposes. In addition, it was reproduced by a group of graduate students from a creditable US university following the documented procedure.**

***Index Terms*—load forecasting, power systems planning.**

I. Introduction

Forecasts are always wrong, but some are useful. To an electric utility, the useful forecasts should at least serve its business needs, such as operations, planning, and energy purchasing, etc. Furthermore, the forecasts are supposed to be accurate, interpretable, and defensible. There is a common misunderstanding in the field of short term load forecasting: overwhelmingly emphasizing on accuracy. In fact, as long as the utilities can be operated with competency in the market, the accuracy of load forecasts is not the most important factor to utilities. In other words, minor improvement of forecasting accuracy beyond certain level may not result in significant economic benefits [10]. Different utilities, depending upon the size, location and electricity consumption behaviors of its customers, may have various comfortable zones of forecasting accuracy, which may range from 2% to 6% with respect to the mean absolute percentage error (MAPE) of one day ahead hourly load forecast.

There are roughly two approaches to improve the accuracy of the forecasts. One is to address and resolve the practical issues, such as forecasting for large utilities with diverse meteorological regions [2], data cleansing problem [1], and developing the forecast for some particular business needs [12], etc, which in turn strengthen the defensibility of the forecasts. The other one is to customize and combine new algorithms and techniques developed from other fields, of which most are still on the theoretical level and of limited practical value to the industry.

Benchmarking has been an issue in short term load forecasting (STLF) for decades. However, it has not received as much attention as it deserves. As a result, there is not an approach well established to produce benchmarking models for comparative assessment. Although regression analysis has been recognized as a legitimate tool to produce benchmarking models, the use of regression analysis varies from one to another [4, 7, 9, 11]. Most of the regression based approach segment the data and build separate models for different seasons, months, days of the week, or hours of the day. Different empirical rules, which may be specific to one utility, have been involved to segment the data or fine tune the model, so procedures are not easy to replicate to other utilities. Therefore, most of them are not eligible to be a benchmark.

This paper presents the benchmarking process, as part of the process to start up the planning department, for a fast growing US utility. A multiple linear regression (MLR) based load forecasting approach is demonstrated to produce the first in-house forecasts for this utility. The resulting model was also used for the utility to determine whether or not to install their own weather stations to collect weather data [6]. Due to the generic set up of the regressors, the proposed model can be served as a naïve MLR benchmark for a wide range of utilities that are accessible to two to three years of hourly temperature and load data.

This paper is organized as follows: Section II introduces the theoretical background of MLR and general linear models (GLM); Section III describes the details of the proposed model; Section IV presents the forecasting accuracy and discuses the interpretation of the model; Section V concludes the paper with discussion of the future work.

II. Multiple Linear Regression Analysis for STLF

A. General Linear Regression Models [8]

The general linear regression model with normal error terms can be defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + e_i \quad (1)$$

where $\beta_0 \cdots \beta_{p-1}$ are the parameters, $X_{i1} \cdots X_{i,p-1}$ are the known constants, e_i is the independent normally distributed random variable $N(0,\sigma^2)$. Then the response function is

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \quad (2)$$

T. Hong is with Quanta Technology, LLC, Raleigh, NC 27607 USA (e-mail: hongtao01@gmail.com).

P. Wang is with SAS Institute, Inc., Cary, NC 27513 USA

H. L. Willis is with Quanta Technology, LLC, Raleigh, NC 27607 USA.

where $X_1 \cdots X_{p-1}$ are $p-1$ predictor variables. Therefore, the definition (1) implies that the observations Y_i are independent normal variables, with mean $E[Y_i]$ as given by (2) and constant variance σ^2.

B. Quantitative and Qualitative Predictor Variables

In many cases, the predictor variables are quantitative. For example, as the customer count (number of customers in the utility's service territory) increases, the load shows an increasing pattern. If we model the load as a linear function of the customer count, the customer count can be considered as a quantitative predictor variable.

However, the definition of (1) does not limit the predictor variables to the quantitative ones. Qualitative predictor variables, sometimes called class variables or dummy variables, such as weekday or weekend, can also be included in the model. Indicator variables with values 0 and 1 can be used to identify the classes of a quantitative variable. For instance, if the load (Y) is predicted based on whether it falls in a weekday or weekend, a qualitative predictor variable X_1 can be defined as follows:

$$\begin{cases} X_1 = 1, & \text{if the day is a weekday} \\ X_1 = 0, & \text{if the day is a weekend} \end{cases} \tag{3}$$

Then the response function is

$$E[Y] = \beta_0 + \beta_1 X_1 \tag{4}$$

For the load in a weekday, $X_1 = 1$, and (4) becomes

$$E[Y] = \beta_0 + \beta_1 \tag{5}$$

For the load in a weekend, $X_1 = 0$, and (4) becomes

$$E[Y] = \beta_0 \tag{6}$$

In general, a qualitative variable with c classes can be represented by c-1 indicator variables. For example, a qualitative variable day of the week with 7 classes (Sunday, Monday, …, Saturday) can be represented as follows by 6 indicator variables:

$$\begin{cases} X_1 = 1, & \text{if the day is Sunday} \\ X_1 = 0, & \text{otherwise} \\ X_2 = 1, & \text{if the day is Monday} \\ X_2 = 0, & \text{otherwise} \\ & \cdots \\ X_6 = 1, & \text{if the day is Friday} \\ X_6 = 0, & \text{otherwise} \end{cases} \tag{7}$$

Then the response function of the regression model with day of the week as the predictor variable is

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 \tag{8}$$

C. Polynomial Regression

Polynomial regression models contain polynomial(s) of the predictor variable(s) making the response function curvilinear. For example, if the load (Y) is predicted by a polynomial regression model with one predictor variable temperature (X_i), and the order of this polynomial is three [3], the following model can be considered:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + e_i \tag{9}$$

It is a special case of (1), because it can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i \tag{10}$$

where $X_{i1} = X_i$, $X_{i2} = X_i^2$, and $X_{i3} = X_i^3$.

Notice that the squared or higher-ordered terms of a qualitative predictor variable is equivalent to the 1st ordered one.

D. Transformed Variables

In the polynomial regression models, the predictor variables can be transformed to reach the standard form of the GLM. The response variable can also be transformed to model some complex, curvilinear response functions. Considering the polynomial regression model of the relationship between the load and temperature, a model with a transformed Y variable can be written as:

$$\ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i \tag{11}$$

Again, it is also a special case of (1). If we let $Y_i' = \ln(Y_i)$, the model (11) can be written as

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i \tag{12}$$

E. Interaction Effects

When the effects of one predictor variables depend on the level(s) of some other predictor variable(s), interaction effects can be included in the GLM. Such models can also be called nonadditive regression models. The interaction terms can be implemented by the multiplying two or more predictor variables. An example of the nonadditive regression model with two predictor variables X_1 and X_2 is the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + e_i \tag{13}$$

It is still a special case of the GLM. By letting $X_{i3} = X_{i1} X_{i2}$, the model (13) can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i \tag{14}$$

F. Linear Model vs. Linear Response Surface

There should be no ambiguousness that the GLM can be used to generate a large variety of nonlinear response surfaces. In other words, linear models are not restricted to linear response surfaces. The term "linear" in GLM refers to the parameters. A regression model is "linear" in the parameters when it can be written as:

$$Y = XB + E \tag{15}$$

where Y is a vector of responses, B is a vector of parameters, X is a matrix of constants, and E is a vector of independent normal random variables with zero expectation and variance-covariance matrix $\sigma^2 I$.

III. Components for the Naïve Benchmark

A. Data

The benchmarking approach proposed in this paper requires two sets of data: hourly load at system level and hourly temperature, which most utilities have access to. Depending upon the maturity of the data management, the data quality may vary. To replicate the benchmarking model, the utility ought to have at least two years of history data with reasonably good quality. In this paper, four years (2005-2008)

of hourly load (in kW) and temperature (in F) data from a medium utility are used in the case study. The first three years are used as the history data, while the last year is the holdout sample. Fig. 1 shows the load series from 2005 to 2008, while Fig. 2 shows the corresponding temperature series.

B. Linear Trend

We define a quantitative variable (Trend) to capture the locally increasing (or decreasing) trend by assigning a natural number to each hour in ascending order. For instance, the Trend variable of the first hour in 2005 is 1, the second hour in 2005 is 2, and the last hour of 2008 is 35064. It should be noticed that such a trend only belongs to the utilities with a stable service territory and local economics. The linear trend can be interpreted as the linear approximation to the load series: when the length of load history is relatively short comparing to the macroeconomic changes, such as recession and booming, the overall trend of the load during the interval of interest can be linearly approximated. The benchmarking model discussed in this paper is not directly applicable to the significant business events, such as merging two utilities and splitting one utility into two.

C. Temperature

The relationship between the load and temperature has been intensively studied during the past several decades. For instance, piecewise linear function was indicated in [2]; piecewise quadratic function was used in [5]; the 3rd ordered polynomials were suggested in [3]. Fig. 3 shows the load temperature scatter plot of the utility in this study. The piecewise functions need the cut-off temperature(s), which may not be exactly the same in different service territories. The southern part of the United States may have a different cut-off temperature from the northern part, because the comfortable temperature zone may be different for people living in different regions. Therefore, the 3rd ordered polynomials of the temperature are used to predict the load for the benchmarking purpose.

D. Calendar Variables

It is well known that there are three seasonal blocks in the load series: day, week, and year. There can be different treatments to each block depending upon the load consumption behavior of the particular service territory. For example, the 7 days of a week can be modeled by a qualitative variable with 2 classes (weekdays and weekends), 3 classes (weekdays, two weekend classes), etc. In different countries, the weekends may be defined differently. For instance, Thursday and Friday are weekends in Iran. As a benchmarking model, the qualitative variables (Hour, Day, and Month) with 24, 7, and 12 classes, are used to model the 24 hours of a day, 7 days of a week, and 12 months of a year respectively. Namely each seasonal block is decomposed into the unit of the highest resolution.

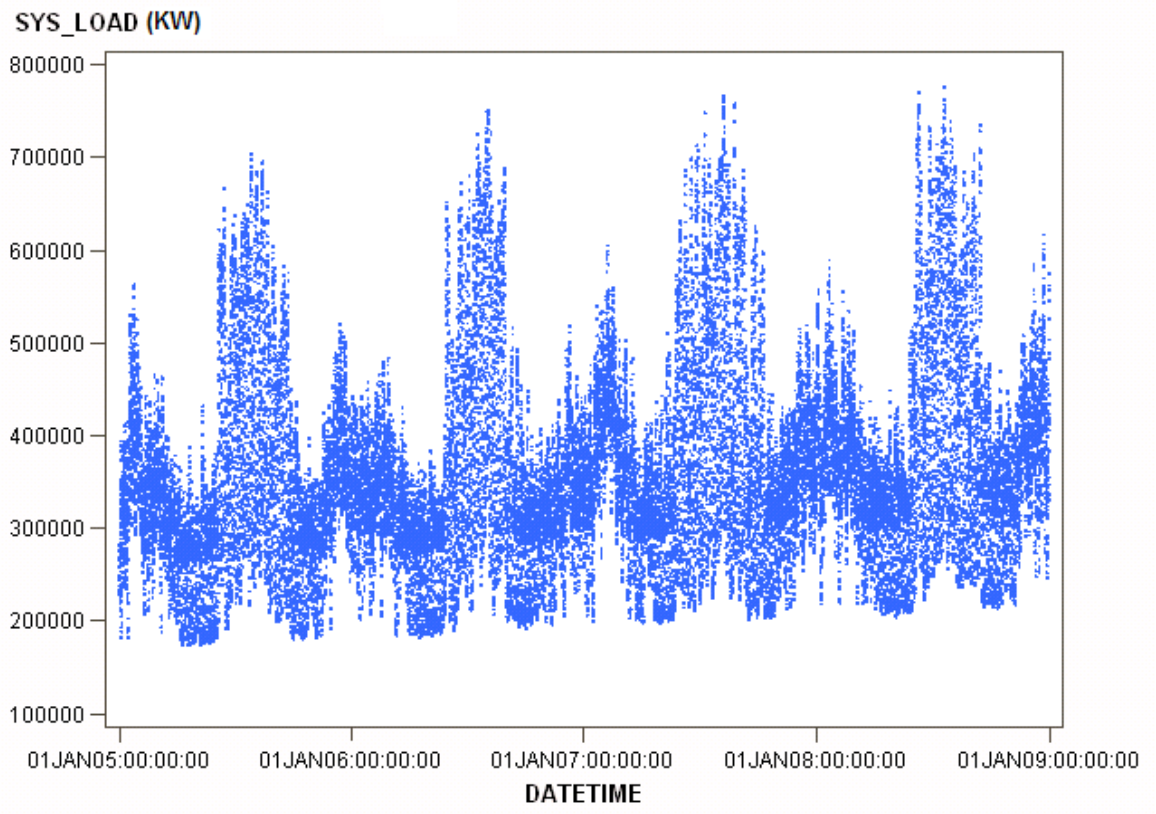

Fig. 1. Load series (2005-2008).

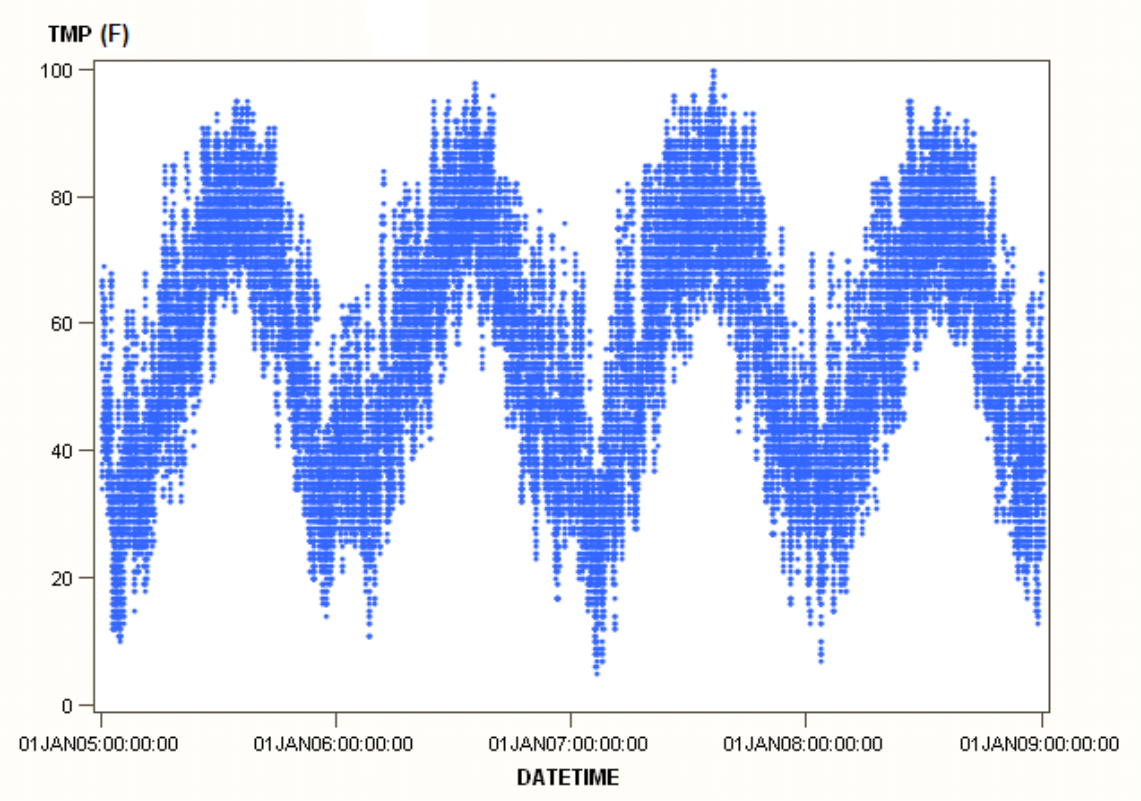

Fig. 2. Temperature series (2005-2008).

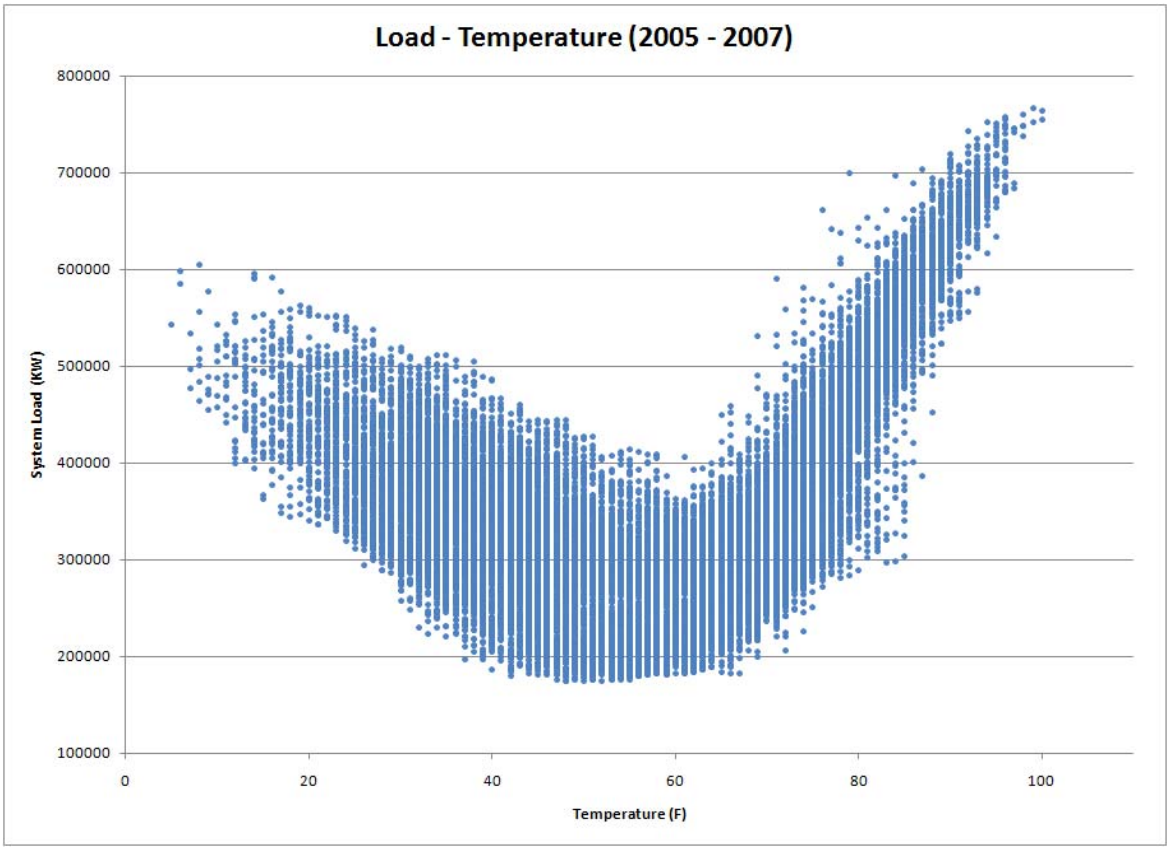

Fig. 3. Load-temperature scatter plot (2005-2007).

E. Interaction Effects

Normally an afternoon is warmer than a midnight; a summer is warmer than a winter. In other words, temperature is not independent of the hour of the day or the month of the year. Therefore, the interaction effects between the temperature (appeared in the form of 3rd ordered polynomial as discussed above) and the calendar variables Hour and Month should be in the model. Since it is not evidential that there is a relationship between days of the week and the temperature, the interaction between temperature and days of the week is not included in the model.

The hours in different days of a week may result in different load due to human activities. For instance, there may be lower load in the morning of the weekends than the other mornings, because people do not have to get up as early as weekdays to go to work, which results in less load at home and office buildings in the weekend mornings. Therefore, the interaction effect between Hour and Day should be included in the benchmark model.

When a qualitative predictor variable interacts with a quantitative one, this quantitative term is not required to appear in the model as a single regressor (main effect). When two qualitative predictor variables interact together, either of them is not required in the model as a main effect. Therefore, the benchmarking model GLMLF-B includes the follows:

1) Quantitative variables: *Trend*, and *TMP* (the current hour temperature);

2) Class variables: *Hour*, *Day*, *Month*;

3) Main effects: *Trend*, *Month*;

4) Interaction effects (also known as cross effects): $Day \times Hour$, $Month \times TMP$, $Month \times TMP^2$, $Month \times TMP^3$, $Hour \times TMP$, $Hour \times TMP^2$, $Hour \times TMP^3$, where the cross sign represents the interaction effect;

5) Intercept.

The model can be written as:

$$E(Load) = \beta_0 + \beta_1 \times Trend + \beta_2 \times Day \times Hour + \beta_3 \times Month + \beta_4 \times Month \times TMP + \beta_5 \times Month \times TMP^2 + \beta_6 \times Month \times TMP^3 + \beta_7 \times Hour \times TMP + \beta_8 \times Hour \times TMP^2 + \beta_9 \times Hour \times TMP^3, \quad (16)$$

With the above information, the model GLMLF-B as specified in (16) can be easily implemented in commercial statistical software packages, such as SAS 9.2 with STAT.

IV. Results and Discussions

This section presents the forecasting performance of the above benchmarking model GLMLF-B. The following error analysis is often used in forecasting:

1) Error, the difference between a quantity and its estimated or measured quantity.

2) Absolute Error, the absolute value of error.

3) Absolute Percentage Error, 100% times the relative error (the error divided by the true value).

The distribution of the above errors can be characterized by mean, standard deviation, minimum, Q1 (first quartile), median, Q3 (third quartile), and maximum values.

Several engineering concepts have been involved when the load forecasts are communicated:

1) Hourly load, the energy consumed in an hour. Sometimes it is calculated by averaging several instantaneous measurements occurred in an hour.

2) Energy, the summation of the hourly load within a specific period.

3) Peak/valley load, the maximum/minimum of the hourly load within a specific period.

4) Peak/valley hour load, the load occurs during the hour where actual peak/valley load occurs.

TABLE I
Results (MAPE, %) of the Naïve Model

	Forecasting horizon (# of days)						
	1	2	3	4	5	6	7
Hourly load	4.98	5.00	5.01	5.01	5.02	5.03	5.04
Daily peak hour	4.27	4.29	4.29	4.29	4.30	4.30	4.30
Daily valley hour	5.44	5.46	5.47	5.49	5.50	5.50	5.52
Daily peak	3.94	3.96	3.97	3.97	3.98	3.99	3.99
Daily valley	4.93	4.95	4.96	4.98	4.99	5.00	5.02
Daily Energy	3.49	3.51	3.52	3.53	3.53	3.54	3.55

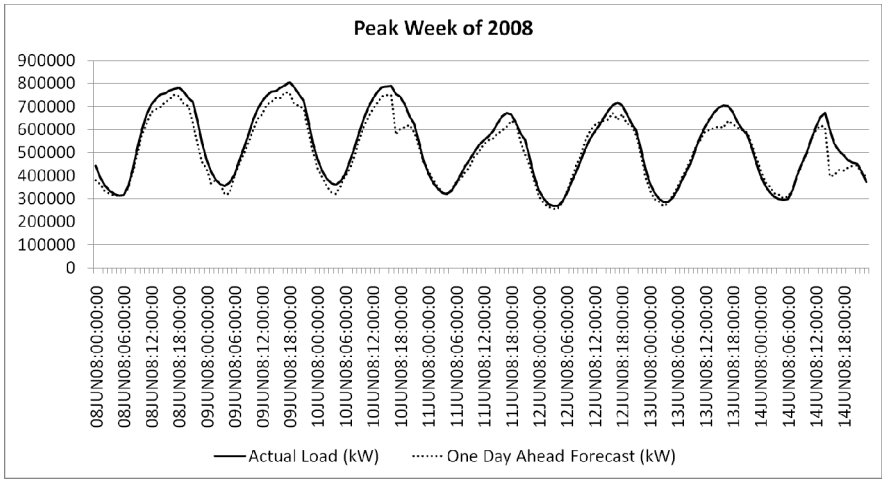

Fig. 4. Forecasting performance during the peak week of 2008.

The various combinations of the above error analysis and engineering concepts, together with the updating cycle and forecast horizon, form the diagnostic statistics for each load forecast. In this paper, to present the performance of a short term load forecast, the mean absolute percentage errors (MAPE) of the hourly load, daily energy, daily peak/valley load, and daily peak/valley hour load can be performed with the updating cycle ranging from one day to seven days, as listed in Table I. Rolling regression was applied to generate the forecasts for different horizons. For instance, when the forecasting horizon is one day, the model is updated every day and forecasts the loads of the next day. In every update, the newly available loads and temperatures are added to the history data. The forecasted and actual loads of the peak week of 2008 are shown in Fig. 4.

The results listed in Table I and Fig. 4 don't seem to be attractive comparing with the ones reported in the literature. In practice, due to the weather forecast errors, the actual MAPEs are even worse. A major reason is that they are produced out of a naïve model, which does not capture the specific local electricity consumption behaviors, such as weekend effect, holiday effect, etc. However, it should be noticed that the predictability of the load series from different utilities varies. For instance, the utility was fairly comfortable with this forecasting performance, which was much superior to the ones produced by several internal and external parties. The proposed model served as a production model in this utility for a year, before it was replaced by a customized one.

V. Conclusion

This paper proposes a naïve MLP based benchmarking approach for STLF. The proposed model is designed to be generic and applicable to a wide range of utilities. The work presented in this paper is based on the experience of helping a US utility develop the first in-house short term load forecasts. The benchmarking process has been implemented in this utility. Future work of this paper is to customize the proposed naïve benchmark to capture the local electricity consumption behavior of this utility.

VI. References

[1] J. Chen, W. Li, A. Lau, J. Cao, and K. Wang, "Automated Load Curve Data Cleansing in Power Systems," *Smart Grid, IEEE Transactions on*, vol. 1, pp. 213-221, 2010.

[2] S. Fan, K. Methaprayoon, and W.-J. Lee, "Multiregion Load Forecasting for System With Large Geographical Area," *IEEE Transactions on Industry Applications*, vol. 45, pp. 1452-1459, 2009.

[3] M. T. Hagan and S. M. Behr, "The Time Series Approach to Short Term Load Forecasting," *IEEE Transactions on Power Systems*, vol. 2, pp. 785-791, 1987.

[4] T. Haida and S. Muto, "Regression based peak load forecasting using a transformation technique," *IEEE Transactions on Power Systems*, vol. 9, pp. 1788-1794, 1994.

[5] T. Hong, M. Gui, M. E. Baran, and H. L. Willis, "Modeling and forecasting hourly electric load by multiple linear regression with interactions," presented at 2010 IEEE Power and Energy Society General Meeting, Minneapolis, Minnesota, USA, July 25-29, 2010.

[6] T. Hong, P. Wang, A. Pahwa, M. Gui, and S. M. Hsiang, "Cost of temperature history data uncertainties in short term electric load forecasting," presented at 2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS2010), Singapore, June 14-17, 2010.

[7] O. Hyde and P. F. Hodnett, "An adaptable automated procedure for short-term electricity load forecasting," *IEEE Transactions on Power Systems*, vol. 12, pp. 84-94, 1997.

[8] H. M. Kutner, J. C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed: McGraw-Hill/Irwin, 2004.

[9] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *IEEE Transactions on Power Systems*, vol. 5, pp. 1535-1547, 1990.

[10] D. K. Ranaweera, G. G. Karady, and R. G. Farmer, "Economic impact analysis of load forecasting," *IEEE Transactions on Power Systems*, vol. 12, pp. 1388-1392, 1997.

[11] S. Ruzic, A. Vuckovic, and N. Nikolic, "Weather sensitive method for short term load forecasting in Electric Power Utility of Serbia," *IEEE Transactions on Power Systems*, vol. 18, pp. 1581-1586, 2003.

[12] T. Saksornchai, W.-J. Lee, K. Methaprayoon, J. R. Liao, and R. J. Ross, "Improve the unit commitment scheduling by using the neural-network-based short-term load forecasting," *Industry Applications, IEEE Transactions on*, vol. 41, pp. 169-179, 2005.

VII. Biographies

Tao Hong is a Principal Engineer in Quanta Technology. His major areas of expertise are in operations research and its related applications in power distribution engineering and T&D planning. He has applied various statistics and optimization techniques to T&D loss studies, development of long term and short term load forecasting algorithms and tools. His work has been applied to many utilities in US and Canada. He received his Bachelor of Engineering degree in Automation from Tsinghua University, Beijing, a Master of Science degree in Electrical Engineering, a Master of Science degree with co-majors in Operations Research and Industrial Engineering, a PhD degree with co-majors in Operations Research and Electrical Engineering, all from North Carolina State University. Dr. Tao Hong is a winner of the poster competition in SAS' 12th Annual Data Mining Conference in 2009 for his work "*Behavior Mining of Electric Load Consumption: A Regression Approach*".

Pu Wang is a research statistician developer in SAS Institute. She is specialized in operations research and statistical analysis with the applications in merchandise intelligence solutions. She has been doing research and development of emerging techniques in market response modeling and demand forecasting for fashion goods retailers. She received her Bachelor of Engineering degree in Industrial Engineering from Tsinghua University, Beijing, where she was awarded with a first-class scholarship for the academic excellence. She received the Master and PhD degrees in Industrial Engineering from North Carolina State University.

H. Lee Willis, PE, (F'92) is a Senior Vice President and Executive Advisor in Quanta Technology. He has more than 35 years of electric T&D systems planning and engineering experience. He has been transmission planning manager at a major investor-owned utility, an executive with a major equipment supplier, and a senior consultant who has directly performed or supervised over 400 system planning and asset management projects for utilities around the world. Lee pioneered many of the modern planning and asset management methods now considered industry best practice, including spatial simulation load forecasting for T&D planning, load-reach voltage planning of feeder system layouts, and options-based asset management prioritization. He is a Fellow of the IEEE and a past chairman of its Power System Planning and Implementation committee. From 1999 to 2005 he served on the US National Research Council, which advises the US Congress and Dept. of Commerce on the nation's civilian technology needs. Lee has published more than 240 papers and articles on power systems, forecasting and planning, and utility practices during his career, including 78 papers in peer-reviewed and refereed technical journals, one theme cover article in the IEEE-wide *Proceedings of the IEEE,* and five cover articles for the *IEEE Computer Applications in Power* journal. He is the author of eight books on power system planning including the *Power Distribution Planning Reference Book* and *Spatial Electric Load Forecasting*.