Targeted Residual Analysis for Improving Electric Load Forecasting

Scott Alfeld, Paul Barford

Department of Computer Sciences
University of Wisconsin - Madison
1210 W. Dayton St, Madison, WI USA
[1] salfeld@cs.wisc.edu
[2] pb@cs.wisc.edu

Abstract*—Management and operation of the electrical grid in the US is handled in large part by regional authorities called *Independent System Operators* (ISO's). One of the key activities of an ISO is *load forecasting* which is critical to short-term energy trading markets and effective operation of the power grid. In this paper, we analyze load forecasts and develop methods for improving forecasts that can be used directly by ISO's or third parties. Specifically, we assess the hourly electrical load forecasts against actual load data provided by Midwest ISO over a two-year period. Residual analysis shows systematic inaccuracies in hourly forecasts that can be caused by a variety of factors including modeling errors and pumped storage in the grid. We utilize machine learning-based methods to improve forecasts over short time horizons. Our methods reduce the mean squared error of forecasts over the entire year by roughly 20%. By shortening the forecast horizon to 1 to 32 hours, we are able to improve by over 90%. These improvements can be important in operational energy market contexts, where even small differences in forecasts can lead to large swings in transmission behavior and market activity.

I. Introduction

The electrical power system in the US is extremely large and complex. At the highest level, it is composed of diverse power generation facilities, transmission (typically overhead) and transformer infrastructure, and a wide range of power consumers. The key objective of the system is to ensure that power supply is sufficient to match demand at any point in time. The main challenges in meeting this objective include the enormous scale of the infrastructure, inherent dynamics and variabilities, and limitations in generation and transmission capabilities.

To address these challenges and help ensure the ready availability of power for the public, the US Federal Energy Regulatory Commission (FERC, which is under the aegis of the US Department of Energy) has established a series of regional organizations that oversee and coordinate activities within the electrical power system. One of the most important aspects of these Regional Transmission Organizations (RTOs) and Independent System Operators (ISOs) is their administration of a region's wholesale electricity market for buying, selling and trading power. A key element of that activity is *load forecasting*. Accurate forecasts are critical for both operations and effective market participation.

Electric load forecasts are produced by ISOs using proprietary models with diverse inputs that include direct projections of power needs from large consumers (*e.g.,* factories or large buildings), weather forecasts, and historical demand among others. Forecasts are generated at a variety of time horizons ranging from forecasting load over the next hour, to estimating annual trends. Forecast load data along with actual hourly load data is openly available from ISOs.

In this paper, we analyze the accuracy of load forecasts produced by the US Midwest ISO (MISO) over a two year period from 2010 and 2011. The goals of our work are to *(i)* understand the accuracy of the forecasts, *(ii)* infer any causes for systematic inaccuracies and *(iii)* improve the accuracy of the forecasts. The primary challenge in our work is that ISOs do not publish the inputs to their forecasting model nor do they disclose any of the details of the model itself. Thus, we are left with only the outputs of the model and actual load values as the basis for our work.

We begin by conducting a residual analysis of forecast versus actual loads. Our results reveal a variety of differences between the two data sets, with forecasts typically underestimating actual load. Over longer time scales, we observe seasonal transitions in residuals. Over shorter time scales, we observe that most of the residual errors are early hours of the day (*i.e.,* between 1:00am and 4:00am). These results and further investigation of MISO's methods lead to our identification of three general sources of error: deliberate distortion, forecast horizon and inherent model error.

Based on these error sources, we develop methods to model and reduce errors in forecasts. Our approach uses a variant of stacked generalization that is adapted for our application. Specifically, we develop linear models which we use to augment existing forecasts from a blackbox model. We then show that for each of the three error sources, our methods improve forecasts. In particular, our methods are able to improve forecasts during times of deliberate distortion (*e.g.,* due to pumped storage activities) and for shorter forecast horizons, reducing the error by over 90% in some instances. Our models can be directly plugged into existing models during some times of the day, reducing the error by 10% to 35% for early morning daytime hours. Our methods of analysis can be used for decision support as well, indicating times of day and year where existing models can be potentially be improved.

II. Related Work

Inaccurate load forecasts lead to various difficulties. Individual entities which consume large amounts of electricity, such as paper mills and data centers, are often legally required to provide load forecasts to utility companies [18]. In addition, Independent System Operators (ISO's) and Regional Transmission Organizations (RTOs) which govern the electricity markets, flow, and generation, must forecast future load. The load forecasts must be accurate to plan the electric dispatch of power plants and ensure stability, reliability, and efficiency [18].

Should the forecast underestimate the actual load, scheduled generation resources may not be sufficient to ensure system reliability. To meet the additional demand (beyond what was forecast), less efficient and more costly generation methods must be used or additional power must be purchased [13]. If instead the forecast underestimates the actual load, generation may have been committed, resulting in unnecessary fuel and maintenance costs. Power may have been unnecessarily purchased, and relative prices may have been set too high [13]. For analyses of the effects of load forecast accuracy, see [1], [11], [20] and references therein. To the best of our knowledge, there is no publicly available assessment of the ramifications (economic or otherwise) of the load forecast data used in this work [23].

Historically, statistical methods such as multiple linear regression, stochastic time series, general exponential smoothing, as well as state-space and knowledge-based methods, have been used to predict future hourly load [19]. Starting in the 1980's, Artificial Neural Networks (ANNs) have been evaluated for use in load forecasting [10]. While it was seen by some as potentially a passing fad in the 1990's [8], by the early 2000's ANNs were widely adopted in load forecasting [12].

ANNs gain some performance improvements over other methods because they are able to detect and account for non-linear relationships between inputs. Today, the prevalence of ANNs continues and many variations and extensions are being explored actively. MISO, as well as *e.g.,* California ISO both use ANNs in their load forecasting[1]. Some examples of extensions being explored in the academic community include combining the wavelet transform with ANNs [5], [24], [22] and Particle Swarm Optimization (PSO) [3].

III. Electricity Ecosystem

Historically in the USA, vertically integrated electric utilities provided power. In 1996, The Federal Energy Regulatory Commission (FERC) issued orders 888 and 889 which opened the doors for competition to the USA's wholesale electricity market, and brought transparency by requiring an open access same-time information system (OASIS) for current operating status and transmission capacities. In 1998 several transmission owners voluntarily came together to establish the not-for-profit company Midwest ISO (MISO). In December of 2001, FERC approved MISO as the first Regional Transmission Organization in the USA. Today, MISO manages over 175 GW of generation capacity, with over 65,000 miles of transmission lines spanning 15 U.S. States and parts of Canada.

Load-serving entities and load-balancing authorities submit their own, local, load forecasts to MISO. The details of this arrangement are found in [18]. MISO incorporates these forecasts, along with other data sources, to create its own load forecasts. The internal workings of MISO's load forecast model, and the data used as input, are proprietary, and unavailable to the public and market participants. It is known, however, that MISO uses an artificial neural network with weather being a primary component [23], [18].

In addition to managing the power grid, MISO also serves as a moderator (much like an auctioneer in a traditional auction) in two wholesale electricity markets. In one of these, the day-ahead market, the price of electricity changes every hour and prices are broadcast for the following day. The annual gross market charges for 2011 were over $23 billion. Detailed descriptions of these markets are found in [15], [16], [17]. The real-time load is a key contributor to the rapidly changing price of electricity.

IV. Residual Analysis

MISO provides midterm load forecasts (MTLF) as well as actual load data for three regions: West, Central, and East. This data is publicly available[2] as the load (in MW) for each hour of each day. MISO uses load forecasts for Reliability Assessment Commitment, Market Participant Estimation of Operating Reserve Obligations, Real Time 5 Minute Dispatch and Look Ahead Commitment Processes. Details of these uses are found in [18]. As described in more detail below, we used total aggregate load (West + Central + East) for the years 2010 and 2011. We chose to analyze the total load in part because taking the sum of the three regions should help reduce random noise.

We examine how the error of the MTLF behaves as a function of time by looking at the following two measures:

1) **Cumulative Error:**
   $c_t = \sum_{i=1}^{t} (M_t - L_t)$.
2) **Cumulative Absolute Error:**
   $a_t = \sum_{i=1}^{t} |M_t - L_t|$

where L_t is the actual load at time t and M_t is the predicted load (the MTLF) at time t. In addition, we consider aggregate statistics as well as the raw signal of errors over time (see Figure 1(a)).

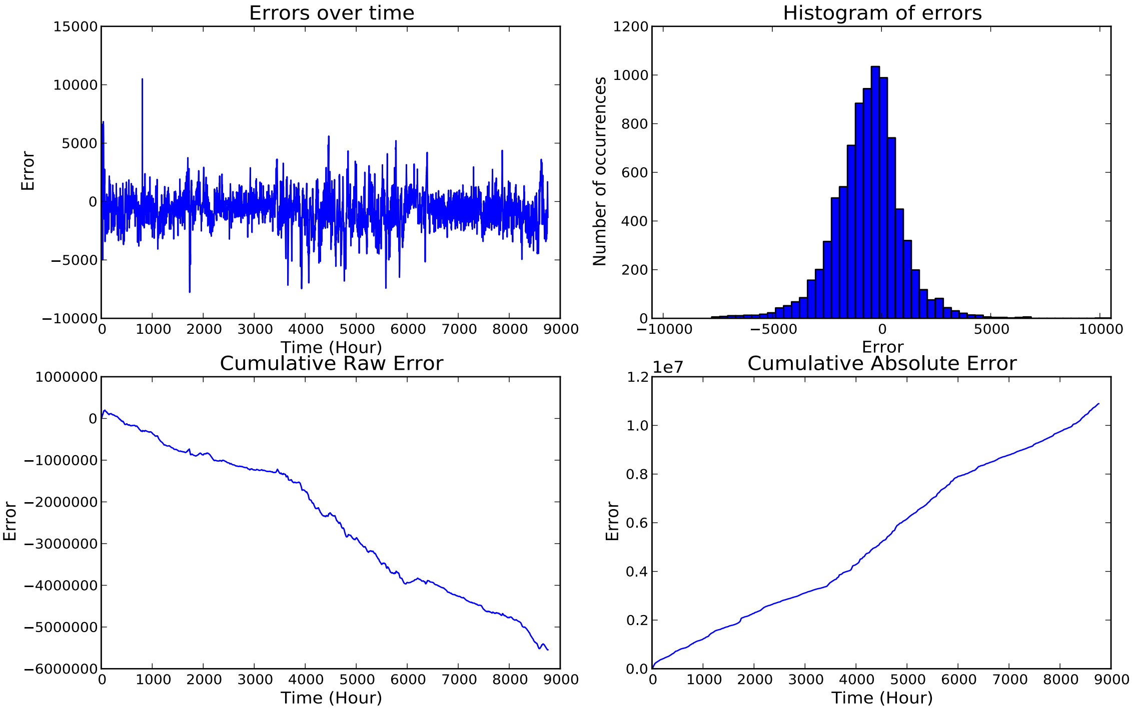
We focus our attention on 2011, noting several features of the errors. The presence of these features is validated by observing similar phenomena in 2010.

First and foremost, we see definite structure, indicating that there is room for model improvement. Second, both the histogram and the cumulative error (upper right and lower left of Figure 1(a), respectively) indicate that the MTLF tends to underestimate the actual load.

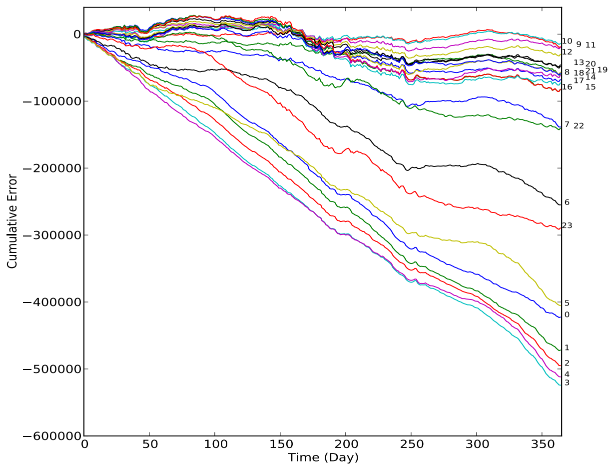
Finally, we note the apparent presence of events occurring where the pattern of errors changes. In the cumulative error

[1](www.caiso.com/1ca5/1ca5a7a026270.pdf)

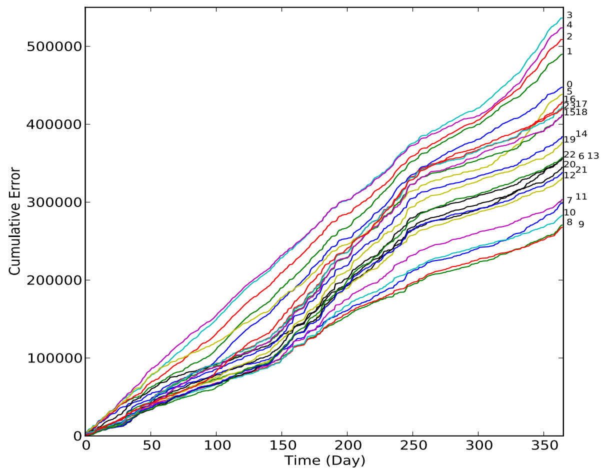
[2](https://www.midwestiso.org/LIBRARY/Pages/Library.aspx)



(a) 2010



(b) 2011



(c) 2011

Fig. 1: **Midterm Load Forecast Errors over all of 2010**. (a): Upper Left: The raw errors, as defined by $M_t - L_t$ over the course of the year. Upper Right: A histogram of the raw error values. Lower Left: The cumulative error ($c_t = \sum_{i=1}^{t} (M_t - L_t)$). Lower Right: The cumulative absolute error (($a_t = \sum_{i=1}^{t} |M_t - L_t|$). (b): **Cumulative error by hour.** Each line represents a single hour of the day. (c): **Cumulative absolute error by hour.** Each line represents a single hour of the day.

(lower left) we see a stabilization (the slope becomes close to zero) from around hour 1100 to roughly hour 2000 (mid February to late March), and again from hour 6000 to hour 7500 (early September to early November). In the cumulative absolute error curve (lower right), we see two pronounced bends. The slope, representative of the average absolute error, is relatively low from the start of the year to roughly hour 3500 (late May). The slope is then high until roughly hour 6000 (early September), before returning back to its more shallow behavior. This behavior follows the intuition of the total load (and thus expected error) being higher in the warmer months and lower in the colder months. The sharp changes, however, may indicate an artifact of MISO's forecasting model.

In a method similar to that used to define the LMP landscape [2], we split the year into 24 segments. We have one segment for each hour of the day, and each segment consists of 365 single-hour periods. Figures 1(b) and 1(c) show the two errors functions for 2011.

Consider Figure 1(b)—the cumulative sum of errors. Each hour of the day is represented by a single line, with the hour marked on the right. Hour i corresponds to the hour from i to $i + 1$ in a 0-based 24-hour clock. The hour from 3:00AM to 4:00AM ($i = 3$) has the most negative cumulative error, as well as the highest cumulative absolute error. The maximum errors at the right end of the horizontal scale correspond to the hours of the day in the order indicated in Table I.

We now consider Figure 1(c). Recall the previously described two bends as shown in Figure 1(a). When we split the error values into individual hours, we observe that a subset of hours exhibit only the first bend, while another subset exhibit only the second and some exhibit both. We believe that these bending points, as well as the behavior of individual hours of the day with respect to them, can be explained by artifacts of the underlying prediction model in combination with natural events (e.g., the changing of the seasons).

V. Sources of Error

In this paper, we identify three sources of error in the MTLF and develop methods to analzye and remedy each.

Deliberate Distortions: MISO does not publish the actual predictions it generates internally. The publicly available MTLF is actually an edited version of MISO's original predictions. The published predictions include alterations due to the use of pumped storage (pumping water up-river to store electricity) in the grid [18].

Forecast Horizon: The *horizon* of a forecast is how far in to the future predictions are made. Each day at 4:00PM, MISO forecasts from midnight to midnight of the following day. This yields a forecast horizon of 32 hours. In general, forecasting a few hours ahead is more accurate than forecasting many hours ahead.

Model Errors: Due to noise and unpredictable circumstances, no model can be completely accurate. We assume that the underlying model used for forecasting is not perfect, and that its deficiencies can be corrected to a degree. Note that this is the most difficult source of error to account for.

Our methods for detecting and adjusting for the first two sources of error are geared towards helping those outside of MISO. If a user aims to forecast actual electricity usage from MISO's publicly available data, our methods for dealing with deliberate distortions yield more accurate results than taking MISO's MTLF at face value. If, instead, a user aims to forecast a short (less than 32 hours) time into the future, our methods for addressing the forecast horizon can be helpful.

For the third source of error (model errors), our models and analyses lend aid directly to MISO in three distinct ways.

1) **A direct plugin**:
Our models may be plugged in directly to improve the accuracy of MISO's predictions. Details of applying our system are found in Section VII.
2) **Decision support**:
Our methods yield new predictions. Targeted residual analysis of these predictions, in combination with MISO's original forecasts, yield insights as to how MISO's model may be improved.
3) **Further data acquisition**:
By incorporating additional sources of data that may or may not be considered by MISO, we may improve their results, which would indicate that their model should be updated with this data. This area is left as future work.

VI. Modeling Methodology

It is important to note that we do not model *actual* load. That is to say that, unlike other work, we are not directly predicting future load. Instead, we explicitly model the discrepancy between the underlying forecast model and the observed load. Using a variant of stacked generalization [26], we use our model to augment the output from the underlying forecast model.

The fact that we model the MTLF's error (specifically the ratio of the actual load over the predicted load) leads to the following phenomenon. Suppose the actual load is 90MWh and the predicted value (MTLF) is 100MHh. This means that the target value for our system is $90/100 = 0.9$. However, if our system predicts a ratio anywhere in the range $(0.8, 1.0)$, then the predicted load will be in the range $(80, 100)$ and thus a more accurate (in terms of absolute deviation) prediction than the original 100. We conjecture that one can, in a method similar to ϵ-insensitive support vector regression, incorporate knowledge of this "butter zone" into the loss function of machine learning algorithms to improve predictions. This is left as future work.

We applied this process and modeled the ratio between the true load and the forecast. We begin with the following notation.

- L_t, the actual load at time t.
- M_t, the predicted (MTLF) load at time t.
- R_t, the ratio at time t.
- $\hat{R}_t$, our model's predicted ratio at time t.
- $\hat{P}_t$, our model's prediction of the load at time t.

To model the discrepancy between MISO's model and reality, we let $R_t = \frac{L_t}{M_t}$. After learning a predictive model of R, we

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Raw	3	4	2	1	0	5	23	6	22	7	15	16	17	14	18	8	21	19	20	13	12	11	9	10
Absolute	3	4	2	1	0	5	16	17	23	18	15	14	19	22	13	6	20	21	12	11	7	10	8	9

TABLE I: **Hour Error Rankings.** On the left end of the rankings (Rank 1), we have the hour with maximal cumulative absolute error, and most extreme (most negative) cumulative raw error. Rank 24 corresponds to the hours with minimal absolute error, and maximal (closest to zero) raw error. Hour i is the hour from i:00 to $(i+1)$:00 on a 24-hour clock.

produce our final predictions as $\hat{P}_t = \hat{R}_t \times M_t$.

For the results presented in this work, we performed prediction via a sliding-window scheme [7]. We chose to model the ratios as a function of the previous d ratios, where we refer to d as the *Markov factor*, in addition to several exogenous variables. We chose $d = 168$ (the number of hours in a week) based on preliminary testing with 2010 data.

We experimented with a variety of models including ϵ-insensitive Support Vector Regression (SVR) with polynomial, sigmoid, and radial basis function kernels [4], [6]. We also tried k Nearest Neighbor (kNN) Regression with a variety of k, and the Holt-Winters [25] algorithm. We compared algorithms based on the mean squared error over all hours in 2010. SVR proved most accurate when emulating MISO's prediction process, and Linear Regression proved best for shorter horizon times (discussed below).

We tried several exogenous variables for predicting R_t.

- The past 24 hourly temperatures (from $t-25$ to $t-1$) for the five most populous cities in MISO's footprint: Chicago, Detroit, Milwaukee, Indianapolis, and Colombus.[3]
- The current (at time t) hour of day (from 0 to 24).
- The current (at time t) day of week (from 0 to 6).
- The past 24 or 168 hourly MTLF values.

Temperature features proved universally unhelpful. We conjecture that this is due to the complex relationship between weather and electricity usage and any benefit from incorporating the additional data was outweighed by the noise introduced. The model used for the results presented here used the hour of day, day of week, and past 168 MTLF values, based again on results for 2010 data.

We standardized the data as a preprocessing step. In general, this is done by rescaling values to be in the range $[0, 1]$. We chose to rescale values over an entire year, based on preliminary testing and standard practice in time series analysis. Note, however, that at the beginning of e.g., 2011, the maximum and minimum values over the year are unknown. Therefore, to maintain the legitimacy of our predicting procedure, we rescaled the values 2011 based on the extreme values found in 2010.

We tuned the SVR hyperparameters via a grid-based coordinate descent method on 2010 data. The best results were obtained by using an rbf kernel with penalization parameter $C = 0.01$, a value of $\epsilon = 0.01$ (regarding ϵ-insensitivity) and $\gamma = 0.01$.

[3]Data obtained from www.wunderground.com.

We used the Python library Scikit-Learn v0.14.1 [21]. Plots presented in this work were created using Matplotlib v1.3.0 [14].

To emulate MISO's prediction process, at 4:00pm every day we forecast the following day. Using a sliding window scheme, we first predict R_t one hour ahead, and then feed it as a feature to our model and predict one hour later.[4] We continue until we have produced 32 ratio predictions (8 hours until midnight, then 24 for the following day). The pre-forecast hours (from 4:00pm to midnight) are then discarded, and we multiply with our ratios to obtain the 24 $\hat{P}_t$'s.

An alternative approach, sometimes used in load forecasting, is to maintain separate models for separate hours of the day.

Note that the 24-model approach may have suffered due to a lack of training data compared to the sliding window. Because individual models predict only one hour of the day, a maximum number of training points for a year is 365. In the sliding window method, however, a typical year contains $365 \times 24 = 8760$ training points.

In addition, the 24-model approach assumes independence among hours of the day. In the sliding window scheme, early predictions may affect later predictions because they are fed back in as features. This allows some encoding of dependence when predicting several hours at a time.

With the goal of decision support in mind, we maintained a consistent training time, training a new model for each set of predictions. That is to say that for day n, we train a single model on days $(n-1, n-2, \ldots)$ and predict all 24 values for day n. To forecast day $n+1$, we train a new model using the same amount of training data, but now ending at day n. The results presented here use a training time of $T = 4032$ hours (24 weeks or approximately six months) and a Markov factor of $d = 168$ hours (one week). All evaluations presented were done by predicting all of 2011. When necessary, part of 2010 was trained on.

VII. Results

In training our models we aimed to minimize Mean Squared Error (MSE). Note that we minimized the MSE of the ratios, not the MSE of the final load predictions. We then generated predictions for every hour of 2011 as described above. We calculated the error of our model, as well as that of the MTLF. The error functions we used are:

[4]We tested a 24-model (one for each hour of the day) approach on the 2010 data. To predict a full day's load, we queried each model separately and concatenated their responses. Using a sliding window outperformed the 24-model approach, and thus it is what we used for the results presented in this work.

1) **Mean Squared Error (MSE)**:
$\frac{1}{n}\sum_t (L_t - \hat{P}_t)^2$.
2) **Mean Absolute Difference (MAD)**:
$\frac{1}{n}\sum_t |L_t - \hat{P}_t|$.
3) **Median Absolute Difference (MED)**:
$\arg\min_\nu |\ |L_t - \hat{P}_t| - \nu|$.

Here, n is the number of times predicted and the MTLF errors are calculated with M_t in place of $\hat{P}_t$.

We list each of the three error functions due to their qualitative differences. MSE is more sensitive to few large errors than to many small errors, while MAD has no such bias. MISO has indicated that improving MSE is a worthwhile goal [23]. While perhaps less intuitive, MED provides a more robust measure of error.

We then computed the ratio of our augmented prediction's error to the MTLF error. A high (above 1.0) ratio means our new predictions do worse than the MTLF; a low ratio (below 1.0) means our methods are improving the forecast performance.

A. *Deliberate Distortion.*

Our method of improving forecasts is agnostic to what errors are resulting from model error, and what errors are deliberate distortions. We apply domain knowledge to distinguish the two when analyzing the behavior of our approach.

We simulated MISO's forecasting procedure, predicting from midnight to midnight at 4:00PM on the preceding day. Over the entire year of 2011, we obtained $\approx$ .80 MISO's MTLF MSE, $\approx$ .85 their MAD, and $\approx$ .79 their MED. Note that therefore, as a direct plug in, we improve on the MTLF's accuracy by comparable amounts in each of the three measures. If instead we improved with respect to MSE, but made no improvement on MAD or MED, it would indicate that we reduced peak errors, but made less general improvement.

Recall that the publicly available MTLF values provided by MISO are not their actual predicted values [18]. MTLF values include alterations due to pumped storage. MISO domain experts have told us that these alterations occur between 10:00PM and 6:00AM [23]. Using this knowledge, we split the 24-hour day into nighttime hours (10:00PM to 6:00AM) and daytime hours (6:00AM to 10:00PM).

	Nighttime	Daytime
MSE Ratio	.441	1.062
MAD Ratio	.611	1.042
MED Ratio	.517	1.050

We do slightly worse than the original MTLF during the daytime hours and substantially better during the nighttime hours. Intuitively, this makes sense because the daytime errors are due to model errors as opposed to deliberate distortions. In summary, in predicting nighttime hours, our method yields $\approx$ .44 the MTLF's MSE averaged over the year (and $\approx$ .61 the MAD, $\approx$ .52 the MED). In predicting daytime hours, further efforts are needed (as described in subsection VII-C) to improve accuracy beyond that of the original MTLF.

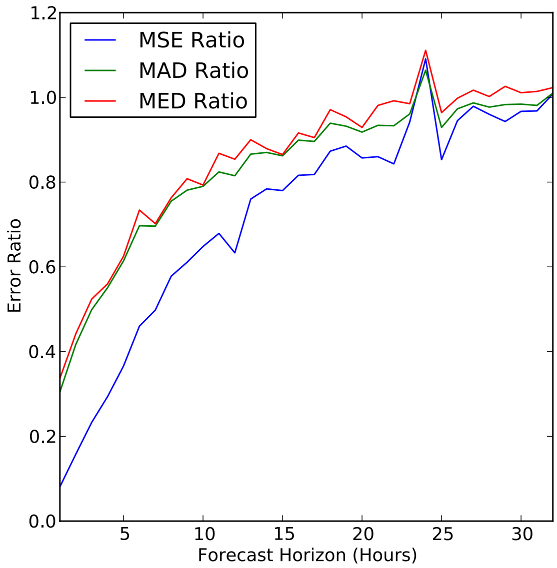


Fig. 2: **Relative daytime error as a function of forecast horizon.** The vertical axis represents the ratio of our method's error to the MTLF error. We consider Mean Squared Error (MSE), Mean Absolute Difference (MAD), and Median error (MED).

B. *Forecast Horizon.*

Forecasting load with a short (less than 32 hours) horizon is beneficial when participating in the wholesale electricity markets. For horizons $h = 1, \ldots, 32$, we predicted load values for 2011. As before, we experimented with a variety of methods including SVR and linear regression. For the results presented here, we used linear regression, which outperformed the other methods based on mean squared error of daytime hours on 2010 data. As before, $d = 168$, however the exogenous variables included only the past 24 hours of the MTLF values.

Recall that MISO predicts for 8 to 32 hours ahead (midnight to midnight the following day at 4:00PM). In our evaluation process, we simulated training one model per day, updating the training set each time. For example, when $h = 2$, we train a model, use it to make $24/2 = 12$ predictions (totaling 24 predicted hours) and then retrain it using the additional day. Note that our models are trained on the original (deliberately distorted) data.

Because we are focused on the forecast horizon, and not deliberate distortions, we measure our improvements only considering the daytime hours. Figure 2 shows the performance of our forecasts relative to the MTLF. Several observations leap out.

First, note that even at as high as 30 hours, *we outperform MISO's forecasts*. This is due to the pre-forecast time of MISO's forecasting procedure. Suppose it is 4:00PM and we provide predictions for the next 30 hours. These 30 hours will include the 24 hour period that MISO is predicting (from midnight to midnight). However, for the pre-forecast hours, from 4:00PM to midnight, we have an advantage over MISO. While MISO produced these forecasts a full 24 hours ago, we are able to forecast them right now, using all of today's most recent data. For this reason, we improve on MISO's predictions even with such lengthy forecast horizons.

Second, there is a pronounced spike at $h = 24$ hours (see Figure 2). We posit that this is due to the following. With a horizon of 24 hours, we predict a full day, every day, at midnight. Because MISO made the predictions at 4:00PM on the previous day, we have an advantage. For example, while MISO predicted noon 20 hours before it occurred, we predicted it only 12 hours ahead. However, over the course of the year this advantage is constant, *i.e.,* we are always using 8 additional hours of data than MISO had.

Consider the 8 hour advantage when predicting 1 hour ahead instead of 9. As shown by Figure 2, this is a greater advantage than that seen when predicting, *e.g.,* 20 hours head instead of 12. Because the early morning hours (midnight to 6:00AM) are not counted in our analysis, the 8 hour advantage of the $h = 24$ model is largely lost.

With the other horizon times, we still have an advantage over the MTLF, but the advantage is not static. With a horizon time of $h = 23$ for example, our first set of predictions is from midnight to 11:00PM. As before, this gives us an advantage of 8 hours. However, as the prediction process continues, we sometimes obtain a very good advantage (for example predicting 1 hour ahead when MISO predicts 8 hours ahead) that we never see in the $h = 24$ model. We conjecture[5] that this syncing up is the cause of the anomaly at $h = 24$.

To account for the inherent advantage of having no pre-forecast time, we ran an additional set of experiments. Here, for a horizon time of $h = 1 \ldots 24$, we compared our results to the MTLF. For each horizon time, we also included a pre-forecast time of 8 hours. That is to say that we forecast a full $h + 8$ hours ahead in a sliding-window scheme, and drop the first 8 predictions.

The results are not as clean as in the previous case. Recall that our approach to identify and account for deliberate distortions (which is faithful to MISO's prediction procedure) uses a horizon of $h = 24$. With the inclusion of an 8-hour pre-forecast time, we only see improvement when h is very low ($h \leq 6$). Even with such a low h, we never see any improvement in terms of MED. This indicates that the pre-forecasting time is a large part of why shortening the forecast horizon produces more accurate forecasts (recall that without the pre-forecast time, the relative MSE, for example is less than 0.1 when $h = 1$).

C. Model Errors.

The ratio of our predictions' MSE to the MTLF's MSE, when considering all daytime hours of the year, is 1.06, meaning ours is roughly 6% worse than MISO's. The ratio

[5]To verify this hypothesis, we tested up to $h = 60$ and found (as expected) a similar, albeit it less pronounced, spike at $h = 48$.

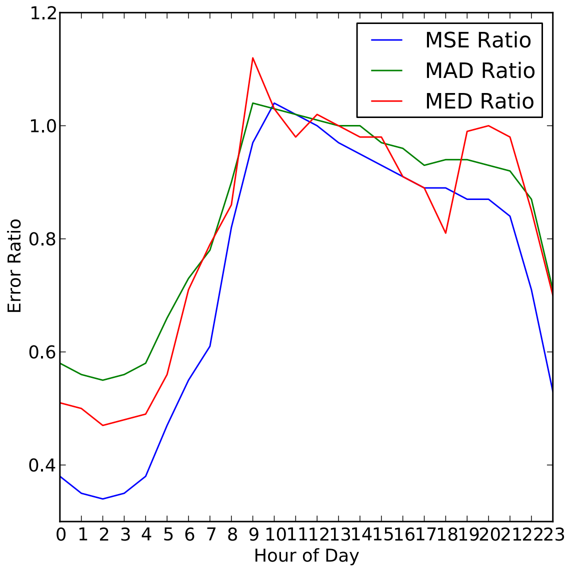


Fig. 3: **Relative error by hour of day.** The horizontal axis is the hour of day in a 0-based 24-hour clock. The vertical axis represents the ratio of our error to the MTLF error.

for mean absolute difference is 1.04 and 1.05 for median absolute difference. As discussed previously, correcting for model errors is the most difficult task. However, with targeted residual analysis we are able to pinpoint when our approach does aid MISO. As we did previously, we split the year into 24 segments and analyze errors.

Figure 3 shows our relative performance across 2011 for each hour of the day. As described above, we reduce the errors during the nighttime hours (where deliberate distortion is taking place). However, note that between 6:00AM and 7:00AM we are yielding roughly .65 the MTLF MSE. In addition, between 7:00AM and 8:00AM we yield $\approx$ .90 the error. This indicates that our system improves on the MTLF beyond accounting for deliberate distortions. Implementing our system as a direct plug-in would help during early morning daytime hours.

VIII. Conclusions and Future Work

Effective operation of the US power grid depends, among other things, on the ability to match supply with demand in real time. To accomplish this, regional power authorities forecast hourly demand on a day ahead basis. These forecasts enable power generation and transmission infrastructures to be configured to meet demand and for power brokers to estimate pricing appropriately. Accurate load forecasts depend on a range of issues – most significantly the weather – and while the details of models used by ISO's are proprietary, both projections and actual loads are publicly available.

In this paper, we examine the accuracy of power load forecasts that were made over a two year period by US Midwest ISO. The goal of our work is to understand the details of errors in forecasts and to develop practical methods for improving forecasts. Examination of differences between forecast and actual loads immediately reveals both longer term seasonal effects as well as diel patterns that suggest three models for sources of errors: deliberate distortions, extended forecast horizons and inherent model deficiencies. To investigate these models and toward our goal of developing methods that can improve forecasts, we develop learning-based methods that seek to minimize forecast error. We demonstrate our methods in the context of our error models to show where and how they can improve forecasts. Specifically we show that we can reduce the mean squared error of exisiting forecasts by over 90% by reducing the forcast horizon. In addition, by detecting and addressing the deliberate distortions of load forecasts, we reduce the mean squared error by roughly 20% over the course of the year and by over 55% for nighttime hours. It is important to note that our method can be immediately applied to load forecasts produced by MISO or other ISO's without requiring details of their models or inputs – we only require historical forecasts and load values.

While we believe that our residual-based methods can be immediately useful in practice, we also find that there are many additional opportunities to improve load forecasting. In general, the ranking and relative performance of different models depends on the residuals evaluated (*c.f.*, [9]). Further investigation using a variety of residual formulations may yield additional information useful for decision support. We used only publicly available forecast and load data and had no information about forecast models for this work. While we considered this a practical approach, access to all inputs along with concrete knowledge of the underlying model, will likely yield further improvements from our system. In addition, recall that instead of modeling actual load, we explictly model the MTLF's error. This leads to the phenomenom of a "butter zone" we discussed, where only predictions outside a certain range result in a worse final prediction. In a method similar to ϵ-insenstive support vector regression, one can incorporate knowledge of the "butter zone" into the loss function of a machine learning algorithm to improve predictions. Finally, we are in the process of considering various ways in which we can expand our model to include additional data sets directly, such as highly targeted weather forecasts. We believe that inclusion of additional input can help to make our model more sensitive to meaningful variations in actual loads that are not currently captured in forecasting models.

This work was supported in part by NSF grant CNS-0905186, ARL/ARO grant W911NF1110227 and the DHS PREDICT Project. Any opinions, findings, conclusions or other recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, ARO or DHS.

References

[1] K.H. Ahlert and C. Block. Assessing the impact of price forecast errors on the economics of distributed storage systems. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[2] S. Alfeld, C. Barford, and P. Barford. Toward an analytic framework for the electrical power grid. In *Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, page 9. ACM, 2012.

[3] ZA Bashir and ME El-Hawary. Applying wavelets to short-term load forecasting using PSO-based neural networks. *Power Systems, IEEE Transactions on*, 24(1):20–27, 2009.

[4] A. Ben-Hur and J. Weston. A user's guide to support vector machines. *Methods in Molecular Biology*, 609:223–239, 2010.

[5] Y. Bi, J. Zhao, and D. Zhang. Power load forecasting algorithm based on wavelet packet analysis. In *Power System Technology, 2004. PowerCon 2004. 2004 International Conference on*, volume 1, pages 987–990. IEEE, 2004.

[6] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[7] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis: forecasting and control*, volume 734. Wiley, 2011.

[8] C. Chatfield. Neural networks: forecasting breakthrough or passing fad? *International Journal of Forecasting*, 9(1):1–3, 1993.

[9] R.A. Clements, F.P. Schoenberg, and D. Schorlemmer. Residual analysis methods for space–time point processes with applications to earthquake forecast models in california. *The Annals of Applied Statistics*, 5(4):2549–2571, 2011.

[10] T. Czernichow, A. Piras, K. Imhof, P. Caire, Y. Jaccard, B. Dorizzi, and A. Germond. Short term electrical load forecasting with artificial neural networks. *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications*, 4:85–100, 1996.

[11] E. Delarue and W. D'haeseleer. Adaptive mixed-integer programming unit commitment strategy for determining the value of forecasting. *Applied Energy*, 85(4):171–181, 2008.

[12] H.S. Hippert, C.E. Pedreira, and R.C. Souza. Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on*, 16(1):44–55, 2001.

[13] B.F. Hobbs, S. Jitprapaikulsarn, S. Konda, V. Chankong, K.A. Loparo, and D.J. Maratukulam. Analysis of the value for unit commitment of improved load forecasts. *Power Systems, IEEE Transactions on*, 14(4):1342–1348, 1999.

[14] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[15] Midwest ISO. BPM-001 market registration. *Business Practice Manual*, 2012.

[16] Midwest ISO. BPM-005 market settlements. *Business Practice Manual*, 2012.

[17] Midwest ISO. BPM-009 market monitoring and mitigation. *Business Practice Manual*, 2012.

[18] Midwest ISO. BPM-025 forecast engineering. *Business Practice Manual*, 2012.

[19] I. Moghram and S. Rahman. Analysis and evaluation of five short-term load forecasting techniques. *Power Systems, IEEE Transactions on*, 4(4):1484–1491, 1989.

[20] MA Ortega-Vazquez and DS Kirschen. Economic impact assessment of load forecast errors considering the cost of interruptions. In *Power Engineering Society General Meeting, 2006. IEEE*, pages 8–pp. IEEE, 2006.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] A.J.R. Reis and A.P.A. da Silva. Feature extraction via multiresolution analysis for short-term load forecasting. *Power Systems, IEEE Transactions on*, 20(1):189–198, 2005.

[23] Midwest ISO Client Relations and Load Forecasting Team. Personal communication, 2012.

[24] D. Tao, W. Xiuli, and W. Xifan. A combined model of wavelet and neural network for short term load forecasting. In *Power System Technology, 2002. Proceedings. PowerCon 2002. International Conference on*, volume 4, pages 2331–2335. IEEE, 2002.

[25] P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.

[26] D.H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.