Improving forecast accuracy using a synthetic weather station

Donaldson, Daniel; Khan, Zafar; Jayaweera, Dilan

DOI:
10.1109/PTC.2019.8810869

Improving Forecast Accuracy Using a Synthetic Weather Station: An Incremental Approach and BFCom2018 Lessons Learned

Daniel L. Donaldson, Zafar A. Khan, and Dilan Jayaweera
Department of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham, UK
Email: {dld818, kzx511, d.jayaweera}@bham.ac.uk

***Abstract*—Selection of the most appropriate forecast model should be governed by the underlying data. This paper investigates the impact of benchmark model selection, recency effect and synthetic weather station selection techniques on load forecast performance and presents a new weighted average based approach to generate a synthetic weather station. Lessons learned from this effort include the criticality of using benchmark models, the need for additional public datasets, and the value of forecasting competitions for learning and model development. The results from this case study validate that addition of recency effect and use of a synthetic weather station, can provide substantial forecast improvements over benchmark models. The results also demonstrate the potential benefit of using a weighted approach for synthetic weather station generation rather than simple averaging.**

***Index Terms*—Load forecasting; Multiple linear regression; Weather station combination;**

I. Introduction

Load Forecasting is an essential component of power system planning and operations. Depending on the application, the forecast timeframe can vary from very short-term (minutes), short-term (hours to weeks), mid-term (weeks to a year), to long-term (decades). Various techniques are employed which seek to produce the best prediction of future load. Examples of such techniques include Multiple Linear Regression (MLR) [1], Exponential Smoothing [2], Artificial Neural Networks [3], semi-parametric additive models [4], and Support Vector Machines (SVM) [5]. The appropriateness of a model depends on the application. Thus, for each use case a model specific to that use case should be developed and validated prior to use. In order to implement these models in the most accurate fashion, understanding of the underlying behavior of the system is required to appropriately tune the model for the specific use case.

As industry practitioners and researchers continue to develop and test a variety of forecasting methodologies, competitions can provide a means to test a range of models on a common set of data. Hence, competitions can serve to facilitate extension of existing forecasting techniques and creation of new ones [6]. One notable example of this is the series of 'M Competitions' which have aimed to "advance the theory and practice of forecasting" over the past 45 years [7]. A recent global forecasting competition was organized by the University of North Carolina at Charlotte entitled the 2018 Big Data and Energy Analytics Forecasting Competition (BFCom2018) [8]. In this paper, the impact of recency effect and weather station selection on electricity load forecasting is investigated using the BFCom2018 qualifier as a case study.

As researchers seek to improve the performance of load forecasting models, often improvements can be made by applying concepts from other disciplines. Recency effect is a term in psychology which refers to the greater influence of recent observations on a person's memory [9]. This term was applied to load forecasting in [10] to refer to the inclusion of lagged temperature terms when forecasting electrical demand. Authors in [10] demonstrated that application of recency effect can provide improvements to forecast accuracy.

When forecasting load across a larger region, weather across the region may differ. Furthermore, the proximity of weather stations to the load of interest may vary. Therefore, combination of data from multiple weather stations may serve to improve forecast accuracy. In [11] temperature data for a fixed number of weather stations was combined and shown to provide forecast improvements. This work was extended in [12] to dynamically identify the optimal number of weather stations to combine for the best forecast. In [11] the approach of using weighted combinations of weather stations based on economic data and load were tested and determined to be less accurate. However, neither [11] or [12] evaluated a weighted combination of weather stations based on relative forecast accuracy.

Contributions from this paper include: 1) Extension of the work in [12] via the development of a new synthetic weather station generation technique based on weighted averaging which can enhance forecast accuracy and 2) Presentation of lessons learned and model development from BFCom2018 which can inform the development of future electric demand forecasting methodologies and facilitate experiential learning.

The rest of the paper is organized as follows: Section II describes the case study data. Section III presents the methodology for each model. Section IV presents the results and lessons learned. Finally, Section V provides the conclusion.

II. Data And Task Description

For the BFCom2018 qualifier, four years of hourly temperature data are provided across 28 individual weather stations as well as three years of hourly load data. The goal was to develop an ex-post load forecast for the fourth year using only the temperature and load provided and the U.S. Federal Holidays. A depiction of the data can be seen in Fig. 1.

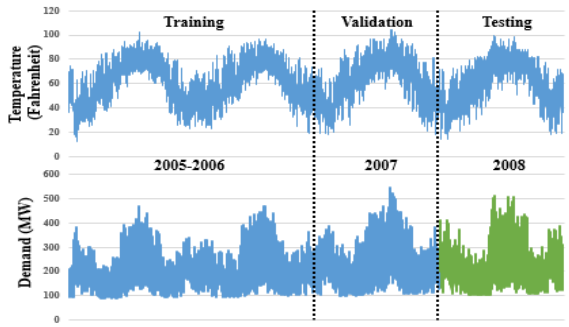

Fig. 1. Case Study Input Data; Data provided for model development is shown in blue with the holdout data shown in green. Note: Temperature data is only depicted for one of the 28 weather stations.

The data provided did not contain missing or otherwise incorrect data and therefore data cleansing techniques were not employed. However, the importance of this step should not be overlooked in data pre-processing.

III. Model Development

This study assessed the efficacy of a seasonal naïve benchmark model [13] and an initial MLR benchmark model proposed in [14]. A sensitivity analysis on each of the individual model variables within the benchmark MLR model was conducted to validate their applicability for this dataset. Then additional models accounting for holiday effect and recency effect (lagged and average temperature) variables were assessed. The best resulting model was then tested on data from 28 different weather stations to identify the best individual and combination of weather stations to use. Two approaches were applied to generate the best synthetic combination of weather stations including the approach in [12] and a new weighted approach. The final forecast was then produced and compared against the benchmark models for the testing dataset.

A. Model Selection Criteria

The error measure used to evaluate each model is the Mean Absolute Percentage Error (MAPE) of the hourly forecasts as in (1):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Actual_t - Forecast_t}{Actual_t}\right| \quad (1)$$

where n is the number of datapoints; t is the point in time; *Actual* is the historical load; and *Forecast* is the projected load.

Each model is trained using the hourly data from 2005-2006 and validated against data from 2007. Model performance is then tested against data from a hold-out period of 2008.

B. Benchmark Models

In order to assess model improvements, a benchmark or series of benchmarks should be incorporated [13]. One basic benchmark when evaluating a forecast with seasonal components is a seasonal naïve benchmark. The seasonal naïve model assumes that the value for a given hour in the forecast horizon is the same as the value for the last historical observation from the same season [13]. The equation for this model is given in (2).

$$Load_{t+h|t} = Load_t \quad (2)$$

where *Load* is the output in MW; t is the point in time; and h is the forecast horizon. For this case study, a horizon value of 8760 was used. This univariate benchmark model does not incorporate the change in explanatory variables such as temperature from one year to the next.

A benchmark model for load forecasting using temperature as an explanatory variable was developed in [14] to provide a consistent reference for testing of forecasting models. This model is further described in [15]. The equation for this benchmark multiple linear regression model (MLR-B) is given by (3).

$$\begin{aligned} Load_t = {} & \beta_0 + \beta_1 \times Trend_t + \beta_2 \times Day_t \times Hour_t + \\ & \beta_3 \times Month_t + \beta_4 \times Month_t \times T_t + \beta_5 \times Month_t \times \\ & T_t^2 + \beta_6 \times Month_t \times T_t^3 + \beta_7 \times Hour_t \times T_t + \beta_8 \times \\ & Hour_t \times T_t^2 + \beta_9 \times Hour_t \times T_t^3 \end{aligned} \quad (3)$$

where *Load* is the output in MW; t is the point in time; *Trend* is a number increasing linearly over the forecast period which seeks to account for any underlying historical trend; *Day* is a categorical variable representing the seven days of the week; *Hour* is a categorical variable representing the 24 hours of the day; *Month* is a categorical variable representing the 12 months of the year; T is a continuous variable representing the temperature in Fahrenheit; and $\beta_0, \beta_1 \ldots \beta_9$ are the regression coefficients.

C. Holiday Effect

Holidays can cause the load to deviate from established patterns. The benchmark model (MLR-B) does not include the impact of holidays, so each of the U.S. Federal Holidays [16] was represented through a separate variable and added to the benchmark model (2) to form MLR-BH. The Holiday effect is modelled as follows:

$$f(H) = \beta_{10}H1_t + \beta_{11}H2_t \ldots + \beta_{19}H10_t \quad (4)$$

where $H1_t, H2_t \ldots H10_t$ represent each of the ten U.S. Federal Holidays.

D. Model with Recency Effect

The impact of recent temperatures on the load forecast was included into the model to form MLR-BHR via evaluation of

hourly lag variables and the moving average of the temperature over the prior 24 hours as described in [10]. In order to account for the lagged temperature effect, variables in (3) which included the temperature T_t were replaced by a lag term, T_{t-h} where h corresponds to the number of hours delay.

$$f(T_{t-h}) = \beta_{20} \times Month_t \times T_{t-h} + \beta_{21} \times Month_t \times T_{t-h}^2 + \beta_{22} \times Month_t \times T_{t-h}^3 + \beta_{23} \times Hour_t \times T_{t-h} + \beta_{24} \times Hour_t \times T_{t-h}^2 + \beta_{25} \times Hour_t \times T_{t-h}^3 \quad (5)$$

The moving mean of the temperature over the previous 24 hours was also included in the model as shown in (6):

$$f(\tilde{T}_{t,d}) = \frac{1}{24} \sum_{h=24d-23}^{24d} T_{t-h} \quad (6)$$

where *t* is the time; *h* is the hour; and *d* is the day.

E. Weather Station Selection

In order to develop an improved weather station for use in the model, several weather stations were combined to form a synthetic weather station. Two approaches were tested to combine these weather stations. The first approach as described in [12], employs a simple average of the temperature across the *k* best weather stations in order to create the synthetic weather station. The algorithm is given in Table I.

TABLE I
ALGORITHM I: WEATHER STATION SELECTION: AVERAGED APPROACH

1: Fit the model developed in (3)-(6) using temperature data from each of the *n* weather stations
2: Calculate the MAPE for each model for the training period
3: Sort the weather stations based on the resulting MAPE in ascending order (from lowest to highest MAPE)
4: Combine the Temperature data for the top *k* weather stations to form a synthetic weather station by taking the average temperature across those *k* stations where *k = 1,2… n*
5: Fit model developed in (3)-(6) using the synthetic weather stations for *k = 1,2… n*
6: Calculate the MAPE for the resulting *n* models
7: Select the value of *k* which provides the lowest MAPE over the validation period

After testing the method described in Table I, an incremental approach for weather station combination was developed which uses the weighted average of the temperature across the k best weather stations. The weights are based on the relative forecast performance using each weather station. The algorithm to identify the weights for each station is given in Table II.

TABLE II
ALGORITHM II: WEATHER STATION SELECTION: WEIGHTED APPROACH

1: Fit the model developed in (3)-(6) using temperature data from each of the *n* weather stations
2: Calculate the MAPE for each model for the training period
3: Sort the weather stations based on the resulting MAPE in ascending order (from lowest to highest MAPE)
4: Calculate the MAPE for each station relative to the most accurate station using the following equation where 1 is the most accurate weather station from prior step

$$RelMAPE_i = \frac{MAPE_1}{MAPE_i}$$

5. Calculate the weights via the following:
for *k* = 1:28
  for *i* = 1:*k*

$$Weight_{i,k} = \frac{RelMAPE_i}{\sum_{j=1}^{k} RelMAPE_j}$$

  end for
end for
6: Calculate the weighted average synthetic temperature by multiplying the individual temperature series for the top *k* weather stations and their corresponding weights for *k = 1,2… n*
7. Fit the model developed in (3)-(6) using the synthetic weather stations for *k = 1,2… n*
8. Select the value of *k* which gives the lowest MAPE over the validation period

IV. RESULTS AND DISCUSSION

A. Evaluation of Benchmark Models

The first benchmark model tested was the seasonal naïve model given in (2). This model inherently assumes that load each year would be identical to load in the prior year. The resulting MAPE employing (2) with a horizon of 8760, representing the same hour from the previous year was 18.9% over the validation period. As expected, this model is inaccurate due to the inability to capture the variation in temperature from year to year.

Next, a Multiple Linear Regression benchmark model (MLR-B) using (3) was evaluated. Each variable in (3) was tested in stepwise fashion to observe the performance improvement it provides. Model 1 (M1) is an initial model reflecting only the polynomial temperature and linear trend variables. Each subsequent model (M2 through M7) uses the prior model as a base and adds an additional variable. Finally, M7 reflects the full benchmark model MLR-B. Results using the first weather station for each of the seven models can be seen in Fig. 2, which depicts the impact of each variable on forecast accuracy.

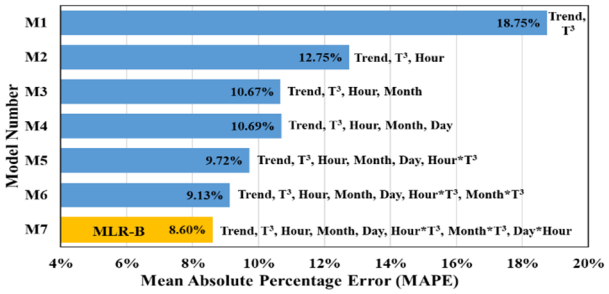

Fig. 2 Benchmark Model Testing Results over validation period

As demonstrated in Fig. 2, each variable incrementally improves the accuracy of the forecast apart from the daily variable (M4). However, when the daily variable is paired with the hourly categorical variable (M7), to model a cross effect, the model improves by 6%. Although Fig. 2 depicts only the results for the first weather station, similar analyses were performed using each of the 28 weather stations. Each result indicated a similar trend of improvements in model accuracy when going from M1 to the full MLR-B.

B. Evaluation of Holiday Effect

Although Holiday Effect was not the primary focus of this paper, a basic holiday model was included by adding the term in (4) to MLR-B. Addition of this term improved the accuracy of the forecast at the 1st weather station from 8.60 to 8.59 over the validation period. The resulting benchmark plus holiday effect model is denoted MLR-BH. However, it should be noted that across the 28 weather stations, the impact of the Holiday Effect term varied. A summary of the impact across each of the weather stations is shown in Table III.

Table III: Evaluation of the Holiday Impact Across all 28 Weather Stations

Performance	Train	Validate
# of stations where Holiday Impact Improves Accuracy	22	7
# of stations where Holiday Impact Reduces Accuracy	6	21

The results in Table III illustrate that the incorporation of the Holiday Effect does not necessarily provide improvements in model accuracy for all weather stations. By modeling each Holiday separately, this means that there are only two observations of each Holiday in the training set. Other means of considering Holidays such as treatment of the Holiday as a Sunday as proposed in [14] are possible avenues of research.

C. Evaluation of Model with Recency Effect

In order to test the improvement in accuracy from inclusion of recency effect, first, lagged temperature terms as described in (5) were added to the MLR-BH model starting with a single hourly lag and increasing the number of hourly lags until limited additional model performance gain is achieved. For the case study, up to four hourly lag terms were tested. The 24-hour average temperature of up to two days was also tested. The first three hourly lags improved the model substantially as did inclusion of the 24-hour average temperature. Addition of the fourth hourly lag only provided a nominal improvement (~0.01%) and resulted in an increase in error across some weather stations. Inclusion of the average temperature over two days made the model worse across most stations. Therefore, for this model (MLR-BHR) only three hourly lag terms and the 24-hour moving average were included. Using the first of the 28 weather stations, the subsequent performance improvements of MLR-BHR in comparison to MLR-B can be seen in Fig. 4.

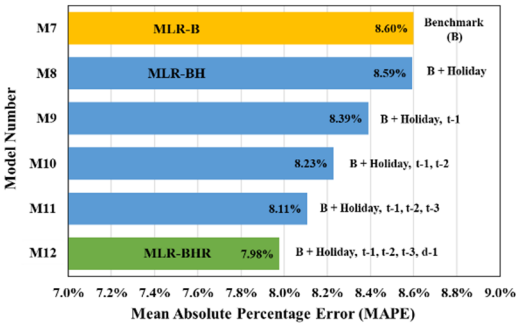

Fig. 4: Performance of Models over the Validation Period

It is notable that the only station that showed improvement with addition of the 48-hour moving average temperature was the most accurate weather station. This suggests that for other datasets, there may be benefit from additional days of averaged temperature.

D. Evaluation of Model with Synthetic Weather Station

After incorporation of recency effect, the impact of a synthetic weather station was assessed using two methods. First the method described in Table I was applied. MLR-BHR was run for each of the 28 weather stations. Fig. 5 depicts the outcome over the training and validation periods.

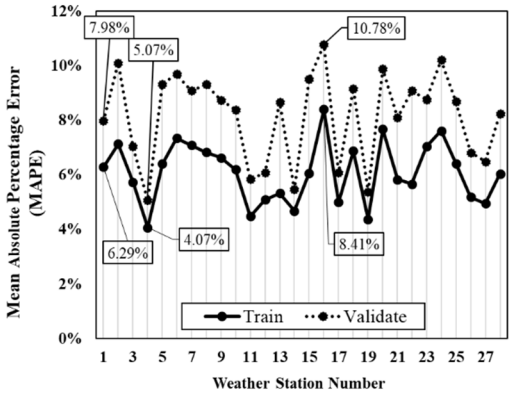

Fig. 5: Weather Station Testing Results indicating clear differences in individual weather station performance

Reliance on a single weather station can result in an inaccurate forecast if a forecaster were to choose a sub-optimal station such as Station 16 or even Station 1. This is demonstrated over the validation period as the best weather station resulted in a MAPE of 5.07% and the worst weather station resulted in a MAPE of 10.78%. The location of each weather station was unknown in BFCom2018, but for additional datasets incorporation of location and other factors might serve to further characterize the best stations. After each weather station was tested, the results were ranked in ascending order and can be seen in Table IV.

Table IV: Weather Stations Ranked by MAPE over the training Period

Rank	Station	MAPE	Rank	Station	MAPE
1	W4	4.07	15	W10	6.20
2	W19	4.38	16	W1	6.29
3	W11	4.49	17	W5	6.41
4	W14	4.69	18	W25	6.42
5	W27	4.95	19	W9	6.63
6	W17	5.00	20	W8	6.83
7	W12	5.10	21	W18	6.88
8	W26	5.18	22	W23	7.05
9	W13	5.33	23	W7	7.09
10	W22	5.67	24	W2	7.14
11	W3	5.74	25	W6	7.34
12	W21	5.83	26	W24	7.62
13	W28	6.05	27	W20	7.67
14	W15	6.07	28	W16	8.41

Next Algorithm I was employed in which the number of weather stations was varied from 1 to 28 and the temperature of these stations combined into a single synthetic weather station using a simple average. The resulting MAPE for this model (MLR-BHR-A) can be seen in Fig. 6. The optimal synthetic weather station for the training period based on this approach is created from the combination of the eight most accurate stations.

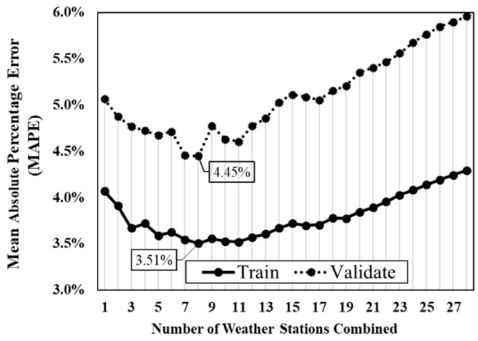

Fig. 6: Optimal Combination of Stations to use for Synthetic Weather Station Generation

A second synthetic weather station development algorithm described in Table II was also tested to identify if further improvement could be gained via weighting of stations relative to their accuracy. This model is referred to as MLR-BHR-W. The improvement from weighting can be seen in Fig.7.

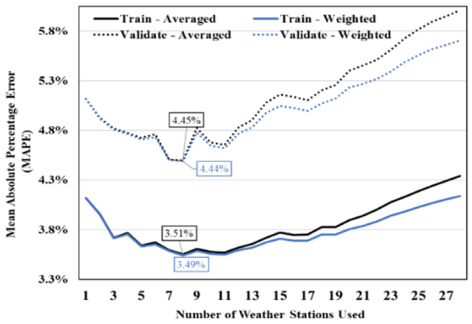

Fig. 7: Comparison of Synthetic Weather Approaches

Results indicate that the weighted approach is better than the simple average across the combinations of stations, and the gap is more significant as less accurate weather stations are included. A comparison of the weights resulting from Algorithm I (simple average) and Algorithm II (weighted average) for the eight station combination is shown in Table V.

Table V: Comparison of Weighting Techniques

Station	% of Synthetic Weather	
	MLR-BHR-A	MLR-BHR-W
W4	0.125	0.144
W19	0.125	0.134
W11	0.125	0.131
W14	0.125	0.125
W27	0.125	0.119
W17	0.125	0.117
W12	0.125	0.115
W26	0.125	0.113

The comparison of performance using MLR-BHR, the model using the simple average based synthetic weather station (MLR-BHR-A) and the model using the weighted average based synthetic weather station (MLR-BHR-W) over the training and validation period can be seen in Table VI.

Table VI: Comparison of model performance

Period	Best Single WS (MLR-BHR)	Best Averaged WS (MLR-BHR-A)	Best Weighted WS (MLR-BHR-W)
Training (2005-6)	4.07%	3.51%	3.49%
Validate (2007)	5.07%	4.45%	**4.44%**

E. Final Model Performance and Lessons Learned

After selecting the most accurate model among those tested (MLR-BHR-W), the model was trained over the entire period from 2005-2007 and the load for the year 2008 was forecast for the case study. The model was then evaluated using the full three years to forecast for the final holdout year reserved for testing. The resulting performance of the final model (MLR-BHR-W) and two other models over the test period is shown in Table VII.

Table VII: Comparison of Model Performance over Full Dataset

Period	Best Single WS (MLR-BHR)	Best Averaged WS (MLR-BHR-A)	Best Weighted WS (MLR-BHR-W)
Training (2005-7)	4.19%	3.67%	3.66%
Test (2008)	6.26%	5.85%	**5.82%**

The best model, MLR-BHR-W, resulted in an overall MAPE of 5.82%. This represents a sizable improvement over the benchmark model, and an improvement over the prior method of synthetic weather station creation via simple averaging. This demonstrates the value of the addition of recency effect and weather station combination. As a result of the experience in BFCom 2018, several other lessons were learned which are applicable to the overall forecasting process.

1) Benchmark Models Serve as a Key Reference

With the regular development of new forecasting models and approaches by academics and practitioners, the development of simple benchmark models provides a valuable starting point in evaluation of a new forecast approach. In this paper, a seasonal naïve benchmark and a MLR benchmark were used as the basis for comparison, however, the development of other standard benchmark models for load forecasting would be valuable.

2) The Impact of Recency and Weather is Dynamic

When building the models to account for recency effect and the best combination of weather stations to form a single synthetic station, it was observed that from one year to the next, the ideal number of lag terms and the ideal number of weather stations changed over the training period, to the validation period and again over the test period. This reinforces observations by researchers in [10] and [12] and points to the need to update models on a regular basis to enable the model to capture recent trends. This also highlights the difficulty in selecting the ideal number of terms without overfitting.

The alternate weighting approach presented demonstrates that there may be a potential ability to further improve performance through additional optimization of the weights.

Further work is needed to evaluate the best practices in this area and test a variety of approaches on additional datasets.

3) Public Datasets Allow for Equivalency in Comparison

One challenge with the comparison of various modelling techniques employed in literature is the lack of a similar dataset used for comparison. The use of consistent reference datasets for load forecasting such as the data provided in forecasting competitions or from sources such as the New England Independent System Operator (NEISO) for regional load data [17] or the Irish Social Science Data Archives customer trials for smart meter data [18] can be beneficial to develop and compare forecasting methods. Although several established datasets exist, the establishment of additional benchmark datasets in other regions would improve the ability to compare different approaches.

4) Competitions Provide a Venue for Experiential Learning

Forecasting competitions can serve as a stimulus to facilitate learning for both researchers and practitioners. In addition to BFCom2018, the series of Global Energy Forecasting Competitions (GEFCom 2012 and 2014) [19], [20], and the series of 'M' competitions [7] have provided a way for new techniques to be evaluated, implemented and reproduced by others. In this way, further crowdsourcing of forecast development techniques can advance the entire field.

5) *Understanding of the Data is Critical*

In order to improve the forecasting accuracy, it is important to understand the causality between forecasting variables and the series being forecasted. Each dataset may be driven by different underlying variables, and a careful examination of the data can lead to development of a more suitable model. Based on the understanding of the weather sensitive nature and the behavior of consumers gained from [10] and [12], two sets of variables, recency and synthetic weather data were added to the model. These models may be suitable for one region of the world but more work to assess their global applicability should be performed. As consumer behavior and preferences change, it will be critical to understand the implications that this has on forecasting models.

V. Conclusions

The detailed development of a forecasting model based on industry best practices can serve to produce an improved forecast and provide a valuable means of experiential learning. An incremental approach to improve the forecast accuracy of the existing models by using a weighted synthetic weather station was proposed and evaluated. The results suggest that the selection of appropriate weights for the weather stations can potentially improve the forecast accuracy. Additional testing should be done to evaluate this method on other datasets and models.

This paper demonstrates that benchmark models can serve as a good starting point for the evaluation and testing of new modeling approaches, but further refinements must be made to fine tune benchmarks for a particular region or use case. As validated in this case study, the incorporation of multiple weather stations into a synthetic station and incorporation of recency effect can provide an improvement in accuracy over established benchmark models. However, gathering accurate temperature data may be complex or expensive, and inaccurate data may limit the performance improvements. Finally, as the ideal number and combination of weather stations may change from year to year, regular analysis is required to ensure the model remains up-to-date.

VI. References

[1] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *IEEE Trans. Power Syst.*, vol. 5, no. 4, pp. 1535–1547, 1990.

[2] A. M. de Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," *J. Am. Stat. Assoc.*, vol. 106, no. 496, pp. 1513–1527, 2011.

[3] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, 2001.

[4] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 134–141, 2012.

[5] B.-J. Chen, M.-W. Chang, and C.-J. Lin, "Load Forecasting Using Support Vector Machines: A Study on EUNITE Competition 2001," *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 1821–1830, Nov. 2004.

[6] M. P. Clements and D. F. Hendry, Eds., *A Companion to Economic Forecasting*. Malden, MA, USA: Blackwell Publishing Ltd, 2004.

[7] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: Results, findings, conclusion and way forward," *Int. J. Forecast.*, vol. 34, no. 4, pp. 802–808, 2018.

[8] U. Charlotte, "BigDEAL energy forecasting competition attracting international interest | Inside UNC Charlotte | UNC Charlotte." [Online]. Available: https://inside.uncc.edu/news-features/2018-10-23/bigdeal-energy-forecasting-competition-attracting-international-interest. [Accessed: 28-Nov-2018].

[9] "recency effect," *American Psychological Association (APA) Dictionary of Psychology*. 2018.

[10] P. Wang, B. Liu, and T. Hong, "Electric load forecasting with recency effect: A big data approach," *Int. J. Forecast.*, vol. 32, no. 3, pp. 585–597, 2016.

[11] S. Lai and T. Hong, "When one size no longer fits all - Electric load forecasting with a geographic hierarchy," *SAS White Pap.*, pp. 1–14, 2013.

[12] T. Hong, P. Wang, and L. White, "Weather station selection for electric load forecasting," *Int. J. Forecast.*, vol. 31, no. 2, pp. 286–295, 2015.

[13] G. Hyndman, Rob J and Athanasopoulos, *Forecasting Principles and Practice*. OTexts, 2018.

[14] T. Hong, "Short Term Electric Load Forecasting," 2010.

[15] T. Hong, P. Wang, and H. L. Willis, "A naïve multiple linear regression benchmark for short term load forecasting," *IEEE Power Energy Soc. Gen. Meet.*, no. 2, pp. 1–6, 2011.

[16] "5 U.S.C. § 6103," 1965.

[17] ISO New England, "ISO New England - Zonal Information." [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info. [Accessed: 12-Dec-2018].

[18] Commission for Energy Regulation (CER), "CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]." Irish Social Science Data Archive. SN: 0012-00., 2012.

[19] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecast.*, vol. 30, no. 2, pp. 357–363, 2014.

[20] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond," Int. J. Forecast., vol. 32, no. 3, pp. 896–913, 2016.