Adaptive sliding window load forecasting

Foster, J., Liu, X., & McLoone, S. (2017). Adaptive sliding window load forecasting. In *2017 28th Irish Signals and Systems Conference, ISSC 2017* Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ISSC.2017.7983638

Published in:
2017 28th Irish Signals and Systems Conference, ISSC 2017

Document Version:
Peer reviewed version

Adaptive Sliding Window Load Forecasting

Judith Foster*, Xueqin Liu, Seán McLoone

Energy, Power and Intelligent Control Research Cluster, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, N. Ireland
{jfoster11, x.liu, s.mcloone}@qub.ac.uk

***Abstract*—Small-scale, renewable generation which is embedded in the distribution network is causing previously unseen fluctuations in demand. In Northern Ireland this new generation, which is not visible to, or controllable by, the system operator, is presenting major challenges for accurate load forecasting. Currently deployed load forecasting methods are struggling to cope due to the rapid growth in this new generation, and its weather dependent nature. In this paper linear load forecasting methods are investigated within a sliding window parameter updating framework, which is adopted to address the non-stationarity of the problem. Initially, models are built using historical load terms selected based on correlation analysis of recorded load data. Then, Forward Selection Regression is used to choose the most important variables from a candidate set, consisting of historical load variables and a range of weather related parameters. Model performance is evaluated on load data for the period 2015-2016. A 7-input model, with parameters updated on the basis of a 5-day sliding window of historical data, is shown to yield optimal results, with a mean absolute percentage error of 2.4%.**

Keywords—electric load forecasting; linear methods; sliding window; forward selection regression

I. Introduction

A European Directive has established a policy for member states to reduce greenhouse gas emissions, setting a target of 40% of electricity generation to come from renewable energy sources by 2020 [1]. In Northern Ireland (NI) this has motivated the local government to introduce attractive incentives for individuals to install residential, small-scale generators such as wind turbines and photovoltaic (PV) panels in order to achieve this target. As a result, the PV capacity in Northern Ireland has increased from almost 0 MW in 2010 to over 100 MW in 2017 (see Fig 1). Small-scale wind in the network is also estimated to be at the same level. This generation cannot be controlled by, and is not visible to, the Transmission System Operator (SONI). Small scale generation capacity is expected to increase above its current levels as we strive to reach the European target.

Network operators must be able to accurately forecast power demand in order to manage supply and thereby maintain grid stability. Short-term forecasting of the day-ahead demand is an important task for operators, enabling day-ahead scheduling of generation. Over-forecasting, i.e. predicting more power than is needed, results in too many generating units being started, leading to unnecessary expenditure. Under-forecasting, on the other hand, is a consequence of having a greater load than predicted. When this occurs the system operator has to purchase expensive peaking power to make up the shortfall at a cost that is much greater than the market price. Both these situations lead to sub-optimal scheduling of generation and create technical challenges for the operator with regard to frequency regulation, voltage control and level of reserve.

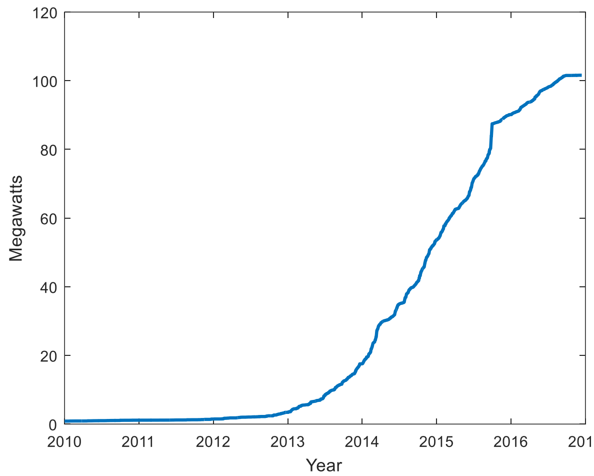

Fig 1 Growth in PV installed capacity on the Northern Ireland power network between 2010 and 2017 (sourced from Ofgem)

A. *Factors affecting the load*

Daily demand is influenced by factors such as calendar variables, weather and economic conditions. Demand peaks in the evening time after the working day is ended and reaches its lowest point at night when most people are sleeping. There are differences between weekend and weekdays with Friday differing from the other weekdays as it leads into the weekend. Holidays will similarly differ in demand to normal weekdays. Weather conditions have an impact on electricity consumption with greater demand when the weather is colder as more electricity is required to heat homes and businesses. Some of these trends are observed in Fig 2, which shows load profiles for various day types. Cost of electricity and the economic climate will also play a role in the load profile.

B. *Recent challenges to load forecasting*

Conventional sources of power are generated according to demand. However, as more and more power is generated from uncontrolled, distributed, renewable sources and used locally or

The authors gratefully acknowledge the financial support of the EPSRC, SONI and NIE networks for this research. We would like to acknowledge staff in SONI and NIE networks for sourcing data, providing secondment opportunities and valuable discussions.

fed into the grid, the character of the load is changing. The number of factors which impact the demand has increased and some of these factors are more unpredictable. The sudden drops and surges of electricity generated from renewable sources cause fluctuations in the demand due to their weather dependent nature. Cloud on a sunny day will reduce the output of PV generation and a temporary change in wind speed on a windy day will affect a wind turbine's output. Fay and Ringwood [2] found that Irish weather forecasting has the added uncertainty of predicting the shift in weather parameters as Atlantic weather fronts reach Ireland. However, the relationship between weather variables and the amount of renewable energy generated is more complex than this. For example, the temperature has been shown to be a factor in the performance of PV modules [3] while humidity and air temperature change the air density, affecting the production of wind power [4]. Further difficulties in predicting how much renewable energy is generated are due to the lack of information on the amount, exact location, orientation and surroundings of small scale generators.

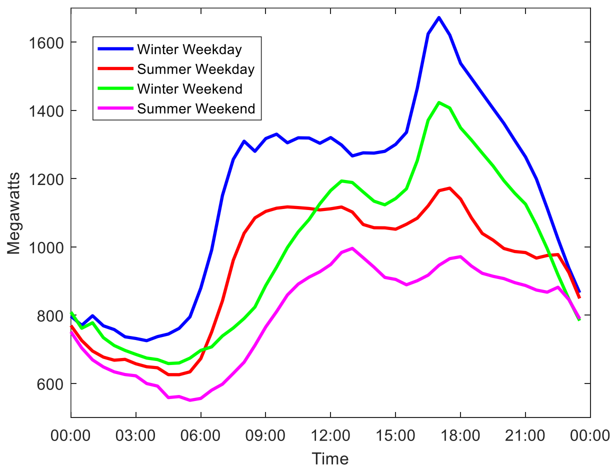

Fig 2 Demand profiles demonstrating the intra-week and intra-year differences

Continually increasing amounts of small-scale, distributed generation (shown in Fig 1) have created additional non-stationarity in the load time series. Traditional forecasting methods are beginning to struggle with this non-stationarity and better techniques must be discovered in order to accurately predict the net demand and maximise the benefits of renewable generation. In this paper, a sliding window model parameter updating methodology is proposed as a means of addressing the challenges with the non-stationarity of the data. Traditional linear models are used, a popular choice for load forecasting [5], with models inputs selected using two approaches. Initially, autoregressive load terms are included, based on a correlation analysis of historical load data. Then, a popular systematic approach to variable selection, known as forward selection regression [6] is employed to select the most appropriate variables from a candidate set consisting of historical load variables, and a range of weather and calendar related parameters.

The remainder of the paper is organised as follows. Section II provides an overview of the Northern Ireland network and data available for analysis. Section III introduces the methodology, setting out how to build models and evaluate performance. In addition, this section describes the sliding window and forward selection regression techniques. Section IV describes several forecasting models and presents results evaluating their load prediction performance. Finally, the conclusions of the study are presented in Section V.

II. System and Data Overview

Northern Ireland is one of the four countries which make up the United Kingdom. It is part of the Single Electricity Market for the whole island of Ireland, which was set up for the purpose of optimising the economic operation of the transmission network and achieving solutions to the technical challenges involved in renewable energy integration. In NI the winter demand peaks at around 1800 MW and the lowest demand can reach 500 MW. The generation capacity is approximately 2500 MW which includes the large-scale wind farms but excludes the interconnectors to Scotland and the Republic of Ireland. Installed small-scale wind and PV generation capacity is currently in excess of 200MW and continuing to grow.

A. Data description

The available data is summarised in Table 1. Thirty minute resolution historical data is available from 2010 onwards, consisting of actual demand data from SONI as well as a limited number of explanatory weather variables. A fuller set of weather variables is also available from the MET office with 9 weather parameters. The weather station used is located at Aldergrove, a central location in Northern Ireland and representative of the country.

B. Variable categories

The data available may be used as raw values or combined to create new variables. Potential explanatory variables are categorised in the following groups:

1) Historic load variables

Some historic loads are highly correlated with the current day's load as shown in Fig 3. The peaks correspond to the same type of weekdays and this strong correlation diminishes with time. The correlation begins to strengthen again in the run up to the same day the previous year. This demonstrates the usefulness of recent data, same weekday data and same season of the year data. Consequently, the previous two weeks' loads and one week either side of the same day last year are considered potential variables.

2) Calendar variables

- Day of the week
- Day of the year
- Yearly cycle

Representing the yearly cycle as a full period of a sine wave is a useful way of accounting for the fact that the start and end of the year are similar.

3) Temperature variables

Actual air, wet bulb and dew point temperature data are provided by the MET office. However, customers may not act immediately to turn on or off heating with a temperature change or a short term variation may not affect customer decisions. Therefore, other means of including the temperature are considered useful [7]. In particular, to account for potential delayed response and accumulative impact, temperature lags and temperature averages of 6, 12, 24 and 36 hours are considered as inputs.

Source	Variables	Time Period
SONI	• Actual load	01/01/10 – 13/12/16
UK MET Office (Alder-grove)	• Air temperature • Wet bulb temperature • Dew point temperature • Wind speed • Wind direction • Cloud base height • Sun duration • Visibility • Humidity	01/01/10 – 13/12/16
Ofgem Renewables and CHP register	• Installed capacity of onshore wind • Installed capacity of wind less than 50kW • Installed capacity of PV	01/01/10 – 13/12/16
Photovoltaic education network [7]	• Potential solar irradiance	Yearly cycle

Table 1 Available data

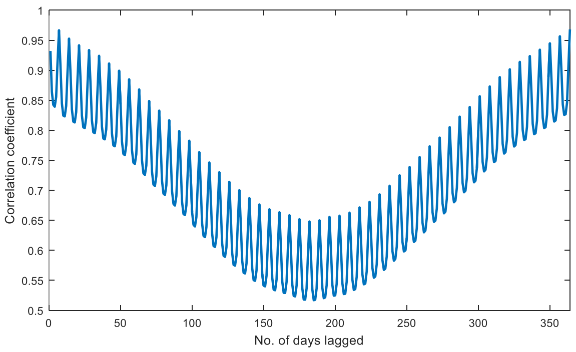

Fig 3 Correlation of the current day's load to the load for previous days up to a year ago

4) Other weather variables

The remaining weather data variables from the MET office, as listed in Table 1, are considered in their raw form. Some of these variables may have a direct impact on the load profile and some may be relevant to weather dependent generation sources.

5) Renewable contribution variables

Renewable energy generation, such as wind and PV, relies on weather conditions. Several variables have been derived from the raw data. Wind power is known to be a function of the wind speed cubed [4], therefore, the cubed and squared wind speed values are also considered as candidate variables. Solar irradiance (I_{sol}) is used in PV output calculations but as this data is not available to us, it is estimated using the potential solar irradiance (I_{pot}) multiplied by the sun duration (S_D). A further option considers the increasing level of installed PV capacity (C_{PV}) giving:

$$I_{sol} \simeq I_{pot} \times S_D \times C_{PV} \qquad (1)$$

To test if a quadratic or cubic relationship exists between potential solar irradiance and PV output, I_{pot}^2, I_{pot}^3, are also considered as candidate variables.

The total number of variables in the candidate set is 72.

III. Methodology

A. Data cleansing

Fig 1 shows how the installed PV capacity has increased dramatically throughout 2014. As the 2015-16 data will include a large amount of renewable generation it will be used as the prediction period to evaluate load forecasting models. This period, which will be denoted as ϕ, includes anomalous days which are not normal working or weekend days; therefore, data cleansing procedures are undertaken to eliminate bank holidays, days which use holidays as inputs to the forecasting model and days which use holidays to build the forecasting model for predicting standard days. The set of days from ϕ which excludes these will be denoted as ϕ_p.

Each model requires different inputs and windows of the data to build the model. Therefore, the proportion of the full dataset covered by the prediction model will be different for each model and is described as the model coverage defined as:

$$\eta_c = \frac{card(\phi_p)}{card(\phi)} \times 100\% \qquad (2)$$

where *card*(.) is the cardinality of a set.

MET office variables for sun duration and wind speed are used to identify the days with the highest renewable energy penetration over the two year period. 60 days with the most sun hours and 73 days with the highest average wind speed are identified for special attention, and denoted as ϕ_s and ϕ_w, respectively.

B. Notation

The following mathematical notation is used in building the models. The actual and predicted load (in MW) will be denoted as y and $\hat{y}$, respectively. Then, the notation y_k, y_{k-7d} and y_{k-364d} will be used to refer to the current sample instant, the value 7 days previously, and the value same day the previous year, respectively, where d represents a full day i.e. 48 samples. The sampling interval is 30 minutes.

C. Performance metric

The Mean Absolute Percentage Error (E_{MAP}), the most widely used measure of performance in load forecasting [9], is used here to evaluate the prediction capability of the different models investigated. It is defined for ϕ_p as:

$$E_{MAP} = \frac{1}{card(\phi_p)} \sum_{k \in \phi_p} \left| \frac{y_k - \hat{y}_k}{y_k} \right| \times 100\% \tag{3}$$

D. Least squares technique

Given a dataset of N samples of a target variable and explanatory variables, a best fit model to the data can be obtained by expressing the problem in matrix form and solving the equation using the least squares technique. Defining,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \tag{4}$$

where, $\mathbf{y}$ is the vector of target load values and $\mathbf{X}$ is the matrix of explanatory variables corresponding to the samples, then $\mathbf{y} = \mathbf{X\theta}$ is the regression model, and the pseudo-inverse of $\mathbf{X}$ multiplied by $\mathbf{y}$:

$$\boldsymbol{\theta}^* = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y} \tag{5}$$

is the least squares solution, i.e. the model that yields:

$$\min_{\theta} \| \mathbf{y} - \hat{\mathbf{y}} \|^2 \tag{6}$$

E. Offline v Sliding Window Methodology

Models trained offline use all the available data, splitting it into training and test sections. A training set of 30% of the dataset selected at random is used to build the model, with model performance scored using the remaining 70% of the data, the test dataset. Monte Carlo simulations are performed to enable statistically robust results to be obtained. Here, presented results are for the average performance over 100 Monte Carlo simulations. In the proposed sliding window methodology, only the most recent data is used to build the model. Model parameters are constantly changing as new historic data becomes available. The diagram in Fig 4 demonstrates the training and test windows for the offline model compared to the sliding window version. One parameter which must be determined for this technique is the optimal number of samples or length of sliding window to use to build the model.

F. Forward Selection Regression (FSR)

Due to the large number of candidate explanatory variables, determining the optimum subset is extremely challenging. Heuristic, approaches based on correlation analysis are not guaranteed to yield optimal results. A more systematic method is necessary. One such method is Forward Selection Regression (FSR) [6], a process of building the model one variable at a time, starting with no variables. Each variable in the set is tested to determine which single one will give the most accurate forecast or the smallest E_{MAP} overall. This variable will be then permanently included in the model. The next step is to test each of the remaining variables in turn to determine which one, combined with the first, improves the performance of the model most. This variable is then added to the model and the process is repeated until the addition of variables no longer improves performance. By design, this method avoids the unnecessary inclusion of two similar variables in the model. The selection process is performed using the 2015-16 dataset with holiday samples or samples affected by historic load variables being holidays, excluded. The portion of data remaining was 56% of the two year period.

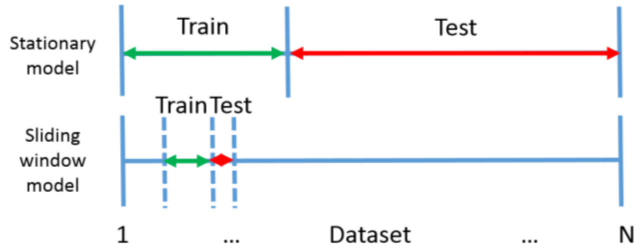

Fig 4 The stationary model has fixed training and testing sections for a dataset containing N samples. The window of data used to build the model for the sliding window model is constantly changing as new historic data becomes available.

IV. Forecasting Models and Results

1) Correlation Analysis Historic Loads Offline Model (CA-HL-O)

The measure of closeness of the relationship between the current load and historic load variables was seen in Fig 3. This suggests a simple model composed of a combination of load variables. Various combinations of the strongly correlated historic loads are tested as input variables to the offline model and the results are presented in Table 2. Overall, the best model in this set is a 5-regessor model using the load from the same time of the day from the previous two weeks and three weeks around the same time the previous year. A 3-regessor model marginally out-performs this model for the sunny day dataset, and is only marginally inferior for the full and windy day datasets. This may be the preferred option if parsimony is a priority.

2) Correlation Analysis Historic Loads Sliding Window Model (CA-HL-SW)

Selecting the optimal subset of data on which to build the model is done by testing a range of window lengths. Forecasting performance for 1 to 30 day window lengths are evaluated as shown in Fig 5. It can be seen that from 1 to 5 days the E_{MAP} reduces by over 0.5%. After this, the improvement is much less significant. Increasing the window of data also increases the probability of the model being built on abnormal days hence, the model coverage decreases, falling from 48% for 5 days to 35% for six days. Consequently, as a compromise between these two competing criteria, 5 days is chosen as the training window size.

Given the results in Table 2, the model which yields the best predictions for the full dataset and for windy days and the model that gives the best performance for sunny days will be investigated within the sliding window parameter updating framework. Table 3 presents the results of this investigation. Improvement in performance is demonstrated overall and for the specially selected day types in comparison to the fixed parameter offline estimated models.

Input parameters	All (ϕ_p) EMAP (%)	Sunny (ϕ_s) EMAP (%)	Windy (ϕ_w) EMAP (%)
y_{k-7d}	3.75	5.61	4.36
y_{k-7d}, y_{k-14d}	3.53	5.31	4.07
y_{k-7d}, y_{k-364d}	3.09	4.56	3.56
y_{k-7d}, y_{k-357d}, y_{k-364d}	2.99	4.32	3.58
y_{k-7d}, y_{k-357d}, y_{k-364d}, y_{k-371d}	2.98	4.34	3.48
y_{k-7d}, y_{k-14d}, y_{k-364d}	3.02	4.62	3.48
y_{k-7d}, y_{k-14d}, y_{k-357d}, y_{k-364d}	2.92	4.35	3.47
y_{k-7d}, y_{k-14d}, y_{k-357d}, y_{k-364d}, y_{k-371d}	2.91	4.39	3.35

Table 2 Forecasting performance of the CA-HL-OL model for different combinations of inputs

Input parameters	All (ϕ_p) EMAP (%)	Sunny (ϕ_s) EMAP (%)	Windy (ϕ_w) EMAP (%)	ηc (%)
y_{k-7d}, y_{k-357d}, y_{k-364d}	2.58	2.83	2.89	57
y_{k-7d}, y_{k-14d}, y_{k-357d}, y_{k-364d}, y_{k-371d}	2.48	3.25	2.71	48

Table 3 Forecasting performance for the CA-HL-SW model

3) Forward Selection Regression Historic Loads and Weather Offline Model (FSR-HLW-OL)

Forward Selection Regression ranks the variables in order of importance with regard to improving the prediction accuracy of the forecasting model. Table 4 provides a list for the top 20 variables selected by this process. As expected, the top variables are the historic load variables. Mean temperature over the past day is the first non-load variable included followed by sun duration and wind speed. These weather variables not only impact the load directly but affect the weather dependent generation. One of the renewable contribution variables, derived to account for the PV component features in 9th place.

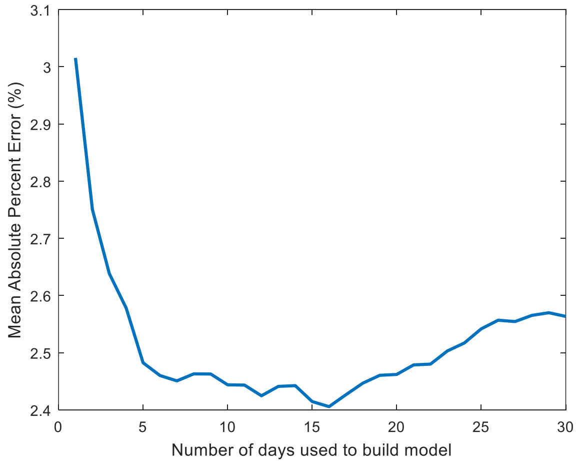

Fig 5 The effect of increasing the number of days used to build the CA-HL-SW model on the EMAP

Rank	Variable	Rank	Variable
1	-7 days	11	Air temp lagged by 6hrs
2	-364 days	12	Humidity
3	-357 days	13	Average wet bulb temp over 24hrs
4	Average air temp over 24hrs	14	-2 days
5	Sun duration	15	-359 days
6	Wind speed	16	-9 days
7	Potential solar irradiance	17	Air temp lagged by 24hrs
8	Yearly cycle	18	Average dew point temp over 36hrs
9	Sun duration × potential solar irradiance	19	Average air temp over 36hrs
10	-14 days	20	Average dew point temp over 24hrs

Table 4 Top 20 variables ranked by FSR

Beginning with the top selected variable and introducing each of the successive rankings in turn, the forecasting performance for the offline model with FSR is shown in Fig 6. Overall performance continues to improve with each additional variable. From this graph, it is difficult to determine the optimum number of variables to use to balance the prediction accuracy with model complexity, but a significant inflection point is notable with 15 variables.

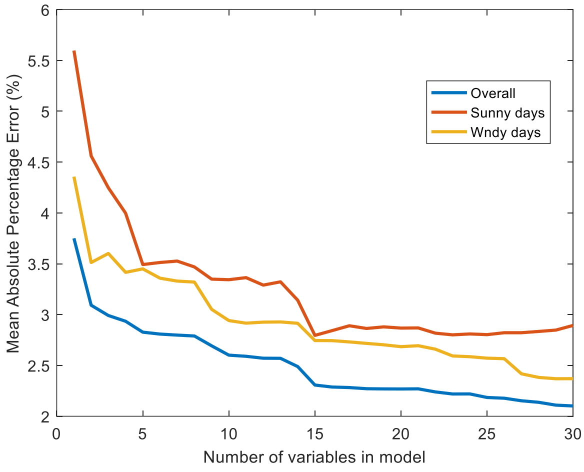

Fig 6 Forecasting performance for the FSR-HL-OL model with an increasing number of model inputs

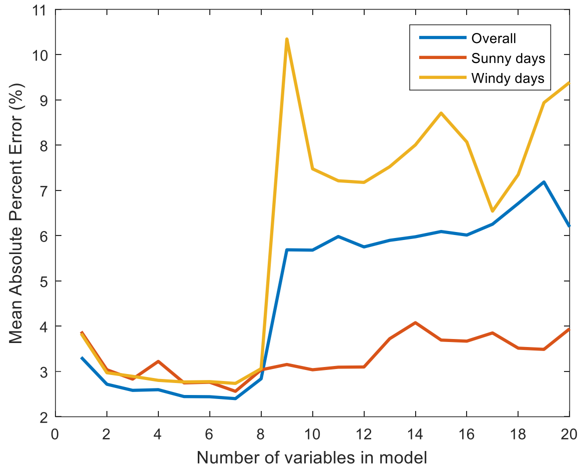

Fig 7 Forecasting performance for the FSR-HLW-SW model with an increasing number of model inputs

4) Forward Selection Regression Historic Loads Sliding Window Model

In a similar fashion the FSR ranked variables are evaluated for the 5-day sliding window prediction models. The results obtained are reported in Fig 7 and clearly show that for the sliding window regime 7 FSR selected variables are optimal.

The best models of each type considered are compared in Table 5. Here, in order to provide a fair comparison 7 FSR selected variables are chosen for both the offline and sliding window FSR models. As can be seen, the sliding window based models are consistently superior to the corresponding offline models.

Model	All (ϕ_p) EMAP (%)	Sunny (ϕ_s) EMAP (%)	Windy (ϕ_w) EMAP (%)	ηc (%)
CA-HL-OL	2.91	4.39	3.35	63
CA-HL-SW	2.48	3.25	2.71	48
FSR-HLW-OL	2.80	3.53	3.33	64
FSR-HLW-SW	2.40	2.56	2.74	57

Table 5 Forecasting performance for each model

V. Discussion and Conclusions

Static offline estimated prediction models are confirmed to be less accurate than those estimated online using a sliding window approach. For both the Correlation Analysis (CA) and FSR regressor selection techniques the EMAP is improved by almost 15%. However, an advantage of the offline models is that they can be used to predict a greater portion of the year.

The introduction of average air temperature, sun duration, wind speed and potential solar irradiance, as selected by FSR, yields a 3% improvement in performance over historic loads only models.

Sunny and windy days prove the most difficult to predict. They are also the days that benefit most from adopting sliding window models, with prediction accuracy improving by more than 25% and 18%, respectively, compared to the static offline models.

In conclusion, with increasing levels of distributed generation contributing to non-stationarity in the load time series, new forecasting approaches are required to overcome the challenges of accurately predicting demand. In this paper a sliding window method is introduced which continuously adapts model parameters to reflect the changing patterns in the load. This approach works well overall and in particular for days where there are high levels of small-scale generation.

VI. References

[1] "Directive 2009/28/EC of the European Parliament and of the Council," European Parliament, 2009.

[2] D. Fay and J. Ringwood, "On the influence of weather forecast errors in short-term load forecasting models," *IEEE Transactions on Power Systems*, vol. 25, no. 3, pp. 1751-1758, 2010.

[3] S. Dubey, J. N. Sarvaiya and B. Seshadri, "Temperature dependent photovoltaic (PV) efficiency and its effect on PV production in the world - a review," *Energy Procedia*, vol. 33, pp. 311-321, 2013.

[4] Z. Sen, "Wind power variations under humid and arid meteorological conditions," *Energy Conversion and Management*, vol. 75, pp. 517-522, 2013.

[5] E. Feinberg, "Load Forecasting," in *Applied Mathematics for Restructured Electric Power Systems*, New York, Springer US, 2012, pp. 269-285.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[7] T. Hong, *Short term load forecasting*, Raleigh: North Carolina State University, 2010.

[8] C. Honsberg and S. Bowden, "Calculation of solar insolation," National Science Foundation, 2013. [Online]. Available: www.pveducation.org/pvcdrom/calculation-of-solar-insolation. [Accessed 22 December 2016].

[9] E. Almeshaiei and H. Soltan, "A methodology for electric power load forecasting," *Alexandria Engineering Journal*, vol. 50(2), pp. 137-144, 2011.