International Journal of Forecasting 30 (2014) 246–256

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A new error measure for forecasts of household-level, high resolution electrical energy consumption

Stephen Haben [a,c,*], Jonathan Ward [b], Danica Vukadinovic Greetham [a], Colin Singleton [a], Peter Grindrod [a,c]

[a] University of Reading, UK
[b] University of Leeds, UK
[c] University of Oxford, UK

ARTICLE INFO

Keywords:
Verification methods
Load forecasting
Volatile data
Smart meter
Error measure

ABSTRACT

As low carbon technologies become more pervasive, distribution network operators are looking to support the expected changes in the demands on the low voltage networks through the smarter control of storage devices. Accurate forecasts of demand at the individual household-level, or of small aggregations of households, can improve the peak demand reduction brought about through such devices by helping to plan the most appropriate charging and discharging cycles. However, before such methods can be developed, validation measures which can assess the accuracy and usefulness of forecasts of the volatile and noisy household-level demand are required. In this paper we introduce a new forecast verification error measure that reduces the so-called "double penalty" effect, incurred by forecasts whose features are displaced in space or time, compared to traditional point-wise metrics, such as the Mean Absolute Error, and p-norms in general. The measure that we propose is based on finding a restricted permutation of the original forecast that minimises the point-wise error, according to a given metric. We illustrate the advantages of our error measure using half-hourly domestic household electrical energy usage data recorded by smart meters, and discuss the effect of the permutation restriction.

1. Introduction

As many countries progress towards a low carbon economy, the increased penetration of low-carbon technologies (LCTs) may produce new risks to the security and robustness of the electricity networks (Combrink & Vaessen, 2006). The decarbonisation of transport and heating (for instance, through the uptake of electric vehicles and heat pumps) is likely to increase the network demand, whilst household microgeneration increases the prospect of a two-way flow of electricity on the network, as consumers become suppliers and feed back into the grid. In short, electricity demand is likely to increase and become more unstable, particularly at the low voltage (LV) level (Combrink & Vaessen, 2006).

In response to these new challenges, the UK government is aiming to help network operators and suppliers prepare for a low carbon economy through initiatives such as the £500m low carbon network fund (LCNF) (Ofgem, 2012), and the roll-out of smart meters to every home in the UK by 2020 (National Grid, 2012). Smart meters are advanced energy meters with a two-way communication capability which record high resolution (typically half-hourly) energy consumption. These detailed patterns of energy demand provide opportunities to improve our understanding of energy consumption habits, to design smarter interventions for energy reductions, and to produce accurate forecasts of energy demand at the LV level. Such accurate forecasts at the level of households, or small aggregations of households, can help distribution network

* Correspondence to: Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Road, Oxford, OX2 6GG, UK. Tel.: +44 01865 611 511.
E-mail address: haben@maths.ox.ac.uk (S. Haben).

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.08.002

operators improve their management and planning of the LV networks. Forecasts can also be combined with network storage devices to improve peak demand reduction. As part of the New Thames Valley Vision[1] LCNF project, storage devices are being considered to help alleviate the high demand on the LV network at peak times. Simple set point control is the simplest and most common way of controlling battery storage, but often fails to reduce the peak demand (Thomas, 2010). However, accurate household-level forecasts could optimise the use of the battery by helping to plan the appropriate charging and discharging of the storage device (Molderink, Bakker, Bosman, Hurink & Smit, 2010; Xu, Xie, & Singh, 2010). Until recently, the majority of load forecasting has been at the medium voltage (MV) to high voltage (HV) substation levels, where the demand is relatively smooth and more regular (for instance, see the review papers by Alfares & Nazeeruddin, 2002; Moghram & Rahman, 1989; Taylor & Espasa, 2008). However, at the LV network to household level, the demand is volatile and noisy, and typically consists of many different types of behaviour, such as frequent but irregular peaks (Brabec, Konár, Pelikán, & Malý, 2008). Hence, forecasting methods developed for the MV and HV levels may not be appropriate for the household level. In order to produce and test the accuracy of household-level forecast demands, appropriate forecast verification methods are required.

Forecast verification hinges on the ability of quantitative measures to assess the similarities between forecasts and observations, what Murphy (1993) refers to as forecast *quality*. Hence, measure-orientated approaches based on point-wise comparisons, such as the mean absolute error (MAE) and root mean square error (RMSE), can often lead to spurious conclusions, see Brooks and Doswell III (1996); Castati et al. (2008), and Hoffman, Liu, Louis, and Grassotti (1995). In particular, an observed feature that is forecasted accurately in terms of size and amplitude, but displaced in time, incurs a "double penalty" (Keil & Craig, 2009). Thus, as we illustrate in this paper, it can be difficult for skilled, plausible forecasts to out-perform even a flat forecast that is of almost no informative value, particularly when the data are volatile and noisy. This problem has long been understood in the meteorology community. Consequently, a large number of alternative verification strategies have been proposed; see Castati et al. (2008) for a review. The class of distribution-oriented approaches (Brooks & Doswell III, 1996; Murphy & Winkler, 1987) offers many insights but requires large quantities of data and is computationally intensive (Brooks & Doswell III, 1996).

One approach to the calculation of displacement errors, which was also pioneered in meteorology, has been to formulate errors using an optimal distortion of the original field, i.e., *smooth* changes in position and amplitude that minimise the misfit between the data and the forecasts (Hoffman et al., 1995). Although such verification methods have been developed widely, they have limited appeal in the setting in which we are interested primarily—volatile, noisy and irregular data. In this case, it may be more appropriate to use verification measures that deform the forecast *discontinuously*. To some extent, such techniques are employed in 'fuzzy' verification techniques for high-resolution weather forecasting (Ebert, 2008). These typically compare the average states of 'events' occurring within a neighbourhood of interest. For real-valued variables, such as the amount of rainfall or the wind intensity, events are defined relative to some threshold. In essence, these methods produce new fields for both the observed and forecasted data, which are then compared using a traditional point-wise metric. Such measures are both scale and threshold dependent, and thus, one must consider a matrix of errors that captures both of these variations.

Many algorithms and metrics have been developed for measuring the similarity of time series, such as Dynamic Time Warping (DTW), longest common subsequence, edit distance on real sequences, and edit distance with real penalty (Chen & Ng, 2004). Often, these algorithms are applied in information retrieval and data mining techniques in order to measure the cost of morphing one time series into another. Dynamic Time Warping is one of the most popular techniques for measuring time series similarity, and has been used successfully in automatic speech recognition algorithms (Muller, 2007). DTW measures the differences between sequences which may vary in time or speed by stretching the time series through the duplication of local points. The difference in the deformed time series is then calculated using a standard L_p metric. A more recent method, called the Move-Split-Merge (MSM) metric, is similar to DTW, except that duplicated and deleted values incur a fixed cost (Stefan, Athitsos, & Das, 2013). For time series matching methods, although suitable for comparing series with the same (but perhaps stretched) shape in time, they are biased toward preserving ordering, and therefore are not flexible enough, in the context of energy demand, to cope with the natural irregularities in household energy usage behaviour. In addition, DTW and MSM will tend to underestimate the costs of repeated peaks by simply merging/duplicating the local peaks, with little or no penalty incurred for the inaccurate repetition. The additional complications and restrictions introduced by such techniques make them unsuitable for measuring the errors of household-level forecasts. This motivates the development of a new forecast error measure, which is the topic of this paper.

Before sophisticated forecasting techniques for household electrical energy usage can be developed, we need to be able to assess their veracity against data quantitatively. However, in this paper we illustrate the fact that the capricious nature of energy usage means that traditional point-wise measure-oriented approaches perform poorly at this task. Our main contribution is to suggest a new approach that allows for some flexibility in the timing of the forecast when computing the error, while retaining some simplicity. Specifically, for each forecast we define the error to be the minimum error (with respect to an appropriate norm) over the set of all restricted spatial/temporal permutations of the forecast. We begin in Section 2 with a formal description of point-wise error measures, particularly the p-norm,

[1] http://www.thamesvalleyvision.co.uk/.

then introduce the "adjusted error" and illustrate its advantages using a simple, synthetic example. In Section 3, we use our new measure to assess the accuracy of a hierarchy of daily forecasts of half-hourly electrical usages, taken from individual household smart meter data. In Section 4, we present a detailed discussion of the effect of the 'adjustment limit', i.e. the maximum allowed permutation displacement. Finally, we draw conclusions and discuss the advantages and disadvantages of our method in Section 5.

2. Measuring errors

2.1. Standard error estimates: the p-norm error

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and $\mathbf{f} = (f_1, f_2, \ldots, f_n)^T$ be the actual and forecasted data vectors respectively, such that each f_i is a prediction of the actual data x_i for $i = 1, \ldots, n$. We focus on one-dimensional data (i.e., time series), but the methods that we describe can be generalised to higher dimensions. Error measures can be described in terms of a vector function

$$E = F(\mathbf{f}, \mathbf{x}), \tag{1}$$

where $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is some metric. In this paper we focus on the absolute p-norm,

$$E_p = \|\mathbf{f} - \mathbf{x}\|_p = \left(\sum_{i=1}^{n} |f_i - x_i|^p \right)^{1/p}, \tag{2}$$

for some $p \geq 1$ (see Golub & Loan, 1996, p. 52). For example, this type of error includes the Mean Absolute Error (MAE) and the root mean square error (RMSE), which are simply constant multiples of the 1-norm and 2-norm errors, respectively.

2.2. The adjusted error

In order to manage and plan the LV networks properly, distribution network operators (DNOs) require an accurate knowledge of the network peak demand. Accurate household-level demand forecasts will help DNOs to understand when and how often network constraints are violated. In addition, the forecasts can be used in conjunction with the smart control algorithms of storage devices to reduce peak demands by helping to create a plan for the battery to charge at times of low demand and then discharge at around the time of the expected peak (Molderink et al., 2010; Rowe, Holderbaum, Potter, & Liu, 2012). Hence, for these applications it is more important that peaks be predicted at approximately the correct times, rather than not at all. However, as was stated in Section 1, such forecasts incur a double penalty from point-wise error measures, and may be judged incorrectly as poor forecasts. This leads to the idea that the error measure should allow for small, possibly discontinuous, displacements of the forecast values in time. We note that there exist many perfect matchings between the forecast values and actuals. Each match can be described by a permutation matrix P. To restrict the magnitude of the displacements of the forecast values, we impose an 'adjustment limit', denoted $w \geq 0$, on the permutations such that $P_{ij} = 0$ for $|i - j| > w$. We define the *adjusted error* as the solution to the minimisation

$$E^w = \min_{P \in \mathcal{P}} F(P\mathbf{f}, \mathbf{x}), \tag{3}$$

for the given metric F, where $\mathcal{P}$ is the complete set of restricted permutations. The *adjusted p-norm error* is then

$$E_p^w = \min_{P \in \mathcal{P}} \|P\mathbf{f} - \mathbf{x}\|_p. \tag{4}$$

The adjusted error is a *semimetric*, not a metric, since in general it does not obey the triangle inequality. The error minimisation is a variant of the assignment problem, a well-known combinatorial optimisation problem that can be solved in polynomial time (Munkres, 1957) using the 'Hungarian method', details of which are provided by Schrijver (2002). To incorporate the adjustment limit into the algorithm, if $|i - j| > w$, we set $|f_i - x_j|^p = \Omega$, where Ω is a large constant that effectively prevents such permutations. The method's time complexity is $O(n(m + n \log n))$ (Tomizawa, 1971), where m is the number of potential error matches, n^2. We note that this method is related to but distinct from the use of the Hungarian algorithm in Monge Type problems (such as the Earth Mover's distance), which redistributes the cumulative mass (Levina & Bickel, 2001). The adjusted error in Eq. (4) does not subdivide or combine separate predictions, but merely reorders them.

The adjustment limit w is a time-scale parameter that is problem dependent and has an important effect on the efficacy of our verification method. If $w = 0$, then we recover the original p-norm error in Eq. (2). Increasing w reduces the adjusted error, but a small error resulting from large displacements is not necessarily indicative of a good forecast. Thus, the mean displacement, which can be obtained from the permutation matrix P, is an additional measure of accuracy that can be used to compare different forecasts. We discuss these points in detail in Section 4.

2.3. A simple example

In this subsection we compare four qualitatively different forecasts of a simple energy load profile using the absolute and adjusted p-norm errors. The synthetic data, illustrated with solid black lines in each panel of Fig. 1, consists of a single peak centred around $t = 5$, with a constant background usage over a 20 time-point domain.

The hypothetical forecasts, illustrated with dashed lines, consist of a flat forecast (F1) (corresponding to the average usage) and a single peak centred around three different times (F2–F4), with the correct background usage. In the context of using the forecasts to reduce the peak demand via a storage device, F2 is a very good forecast, F3 is reasonable, and both F1 and F4 are poor. Planning the control of a storage device using the F2 forecast will enable a large reduction in peak demand, and F3 should still facilitate moderate peak load shedding, due to the expectation of a peak at approximately the correct time. However, F1 and F4 would provide no peak load shedding due to the inaccuracy in forecasting the peak demand. The absolute and adjusted p-norm errors, for $p = 4$, of each of the forecasts illustrated in Fig. 1 are presented in Table 1.

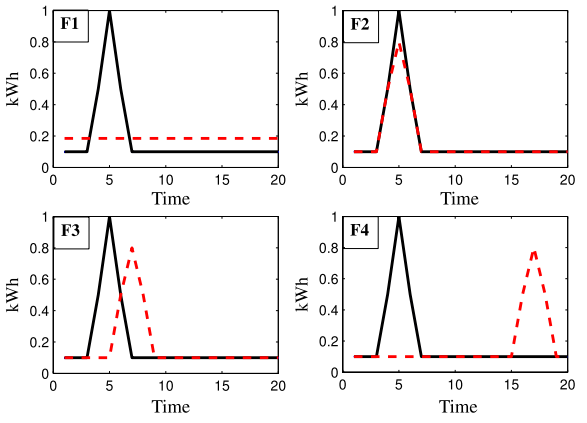

Fig. 1. Four 'forecasts', F1, F2, F3 and F4 (dashed lines), together with the actual data (solid lines), for a simplified example.

Table 1
Comparison of the error measurements given by the different norms for the four different forecasts, F1–F4, described in the main text.

Error	Forecast			
	F1	F2	F3	F4
Absolute error	0.82	0.20	0.99	1.00
Adjusted error ($w = 1$)	0.82	0.20	0.79	1.00
Adjusted error ($w = 2$)	0.82	0.20	0.48	1.00
Adjusted error ($w = 3$)	0.82	0.20	0.20	1.00

We have used the 4-norm rather than the more common 2-norm because we want to penalise large errors (i.e., missed peaks) much more than small errors. Different values of p yield qualitatively similar results. Table 1 illustrates the following:

- *Absolute* 4*-norm error.* The good forecast, F2, has the smallest error, while the flat forecast, F1, has a smaller error than either the poor forecast, F4, or the reasonable forecast, F3. This illustrates the double penalty effect which is present in point-wise error measures.
- *Adjusted* 4*-norm error*, $w = 1$ *and* $w = 2$. The reasonable (F3) forecast error is reduced to about 95% and 58% of the flat (F1) forecast error for the adjustment limits $w = 1$ and $w = 2$ respectively. The F1, F2 and F4 forecast errors are the same for both the adjusted and absolute measures—displacing the forecast values does not change the errors.
- *Adjusted* 4*-norm error*, $w = 3$. The good (F2) and reasonable (F3) forecast errors are equal. However, we can still distinguish F2 as being the better forecast with this method by considering the mean displacement. F2 has a zero mean displacement of the forecast values (implying that the minimum permutation is achieved by the forecast), whereas F3 has a mean displacement of 0.6 grid points over the 20 forecasted values.

In summary, the synthetic example illustrates that the adjusted p-norm error can give a more accurate representation of the forecast usefulness than the standard p-norm error.

3. Application to household energy load forecasting

As was shown in the previous section, standard point-wise measures may not be adequate for assessing the accuracy of a forecast. Although many forecast methods have been developed and calibrated for smoother higher voltage demands (see for instance the review paper by Alfares & Nazeeruddin, 2002), their accuracies when applied to household- or LV-level demand cannot be assessed until an appropriate error measure has been established. Once suitable benchmarks have been developed, both old and new forecasting methods can be tested and compared, and other techniques, such as clustering, can be applied to improve the forecasts. In this section, we consider the standard and adjusted 4-norm errors in order to compare the performances of three simple forecasting methods applied to half-hourly domestic household electrical energy usage data. The data were collected by household smart meters as part of the Ofgem-managed Energy Demand Research Project (EDRP) trial run by Scottish and Southern Energy (SSE).[2] A wide variety of energy usage behaviours are observed between households, and the individual household demand is both volatile and noisy. However, there are daily, weekly and seasonal patterns that could potentially be exploited by forecasting methods. Such forecasts can have a positive impact on network operations and planning.

3.1. An example with three households

Fig. 2(a)–(c) illustrate a week's worth of half-hourly electrical energy usage profiles, in kilowatt-hours (kWh), for three representative UK households. Household A consumes most of their energy during one or two peak periods at regular daily intervals. Thus, we would hope to be able to forecast their usage fairly accurately. Household B has irregular peak demands that are smaller than those of the other households, but they maintain a fairly constant background usage. Household C is the most volatile, having large irregular peak demands and periods of low usage. We would expect this household's energy usage to be difficult to forecast. The average daily energy usages for households A, B and C are 5.51 kWh, 9.89 kWh and 18.12 kWh, respectively.

Our household energy usage dataset consists of 10 weeks of half-hourly kWh records (3360 in total) for each of the three households. Each forecast generates an unsupervised rolling daily prediction from midnight over the course of the 10th week, with access to the full data-history of each household separately. Our aim is to assess the validation techniques, and as a consequence, the forecast methods that we implement are chosen to form a clear hierarchy. The three methods by which each of the daily forecasts are generated are as follows:

1. *Flat* forecast: The average usage over the previous 7 days, used as the forecast for all time periods.

[2] See http://www.ofgem.gov.uk/sustainability/edrp/Pages/EDRP.aspx for further details.

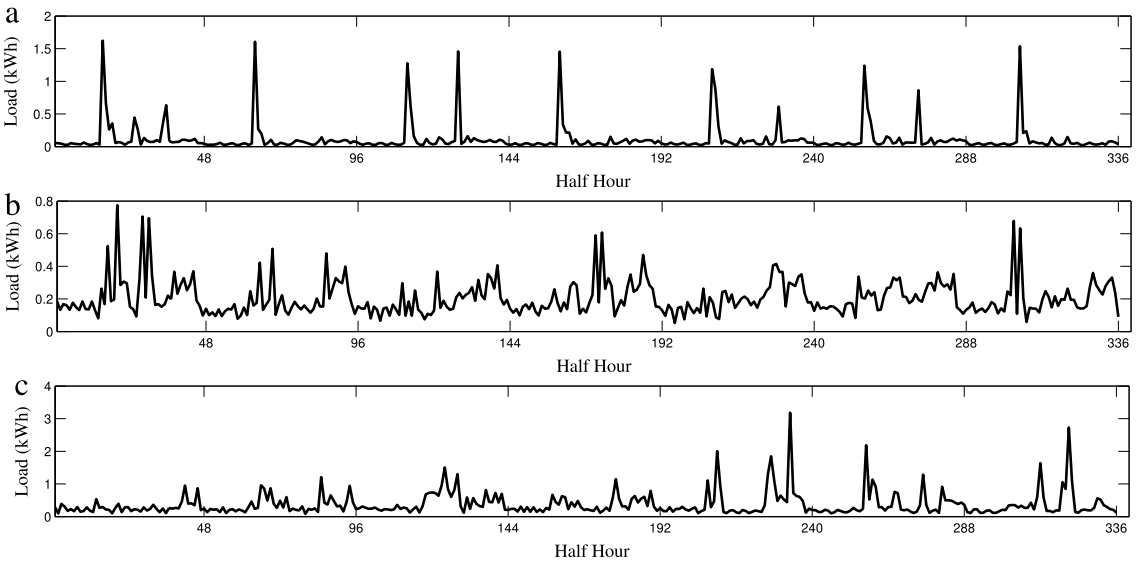

Fig. 2. Example of half-hourly smart meter electrical energy usage (in kWh) for (a) household A, (b) household B and (c) household C, as described in the main text.

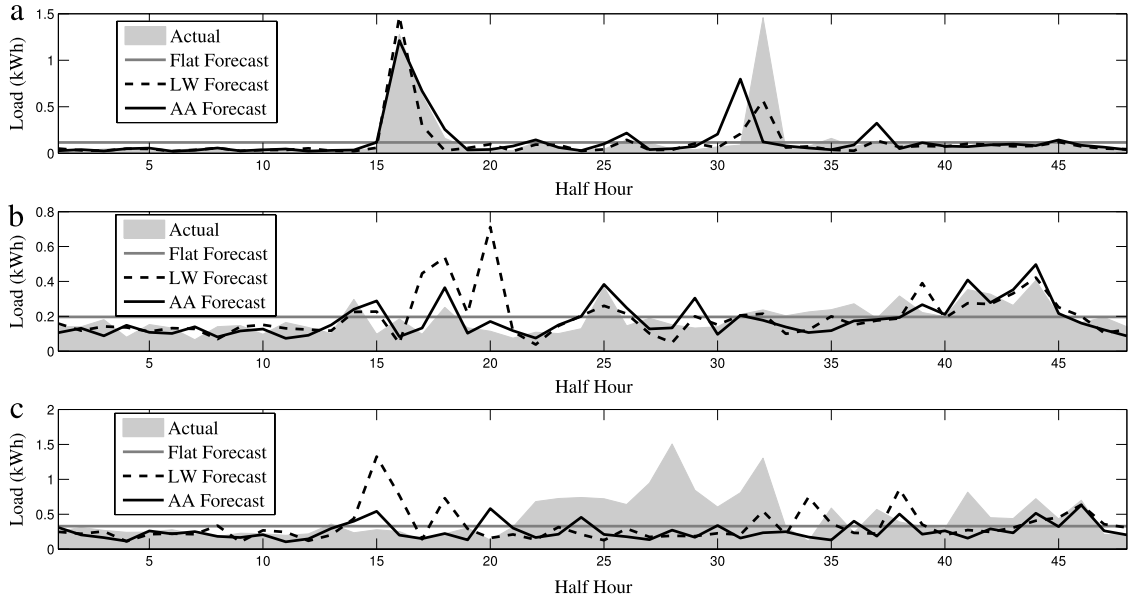

Fig. 3. Forecasted usage of each household described in the text on Wednesday of the final week of the data set. The plots show the actual usage (shaded area) and the forecasts for (a) household A, (b) household B and (c) household C, using the AA (black line), LW (dashed line) and Flat (gray line) forecast methods.

2. *Last week* (LW) forecast: The usage on the same day of the previous week.
3. *Averaged adjustment* (AA) forecast: A combination of a historic average and baseline usage. A detailed description can be found in Appendix.

Snapshots of a single day's data from each household, with the corresponding forecasts, are illustrated in Fig. 3.

Clearly, the flat forecast provides little informative value, while the LW forecast is innately realistic but performs poorly for irregularities in the week-to-week behaviour. The AA forecast is subjectively better than the other forecasts, but volatility still reduces its performance.

As in the simple example described in Section 2.3, we compare the absolute and adjusted p-norm errors with $p = 4$, in order to penalise larger peaks to a greater extent than smaller peaks. We use $w = 3$ as the adjustment limit, and hence the forecasts can be displaced by up to one and a half hours *either side* of their original forecast time. The effects of w are considered in more detail in Section 4. Because the forecasts produce rolling daily predictions, we

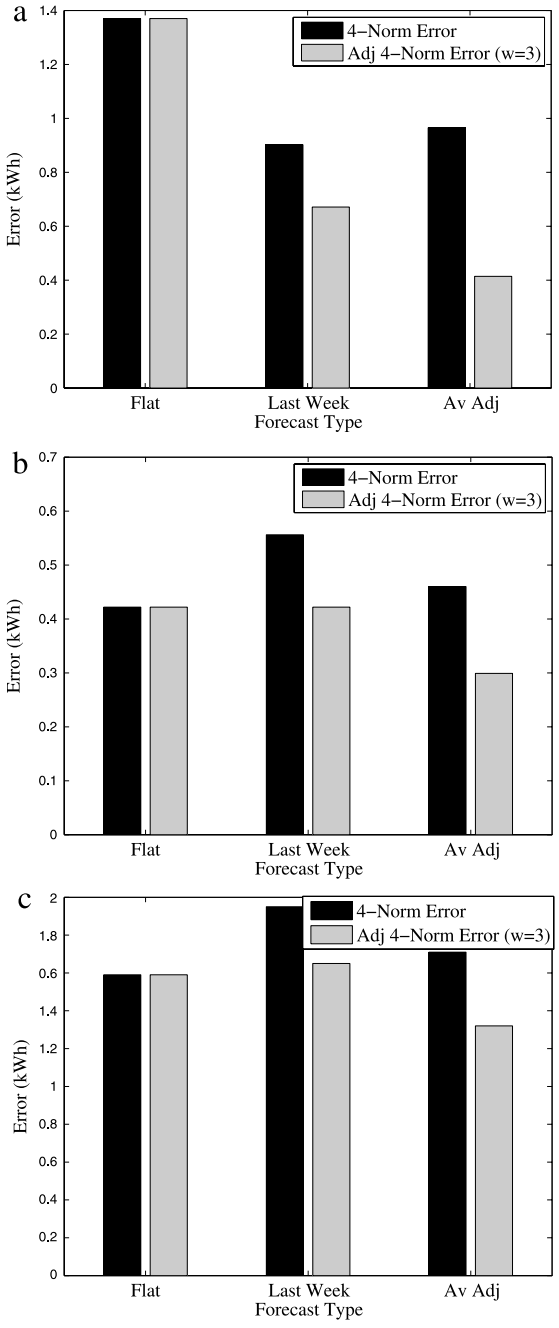

Fig. 4. Panels (a)–(c) correspond to households A–C respectively. Each panel depicts the daily averages of the 4-norm (black) and adjusted 4-norm (gray) errors for the three forecasts.

calculate the ith day's errors for each measure, e_i, and use the mean absolute error,

$$\langle E \rangle = \frac{1}{7} \sum_{i=1}^{7} e_i, \tag{5}$$

to compare forecasts.

The daily mean errors of each forecast method are shown in Fig. 4(a)–(c) for households A–C, respectively. The black bars show the daily mean 4-norm error and the gray bars show the daily mean adjusted 4-norm error. Focusing first on the 4-norm errors, we note that the flat forecast out-performs the other forecasts for both households B and C. In addition, the AA forecast is beaten by the LW forecast for household A. Clearly, these results do not agree with the proposed forecasting hierarchy. In particular, we know that the flat forecast reproduces none of the daily household usage patterns. By ignoring peaks altogether, the flat forecast avoids the double penalty and can appear to be better than more sophisticated forecasts, but it is clearly of no use for control or scheduling purposes.

We now consider the 4-norm adjusted errors, illustrated with gray bars in panels (a)–(c) of Fig. 4. We note that the adjusted norm does not change the flat forecast errors, but reduces all of the LW and AA errors. The AA forecast is now the most successful forecast for all households, with a marked improvement for household A in particular. This can be attributed to the regular peak demands which are observed in the data being forecasted close to when they actually occur, and the absence of the double penalty in the adjusted error measure. Relative to the flat forecast errors, the improvement in the errors for the AA forecast decreases from households A to C, owing to the relative increases in volatility. The errors for household C are by far the largest in magnitude, and the relative differences between methods are the smallest, indicating that forecast sophistication only introduces marginal relative improvements as the volatility increases.

3.2. An example with six hundred households

To illustrate that our results hold more generally, we consider the 4-norm and adjusted 4-norm errors of the three forecasting methods, applied to the usage data of 600 individual domestic households. As in the example above, the dataset for each household consists of half hourly electrical energy usage over a 10-week period, collected by smart meters during the EDRP trial. Using the Flat, LW and AA methods, rolling daily forecasts of each household's energy usage were produced for the final week of each dataset. Fig. 5 shows the mean daily difference between the flat forecast errors and the 4-norm and 4-norm adjusted errors (with $w = 3$) for both the LW and AA forecasts. The horizontal axis represents the mean daily difference between the 4-norm errors of the Flat forecast and the LW (or AA) forecast, and the vertical axis represents the mean daily difference between the adjusted 4-norm errors of the Flat forecast and the adjusted 4-norm errors of the LW (or AA) forecast. The diagonal line indicates where the mean 4-norm and mean adjusted 4-norm errors are equal. Since the adjusted 4-norm error is always smaller than (or at most equal to) the 4-norm error, no forecasts can occupy the area below the line.

The three occupied quadrants of the graph establish a three-cluster segmentation of the forecasts in terms of their accuracy:

1. Points in the lower-left quadrant represent forecasts whose mean 4-norm and mean adjusted 4-norm errors are larger than or equal to the mean flat forecast errors. We refer to these forecasts as *Poor*.

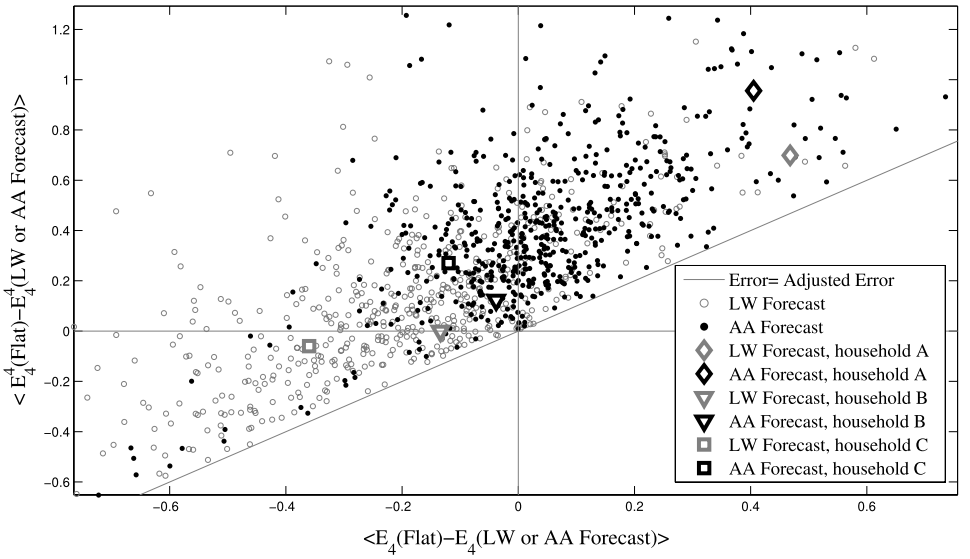

Fig. 5. Mean daily difference in the adjusted 4-norm forecast errors of the flat and LW (unfilled circles) or AA (filled) forecasts versus the mean daily difference in 4-norm forecast errors of the flat and LW (unfilled circles) or AA (filled) forecasts. Also included are the data for households A (diamonds), B (triangles) and C (squares).

2. Points in the upper-left quadrant represent forecasts whose mean flat forecast error is smaller than the mean 4-norm error forecast but larger than the mean adjusted 4-norm error. Since the small temporal re-alignment has reduced the error compared to the 4-norm error, we refer to these forecasts as *Good after adjustment*.
3. Points in the top right quadrant represent forecasts whose mean 4-norm and mean adjusted 4-norm errors are both smaller than the mean flat forecast errors. We refer to these as *Good* forecasts.

The plot shows that, in general, the AA forecasts (filled circles) are superior to the LW forecasts (unfilled circles). The majority of the AA forecasts are either good (360) or good after adjustment (208). Only 32 of the AA forecasts are poor, whereas 225 of the LW forecasts are poor. For the LW method, only 105 are good forecasts, and just under half (270) are good after adjustments. Of the 600 households, the LW forecasts only out-perform the AA forecasts for 30 households in the 4-norm, but for 46 households in the adjusted 4-norm. In Fig. 5, we also include the data for the LW and AA forecasts of households A, B and C. In terms of our accuracy classification, both the LW and AA are good forecasts for household A, whereas for households B and C, the AA forecast is only good after adjustment and the LW forecast is poor. The large proportion of forecasts that are good after adjustment are particularly important. If only the 4-norm is used as an accuracy measure, then these forecast methods could potentially be rejected mistakenly, despite their improved score relative to the adjusted norm.

4. The adjustment limit

The choice of the adjustment limit, w, is largely subjective and application-specific, but can have important implications for the problem being investigated. For instance, the adjusted norm can be used to identify households whose peaks can be forecasted accurately within w of the actual peak. Such forecasts can then be used to create a charging/discharging plan of a storage device, in order to reduce the anticipated peak demand on the LV network (Molderink et al., 2010; Rowe et al., 2012). Different sizes of the adjustment window therefore change the potential reduction in the network peak demand. For small windows (e.g., $w = 0$), many forecasts will incur a double penalty, and battery storage could be discarded mistakenly as an inappropriate solution for these networks. Similarly, a large adjustment window means that very inaccurate forecasts will have small errors, but will provide little information about the size and timing of the actual peak demand, and provide little if any reduction in peak demand. Hence, in this application, the choice of w should be made with the aim of maximising the potential peak demand reduction using the battery. The specific choice of w based on a particular application is beyond the scope of this work, but will be considered in future work in the context of smart storage. In this section we analyse the properties of the adjusted error in more detail by considering the measure as a function of w for forecasts of household smart meter data, and investigate how this can inform us of the predictability properties of different households.

Fig. 6 displays the mean adjusted 4-norm error for each of the households introduced in Section 3 for the AA and LW forecasts, illustrated in panels (a) and (b) respectively, for different values of w. Each curve is a monotonically decreasing function of the adjustment limit. The black marker on each line shows where the forecast error equals the error of the flat forecast (the forecasts for household A have smaller errors than the flat forecast in these examples, hence the absence of a marker). For all households, in order

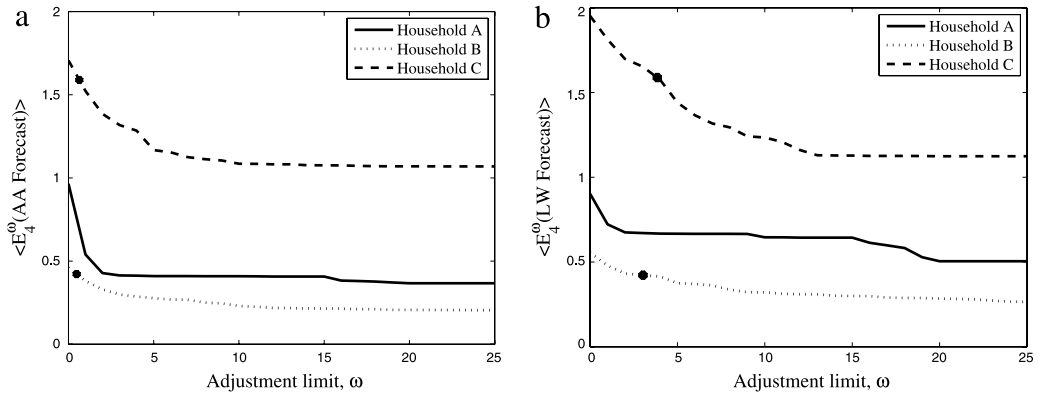

Fig. 6. The mean adjusted errors for (a) the AA forecast and (b) the LW forecast for the usage of households A (solid line), B (dotted line) and C (dashed line) as a function of w. The black marker on each line shows where the forecast errors equal the errors of the flat forecast.

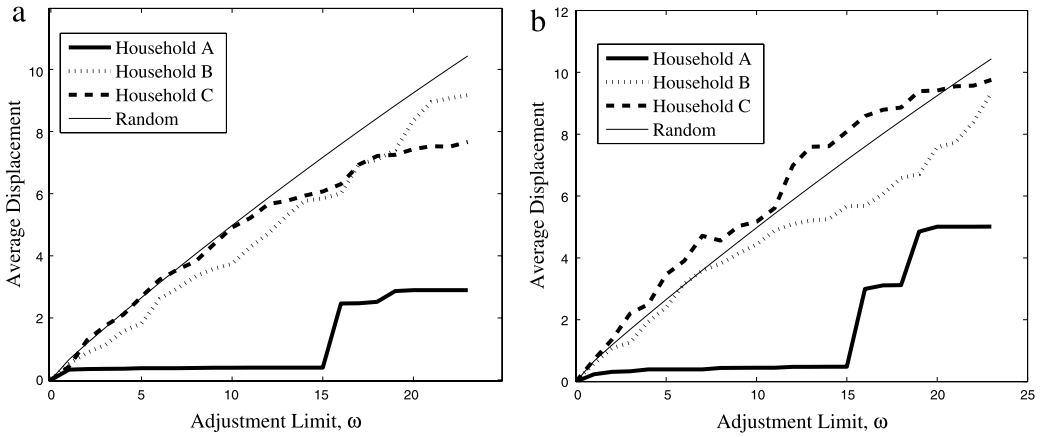

Fig. 7. Average weighted displacement of forecasted points for different adjustment limits for (a) the AA forecast and (b) the LW forecast for the three different households A (solid line), B (dotted) and C (dashed). Also included is the expected displacement if forecast points were assigned randomly.

to outperform the flat forecast, the AA forecast must use $w \geq 1$, whereas the LW forecast must use $w \geq 4$. As we increase w, large reductions in the adjusted error indicate that large peaks in the forecast are being matched to the actuals. We focus on the AA forecast for our analysis, but similar results hold for the LW forecast. As we increase w from 0 to 2, there are large decreases in the adjusted error of the forecast for household A due to the closeness (within 3 half-hours) of the peaks in the forecasts and the actual usage. Moderate decreases in the forecast errors are also observed for household C, although the errors are still relatively large compared with the errors in the forecasts for households A and B, even with $w = 20$ (shifts of ± 10 h). Household B has a slow rate of reduction as w increases. As is shown in Section 3, the general behaviour of household B can be forecasted accurately, and so the slow reduction is likely to be due to the matching of the small daily irregularities.

The adjusted error decreases with an increasing w, but this is likely to increase the mean displacement of the forecast positions simultaneously. Smaller displacements are more desirable, as they indicate a closer proximity of the features of the forecast to the actuals. To fully describe the accuracy of a forecast, we must consider both the mean displacement and the adjusted error of the forecast. As was shown for the synthetic example in Section 2.3, the mean displacement can be used to distinguish between the accuracies of two forecasts with the same adjusted error. Since we are interested primarily in the displacement of the peak loads, we consider a weighted mean displacement. Suppose that the forecast at point i, f_i, is matched to the actual at j, and $d_i = |i - j|$ is the forecast displacement; then, we define the average displacement for each day as

$$\hat{D} = \frac{\sum_{i=1}^{48} f_i^4 d_i}{\sum f_i^4}. \tag{6}$$

The power of 4 ensures that our measure is representative of larger peaks.

Fig. 7 shows the mean displacement of the AA and LW forecasts over the final week as a function of w for each household, together with a plot of the expected average displacement if the forecast had been assigned randomly. (The random displacement is found by calculating the

expected displacement for each of the 48 daily points within the adjustment limit, assuming that any displacement is equally likely. The mean over the 48 daily points is then calculated.) We present the results for the AA forecast, but the LW forecast results are similar. The mean displacements of the forecasts for households B and C closely match the random displacement curve when $w < 10$. It is likely that the features of the forecasts are being matched to the irregular week-to-week behaviours of the households. As we showed in Section 3, the regular behaviour of household B is forecasted accurately, but the small, irregular demands are forecasted poorly. Household C has no regular week-to-week behaviour, and is largely unpredictable. In contrast, household A has a regular weekly behaviour and the peaks are forecasted accurately, and therefore the mean displacement remains small for all w values. As the adjustment limit is increased beyond $w = 15$, some of the afternoon and morning peaks are matched, resulting in a small increase in the size of the average displacement.

Figs. 6 and 7 together reveal extra information about the usage patterns and forecast accuracies of each of the different households. In particular, for household A, sharp drops in the forecast error as w is increased from 0 to 2 indicate that the forecasts approximate the large features in the data closely. The small average displacements confirm that the regular peaks are being matched. In contrast, for household C, the large reduction in forecast errors is likely to be the result of matching the random, irregular behaviour, as is shown by the mean displacement being similar to a random assignment in Fig. 7. Similarly, we find that the small reductions in the adjusted error for household B as we increase w are mainly a consequence of matching the small irregular behaviours which are missed by the forecast.

5. Discussion

As low carbon technologies become ubiquitous, there are increased risks to the robustness and security of low voltage (LV) electricity networks. The electrification of heating and transport is expected to increase network peak demand, while the increased uptake of more intermittent forms of generation such as photovoltaics is likely to increase network volatility. In order to be able to manage the local networks effectively, it is vital that distribution network operators understand how demand is changing and what practical solutions are available. Household smart meters are becoming an integral part of many governments' low carbon agendas, and many countries aim to have a meter in every home within the next decade. Smart meters provide a valuable opportunity for detailed data analytics, and in particular for forecasts at the individual and low voltage substation levels. Accurate household-level forecasts can also be utilised for planning the smart control of storage devices so as to reduce peak demands, and for understanding how often network constraints are violated. However, before useful household-level forecasts can be developed, an appropriate verification measure must be established for assessing the accuracy of such forecasts.

In this paper we suggest such a measure for assessing the success of forecasts of volatile and noisy data. A standard treatment of forecast accuracy is to consider the p-norm of the error, but, due to the "double penalty" effect, such measures are inadequate, especially when attempting to forecast peaks and troughs in the data. Any successful forecast method requires a degree of flexibility in the spatial/temporal positioning of the peaks. Our proposed solution, the adjusted p-norm error, allows for limited permutations of the forecasted data, which reduces the penalty imposed on shifted peaks. This is first illustrated with a simple synthetic example, then demonstrated on forecasts of real, high resolution household electrical energy usage.

To test the forecast measure, three forecast methods were applied to three separate households' energy demand data, with varying degrees of week to week regularity, and hence, forecastability. The forecasts varied in skill, with a clear hierarchy: an innately poor flat forecast, a poor, yet realistic 'last week as this week' forecast, and an adjusted-average of the previous week's behaviour. We found that, with respect to a point-wise metric, the flat forecast could outperform many of the more realistic, informative forecasts. This was not the case with our new error measure. In Section 3.2, we also applied the measure to forecasts of 600 independent households, which confirmed the ability of the new measure to distinguish successfully between the accuracies of the three forecast methods. In addition, we also considered the effect of changing w on the adjusted error and the average displacement of the matched forecasts. This offered further insights into the accuracies of the forecasts. In summary, in this paper we have presented a new method for verifying the forecast accuracy which has been shown to be effective and efficient for assessing the accuracy of shifted features of volatile and noisy data sets.

The new measure presented in this paper deforms the forecast in a discontinuous way, which may not be appropriate for all applications. For high voltage level demand, which is more smooth and regular, the standard point-wise measures are adequate. In contrast, for volatile and irregular data, the smoothness of the deformation may be less significant and the measure presented here may be suitable. An additional advantage of the adjusted norm is that it can be applied using any standard norm and requires only a single control parameter, w.

Accurate household-level forecasts can be utilised in smart control algorithms to plan the charging/discharging of a battery on LV networks which are close to maximum capacity (Molderink et al., 2010; Rowe et al., 2012). Hence, a principal motivation for the new forecast measure presented in this paper is to identify those LV networks in which smart storage could be an effective solution for peak demand reduction. In addition, it is arguably more appropriate for the forecast measure to impose heavier penalties on peaks which are forecasted too late than on those which are forecasted too early, to ensure that the battery is charged sufficiently before the anticipated peaks, in order to maximise peak reductions. In future work, we consider how the size of the adjustment window and allowing for biases in timing can affect the potential

peak reduction via the smart control of storage devices at the household to LV substation-level. In addition, we will also consider the accuracy of more traditional forecast methodologies which are used in higher voltage load forecasting relative to that of our new measure, to test their suitability for forecasting electricity demand at the household level.

Acknowledgments

We wish to thank Scottish and Southern Energy Power Distribution (SSEPD) for their support, and Scottish and Southern Energy (SSE) for providing the EDRP data for use in this project. SH, PG and CS would like to thank SSEPD for the funding of this project via the Ofgem 'Innovation Funding Incentive' (http://www.ofgem.gov.uk/Networks/Techn/NetwrkSupp/Innovat/ifi/Pages/ifi.aspx) for Innovation Funding Incentive reports (2011_01 LC Smart Analytics). SH, PG, DVG and CS would also like to thank SSEPD for support via the New Thames Valley Vision Project (SSET203 New Thames Valley Vision), funded through the Low Carbon Network Fund. JAW acknowledges the EPSRC for support of MOLTEN (EP/I016058/1).

Appendix. The averaged adjustment forecast

In this Appendix we briefly describe the Averaged Adjustment (AA) forecast, as implemented in this report. For clarity, we show how we forecast for one particular day; the other days of the week are forecasted in an analogous way. We assume that we have N daily usage profiles at a half-hourly resolution of the dth day of the week ($d = 1, \ldots, 7$), which we write as $\mathbf{G}^{(k)} = (g_1^{(k)}, g_2^{(k)}, \ldots, g_{48}^{(k)})^T$ for $k = 1, 2, \ldots, N$, where $\mathbf{G}^{(1)}$ is the previous week's usage on the dth day, $\mathbf{G}^{(2)}$ is the usage on the dth day from two weeks before, etc. We create a base profile $\mathbf{F}^{(1)} = (f_1^{(1)}, f_2^{(1)}, \ldots, f_{48}^{(1)})^T$, where each half hour is defined as the median value over all N half hours. We update the baseline profile iteratively using matching with each successive previous week's data. This is performed as follows. Suppose that $\mathbf{F}^{(k)}$ is the current baseline for the kth iteration ($1 \leq k \leq N-1$). We define $\hat{\mathbf{G}}^{(k)} = \hat{P}\mathbf{G}^{(k)}$, where $\hat{P} \in \mathcal{P}$ is a permutation matrix such that

$$\|\hat{P}\mathbf{G}^{(k)} - \mathbf{F}^{(k)}\|_4 = \min_{P \in \mathcal{P}} \|P\mathbf{G}^{(k)} - \mathbf{F}^{(k)}\|_4, \tag{7}$$

where $\mathcal{P}$ represents the set of restricted permutations of the half hour loads (i.e., each half hour i moved to some half hour j, where $|i-j| \leq w$, and w is the deformation limit, as described in Section 2.2). In other words, $\hat{\mathbf{G}}^{(k)}$ is the usage from the previous week that minimises the deformed norm error between the baseline load usage and the usage of the current week $\mathbf{G}^{(k)}$. The new baseline is defined to be

$$\mathbf{F}^{(k+1)} = \frac{1}{k+1}(\hat{\mathbf{G}}^{(k)} + k\mathbf{F}^{(k)}). \tag{8}$$

This process is repeated for each of the remaining weeks, to give the final forecast $\mathbf{F}^{(N)}$. Hence, the forecast is defined to be an average of the initial baseline and permutations of the previous weeks:

$$\mathbf{F}^{(N)} = \frac{1}{N+1}\left(\sum_{k=1}^{N} \hat{\mathbf{G}}^{(k)} + \mathbf{F}^{(1)}\right). \tag{9}$$

References

Alfares, H., & Nazeeruddin, M. (2002). Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science*, *33*, 23–34.

Brabec, M., Konár, O., Pelikán, E., & Malý, M. (2008). A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. *International Journal of Forecasting*, *24*, 659–678.

Brooks, H., & Doswell III, C. (1996). A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather and Forecasting*, *11*, 288–303.

Castati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., et al. (2008). Review forecast verification: current status and future directions. *Meteorological Applications*, *15*, 3–18.

Chen, L., & Ng, R. (2004). On the marriage of Lp-norms and edit distance. *Proceedings of the thirtieth international conference on very large databases: vol. 30* (pp. 792–803).

Combrink, F.M., & Vaessen, P.T.M. (2006). Low voltage, but high tension. https://www.idc-online.com/technical_references/pdfs/electrical_engineering/LV-HT.pdf (accessed 20.01.2013).

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, *15*, 51–64.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations* (3rd ed.). Johns Hopkins University Press.

Hoffman, R. N., Liu, Z., Louis, J.-F., & Grassotti, C. (1995). Distortion representation of forecast errors. *Monthly Weather Review*, *123*, 2758–2770.

Keil, C., & Craig, G. C. (2009). A displacement and amplitude score employing an optical flow technique. *Weather and Forecasting*, *24*, 1297–1308.

Levina, E., & Bickel, P. (2001). The Earth Mover's distance is the Mallows distance: some insights from statistics. In *Proceedings of the eighth IEEE international conference on computer vision: vol. 2* (pp. 251–256).

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

Molderink, A., Bakker, V., Bosman, M. G. C., Hurink, J. L., & Smit, G. J. M. (2010). A three-step methodology to improve domestic energy efficiency. *IEEE innovative smart grid technologies conference, 19–21 Jan. 2010*.

Muller, M. (2007). *Information retrieval for music and motion*. Springer.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, *5*, 32–38.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*, 281–293.

Murphy, A., & Winkler, R. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*, 1330–1338.

National Grid (2012). UK Future Energy Scenarios 2012. http://www.nationalgrid.com/NR/rdonlyres/C7B6B544-3E76-4773-AE79-9124DDBE5CBB/56766/UKFutureEnergyScenarios2012.pdf (accessed 20.01.2013).

Ofgem (2012). Low carbon networks fund governance document v.5. http://www.ofgem.gov.uk/Networks/ElecDist/lcnf/Pages/lcnf.aspx (accessed 10.02.2013).

Rowe, M., Holderbaum, W., Potter, B., & Liu, Y. (2012). The scheduling and control of storage devices on the low voltage network using forecasted energy demand. https://www.reading.ac.uk/CMOHB/resources/cmohb-resources.aspx. *The low carbon network consumer behaviour workshop*, University of Reading, 7th November 2012.

Schrijver, A. (2002). *Combinatorial optimization: polyhedra and efficiency*. Springer.

Stefan, A., Athitsos, V., & Das, G. (2013). The move-split-merge metric for time series. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 1425–1438.

Taylor, J. W., & Espasa, A. (2008). Energy forecasting. *International Journal of Forecasting*, *24*, 561–565.

Thomas, P. R. (2010). American electric power's community energy storage. *EPRI 4th international conference on the integration of renewable and distributed energy resources*.
Tomizawa, N. (1971). On some techniques useful for solution of transportation network problems. *Networks*, *1*, 33–34.
Xu, Y., Xie, L., & Singh, C. (2010). Optimal scheduling and operation of load aggregator with electric energy storage in power markets. *North America power symposium*.

Stephen Haben is a PDRA in the Mathematical Institute at the University of Oxford and an academic partner on the New Thames Valley Vision Project, as part Ofgem's Low Carbon Network Fund (LCNF). His interests include forecasting, clustering methods, data assimilation and large data analytics.

Jonathan Ward is a lecturer in the School of Mathematics at the University of Leeds, UK. He obtained his Ph.D. from the Engineering Mathematics department at the University of Bristol under the supervision of R. Eddie Wilson. He has worked in a wide range of areas, including traffic modelling, complex systems, graph theory and mathematical applications in the social sciences.

Danica Vukadinovic Greetham is Lecturer in Mathematics and Co-director of the Centre for the Mathematics of Human Behaviour. She worked as a PDRA on a Digital Economy EPSRC granted project focusing on developing algorithms for network optimisation and scaling up existing methods to very large networks. She worked in industry for five years as a research scientist on the agent-based modelling of markets and consumers, simulating markets' and people's behaviours. Her main interests include algorithmic graph theory, the analysis of online and offline social networks, the modelling and analysis of large networks, and social network based behaviour change interventions.

Colin Singleton is the Technical Director at CountingLab Limited, a spinoff from the University of Reading's Department of Mathematics and Statistics. He is a mathematical analyst and his expertise is in modelling, profiling and forecasting customer behaviour from large and very large data sets. Having worked for four years as a senior consultant at Numbercraft Limited providing customer insight, he subsequently worked as a consultant/software developer at Analysys on business models in the telecoms sector. He then worked for two years as a postdoc at the University of Reading, researching and developing customer profiling models, before helping set up CountingLab.

Peter Grindrod is a Professor of Mathematics within the Mathematical Institute at the University of Oxford. He is a member of BBSRC Council, a former member of the EPSRC Council, a former President of the IMA and a member of the Ministry of Defense DSAC, and he was awarded a CBE in 2005 for services to mathematics. He has founded three successful analytics companies, working in sectors such as retail, consumer goods, energy supply, mobile telephony, credit risk, digital media and digital marketing.