Improving short term load forecast accuracy via combining sister forecasts

Jakub Nowotarski¹,²
Bidong Liu²
Rafał Weron¹
Tao Hong²

¹ Department of Operations Research, Wrocław University of Technology, Poland
² Energy Production and Infrastructure Center, University of North Carolina at Charlotte, USA
Improving Short Term Load Forecast Accuracy via Combining Sister Forecasts

Jakub Nowotarski\textsuperscript{a,b}, Bidong Liu\textsuperscript{b}, Rafał Weron\textsuperscript{a}, Tao Hong\textsuperscript{b}

\textsuperscript{a}Department of Operations Research, Wrocław University of Technology, Wrocław, Poland
\textsuperscript{b}Energy Production and Infrastructure Center, University of North Carolina at Charlotte, USA

Abstract

Although combining forecasts is well-known to be an effective approach to improving forecast accuracy, the literature and case studies on combining load forecasts are relatively limited. In this paper, we investigate the performance of combining so-called sister load forecasts, i.e., predictions generated from a family of models which share similar model structure but are built based on different variable selection processes. We consider eight combination schemes: three variants of arithmetic averaging, four regression based and one performance based method. Through comprehensive analysis of two case studies developed from public data (Global Energy Forecasting Competition 2014 and ISO New England), we demonstrate that combining sister forecasts outperforms the benchmark methods significantly in terms of forecasting accuracy measured by Mean Absolute Percentage Error. With the power to improve accuracy of individual forecasts and the advantage of easy generation, combining sister load forecasts has a high academic and practical value for researchers and practitioners alike.

Keywords: Electric load forecasting, Forecast combination, Sister forecasts.

1. Introduction

Short term load forecasting is a critical function for power system operations and energy trading. The increased penetration of renewables and the introduction of various demand response programs in today’s energy markets has contributed to higher load volatility, making forecasting more difficult than ever before (Motamedi et al., 2012; Pinson, 2013; Morales et al., 2014; Hong and Shahidehpour, 2015). Over the past few decades, many techniques have been tried for load forecasting, of which the popular ones are artificial neural networks, regression analysis and time series analysis (for reviews see e.g. Weron, 2006; Chan et al., 2012; Hong, 2014). The deployment of smart grid technologies has brought large amount of data with increasing resolution both temporally and spatially, which motivates the development of hierarchical load forecasting methodologies. The Global Energy Forecasting Competition 2012 (GEFCom2012) stimulated
many novel ideas in this context (the techniques and methodologies from the winning entries are summarized in Hong et al., 2014).

In the forecasting community, combination is a well-known approach to improving accuracy of individual forecasts (Armstrong, 2001). Many combination methods have been proposed over the past five decades, including simple average, Ordinary Least Squares (OLS) averaging, Bayesian methods, and so forth (for a review see Wallis, 2011). Simple average is the most commonly used method shown to be quite effective in practice (Genre et al., 2013).

Although forecast combination has recently received considerable interest in the electricity price forecasting literature (Bordignon et al., 2013; Nowotarski et al., 2014; Weron, 2014; Raviv et al., 2015) and despite the early applications in load forecasting (see e.g. Bunn, 1985; Smith, 1989), load forecast combination is still an under-developed area. Since weather is a major driving factor of electricity demand, some research efforts were devoted to combining weather forecasts (Fan et al., 2009; Fan and Hyndman, 2012) and combining load forecasts from different weather forecasts (Fay and Ringwood, 2010; Charlton and Singleton, 2014). There are also some studies on combining forecasts from wavelet decomposed series (Amjady and Keynia, 2009) or independent models (Wang et al., 2010; Taylor, 2012; Matijaš et al., 2013). However, to our best knowledge, there is no comprehensive study on the use of different combination schemes in load forecasting.

The fundamental idea of using forecast combination is to take advantage of the information which is underlying the individual forecasts and often unobservable to forecasters. The general advice is to combine forecasts from diverse and independent sources (Batchelor and Dua, 1995; Armstrong, 2001), which has also been followed by the aforementioned load forecasting papers. In practice, however, combining independent forecasts has its own challenge. If the independent forecasts were produced by different experts, the cost of implementing forecast combination is often unaffordable to utilities. On the other hand, if the independent forecasts were produced by the same forecaster using different techniques, the individual forecasts often present varying degrees of accuracy (for a discussion see Weron, 2014), which may eventually affect the quality of forecast combination.

This paper examines a novel approach to load forecast combination: combining sister load forecasts. The sister forecasts are predictions generated from a family of models, or sister models, which share similar model structure but are built based on different variable selection processes, such as different lengths of calibration window and different group analysis settings. The idea of sister forecasts was first proposed and used by Liu et al. (2015), where the authors combined sister load forecasts to generate probabilistic (interval) load forecasts rather than point forecasts as done in this paper. In the forecast combination literature, a similar but less general idea was proposed by Pesaran and Pick (2011), where the authors combined forecasts from the same model calibrated from different lengths of calibration window.

The contribution of this paper is fourfold. Firstly, this is the first empirical study of combining sister forecasts in point load forecasting literature. Secondly, the proposed method is easy to implement compared to combining independent expert forecasts. Thirdly, to our best knowledge, this is the most extensive study so far on combining point load forecasts, considering eight combination and two selection schemes. Finally, the two presented case studies are based on publicly available data (GEFCom2014 and ISO New England), which enhances the reproducibility of our work by other researchers.
The rest of this paper is organized as follows. Section 2 introduces the sister load forecasts, eight combination methods to be tested, and two benchmark methods to be compared with. Section 3 describes the setup of the two case studies. Section 4 discusses the forecasting results, while Section 5 wraps up the results and concludes the paper.

2. Combining Sister Load Forecasts

2.1. Sister models and sister forecasts

When developing a model for load forecasting, a crucial step is variable selection. Given a large number of candidate variables and their different functional forms, we have to select a subset of them to construct the model. The variable selection process may include several components, in particular data partitioning, the selection of error measures and the choice of the threshold to stop the estimation process. Applying the same variable selection process to the same dataset, we should get the same subset of variables. On the other hand, different variable selection processes may lead to different subsets of variables being selected. Following Liu et al. (2015), we call the models constructed by different (but overlapping) subsets of variables sister models and forecasts generated from these models – sister forecasts.

In this study we use a relatively rich family of regression models to yield the sister forecasts. The rationale behind this choice is twofold. Firstly, regression analysis is a load forecasting technique widely used in the industry (Weron, 2006; Hong, 2010; Hyndman and Fan, 2010; Charlton and Singleton, 2014; Goude et al., 2014; Hong, 2014). Secondly, in the load forecasting track of the GEFCom2012 competition the top four winning entries used regression-type models (Hong et al., 2014). Nevertheless, other techniques – such as neural networks, support vector machines or fuzzy logic – can also fit in the proposed framework to generate sister forecasts.

We start from a generic regression model that served as the benchmark in the GEFCom2012 competition:

\[ \hat{y}_t = \beta_0 + \beta_1 M_t + \beta_2 W_t + \beta_3 H_t + \beta_4 W_t H_t + f(T_t), \]  
(1)

where \( \hat{y}_t \) is the load forecast for time (hour) \( t \), \( \beta_i \) are the coefficients, \( M_t, W_t \) and \( H_t \) are the month-of-the-year, day-of-the-week, and hour-of-the-day classification variables corresponding to time \( t \), respectively, \( T_t \) is the temperature at time \( t \), and

\[ f(T_t) = \beta_3 T_t + \beta_5 T_t^2 + \beta_7 T_t^3 + \beta_8 T_t M_t + \beta_9 T_t^2 M_t + \beta_{10} T_t^3 M_t + \beta_{11} T_t H_t + \beta_{12} T_t^2 H_t + \beta_{13} T_t^3 H_t. \]  
(2)

Note that to improve the load forecasts we could apply further refinements, such as processing holiday effects and weekday grouping (see e.g. Hong, 2010). However, the focus of this paper is not on finding the optimal forecasting models for the datasets at hand. Rather on presenting a general framework that lets the forecaster improve prediction accuracy via combining sister forecasts, starting from a basic model, be it regression, an ARMA process or a neural network.

Like in Liu et al. (2015), the differences between the sister models built on the generic regression defined by Eq. (1) and (2) are the amount of lagged temperature variables \( \sum_{\text{lag}} f(T_{t-lag}) \), \( lag = 1, 2, \ldots \), and lagged daily moving average temperature variables \( \sum_d f(T_{t,d}) \), \( d = 1, 2, \ldots \),
where $\tilde{T}_{t,d} = \frac{1}{\Delta T} \sum_{k=24d-23}^{24d} T_{t-k}$ is the daily moving average temperature of day $d$, added to Eq. (1). Hence the whole family of models used here can be written as:

$$\hat{y}_i = \beta_0 + \beta_1 M_t + \beta_2 W_t + \beta_3 H_t + \beta_4 W_t H_t + f(T_t) + \sum_d f(\tilde{T}_{t,d}) + \sum_{\text{lag}} f(T_{t-\text{lag}}).$$

By adjusting the length of the training dataset (here: two or three years) and the partition of the training and validation datasets (here: using the same four calibration schemes as in Hong et al., 2015, that either treat all hourly values as one time series or as 24 independent series), we can obtain different ‘average–lag’ (or $d$–lag) pairs, leading to different sister models. In this paper, we use 8 sister models as in Liu et al. (2015), though the proposed framework is not limited to 8 models only.

2.2. Forecast averaging techniques

As mentioned above, we are interested in the possible accuracy gains generated by combining sister load forecasts. For $M$ individual forecasts $\hat{y}_{1t}, \ldots, \hat{y}_{M_t}$ of load $y_t$ at time $t$, the combined load forecast is given by:

$$\hat{y}_c = \frac{M}{\sum_{i=1}^{M} w_{it}} \hat{y}_{it},$$

where $w_{it}$ is the weight assigned at time $t$ to sister forecast $i$. The weights are computed recursively. The combined forecast for day $t$ utilizes individual (in our case – sister) forecasts ranging from the forecast origin to hour 24 of day $t-1$. Hence, the forecasting setup for the combined model is the same as for the sister models.

2.2.1. Simple, trimmed and Windsorized averaging

The most natural approach to forecast averaging utilizes the arithmetic mean of all forecasts of the different (individual) models. It is highly robust and is widely used in business and economic forecasting (Genre et al., 2013; Weron, 2014). We call this approach Simple averaging.

In this study we introduce two – robust to outliers – extensions of simple averaging. Trimmed averaging (denoted by TA) discards two extreme forecasts for a particular hour of the target day. The arithmetic mean is therefore taken over the remaining 6 models. On the other hand, Windsorized averaging (denoted by WA) replaces the two extreme individual forecasts by the second largest and the second smallest individual forecasts. Hence, the arithmetic mean is taken over 8 models, but forecasts of two models are used twice.

2.2.2. OLS and LAD averaging

Another relatively simple but effective method is Ordinary Least Squares (OLS) averaging. The method was introduced by Crane and Crotty (1967), but its popularity was trigged by Granger and Ramanathan (1984). Since then, numerous variants of OLS averaging have been considered in the literature.

In the original proposal the combined forecast is obtained from the following regression:

$$y_t = w_{0t} + \sum_{i=1}^{M} w_{it} \hat{y}_{it} + e_t,$$
and the corresponding load forecast combination $\hat{P}_t^c$ at time $t$ using $M$ models is calculated as

$$\hat{y}_t^c = \hat{w}_{0t} + \sum_{i=1}^{M} \hat{w}_{it}\hat{y}_{it}. \quad (6)$$

This approach has the advantage of generating unbiased combined forecasts. However, the vector of estimated weights $\{\hat{w}_1, \ldots, \hat{w}_M\}$ is likely to exhibit an unstable behavior – a problem sometimes dubbed ‘bouncing betas’ or collinearity – due to the fact that different forecasts for the same target tend to be correlated. As a result, minor fluctuations in the sample can cause major shifts of the weight vector.

To address this issue, Nowotarski et al. (2014) have proposed to use a more robust version of linear regression with the absolute loss function $\sum_i |e_i|$ in (5), instead of the quadratic function $\sum_i e_i^2$. The resulting model is called Least Absolute Deviation (LAD) regression. Note that LAD regression is a special case of Quantile Regression Averaging (QRA) introduced in Nowotarski and Weron (2015) and for the first time applied to load forecasting in Liu et al. (2015), i.e. considering the median in quantile regression yields LAD regression.

2.2.3. PW and CLS constrained averaging

The original formulation of OLS averaging may lead to combinations with negative weights, which are hard to interpret. To address this issue we may constrain the parameter space. For instance, we may admit only positive weights (denoted later in the text by PW):

$$w_{0t} = 0 \quad \text{and} \quad w_{it} \geq 0, \quad \forall i, t. \quad (7)$$

Aksu and Gunter (1992) found PW averaging to be a strong competitor to simple averaging and to almost always outperform (unconstrained) OLS averaging.

The second variant considered in this study, called constrained regression or Constrained Least Squares (CLS), restricts the model even more and admits only positive weights that sum up to one:

$$w_{0t} = 0 \quad \text{and} \quad \sum_{i=1}^{M} w_{it} = 1, \quad \forall t. \quad (8)$$

CLS averaging yields a natural interpretation of the coefficients, $w_{it}$, which can be viewed as relative importance of each model in comparison to all other models. Note that there are no closed form solutions for the PW and CLS averaging schemes. However, they can be solved using quadratic programming (Nowotarski et al., 2014; Raviv et al., 2015).

2.2.4. IRMSE averaging

A simple performance-based approach has been suggested by Diebold and Pauly (1987). IRMSE averaging computes the weights for each individual model based on their past forecasting accuracy. Namely, the weight for each model is equal to the inverse of its Root Mean Squared Error (RMSE). This is a very intuitive approach – the smaller a method’s error in the calibration
period, the greater its weight:

\[ w_{it} = \frac{\left( \frac{\text{RMS} E_{it}}{\sum_{m=1}^{M} \text{RMS} E_{im}} \right)^{-1}}{\sum_{i=1}^{M} \left( \frac{\text{RMS} E_{it}}{\sum_{m=1}^{M} \text{RMS} E_{im}} \right)^{-1}} = \frac{1}{\text{RMS} E_{it}} \sum_{i=1}^{M} \frac{1}{\text{RMS} E_{it}}. \]  

(9)

Here, \( \text{RMS} E_{it} \) denotes the out-of-sample performance for model \( i \) and is computed in a recursive manner using forecast errors from the whole calibration period.

3. Case Study Setup

3.1. Data description

The first case study is based on data released from the probabilistic load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L, see Fig. 1). The original GEFCom2014-L data includes 6 years (2006-2011) of hourly load data and 11 years (2001-2011) of hourly temperature data from a U.S. utility. Six years of load and temperature data is used for the case study, where the temperature is the average of 25 weather stations. Based on two training datasets (2006-2008 and 2007-2008) and four data selection schemes proposed in Hong et al. (2015), we identify 8 sister models using year 2009 as the validation data. Then each sister model is estimated using 2 years (2008-2009) and 3 years (2007-2009) of training data to generate 24-hour ahead forecasts on a rolling basis for 2010. Following the same steps, we also generate eight 24-hour ahead forecasts on a rolling basis for 2011, with the training data of two (2009-2010) and three years (2008-2010).

While the GEFCom2014-L data includes only one load series, we would like to extend the experiment to additional zones from other locations. Therefore, we develop the second case study using data published by ISO New England, see Fig. 2. The territory of ISO New England can
be divided into 8 zones, including Connecticut, Main, New Hampshire, Rhode Island, Vermont, North central Massachusetts, Southeast Massachusetts, and Northeast Massachusetts. We aggregate three zones (Zone 6 to Zone 8) in Massachusetts to Zone 9. We aggregate all 8 zones (Zone 1 to Zone 8) to get Zone 10, representing the total demand of ISO New England. 7 years of hourly data (2007-2013) from 10 load zones are used for the second case study. We generate forecasts for ISO New England following similar steps as for GEFCom2014-L data. In other words, we generate eight 24-hour ahead forecasts on a rolling basis for 2012 with the training data of two (2010-2011) and three years (2009-2011), as well as eight 24-hour ahead forecasts on a rolling basis for 2013 with the training data of two (2011-2012) and three years (2010-2012).

3.2. Two Benchmark models

We use two benchmarks, both based on the concept of the ‘best individual model’. In a straightforward manner the best individual model could be defined as the best performing individual (in our study – sister) model from an ex-post perspective. Although conceptually pleasing, an ex-post analysis is not feasible in practice – one cannot use information about the quality of a model’s
predictions in the future for forecasting conducted at an earlier moment in time. Hence, like in Nowotarski et al. (2014), we evaluate the performance of combining schemes against that of the realistic alternative of selecting a single model specification beforehand.

We allow for two choices of the best individual model – a static and a dynamic one. The Best Individual ex-ante model selection in the Validation period (BI-V) picks an individual model only once, on the basis of its Mean Absolute Error (MAE) in the validation period (see Section 3.1). For each of the 10 considered zones in the ISO New England case study, one benchmark BI-V model is selected for all 24 hours, the one with the smallest MAE.

The Best Individual ex-ante model selection in the Calibration window (BI-C) picks a sister model in a rolling manner. We choose the individual model that yields the best forecasts in terms of MAE for the data covering the first prediction point up hour 24 of the day in which the prediction is made, like for the forecast averaging schemes. Note that BI-C is essentially a model (or forecast) selection scheme, but we can view it also as a special case of forecast averaging with degenerate weights given by the vector:

\[ w_{it} = \begin{cases} 
1 & \text{if model } i \text{ has lowest MAE}, \\
0 & \text{otherwise}. 
\end{cases} \quad (10) \]

Note also that in what follows we use a weekly evaluation metric as in Weron and Misiorek (2008) and Nowotarski et al. (2014), while the weights for the individual forecasts and in particular the choice of the BI-C model are determined on a daily basis.

3.3. Forecasts evaluation methods

In this Section we evaluate the quality of 24-hour ahead load forecasts in two one-year long out-of-sample periods: (i) **year 2011** for GEFCom2014-L data (excluding three days, August 27-29, with extremely low loads) and (ii) **year 2013** for ISO New England data, see Figs. 1 and 2, respectively. Forecasts for all considered models are determined in a rolling manner: models (as well as model parameters and combination weights) are reestimated on a daily basis and a forecast for all 24 hours of the next day is determined at the same point in time. Forecasts are first calculated for each of the eight sister models. Then they are combined according to estimated weights for each of the eight forecast combination methods and two model selection schemes:

1. Simple – a simple (arithmetic) average of the forecasts provided by all eight sister models,
2. TA – a trimmed mean of the sister models, i.e. an arithmetic average of the six central sister forecasts (the two sister models with the extreme predictions – one at the high and one at the low end – are discarded),
3. WA – a Winsorized mean of the sister models, i.e. an arithmetic average of eight sister forecasts after replacing the two sister models with the extreme predictions by the extreme remaining values,
4. OLS – forecast combination with weights determined by Eq. (5) using standard OLS,
5. LAD – forecast combination with weights determined by Eq. (5) using least-absolute-deviation regression,
6. PW – forecast combination with weights determined by Eq. (5), only allowing for positive weights \( w_{it} \geq 0 \),
7. CLS – forecast combination with weights determined by Eq. (5) with constraints $w_i \geq 0$ and $\sum_{i=1}^{M} w_i = 1$.
8. IRMSE – forecast combination with weights determined by Eq. (9).
9. BI-V – the sister method that would have been chosen ex-ante, based on its forecasting performance in the validation period.
10. BI-C – the sister method that would have been chosen ex-ante, based on its forecasting performance in the calibration period (i.e. from the first prediction point until hour for the day the prediction is made).

We compare model performance in terms of the Mean Absolute Percentage Error (MAPE). Additionally, we conduct the Diebold and Mariano (1995) test (DM) to formally assess the significance of the outperformance of the forecasts of one model by those of another model. As noted above, predictions for all 24 hours of the next day are made at the same time using the same information set. Therefore, forecast errors for a particular day will typically exhibit high serial correlation as they are all affected by the same-day conditions. Therefore, we conduct the DM tests for each of the $h = 1, \ldots, 24$ hourly time series separately, using absolute error losses of the model forecast:

$$L(\epsilon_{h,t}) = |\epsilon_{h,t}| = |P_{h,t} - \hat{P}_{h,t}|.$$  (11)

Note that Bordignon et al. (2013) and Nowotarski et al. (2014) used a similar approach, i.e. performed DM tests independently for each of the load periods considered in their studies. Further note that we conducted additional DM tests for the quadratic loss function. Since the results were qualitatively similar, we omitted them here to avoid verbose presentation.

For each forecast averaging technique and each hour we calculate the loss differential series $d_t = L(\epsilon_{FA,t}) - L(\epsilon_{benchmark,t})$ versus each of the benchmark models (BI-V and BI-C). We perform two one-sided DM tests at the 5% significance level:

- a standard test with the null hypothesis $H_0 : E(d_t) \leq 0$, i.e. the outperformance of the benchmark by a given forecast averaging method,
- the complementary test with the reverse null $H_0 : E(d_t) \geq 0$, i.e. the outperformance of a given forecast averaging method by the benchmark.

4. Results and Comparison

4.1. GEFCom2014

Let us first discuss the results for the GEFCom2014 dataset. The MAPE values for all considered methods are summarized in the second column of Table 1. Obviously, all combining schemes outperform both benchmarks: the BI-C model as well as the first sister model (Ind1), which was the best performing individual method in the validation period (year 2010), i.e. the BI-V model. Overall, the most accurate model is trimmed averaging (TA), followed by Windsorized averaging (WA), Simple and IRMSE. They outperform the BI-C model by ca. 0.2%, which corresponds to a 4% error reduction.
As mentioned above, we also formally investigate the possible advantages from combining over model selection. The DM test results for the GEFCom2014 dataset are presented in Table 2. When tested against Ind1 (=BI-V), we note that Simple, TA and IRMSE are significantly better (at the commonly used 5% level) for 22 out of 24 hours, which is an excellent result. The combining approaches with the relatively worst performance are PW and OLS, both significantly beating the BI-V benchmark 8 times. However, their test statistics still have majority of positive values, 17 times for PW and 15 for OLS.

The test against the BI-C model provides slightly different and less clear cut results. This time out of all combining models the best one is CLS, which is significantly better than the BI-C benchmark for 20 hours. This model is followed by IRMSE and TA (17) and Simple (16). Finally, we should mention that for none of the hours a combining model was significantly worse than any of the two benchmarks. This clearly points to the advantages of combining sister load forecasts.

### 4.2. ISO New England

The MAPE values for all considered methods and all zones (summarized in columns 3-12 of Table 1) confirms our conclusions from Section 4.1 for the GEFCom2014 dataset. In general the combined models perform better than the individual models. This has just one exception – for Zone 2, sister model Ind3 performed better than five combination methods (Simple, TA, WA, CLS and IRMSE), but still worse than the remaining three methods (LAD, PW and OLS). Note, however, that Ind3 performed so well only in the test period (year 2013). In the validation period (year 2012) it was outperformed by Ind2, i.e. the BI-V model.

Let us now focus on zone 10, as it is the aggregated zone that measures the total load in the ISO NE market. Again, the results support of the idea of combining. The combined models yield very similar results, all being clearly more accurate than the sister models – the worst combining
models for zone 10, i.e. LAD and OLS, have MAPE of 2.14% which is better by 0.1% than that of
the best individual model (Ind2 with MAPE of 2.24%) and by nearly 0.2% than that of the BI-V
model (Ind4 with MAPE of 2.32%).

In Table 3 we summarize the Diebold-Mariano test results for zone 10. The overall the conclu-
sions are essentially the same as those for the GEFCom2014 dataset, only this time we can observe
that during late night/early morning hours (3am–6am) the BI-V benchmark (i.e. Ind4 sister model)
is extremely competitive and impossible to beat by a large margin. Also, contrary to the results for
the GEFCom2014 dataset, the BI-C model is significantly worse than the BI-V benchmark. This
is, however, the only combining model that is found to be significantly worse than BI-V.

Overall, models with the largest number of hours (20 out of 24 hours) during which they
significantly outperform the BI-V benchmark are Simple, TA and IRMSE. And again, this fact is
similar to what we have observed for the GEFCom2014 dataset. The latter conclusion has very
important implications, especially for practitioners. These three models are easy to implement
and do not require numerical optimization (hence are fast to compute). Moreover, Simple and
trimmed averaging (TA) work just on future predictions of the individual models, meaning that
even no calibration of weights is required.

Finally, the lower part of Table 3 presents DM test results versus the BI-C benchmark. The
advantage the combining schemes have over model selection is even more striking here. The two
models with the smallest number of hours during which they outperform the BI-C benchmark,
namely LAD and OLS, are better for as many as 19 out of 24 hours.

5. Conclusions

Even though the combination approach is very simple, it is powerful enough to improve accu-
curacy of individual forecasts. In this paper, we investigate the performance of multiple methods to
combine sister forecasts, such as three variants of arithmetic averaging, four regression based and
one performance based method. In the two case studies of GEFCom2014 and ISO New England,
combing sister forecasts beats benchmark methods significantly in terms of forecasting accuracy,
as measured by MAPE and further evaluated by the DM test, which assesses the significance of
the outperformance of the forecasts of one model by those of another model in statistical terms.

Overall, two averaging schemes – Simple and trimmed averaging (TA) – and the performance
based method – IRMSE – stand out as the best performers. All three methods are easy to imple-
ment and fast to compute, the former two do not even require calibration of weights. Given that
sister models are easy to construct and sister forecasts are convenient to generate, our study has
important implications for researchers and practitioners alike.

Acknowledgments

This work was partially supported by the Ministry of Science and Higher Education (MNiSW,
Poland) core funding for statutory R&D activities and by the National Science Center (NCN,
Poland) through grant no. 2013/11/N/HS4/03649.
Table 2: Results for conducted one-sided Diebold-Mariano tests for the GEFCom2014 dataset. Positive numbers indicate the outperformance of the benchmark by a given forecast averaging method: BI-V (top) and BI-C (bottom). Negative numbers – the opposite situation. Numbers in bold indicate significance at the 5% level.

| Method | Hour 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|--------|--------|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Simple |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| TA     |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| W A    |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| OLS    |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| LAD    |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| PW     |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| CLS    |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| IRMSE  |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| BI-C   |       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

Significance at the 5% level.
Table 3: Results for conducted one-sided Diebold-Mariano tests for the ISO NE dataset. Positive numbers indicate the outperformance of the benchmark by a given forecast averaging method: BI-V (top) and BI-C (bottom), negative numbers – the opposite situation. Numbers in bold indicate significance at the 5% level.

<table>
<thead>
<tr>
<th>ISO New England Zone 10 (aggregated), vs BI-V (=Ind4)</th>
<th>Hour</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simple</strong></td>
<td>2.20</td>
<td>2.08</td>
<td>1.47</td>
<td>1.00</td>
<td>0.62</td>
<td>1.19</td>
<td>1.20</td>
<td>2.33</td>
<td>3.58</td>
<td>5.40</td>
<td>6.60</td>
<td>6.12</td>
<td>5.79</td>
<td>5.84</td>
<td>5.53</td>
<td>5.23</td>
<td>4.21</td>
<td>5.30</td>
<td>4.12</td>
<td>4.31</td>
<td>3.96</td>
<td>2.09</td>
<td>1.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>WA</strong></td>
<td>3.38</td>
<td>2.88</td>
<td>1.49</td>
<td>0.94</td>
<td>0.81</td>
<td>1.37</td>
<td>2.13</td>
<td>3.40</td>
<td>3.59</td>
<td>5.34</td>
<td>6.45</td>
<td>6.24</td>
<td>5.66</td>
<td>5.25</td>
<td>5.32</td>
<td>5.29</td>
<td>4.19</td>
<td>3.19</td>
<td>2.28</td>
<td>4.10</td>
<td>4.35</td>
<td>3.97</td>
<td>1.97</td>
<td>1.76</td>
<td></td>
</tr>
<tr>
<td><strong>OLS</strong></td>
<td>2.96</td>
<td>2.51</td>
<td>0.81</td>
<td>0.66</td>
<td>1.73</td>
<td>3.23</td>
<td>3.28</td>
<td>4.30</td>
<td>5.00</td>
<td>4.34</td>
<td>4.05</td>
<td>4.00</td>
<td>4.23</td>
<td>4.56</td>
<td>3.72</td>
<td>3.31</td>
<td>1.82</td>
<td>3.01</td>
<td>3.57</td>
<td>2.66</td>
<td>1.89</td>
<td>1.28</td>
<td>0.57</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td><strong>LAD</strong></td>
<td>2.86</td>
<td>2.70</td>
<td>1.32</td>
<td>1.01</td>
<td>0.90</td>
<td>2.18</td>
<td>3.70</td>
<td>3.49</td>
<td>4.19</td>
<td>4.56</td>
<td>3.84</td>
<td>3.54</td>
<td>3.56</td>
<td>3.69</td>
<td>3.91</td>
<td>3.31</td>
<td>2.91</td>
<td>1.35</td>
<td>2.71</td>
<td>3.54</td>
<td>2.60</td>
<td>1.45</td>
<td>1.43</td>
<td>0.85</td>
<td>1.03</td>
</tr>
<tr>
<td><strong>PW</strong></td>
<td>2.84</td>
<td>2.63</td>
<td>1.18</td>
<td>0.96</td>
<td>1.06</td>
<td>0.97</td>
<td>2.18</td>
<td>3.50</td>
<td>3.62</td>
<td>4.76</td>
<td>5.40</td>
<td>4.89</td>
<td>4.83</td>
<td>4.38</td>
<td>4.55</td>
<td>4.78</td>
<td>4.16</td>
<td>3.33</td>
<td>2.02</td>
<td>3.49</td>
<td>4.35</td>
<td>2.88</td>
<td>0.85</td>
<td>1.03</td>
<td>0.85</td>
</tr>
<tr>
<td><strong>CLS</strong></td>
<td>3.11</td>
<td>2.48</td>
<td>0.78</td>
<td>0.29</td>
<td>0.19</td>
<td>0.68</td>
<td>1.68</td>
<td>3.11</td>
<td>3.50</td>
<td>4.17</td>
<td>6.10</td>
<td>5.86</td>
<td>5.32</td>
<td>5.40</td>
<td>5.17</td>
<td>4.07</td>
<td>3.21</td>
<td>1.62</td>
<td>3.37</td>
<td>4.00</td>
<td>3.19</td>
<td>0.98</td>
<td>1.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>3.16</td>
<td>2.63</td>
<td>1.40</td>
<td>0.93</td>
<td>0.58</td>
<td>1.16</td>
<td>1.86</td>
<td>3.35</td>
<td>3.61</td>
<td>5.46</td>
<td>6.60</td>
<td>6.11</td>
<td>5.70</td>
<td>5.42</td>
<td>5.37</td>
<td>5.23</td>
<td>4.21</td>
<td>3.15</td>
<td>4.11</td>
<td>4.33</td>
<td>3.92</td>
<td>1.97</td>
<td>1.72</td>
<td>1.39</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>BI-C</strong></td>
<td>-0.97</td>
<td>-1.27</td>
<td>-2.00</td>
<td>-2.82</td>
<td>-2.22</td>
<td>0.73</td>
<td>1.13</td>
<td>2.44</td>
<td>1.87</td>
<td>2.34</td>
<td>3.30</td>
<td>3.46</td>
<td>2.35</td>
<td>2.73</td>
<td>2.35</td>
<td>2.74</td>
<td>2.50</td>
<td>2.38</td>
<td>0.07</td>
<td>-0.81</td>
<td>1.57</td>
<td>0.79</td>
<td>-2.66</td>
<td>-2.54</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISO New England Zone 10 (aggregated), vs BI-C (=Ind4)</th>
<th>Hour</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simple</strong></td>
<td>4.51</td>
<td>4.19</td>
<td>4.07</td>
<td>4.82</td>
<td>3.25</td>
<td>2.74</td>
<td>0.62</td>
<td>1.16</td>
<td>1.89</td>
<td>2.04</td>
<td>2.12</td>
<td>1.71</td>
<td>3.36</td>
<td>2.71</td>
<td>3.12</td>
<td>1.89</td>
<td>1.99</td>
<td>1.11</td>
<td>2.83</td>
<td>2.56</td>
<td>3.34</td>
<td>3.52</td>
<td>5.27</td>
<td>5.69</td>
<td></td>
</tr>
<tr>
<td><strong>WA</strong></td>
<td>4.72</td>
<td>4.39</td>
<td>4.15</td>
<td>4.90</td>
<td>3.50</td>
<td>3.13</td>
<td>1.25</td>
<td>2.03</td>
<td>2.01</td>
<td>2.08</td>
<td>2.12</td>
<td>1.93</td>
<td>3.50</td>
<td>2.45</td>
<td>3.09</td>
<td>1.94</td>
<td>1.79</td>
<td>0.90</td>
<td>2.94</td>
<td>2.41</td>
<td>3.36</td>
<td>3.48</td>
<td>5.20</td>
<td>5.14</td>
<td></td>
</tr>
<tr>
<td><strong>OLS</strong></td>
<td>3.58</td>
<td>3.25</td>
<td>1.90</td>
<td>2.64</td>
<td>1.91</td>
<td>1.78</td>
<td>0.31</td>
<td>1.02</td>
<td>2.43</td>
<td>3.34</td>
<td>2.64</td>
<td>2.43</td>
<td>4.17</td>
<td>3.22</td>
<td>3.88</td>
<td>2.64</td>
<td>1.66</td>
<td>1.13</td>
<td>2.40</td>
<td>3.57</td>
<td>3.18</td>
<td>2.41</td>
<td>4.29</td>
<td>4.05</td>
<td></td>
</tr>
<tr>
<td><strong>LAD</strong></td>
<td>4.53</td>
<td>4.17</td>
<td>4.17</td>
<td>4.94</td>
<td>4.17</td>
<td>2.45</td>
<td>2.16</td>
<td>2.66</td>
<td>2.57</td>
<td>2.83</td>
<td>1.72</td>
<td>0.75</td>
<td>1.90</td>
<td>1.38</td>
<td>1.89</td>
<td>0.77</td>
<td>1.09</td>
<td>0.96</td>
<td>1.81</td>
<td>2.48</td>
<td>2.77</td>
<td>2.35</td>
<td>4.75</td>
<td>5.27</td>
<td></td>
</tr>
<tr>
<td><strong>PW</strong></td>
<td>4.52</td>
<td>4.45</td>
<td>3.89</td>
<td>4.79</td>
<td>3.93</td>
<td>2.40</td>
<td>1.97</td>
<td>2.60</td>
<td>2.03</td>
<td>3.66</td>
<td>2.12</td>
<td>1.83</td>
<td>3.28</td>
<td>2.28</td>
<td>2.89</td>
<td>1.70</td>
<td>2.28</td>
<td>1.53</td>
<td>2.97</td>
<td>3.21</td>
<td>3.48</td>
<td>2.75</td>
<td>4.65</td>
<td>5.30</td>
<td></td>
</tr>
<tr>
<td><strong>CLS</strong></td>
<td>5.34</td>
<td>4.75</td>
<td>3.97</td>
<td>4.62</td>
<td>3.16</td>
<td>2.40</td>
<td>1.16</td>
<td>1.86</td>
<td>2.48</td>
<td>3.63</td>
<td>2.89</td>
<td>2.28</td>
<td>3.73</td>
<td>3.16</td>
<td>3.61</td>
<td>2.32</td>
<td>2.30</td>
<td>1.53</td>
<td>2.59</td>
<td>3.22</td>
<td>3.30</td>
<td>3.15</td>
<td>5.40</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>4.54</td>
<td>4.20</td>
<td>4.06</td>
<td>4.82</td>
<td>3.25</td>
<td>2.76</td>
<td>0.69</td>
<td>1.61</td>
<td>1.95</td>
<td>3.03</td>
<td>2.20</td>
<td>1.78</td>
<td>3.44</td>
<td>2.80</td>
<td>3.21</td>
<td>1.98</td>
<td>2.08</td>
<td>1.19</td>
<td>2.84</td>
<td>3.52</td>
<td>3.38</td>
<td>3.52</td>
<td>5.19</td>
<td>5.10</td>
<td></td>
</tr>
</tbody>
</table>
References


01 Probabilistic load forecasting via Quantile Regression Averaging on sister forecasts by Bidong Liu, Jakub Nowotarski, Tao Hong and Rafał Weron
02 Sister models for load forecast combination by Bidong Liu, Jiali Liu and Tao Hong
03 Convenience yields and risk premiums in the EU-ETS - Evidence from the Kyoto commitment period by Stefan Trück and Rafał Weron
04 Short- and mid-term forecasting of baseload electricity prices in the UK: The impact of intra-day price relationships and market fundamentals by Katarzyna Maciejowska and Rafał Weron
05 Improving short term load forecast accuracy via combining sister forecasts by Jakub Nowotarski, Bidong Liu, Rafał Weron and Tao Hong