Working Paper Series, N. 1, January 2011

Combining forecasts for electricity prices

Silvano Bordignon[1], **Derek W. Bunn**[2], **Francesco Lisi**[1], **Fany Nan**[1]

[1] Department of Statistical Sciences, University of Padua,
Via Cesare Battisti 241, 35121 Padova, Italy.

[2] Department of Management Science and Operations,
London Business School, Regent's Park, London NW1 4SA UK

Department of Statistical Sciences
University of Padua
Italy

Abstract: This paper considers how well the approach of combining forecasts extends to the context of electricity prices. With the increasing popularity of regime switching and time-varying parameter models for predicting power prices, the multi model and evolutionary considerations that usually support the combining of simpler time series methods may be less applicable when the individual models incorporate these features. We address this question with a backtesting analysis on British day-ahead prices. Furthermore, given the volatility of power prices and concerns about accurate forecasting under extreme price excursions, we evaluate the results using various error metrics including expected shortfall. The comparisons are furthermore carefully simulated to consider model selection uncertainty in order to realistically test the value of combining as an ex ante policy. Overall, our results support combining for both accurate operational planning and risk management.

Keywords: Forecasts combination, Prediction accuracy, ARMAX, Time-varying parameter regression, Markov regime switching, Electricity price forecasting.

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Final version (2010-12-28)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Fany Nan
tel: +39 049 827 4124
fany.nan@stat.unipd.it
http://homes.stat.unipd.it/fany

Combining forecasts for electricity prices

Silvano Bordignon[1], **Derek W. Bunn**[2], **Francesco Lisi**[1], **Fany Nan**[1]

[1] Department of Statistical Sciences, University of Padua,
Via Cesare Battisti 241, 35121 Padova, Italy.

[2] Department of Management Science and Operations,
London Business School, Regent's Park, London NW1 4SA UK

Abstract: This paper considers how well the approach of combining forecasts extends to the context of electricity prices. With the increasing popularity of regime switching and time-varying parameter models for predicting power prices, the multi model and evolutionary considerations that usually support the combining of simpler time series methods may be less applicable when the individual models incorporate these features. We address this question with a backtesting analysis on British day-ahead prices. Furthermore, given the volatility of power prices and concerns about accurate forecasting under extreme price excursions, we evaluate the results using various error metrics including expected shortfall. The comparisons are furthermore carefully simulated to consider model selection uncertainty in order to realistically test the value of combining as an ex ante policy. Overall, our results support combining for both accurate operational planning and risk management.

Keywords: Forecasts combination, Prediction accuracy, ARMAX, Time-varying parameter regression, Markov regime switching, Electricity price forecasting.

1 Introduction

The value of combining forecasts to achieve accurate predictions is now well-established, with extensive research and convincing applications extending back over 50 years to the work of Granger and his colleagues at Nottingham, Reid (1968, 1969), Bates and Granger (1969) and Newbold and Granger (1974). Despite this body of knowledge, it is quite surprising to observe the absence of substantial research on combining in the context of forecasting electricity prices. Since the established research on electricity markets suggests a wide variety of candidate methods for price forecasting (see, for example, Bunn, 2004; Weron, 2006; Serati et al., 2008) but without any predominant method having emerged, and with model selection varying over time (Chen and Bunn, 2010), the benefits of combining would appear to be very propitious. However, given that the approaches of regime switching, which has an implicit multimodel structure, and time-varying parameter models, which capture model evolutions, have become widely advocated to represent power price dynamics, it is possible that these specifications, to the extent that such models are included in the candidate set of predictive models, may encapsulate and thereby preclude any benefits of simple combinations. We therefore investigate this open question through a detailed study of the effectiveness of combining a set of four carefully specified models, ARMAX, linear regression, Markov regime switching and time-varying regressions, as applied to day-ahead forecasting of British half-hourly power prices.

Methods of increasing sophistication followed the simple adaptive time series approach of Bates and Granger (1969), including Bayesian (Bunn, 1975, 1977), and econometric (Granger and Ramanathan, 1984), as well as extensions to large data sets (Stock and Watson, 2001, 2004), but, for robust forecasting, it has appeared hard to improve upon simple averaging (Makridakis and Winkler, 1983; Clemen, 1989; Stock and Watson, 2001, 2004; Smith and Wallis, 2009). We therefore do not address the question of developing combining methods to improve on simple averaging. We do, however, consider the less commonly addressed question of effectiveness at extreme outcomes. Because the spiky nature of power prices has been one of the motivations for regime switching methods, it seems appropriate that, when combinations include regime switching methods, the accuracy of the combination should be assessed not only in terms of the expected value, but also on a quantile defined value-at-risk ("expected shortfall") measure. In this research, we are therefore motivated to analyse the results using a number of error metrics including expected shortfall.

Many research papers have suggested that combining will perform better than individual methods (Clemen, 1989; Clements and Hendry, 1998; de Menezes et al., 2000; Riedel and Gabrys, 2005; Altavilla and De Grauwe, 2006; Timmermann, 2006; Chen and Yang, 2007; Clark and McCracken, 2009), including some applications to electricity demand forecasting (see Taylor and Majithia, 2000; Taylor, 2010). In the context of electricity prices, García-Martos et al. (2007) similarly advocate combining, but within a single model class (ARIMA), to deal with specification uncertainty. Despite the volume of comparisons published, it is an open question how many of the results in favour of combining are actually statistically significant. Moreover, in addition to this question, we are careful in our comparisons to consider, not simply the usual ex post evaluation of whether combining would have outperformed the best individual methods, but the more realistic setting of whether combining would have performed better than the individual method which would have be chosen ex ante. Given that part of the motivation for combining is that individual model performances are unstable, it is important to evaluate the procedures with a backtesting experiment that incorporates this unstable model selection aspect in a simulated ex ante way.
The paper is organized as follows. In Section 2 we present the price data from the UK Power Exchange (UKPX). The individual models and price drivers included therein as regressors are described in Section 3. Section 4 introduces the combination methodology and explains how the forecasts are evaluated. Section 5 contains the experimental design and the results of our work. Section 6 concludes.

2 The data

This work considers price data from the UK Power Exchange (UKPX) for the period April 1st, 2005 - September 30th, 2006: the choice of the starting date is important because it refers to the market that had just been extended to include Scotland.
The British power market is considered to be a fully competitive market and one of the most mature in the world (see Karakatsani and Bunn, 2008b for a detailed exposition).
The price series have half-hourly frequency, so that each day consists of 48 observations, one for each load period. We denote by P_{jt} the spot price at day t and load period j ($t = 1, 2, ..., N$, $j = 1, 2, ..., 48$). Since our interest lies mainly in price modelling and prediction during working days, weekends and holidays were removed from the data following the approach used by Ramanathan et al. (1997) and Karakatsani and Bunn (2008a), among others. Moreover, in adopting an intradaily approach, we consider separately each load period, according to a well-established precedent for electricity loads and prices (Ramanathan et al., 1997; Bunn, 2000; Bunn and Karakatsani, 2003). Results were analysed in detail for five representative periods of the day: load periods 6 (02:30-03:00am), 18 (08:30-09:00am), 28 (13:30-14:00pm), 38 (18:30-19:00pm) and 44 (21:30-22:00pm). The night-time load period 6 is the least volatile; periods 18, 28 and 38 represent peak hours, and show a high volatility with sudden peaks during winter and summer in both 2005 and 2006. Finally, period 44 is relatively stable, with moderate volatility. These characteristics are common in electricity price dynamics as indicated, amongst others, in Huisman and Mahieu (2003) and Knittel and Roberts (2005).
Each series has length $n = 380$. Figure 1 contains the plots of the five log-price time series considered; the logarithmic transformation was used to stabilize variance. The log-price series show neither a well-defined long-run behaviour nor a clear seasonal dynamics. However, levels change with the seasons, with an increase during the winter season. Moreover, the application of unit root tests indicates that the series are not stationary. In fact, the Augmented Dickey-Fuller test (Said and Dickey, 1984) rejects the null hypothesis of unit root only for period 28 and KPSS test (Kwiatkowski et al., 1992) always rejects the null hypothesis of stationarity (see Table 1).

Since some of the models considered or analysis require stationarity, in order to meet this requirement we assume that each series is the sum of a non stationary level component D_{jt}, describing level changes and/or long term and/or semi-periodic behaviour, and a residual stationary stochastic component p_{jt}, formally $\log P_{jt} = D_{jt} + p_{jt}$.
In the present work, the D_{jt} component has been estimated once for all by using a nonparametric technique based on the nearest neighbors method, also known as Friedman supersmoother (Friedman, 1984). The resulting series $p_{jt} = \log P_{jt} - D_{jt}$ are clearly stationary as can be seen in the right panel of Figure 1 and confirmed by both the ADF test and the KPSS test (see Table 1). In the following they will be referred as adjusted series.
Moreover, since here we are mainly interested in the relative predictive performance among a set of models and their combinations, we will focus on the prediction of

Table 1: *Unit root tests for* $logP_{jt}$ *and* p_{jt}*. Symbols* *, ** *mean that the null hypothesis is rejected at* 1% *and* 5% *significance level respectively. In the ADF test, lag lengths are chosen following Ng and Perron (1995) method.*

	$\log P_{jt}$		p_{jt}	
Load Period	ADF	KPSS	ADF	KPSS
6 (02:30-03:00am)	-1.981	0.958^{*}	-7.795^{*}	0.015
18 (08:30-09:00am)	-2.973	0.829^{*}	-6.917^{*}	0.017
28 (13:30-14:00pm)	-3.537^{**}	0.417^{*}	-6.372^{*}	0.015
38 (18:30-19:00pm)	-2.442	1.002^{*}	-7.309^{*}	0.014
44 (21:30-22:00pm)	-2.455	0.914^{*}	-7.555^{*}	0.016

p_{jt}, whereas the D_{jt} component is fixed and equal for all models and combinations.

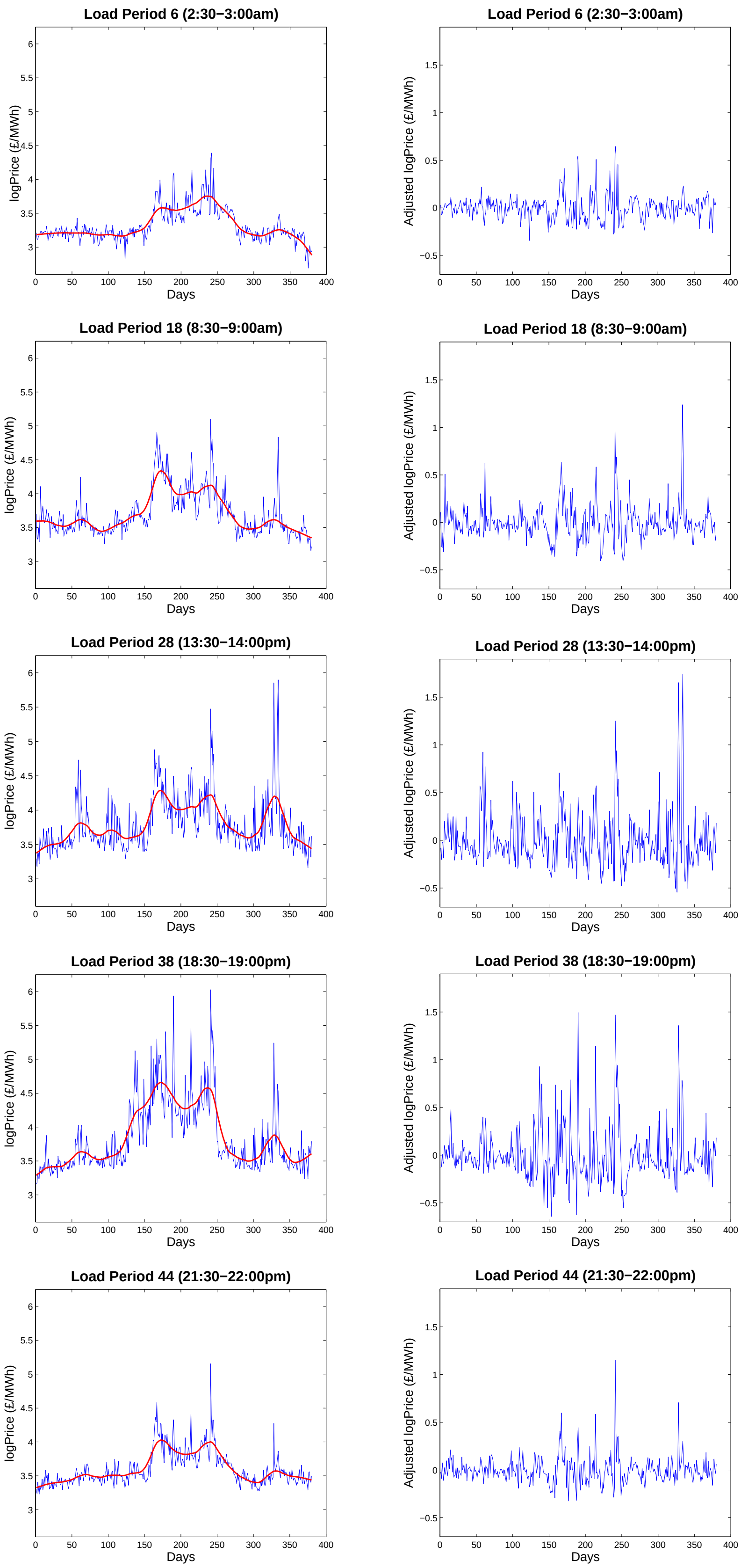
3 Individual forecasts

The individual models involved in this study are chosen because each of them is, potentially, very suitable to describe some specific features of the price dynamics. All models are based on a set of explanatory variables (in the log scale) that are strongly linked with the price evolution (see, Karakatsani and Bunn, 2008a among others), namely:
- the *Demand Forecast*, the national day-ahead demand forecast published by the system operator for each load period at time $t-1$;
- the *Indicated Margin*, the available capacity margin, defined as the difference between the sum of the maximum export limits nominated by each generator prior to each trading period, as its maximum available output capacity, and the demand forecast;
- the *Gas Price*, the daily UK natural gas one-day forward price, from the main National Balancing Point (NBP) hub. This is included because of its strong relation with power prices, especially during winter spikes. In particular, the series of deviations of gas prices from its deterministic component was considered;
- *Past Prices*, in particular, lags 1 and 5, corresponding to the previous day price and to the previous week price;
- *Volatility*, an indicator of instability and risk for both the electricity price series and for the demand forecast series. It is defined as the coefficient of variation computed on a rolling windows of the last 5 days.

The values at time $t-1$ of the first three variables represent forecasts for the next day. To face possible non linear relations between price and demand, and price and margin, quadratic polynomials of demand and margin were introduced. The individual forecasting models used in this study are:

- an ARMAX(p, q, r) model, where p and q are respectively the orders of the autoregressive and moving average parts, r is the order of the exogenous variable.

Figure 1: *Left panel: log-price time series,* $logP_{jt}$*, with superimposed* D_{jt} *for the period April 2005 - September 2006. Right panel: the adjusted series* p_{jt}*.*



In particular, for our dataset the identified model is the ARMAX(1,1,1).

$$p_{jt} = \phi_j p_{j(t-1)} + \varepsilon_{jt} + \theta_j \varepsilon_{j(t-1)} + \beta_j z_{j(t-1)}, \quad \varepsilon_{jt} \sim WN(0, \sigma_j^2), \tag{1}$$

where $z_{j(t-1)}$ is the indicated margin representing the exogenous variable, ε_{jt} is the error term and $\phi_j, \theta_j, \beta_j$ are constant coefficients. This model captures gradual adaptation through the the serial correlation in the adjusted log price series and immediate shocks in pricing caused by scarcity. It was estimated through maximum likelihood methods.

- a conventional constant parameter regression model (LR), which accounts for relations between prices and the various price drivers. The model is specified as:

$$p_{jt} = \boldsymbol{\beta}_j' \mathbf{X}_{jt} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim WN(0, \sigma_j^2) \tag{2}$$

where $\boldsymbol{\beta}_j$ is a $k \times 1$ vector of constant coefficients, $\mathbf{X}_{jt}$ is the $k \times 1$ vector of regressors and ε_{jt} is an error term. The regressors are selected with stepwise backward techniques (AIC criterion) among the variables described above. The estimation was performed through maximum likelihood methods.

- a time-varying parameter regression model (TVR), with random walk parameters, allowing for price driver effects that continuously evolve:

$$p_{jt} = \boldsymbol{\beta}_{jt}' \mathbf{X}_{jt} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim WN(0, \sigma_{\varepsilon_j}^2), \tag{3}$$

$$\boldsymbol{\beta}_{j(t+1)} = \boldsymbol{\beta}_{jt} + \boldsymbol{\nu}_{jt}, \quad \boldsymbol{\nu}_{jt} \sim WN_k(0, \mathbf{H}_j), \tag{4}$$

where $\boldsymbol{\beta}_{jt}$ is a vector of time-varying coefficients, $\mathbf{X}_{jt}$ is the vector of regressors, ε_{jt} is the error term of the measurement equation and $\boldsymbol{\nu}_{jt}$ is the error term vector of the transition equation. It is assumed that $\mathrm{E}(\varepsilon_{jt}\boldsymbol{\nu}_{jt}) = 0$ and $\mathbf{H}_j = \mathrm{diag}\{\sigma^2_{\nu_{jk}}\}$. For this model parameters were estimated using state space methods and the Kalman filter (Hamilton, 1994 and Durbin and Koopman, 2001).

- a Markov regime switching model (MS) which should capture spikes and discontinuities in price series, distinguishing between normal and high-price regimes. It is defined as:

$$p_{jt} = \boldsymbol{\beta}_{jS_t}' \mathbf{X}_{jt} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim WN(0, \sigma_{jS_t}^2), \tag{5}$$

$$\Pr(S_t = i | S_{t-1} = h) = \pi_{ih}, \quad \forall i, h \in S \tag{6}$$

where S_t is the latent regime at time t, $S = \{1, 2\}$ the set of possible states (say, base and peak), $\boldsymbol{\beta}_{jS_t}$ is the vector of coefficients in regime S_t, $\mathbf{X}_{jt}$ is the vector of regressors, $\sigma^2_{jS_t}$ the error variance in regime S_t and π_{ih} the transition probability between states i and h.
Maximum likelihood estimates of β_{jS_t} and $\sigma^2_{jS_t}$ are performed using the EM algorithm while for smoothed inferences of regimes, Kim's algorithm was used (Hamilton, 1994; Kim, 1994). The estimation procedure was applied referring both to the expanding dataset case (MS) and to the 6 month rolling windows

case (MS6). Once a MS model has been estimated, price forecasts are calculated as the linear combination of predicted prices across regimes weighted by predicted regime probabilities.

The regressors that were significant, at the 5% level, in the five different load periods are listed in Table 2. As can be seen, different periods have different significant specifications.

Table 2: *Final sets of regressors obtained with stepwise backward techniques.*

	Period 6	Period 18	Period 28	Period 38	Period 44
$intercept$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
p_{t-1}	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
$demF_{t-1}$	—	$\surd$	$\surd$	$\surd$	—
$demF^2_{t-1}$	—	$\surd$	—	$\surd$	—
$margin_{t-1}$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
$margin^2_{t-1}$	$\surd$	—	—	$\surd$	—
$gasF.res_{t-1}$	$\surd$	$\surd$	—	—	$\surd$
$demVol_t$	$\surd$	—	—	—	—
$priceVol_t$	—	—	—	—	$\surd$

4 Combining forecasts

In general, a forecast combination based upon a set of K competing spot price predictors producing forecasts $\hat{P}_t^{(1)}, ..., \hat{P}_t^{(K)}$ of P_t, based on the information available up to time $t-1$, is given by:

$$\hat{P}_t^C = f\left(\hat{P}_t^{(1)}, ..., \hat{P}_t^{(K)}; \boldsymbol{\theta}\right) \tag{7}$$

with f a generic function, possibly nonlinear, and $\boldsymbol{\theta}$ a parameter vector. Using linear functions, expression (7) becomes

$$\hat{P}_t^C = \sum_{k=1}^{K} \theta_k \hat{P}_t^{(k)}. \tag{8}$$

where the vector $\boldsymbol{\theta}$ optimizes some criterion. Several studies have shown that, due to the effect of finite-sample error in estimating the combining weights, an equally weighted mean is often the best choice (Makridakis and Winkler, 1983; Clemen, 1989; Stock and Watson, 2001, 2004; Smith and Wallis, 2009). We follow this conclusion and in the rest of the paper we assume $\theta_k = 1/K$.

In our case, the forecasts derive from the models described in the previous section[1], and thus, for each trading period there are five forecasts of the same spot

[1] Here we consider as different predictive models, the Markov switching models based on the expanding dataset (MS) and the 6 months rolling windows (MS6)

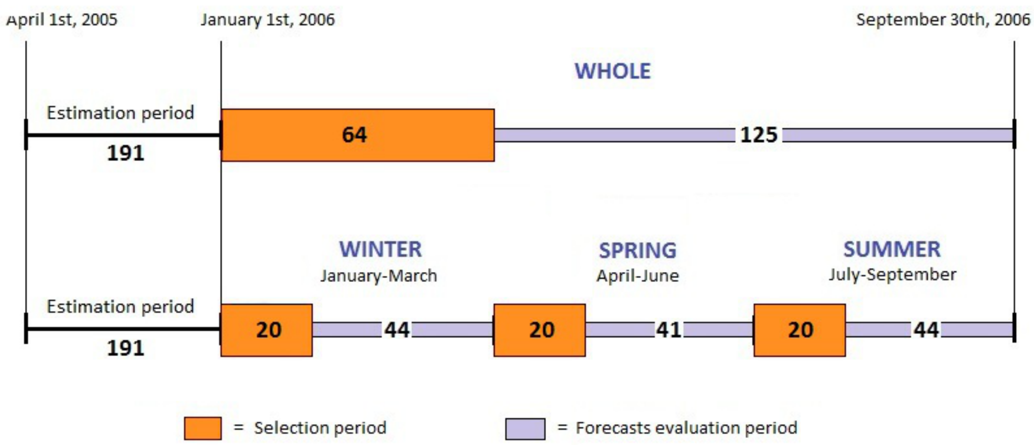
price, P_{jt} that can be considered singularly or combined.
Although the final price predictions would be given by

$$\hat{P}_{jt} = \exp(D_{jt} + \hat{p}_{jt}) \tag{9}$$

with $\hat{p}_{jt}$ the prediction of p_{jt}, when we refer to out-of-sample predictions we mean that we are considering out-of-sample forecasts of p_{jt}. Note that, although this is not a real out-of-sample prediction of P_{jt} because D_{jt} has been estimated with a smoother and not predicted, in our context this approach does not affect relative conclusions because all models are equally favoured or penalized by D_{jt}.
The whole dataset (April 1st, 2005 - September 30th, 2006) was divided into three parts. The first part, covering the period April 1st, 2005 - December 31th, 2005, is used only for individual model estimation. The remaining period (January 1st, 2006 - September 30th, 2006, 189 data) has been divided in further two parts: 1/3 is used to calibrate combined forecasts, i.e. to select the constituents of the combination, and 2/3 to out-of-sample forecasts evaluation (see Figure 2). Moreover, to compare the relative forecasting performances between individual models and combinations of the forecasts, 4 forecasting (sub-)periods were considered: the first three are associated with the different seasons (January-March, 44 data; April-June, 41 data; and July-September, 44 data) while the fourth includes the three seasons (January 1st, 2006 - September 30th, 2006, 125 data). The reason is to detect how much the forecasting accuracy of the predictions is influenced by the period of the year as well as by the considered trading period.

In our analyses comparisons are made on two levels: firstly we considered the

Figure 2: The framework of the prediction experiment (numbers in bold are sample sizes).



forecasting performance with respect to the following four statistics

$$\text{MSE} = \frac{1}{m}\sum_{t=1}^{m}(P_{jt} - \hat{P}_{jt})^2 \qquad \text{MSPE} = \frac{1}{m}\sum_{t=1}^{m}\left(100 \times \frac{P_{jt} - \hat{P}_{jt}}{P_{jt}}\right)^2$$

$$\text{MAE} = \frac{1}{m}\sum_{t=1}^{m}\left|P_{jt} - \hat{P}_{jt}\right| \qquad \text{MAPE} = \frac{1}{m}\sum_{t=1}^{m}\left|100 \times \frac{P_{jt} - \hat{P}_{jt}}{P_{jt}}\right|$$

with m the length of the forecasting period. We considered the significance of the difference in forecasting accuracy by means various tests, i.e. the Diebold and Mariano test (Diebold and Mariano, 1995), whose null hypothesis is that of no difference in the accuracy of two competing forecasters; a test based on the MCS (Model Confidence Set) procedure of Hansen et al. (2003, 2005) that, for two models, is similar to the Diebold and Mariano test but it estimates the distribution of the test statistic by a bootstrap procedure; and a test of forecast encompassing, whose null hypothesis is that predictions based on a model (for example CC) do not contain additional information with respect to those based on a second model (for example CI; in this case we say that CI encompasses CC). In the research literature, several formulations of encompassing test have been suggested (Newbold and Harvey, 2004; Clements and Harvey, 2007); here we adopted the specification given by Harvey et al. (1998), i.e. the modified Diebold and Mariano test statistic with demeaned forecasting errors. In the first two tests the equivalence between predictors is assessed with respect to some specified loss functions: here we considered mean square error (MSE) and mean absolute error (MAE). All tests were reported at the 5% significance level.

5 Comparing individual model forecasts and combinations of forecasts

Forecasting performances of the individual models and combinations are evaluated distinguishing among the 5 load periods ($j = 6, 18, 28, 38, 44$) referring to the trading hour of the day, 4 forecasting 'seasons' (3 'seasons' and the whole period) 4 prediction error statistics (MSE, MSPE, MAE, MAPE) and, when the Diebold and Mariano and/or the MCS tests are involved 2 loss functions (squared errors and absolute errors).

According to the approach followed by Hibon and Evgeniou (2005), all comparisons are performed from two different perspectives. Firstly we compare ex post the predictive performance of the best individual model (BI) with that of the best combination (BC). Since the evaluation is made ex post, this is not an out-of-sample prediction and it only allows us to check if there exists a combination giving better predictive performance than individual forecasts. Obviously, results are related to the specific models we considered.

In a second step, the comparisons are made considering models that have been selected in-sample and, thus, they account for possible misspecifications and/or estimation errors. We denote by CI the chosen individual model and by CC the chosen combination. In this case, out-of-sample predictions are involved.

The model selection is performed minimizing, in the validation period, one of the prediction error statistics described above and thus the models selected with respect to different indicators are not necessarily the same and, indeed, usually differ. When the descriptive indicators are involved, our study involves 80 cases (5 load periods $\times$ 4 'seasons' $\times$ 4 indicators). The number of cases scales consequently if some element (load period, 'season' or indicator) is kept fixed.
The results are graphically summarized, for the whole period case, in Figures 3-4. For example, the panel in position (1,1) of Figure 3 shows for the load period 6 and the MSE indicator the predictive performances in the out-of-sample forecasting period. The five points on the left represent the values of MSE corresponding to our five models, while the 26 points on the right relate to the MSE associated to the 26 possible combinations of 2, 3, 4 or 5 individual forecasts. The best/worst ex post individual model and combination, corresponding to the minimum/maximum value of the indicator, are reported in the figure. In this case the best performance is obtained with the forecasts combination of three models TVR, MS and ARMAX, which outperforms the best individual model MS. The arrows denote the MSE associated with the model/combination chosen in-sample. Note that, although there are 26 possible combinations and only 5 models, the comparison is fair because, in both categories, we consider only the model selected in-sample. The range of the MSE values can be interpreted as a measure of selection risk among individual forecasts or among combinations.
Detailed results are given, for all cases, in Tables 3-7, where we list the exact prediction error indicators and the p-values i) of the one-sided Diebold and Mariano test for the null hypothesis that best (chosen) individual forecasts have the same accuracy of the best (chosen) combined forecasts; ii) of the MCS test for the same hypothesis and iii) of the forecast encompassing for the null hypothesis that individual model predictions contain all the information contained in the combined predictions. Diebold and Mariano and MCS tests are performed with respect to loss functions based both on squared (rows MSE) and absolute errors (row MAE). This implies that the total number of comparisons is 160. Since the chosen models are different for different indicators, we have different p-values corresponding to different indicators. Table 8 lists a summary of the comparisons.
Table 9 contains the differences of performances of individual and combined forecasts with respect to the best possible performance (B), that is the minimum value of the prediction error statistics chosen ex post among all individual and combined forecasts. In particular, it lists the difference of performance, with respect to the best case, of the worst and of the chosen individual and combined forecasts. This gives us information about the riskiness of the two approaches.

5.1 Ex post analyses

In this first battery of analyses we compare, ex post, the best individual forecasts, among our five models, and the best combination of the predictions based on these models. The findings (see Figures 3-4 and Tables 3-7) highlight that, in general, combined models show better prediction ability in terms of prediction error statis-

tics. If we consider all the 80 comparisons[2], in 76% of them, the best possible forecasting model, obtained among all the individual models and all the combinations for each measure, is a combination (see also Table 8). Moreover, the worst performance - among all individual and combined forecasts - is always given by an individual model, so that selecting among combinations seems to be less risky than among individual models.
However, when we analyze the significance of the forecasting performance by means of tests (DM, MCS, encompassing), the predictive accuracy of the best combination is significantly better than that of the best individual model in only 8.75% of the 160 comparisons[3], according to the DM test and in 3.75% according to MCS test. On the contrary, however, for both tests the individual model accuracy never significantly outperforms that of the best combination (see also Table 8).
In general, our analyses indicate that the best performances are obtained combining predictions of only two or three models. For example, considering the MAPE indicator in Figure 4, the best performing combination for the least volatile load period 6 and for the peak load period 38 is obtained with the models TVR, MS and ARMAX. This agrees with previous research: it has been argued that, rather than combining the full set of forecasts, it is often advantageous to discard the models with the worst performance (see, for instance, Aiolfi and Favero, 2005; Granger and Jeon, 2004; Marcellino, 2004; Stock and Watson, 2001, 2004). However, in our study some exceptions emerge when the worst predictive model is the TVR. In 7 cases, for the whole forecasting period (load periods 6, 18 and 44), and in 2 cases, during summer (load period 6), the best combination contains this (the worst performing) model.

5.2 Ex ante analyses

We focus now on the forecasting comparison of models chosen ex ante, as it might happen in practice. Thus, when models have to be selected, there is the risk that the chosen model is much worse than the best possible choice in terms of out-of-sample accuracy. For each period, the ex ante selection process considers individual methods and combinations.
For these analyses the series have been divided into three parts (see also Figure 2): an estimation period, coinciding with the in-sample period for the ex post analysis; a validation period, of length 1/3 of the remaining data[4], used to enable the selection of the best individual model and combination ex ante and a forecasting period given by the last 2/3 of data[5], used for out-of-sample comparisons among models.
With respect to the indicators, the results are similar to those of the ex post case: the selected combined predictions produce forecasting error statistics lower than the selected individual model predictions in about 79% of cases (for detailed results see Tables 3-8).
However, the situation is quite different from the corresponding ex post case when

[2] 5 load periods × 4 'seasons' × 4 indicators
[3] 5 load periods × 4 'seasons' × 4 indicators × 2 loss functions
[4] 64 data for the whole period and 20 data for the subperiods
[5] 125 data for the whole period and 44 or 41 data for the subperiods

we consider the statistical significance of the difference in out-of-sample forecasting accuracy. Indeed, combined predictions are significantly more accurate than individual model predictions in 33.13% of cases for D-M test and 18.13% for MCS test. The contrary is true only in 1.25% of cases for DM test and only in 0.63% of cases for MCS test (for detailed results see Tables 3-8). This points out the benefit in choosing among combinations in ex ante situations: our findings indicate that, in general, we obtain forecasts that are more accurate than selecting among the individual models, and when they are not more accurate, they are almost always not worse. Similar conclusions can be drawn with respect to the encompassing test: globally, the hypothesis that the chosen single forecasts contain the same information as the chosen combined forecasts is rejected 1/3 of times.

5.3 Risk analysis

Our third way to compare individual forecasts and combined forecasts is through the analysis of risks. In this regard, two interpretations of risk were considered. The first one refers to the risk of an incorrect individual model or combination selection, that is the risk of choosing a model or a combination that is not the best. We call this selection risk. The second kind of risk is that related to the probability of incurring in large prediction error and we call it prediction risk.

With respect to the selection risk, Table 9 shows that - in terms of performance indicators - the distance from the globally best predictor (that is, the best predictor among combinations and individual models, B) is generally smaller for the combination (compare column "CC-B" of Table 9 with respect to column "CI-B"). This suggests that combining forecasts is less risky.

As a measure of prediction risk the so-called Expected Shortfall (ES), the average forecasting error exceeding a specified quantile of the forecasting error distribution, was considered. To have reliable results, this kind of analysis was performed only for the whole period and for the quantiles, 95% and 97.5%. Moreover, in order to compare the Expected Shortfalls a simple rule was adopted: we say that the forecast combination is better than individual forecasts if the reduction in the ES is at least 5% (and viceversa). Interpreting our results, although in most of cases the differences are smaller than 5%, the combination led to improvements which are larger than 5% in about 35% of cases, while improvements larger than 5% for individual models occur only in about 7.5% of case.

6 Summary and conclusions

We have compared the relative forecasting performances of five individual models and simple average combinations. The summary findings are as follows:

- in ex post comparisons, although the combined forecasts perform better than individual forecasts in 76% of cases, only in a few cases they are also significantly more accurate at the 5% level;

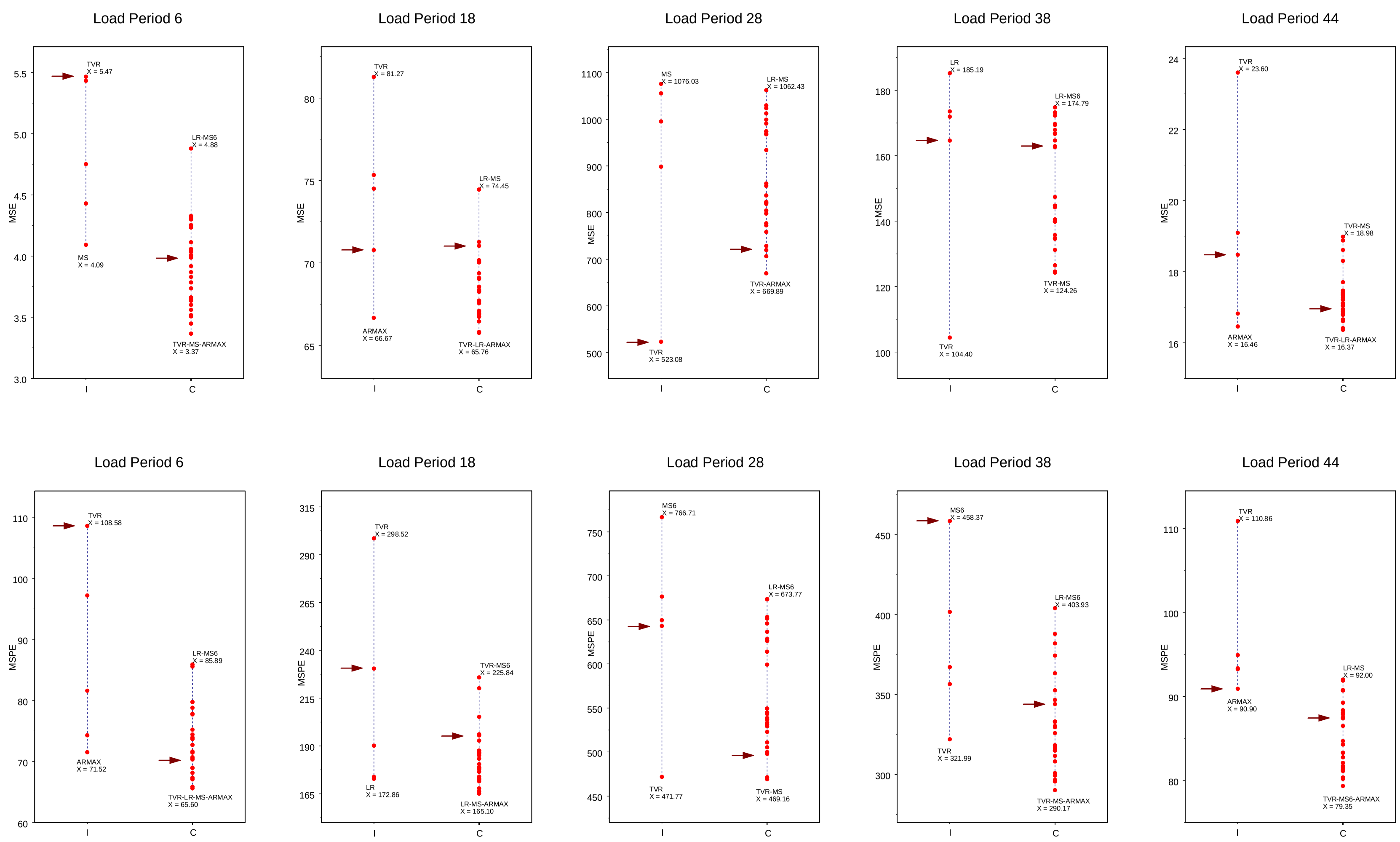
- in ex ante comparisons, when out-of-sample predictions are involved, the general indications are not very different but quite different in terms of the significance of the improvements. Indeed, when the analyses are based on individual and combined forecasts obtained through in-sample selection, the latter is significantly more accurate than individual forecasts in about 33% of cases. On the contrary, individual forecasts are more accurate in only 1% of cases. Thus, within the limit of our data and of the considered models, we can conclude that in about 99% of cases, seeking a combination of forecasts leads to predictions more accurate than or equivalent to those obtained through seeking to identify the individually best forecasts;

- our study stresses also that choosing an individual model out of a set of models is more risky than choosing among combinations of their forecasts and that combining is effective under value at risk criteria as well as for average accuracy.

In terms of the sensitivity of these results, it is worth noting that very similar results were obtained by considering adaptive weights, following Bates and Granger (1969), rather than simple averaging. Interestingly, similar results can be obtained by using all five methods in the combination rather than a chosen subset, but only if the adaptive weights are used instead of simple averaging. It is intuitive that if the task of optimising a subset is avoided, there is a compensating need to use optimal weights.

Finally, these analyses provide further indications of the specification difficulties in modelling electricity prices. The fact that a simple combination of a subset of quite sophisticated methods such as Markov regime switching and time varying regressions, as well as ARMAX and linear regression, provides a more accurate forecasting procedure, points to the inadequacies in each of these methods and/or the ability to select the best performing one reliably.

Figure 3: *Forecasting performances of the individual models (I, on the left inside each figure) and of all the combinations (C, on the right). The arrows indicate the value of the indicator (MSE in the first row and MSPE in the second row) for the models chosen in-sample. Results refer to the whole out-of-sample period (125 data).*



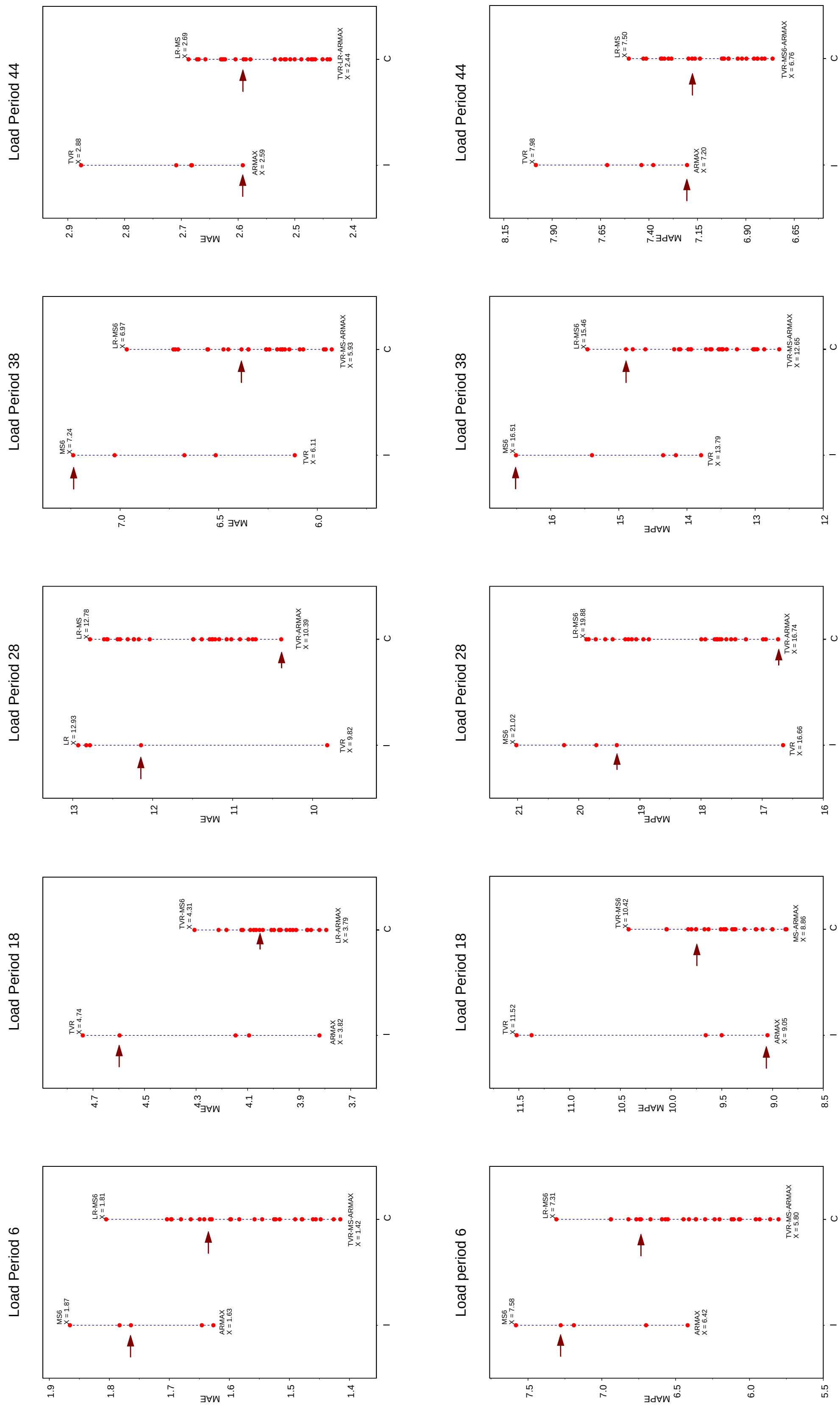


Figure 4: *Forecasting performances of the individual models (I, on the left inside each figure) and of all the combinations (C, on the right). The arrows indicate the value of the indicator (MAE in the first row and MAPE in the second row) for the models chosen in-sample. Results refer to the whole out-of-sample period (125 data).*

Table 3: *Load period 6. Prediction error statistics values and p-values for the Diebold-Mariano, MCS and encompassing tests. BI = best individual model (ex post); BC = best combination (ex post); CI = chosen (ex ante) individual model; CC = chosen (ex ante) combination.*

	Whole				**Winter**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	4.092	71.521	1.627	6.419	56.454	266.751	5.291	12.089
BC	3.366	65.598	1.415	5.803	56.140	236.093	5.222	11.944
CI	5.466	108.583	1.764	7.278	69.946	368.523	5.972	14.267
CC	3.987	70.496	1.632	6.735	68.598	318.663	5.599	12.736
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.014	0.051	0.034	0.034	0.480	0.408	0.480	0.480
MAE	< 0.001	0.057	0.019	0.019	0.435	0.321	0.435	0.435
	MCS test p-values							
MSE	0.025	0.108	0.072	0.083	0.961	0.785	0.962	0.963
MAE	0.001	0.111	0.044	0.041	0.866	0.650	0.865	0.856
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.006	< 0.001	0.018	0.018	0.341	0.438	0.155	0.155
MAE	0.015	< 0.001	0.078	0.078	0.386	0.458	0.104	0.104
	MCS test p-values							
MSE	0.008	0.001	0.034	0.032	0.702	0.777	0.257	0.248
MAE	0.031	0.003	0.173	0.179	0.780	0.918	0.236	0.228
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.001	< 0.001	0.002	0.002	0.419	0.479	0.097	0.097

	Spring				**Summer**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	2.401	37.200	1.170	4.724	4.054	94.443	1.549	7.128
BC	2.419	40.024	1.280	5.280	3.755	94.321	1.391	6.558
CI	6.163	107.967	2.153	8.949	4.363	104.202	1.681	7.707
CC	3.813	68.011	1.662	6.956	4.099	101.553	1.391	6.558
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.463	0.463	0.463	0.463	0.121	0.234	0.301	0.301
MAE	0.058	0.058	0.058	0.058	0.039	0.229	0.037	0.037
	MCS test p-values							
MSE	0.920	0.921	0.923	0.924	0.204	0.465	0.578	0.576
MAE	0.226	0.215	0.219	0.219	0.108	0.421	0.078	0.067
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	< 0.001	< 0.001	< 0.001	< 0.001	0.185	0.065	0.149	0.149
MAE	< 0.001	< 0.001	< 0.001	< 0.001	0.290	< 0.001	0.002	0.002
	MCS test p-values							
MSE	< 0.001	< 0.001	< 0.001	< 0.001	0.302	0.101	0.342	0.358
MAE	< 0.001	< 0.001	< 0.001	< 0.001	0.540	0.001	0.004	0.005
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	< 0.001	< 0.001	< 0.001	< 0.001	0.037	0.081	0.196	0.196

Table 4: *Load period 18. Prediction error statistics values and p-values for the Diebold-Mariano, MCS and encompassing tests. BI = best individual model (ex post); BC = best combination (ex post); CI = chosen (ex ante) individual model; CC = chosen (ex ante) combination.*

	Whole				**Winter**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	66.670	172.855	3.822	9.050	280.442	330.798	8.908	13.517
BC	65.758	165.096	3.795	8.863	276.851	313.603	9.099	13.237
CI	70.780	230.445	4.597	9.050	355.069	378.061	10.476	14.738
CC	71.037	195.548	4.054	9.753	321.389	321.217	9.369	13.530
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.182	0.119	0.392	0.343	0.377	0.284	0.495	0.181
MAE	0.312	0.005	0.437	0.499	0.236	0.305	0.355	0.323
	MCS test p-values							
MSE	0.609	0.159	0.711	0.540	0.756	0.520	0.991	0.277
MAE	0.668	0.007	0.853	0.998	0.400	0.587	0.671	0.619
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.482	0.011	0.011	0.449	0.070	0.081	0.081	0.081
MAE	< 0.001	< 0.001	< 0.001	0.021	0.070	0.071	0.071	0.071
	MCS test p-values							
MSE	0.954	0.028	0.030	0.893	0.089	0.091	0.091	0.086
MAE	< 0.001	< 0.001	< 0.001	0.039	0.152	0.092	0.097	0.092
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.740	0.003	0.003	0.479	0.034	0.097	0.097	0.097

	Spring				**Summer**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	11.382	91.653	2.509	7.453	9.407	93.710	2.136	6.817
BC	11.396	83.109	2.543	7.439	9.740	95.559	2.031	6.499
CI	17.814	91.653	2.509	7.453	14.503	101.057	2.136	6.817
CC	13.169	92.395	2.578	7.598	15.219	95.797	2.146	6.874
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.497	0.482	0.497	0.497	0.335	0.172	0.300	0.300
MAE	0.433	0.423	0.433	0.433	0.264	0.395	0.239	0.239
	MCS test p-values							
MSE	0.993	0.965	0.995	0.995	0.456	0.345	0.601	0.598
MAE	0.863	0.836	0.868	0.855	0.480	0.747	0.415	0.411
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.003	0.314	0.314	0.314	0.301	0.191	0.191	0.191
MAE	0.001	0.365	0.365	0.365	0.454	0.442	0.442	0.442
	MCS test p-values							
MSE	0.003	0.655	0.664	0.673	0.608	0.415	0.417	0.420
MAE	< 0.001	0.721	0.708	0.722	0.897	0.879	0.874	0.868
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.005	0.640	0.640	0.640	0.797	0.326	0.326	0.326

Table 5: *Load period 28. Prediction error statistics values and p-values for the Diebold-Mariano, MCS and encompassing tests. BI = best individual model (ex post); BC = best combination (ex post); CI = chosen (ex ante) individual model; CC = chosen (ex ante) combination.*

	Whole				**Winter**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	523.078	471.771	9.816	16.658	924.282	590.634	15.881	19.254
BC	669.886	469.162	10.392	16.740	913.736	617.322	15.842	19.328
CI	523.078	643.093	12.145	19.377	934.600	590.634	15.881	19.254
CC	719.774	497.826	10.392	16.740	936.435	621.086	16.690	20.626
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.095	0.095	0.095	0.095	0.432	0.386	0.424	0.424
MAE	0.190	0.141	0.190	0.190	0.455	0.291	0.466	0.466
	MCS test p-values							
MSE	0.259	0.268	0.264	0.266	0.843	0.731	0.839	0.835
MAE	0.378	0.218	0.370	0.388	0.907	0.593	0.918	0.925
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.101	0.082	0.082	0.082	0.434	0.424	0.157	0.157
MAE	0.152	0.010	0.010	0.010	0.130	0.466	0.074	0.074
	MCS test p-values							
MSE	0.282	0.133	0.127	0.121	0.896	0.839	0.395	0.410
MAE	0.235	0.028	0.032	0.029	0.424	0.928	0.135	0.139
$\mathbf{H_0}$	**Encompassing test p-values**							
CI encompasses CC	0.228	0.151	0.151	0.151	0.990	0.765	0.449	0.449
	Spring				**Summer**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	68.973	433.676	6.382	16.075	60.680	382.361	5.513	14.849
BC	73.274	406.469	6.140	15.318	55.241	369.110	5.279	14.363
CI	88.021	448.339	6.382	16.509	81.937	545.883	6.561	17.682
CC	73.274	408.630	6.140	15.318	79.430	433.093	6.318	16.971
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.207	0.191	0.207	0.005	0.256	0.297	0.297	0.385
MAE	0.139	0.411	0.139	0.030	0.139	0.266	0.266	0.281
	MCS test p-values							
MSE	0.558	0.449	0.541	0.050	0.422	0.584	0.571	0.746
MAE	0.417	0.855	0.417	0.151	0.354	0.491	0.487	0.580
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.005	0.207	0.207	0.207	0.395	0.099	0.395	0.395
MAE	0.030	0.139	0.139	0.139	0.221	0.063	0.221	0.221
	MCS test p-values							
MSE	0.047	0.543	0.544	0.539	0.744	0.032	0.743	0.746
MAE	0.140	0.420	0.418	0.416	0.424	0.101	0.433	0.440
$\mathbf{H_0}$	**Encompassing test p-values**							
CI encompasses CC	0.001	0.998	0.998	0.998	0.323	0.023	0.323	0.323

Table 6: *Load period 38. Prediction error statistics values and p-values for the Diebold-Mariano, MCS and encompassing tests. BI = best individual model (ex post); BC = best combination (ex post); CI = chosen (ex ante) individual model; CC = chosen (ex ante) combination.*

	Whole				**Winter**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	104.403	321.991	6.114	13.793	2373.833	622.140	23.497	19.307
BC	124.260	290.172	5.927	12.646	2316.932	614.194	23.339	19.126
CI	164.629	458.370	7.238	16.509	2795.410	903.622	26.675	22.443
CC	162.906	343.986	6.385	14.896	2431.601	696.829	24.507	20.160
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.170	0.173	0.173	0.173	0.403	0.460	0.460	0.460
MAE	0.282	0.282	0.282	0.282	0.130	0.444	0.444	0.444
	MCS test p-values							
MSE	0.452	0.551	0.560	0.560	0.783	0.920	0.923	0.924
MAE	0.596	0.606	0.597	0.600	0.207	0.896	0.893	0.895
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.377	0.002	0.002	< 0.001	0.014	0.027	0.424	0.424
MAE	0.256	< 0.001	< 0.001	0.001	0.004	0.021	0.035	0.035
	MCS test p-values							
MSE	0.744	0.070	0.075	0.030	0.017	0.035	0.841	0.841
MAE	0.515	0.004	0.003	0.010	0.007	0.042	0.035	0.035
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.131	< 0.001	< 0.001	< 0.001	0.012	0.012	0.598	0.598

	Spring				**Summer**			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	39.878	236.531	4.298	11.162	29.902	254.507	4.023	12.019
BC	39.751	216.672	4.063	10.533	25.076	216.955	3.760	11.092
CI	55.925	314.892	4.872	12.845	36.825	292.681	4.673	13.588
CC	39.751	216.672	4.063	10.643	32.603	292.619	4.465	13.387
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.480	0.032	0.032	0.053	0.101	0.167	0.101	0.101
MAE	0.178	0.154	0.154	0.107	0.076	0.055	0.076	0.076
	MCS test p-values							
MSE	0.963	0.091	0.105	0.244	0.130	0.250	0.118	0.126
MAE	0.300	0.294	0.300	0.233	0.315	0.071	0.322	0.310
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.031	0.031	0.031	0.031	0.269	0.269	0.269	0.231
MAE	0.047	0.047	0.047	0.047	0.327	0.327	0.327	0.379
	MCS test p-values							
MSE	0.038	0.039	0.043	0.044	0.459	0.452	0.452	0.400
MAE	0.072	0.069	0.070	0.076	0.622	0.601	0.608	0.713
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.020	0.020	0.020	0.020	0.081	0.081	0.081	0.066

Table 7: *Load period 44. Prediction error statistics values and p-values for the Diebold-Mariano, MCS and encompassing tests. BI = best individual model (ex post); BC = best combination (ex post); CI = chosen (ex ante) individual model; CC = chosen (ex ante) combination.*

	Whole				Winter			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	16.462	90.904	2.592	7.203	309.909	276.057	7.840	11.983
BC	16.366	79.345	2.438	6.762	329.600	266.004	7.609	11.472
CI	18.478	90.904	2.592	7.203	360.028	284.415	8.885	12.126
CC	16.944	87.873	2.590	7.177	355.264	287.700	8.026	12.159
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.480	0.417	0.480	0.417	0.183	0.183	0.183	0.183
MAE	0.130	0.170	0.130	0.170	0.152	0.152	0.152	0.152
	MCS test p-values							
MSE	0.960	0.803	0.956	0.813	0.564	0.569	0.571	0.569
MAE	0.320	0.343	0.320	0.361	0.424	0.420	0.418	0.422
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.148	0.356	0.356	0.356	0.270	0.270	0.090	0.270
MAE	0.127	0.494	0.494	0.494	0.417	0.417	0.017	0.417
	MCS test p-values							
MSE	0.198	0.669	0.677	0.682	0.490	0.482	0.245	0.489
MAE	0.199	0.986	0.985	0.987	0.820	0.819	0.020	0.816
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.188	0.964	0.964	0.964	0.304	0.304	0.090	0.304

	Spring				Summer			
	MSE	MSPE	MAE	MAPE	MSE	MSPE	MAE	MAPE
Models	**Prediction error statistics values**							
BI	5.238	54.940	1.842	5.963	5.103	46.259	1.710	5.186
BC	5.136	52.830	1.905	6.170	5.072	45.764	1.684	5.089
CI	6.137	65.836	2.169	7.102	5.256	47.856	2.034	6.201
CC	6.144	64.316	2.064	6.682	5.072	49.075	1.845	5.817
	BI vs. BC							
Loss Function	**D-M test p-values**							
MSE	0.413	0.413	0.413	0.413	0.431	0.424	0.424	0.424
MAE	0.229	0.229	0.229	0.229	0.337	0.320	0.320	0.320
	MCS test p-values							
MSE	0.826	0.827	0.822	0.822	0.777	0.861	0.863	0.860
MAE	0.488	0.487	0.493	0.479	0.709	0.644	0.647	0.651
	CI vs. CC							
Loss Function	**D-M test p-values**							
MSE	0.496	0.496	0.496	0.496	0.154	0.292	0.009	0.112
MAE	0.231	0.231	0.231	0.231	0.061	0.040	< 0.001	0.032
	MCS test p-values							
MSE	0.992	0.992	0.994	0.991	0.136	0.581	0.010	0.153
MAE	0.475	0.464	0.465	0.467	0.175	0.105	0.003	0.080
$\mathbf{H}_0$	**Encompassing test p-values**							
CI encompasses CC	0.139	0.139	0.139	0.139	0.114	0.508	0.004	0.058

Table 8: *Summary of comparisons on the whole: percentage and, in brackets, number of cases. BI = best individual model (ex post); BC = best combination (ex post); CI = chosen (ex ante) individual model; CC = chosen (ex ante) combination.*

Prediction error statistics values					
	Whole	Winter	Spring	Summer	Totals
BC better than BI	80.00% (20)	80.00% (20)	55.00% (20)	90.00% (20)	76.25% (80)
BI better than BC	20.00% (20)	20.00% (20)	45.00% (20)	10.00% (20)	23.75% (80)
CC better than CI	85.00% (20)	70.00% (20)	80.00% (20)	80.00% (20)	78.75% (80)
CI better than CC	15.00% (20)	30.00% (20)	20.00% (20)	20.00% (20)	21.25% (80)
Significance of differences with D-M test (MSE and MAE loss functions)					
	Whole	Winter	Spring	Summer	Totals
BC better than BI	17.50% (40)	0.00% (40)	10.00% (40)	7.50% (40)	8.75% (160)
BI better than BC	0.00% (40)	0.00% (40)	0.00% (40)	0.00% (40)	0.00% (160)
CC better than CI	50.00% (40)	17.50% (40)	50.00% (40)	15.00% (40)	33.13% (160)
CI better than CC	2.50% (40)	0.00% (40)	0.00% (40)	2.50% (40)	1.25% (160)
Significance of differences with MCS test (MSE and MAE loss functions)					
	Whole	Winter	Spring	Summer	Totals
BC better than BI	12.50% (40)	0.00% (40)	2.50% (40)	0.00% (40)	3.75% (160)
BI better than BC	0.00% (40)	0.00% (40)	0.00% (40)	0.00% (40)	0.00% (160)
CC better than CI	47.50% (40)	15.00% (40)	37.50% (40)	12.50% (40)	28.13% (160)
CI better than CC	2.50% (40)	0.00% (40)	0.00% (40)	0.00% (40)	0.63% (160)
Encompassing test					
	Whole	Winter	Spring	Summer	Totals
CI encompasses CC	55.00% (20)	85.00% (20)	50.00% (20)	85.50% (20)	67.50% (80)

Table 9: *Differences of prediction error statistics values. Out-of-sample periods (125, 44, 41 and 44 data). B = best possible statistics value obtained among all individual models and all combinations, WI = worst value obtained among the individual models, WC = worst value obtained among the combinations, CI = value obtained with the chosen individual model, CC = value obtained with the chosen combination.*

	Whole				Winter				Spring				Summer			
Statistics	WI-B	WC-B	CI-B	CC-B	WI-B	WC-B	CI-B	CC-B	WI-B	WC-B	CI-B	CC-B	WI-B	WC-B	CI-B	CC-B
Load Period 6																
MSE	2.101	1.513	2.101	0.621	26.101	15.655	13.806	12.458	3.762	3.012	3.762	1.412	1.935	0.467	0.608	0.344
MSPE	42.985	20.289	42.985	4.897	132.430	118.312	132.430	82.570	70.767	58.402	70.767	30.812	74.579	25.749	9.881	7.232
MAE	0.450	0.390	0.349	0.217	1.338	1.023	0.750	0.377	0.983	0.820	0.983	0.491	0.299	0.266	0.290	0.000
MAPE	1.779	1.504	1.475	0.932	2.810	2.465	2.323	0.792	4.224	3.571	4.224	2.232	1.276	1.035	1.149	0.000
Load Period 18																
MSE	15.514	8.691	5.022	5.279	78.218	46.751	78.218	44.538	8.462	5.035	6.432	1.787	20.543	11.211	5.096	5.812
MSPE	133.427	60.748	65.349	30.452	64.458	31.624	64.458	7.614	103.256	63.95	8.544	9.285	174.625	94.032	7.348	2.087
MAE	0.945	0.511	0.802	0.259	1.568	1.060	1.568	0.461	1.051	0.581	0.000	0.069	0.924	0.449	0.106	0.115
MAPE	2.659	1.555	0.187	0.890	1.735	1.174	1.500	0.293	3.640	2.044	0.014	0.159	2.823	1.360	0.318	0.375
Load Period 28																
MSE	552.949	539.349	0.000	196.696	162.911	114.464	20.864	22.699	61.415	34.114	19.047	4.300	34.930	24.189	26.696	24.189
MSPE	297.548	204.611	173.932	28.665	169.797	158.353	0.000	30.452	561.288	274.211	41.869	2.160	235.200	162.703	176.773	63.983
MAE	3.115	2.968	2.329	0.576	2.062	1.941	0.039	0.848	2.615	1.439	0.241	0.000	1.639	1.039	1.282	1.039
MAPE	4.363	3.218	2.719	0.082	3.998	3.768	0.000	1.373	8.348	4.365	1.190	0.000	4.153	2.608	3.319	2.608
Load Period 38																
MSE	80.792	70.391	60.226	58.504	552.739	282.417	478.478	114.669	27.375	20.026	16.174	0.000	15.855	11.770	11.749	7.527
MSPE	168.198	113.762	168.198	53.813	578.899	211.887	289.429	82.635	192.962	120.513	98.22	0.000	179.707	139.799	75.726	75.664
MAE	1.311	1.040	1.311	0.458	6.358	3.167	3.336	1.168	1.691	1.166	0.808	0.000	1.253	0.798	0.913	0.705
MAPE	3.863	2.814	3.863	2.250	7.606	3.277	3.317	1.035	4.937	3.256	2.311	0.110	3.831	2.693	2.496	2.294
Load Period 44																
MSE	7.234	2.616	2.112	0.578	167.503	88.346	50.119	45.355	3.313	1.493	1.001	1.007	1.023	0.576	0.184	0.000
MSPE	31.520	12.654	11.559	8.527	348.45	133.603	18.411	21.695	35.261	17.380	13.006	11.486	10.173	5.026	2.093	3.311
MAE	0.438	0.249	0.153	0.152	2.779	1.515	1.276	0.417	0.544	0.381	0.327	0.222	0.350	0.230	0.350	0.161
MAPE	1.222	0.742	0.442	0.416	4.550	2.409	0.654	0.688	1.758	1.276	1.139	0.719	1.112	0.728	1.112	0.728

References

Aiolfi, M. and Favero, C. A. (2005). Model uncertainty, thick modeling and the predictability of stock returns, *Journal of Forecasting* **24**: 233–254.

Altavilla, C. and De Grauwe, P. (2006). Forecasting and combining competing models of exchange rate determination, *CESifo Working Paper Series No.5*, CESifo GmbH.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts, *Operational Research Quarterly* **20**: 451–468.

Bunn, D. W. (1975). A bayesian approach to the linear combination of forecasts, *Operational Research Quarterly (1970-1977)* **26**(2): 325–329.

Bunn, D. W. (1977). A comparative evaluation of the outperformance and minimum variance procedures for the linear synthesis of forecasts, *Operational Research Quarterly (1970-1977)* **28**(3): 653–662.

Bunn, D. W. (2000). Forecasting loads and prices in competitive power markets, *Proceedings of the IEEE* **88**(2): 163–169.

Bunn, D. W. (2004). *Modelling prices in competitive electricity markets*, Wiley.

Bunn, D. W. and Karakatsani, N. (2003). Forecasting electricity prices. Working Paper, London Business School.

Chen, D. and Bunn, D. W. (2010). Analysis of the nonlinear response of electricity prices to fundamental and strategic factors, *IEEE Transactions on Power Systems* **25**(4): 595–606.

Chen, Z. and Yang, Y. (2007). Time series models for forecasting: Testing or combining?, *Studies in Nonlinear Dynamics & Econometrics* **11**(1): Article 3.

Clark, T. E. and McCracken, M. W. (2009). Combining forecasts from nested models, *Oxford Bulletin of Economics and Statistics* **71**(3): 303–329.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting* **5**: 559–583.

Clements, M. P. and Harvey, D. I. (2007). *Palgrave Handbook of Econometrics. Volume 2: Applied Econometrics*, Basingstoke, New York: Palgrave MacMillan, 2006-2009, chapter 4 - Forecast combination and encompassing, p. 169.

Clements, M. P. and Hendry, D. F. (1998). *Forecasting Economic Time Series*, Cambridge University Press.

de Menezes, L. M., Bunn, D. W. and Taylor, L. W. (2000). Review of guidelines for the use of combined forecasts, *European Journal of Operational Research* **120**: 190–204.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**: 253–263.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*, Oxford University Press.

Friedman, J. H. (1984). A variable span scatterplot smoother, *Tecnical Report 5*, Laboratory for Computational Statistics, Stanford University.

García-Martos, C., Rodríguez, J. and Sánchez, M. J. (2007). Mixed models for short-run forecasting of electricity prices: Application for the spanish market, *IEEE Transactions on Power Systems* **22**(2): 544–552.

Granger, C. W. J. and Jeon, Y. (2004). Thick modeling, *Economic Modelling* **21**: 323–343.

Granger, C. W. J. and Ramanathan, R. (1984). Improved methods of combining forecasts, *Journal of Forecasting* **3**: 197–204.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton.

Hansen, P. R., Lunde, A. and Nason, L. M. (2003). Choosing the best volatility models: The model confidence set approach, *Oxford Bulletin of Economics and Statistics* **65**: 839–861. Supplement.

Hansen, P. R., Lunde, A. and Nason, L. M. (2005). Model confidence sets for forecasting models, *Working Paper 2005-7*, Federal Reserve Bank of Atlanta.

Harvey, D., Leybourne, S. and Newbold, P. (1998). Tests for forecast encompassing, *Journal of Business & Economic Statistics* **16**(2): 254–259.

Hibon, M. and Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations, *International Journal of Forecasting* **21**: 15–24.

Huisman, R. and Mahieu, R. (2003). Regime jumps in electricity prices, *Energy Economics* **25**: 425–34.

Karakatsani, N. and Bunn, D. W. (2008a). Forecasting electricity prices: The impact of fundamentals and time-varying coefficients., *International Journal of Forecasting* **24**: 764–785.

Karakatsani, N. and Bunn, D. W. (2008b). Intra-day and regime-switching dynamics in electricity price formation, *Energy Economics* **30**: 1776–1797.

Kim, C. J. (1994). Dynamic linear models with markov-switching, *Journal of Econometrics* **60**: 1–22.

Knittel, C. R. and Roberts, M. R. (2005). An empirical examination of restructured electricity prices, *Energy Economics* **27**(5): 791–817.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics* **54**: 159–178.

Makridakis, S. and Winkler, R. L. (1983). Averages of forecasts: Some empirical results, *Management Science* **29**: 987–996.

Marcellino, M. (2004). Forecast pooling for short time series of macroeconomic variables, *Oxford Bulletin of Economic and Statistics* **6**: 91–112.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion), *Journal of the Royal Statistical Society, Series A* **137**: 131–164.

Newbold, P. and Harvey, D. I. (2004). Forecast combination and encompassing, *in* M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, Blackwell Publishing, chapter 12, pp. 268–293.

Ng, S. and Perron, P. (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag, *Journal of the American Statistical Association* **90**: 268–281.

Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F. and Brace, C. (1997). Short-run forecasting of electricity loads and peaks, *International Journal of Forecasting* **13**: 161–174.

Reid, D. J. (1968). Combining three estimates of gross domestic product, *Economica* **35**: 431–444.

Reid, D. J. (1969). *A comparative study of time series prediction techniques on economic data*, PhD thesis, University of Nottingham, Nottingham.

Riedel, S. and Gabrys, B. (2005). Evolving multilevel forecast combination models - an experimental study, *Proceedings of NiSIS'2005 Symposium*, Albufeira, Portugal.

Said, S. E. and Dickey, D. (1984). Testing for unit roots in autoregressive moving-average models with unknown order, *Biometrika* **71**: 599–607.

Serati, M., Manera, M. and Plotegher, M. (2008). Modelling electricity prices: from the state of the art to a draft of a new proposal, *Working paper*, Fondazione ENI Enrico Mattei.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle, *Oxford Bulletin of Economics and Statistics* **71**(3): 331–355.

Stock, J. H. and Watson, M. W. (2001). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, *in* R. F. Engle and H. White (eds), *Festschrift in Honour of Clive Granger*, Cambridge University Press, Cambridge, pp. 1–44.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set, *Journal of Forecasting* **23**: 405–430.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting, *European Journal of Operational Research* **204**: 139–152.

Taylor, J. W. and Majithia, S. (2000). Using combined forecasts with changing weights for electricity demand profiling, *The Journal of the Operational Research Society* **51**: 72–82. Part Special Issue: OR and Strategy.

Timmermann, A. G. (2006). Forecast combinations, *in* G. C. W. J. Elliot, G and A. Timmermann (eds), *Handbook of Economic Forecasting*, Vol. 1, Amsterdam: Elsevier, pp. 135–196.

Weron, R. (2006). *Modelling and Forecasting Electricity Loads and Prices: A Statistical Approach*, Wiley, Chichester.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITA DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy