

Density Forecasting for Weather Derivative Pricing

James W. Taylor

and

Roberto Buizza

International Journal of Forecasting, 2006, Vol. 22, pp. 29-42.

Address for Correspondence:

James W. Taylor
Saïd Business School
University of Oxford
Park End Street
Oxford OX1 1HP, UK

Tel: +44 (0)1865 288927

Fax: +44 (0)1865 288805

Email: james.taylor@sbs.ox.ac.uk

Density Forecasting for Weather Derivative Pricing

Abstract

Weather derivatives enable energy companies to protect themselves against weather risk. Weather ensemble predictions are generated from atmospheric models and consist of multiple future scenarios for a weather variable. They can be used to forecast the density of the payoff from a weather derivative. The mean of the density is the fair price of the derivative, and the distribution about the mean is important for risk management tools, such as value-at-risk models. In this empirical paper, we use one to 10 day-ahead temperature ensemble predictions to forecast the mean and quantiles of the density of the payoff from a 10-day heating degree day put option. The ensemble-based forecasts compare favourably with those based on a univariate time series GARCH model. Promising quantile forecasts are also produced using quantile autoregression to model the forecast error of an ensemble-based forecast for the expected payoff.

Keywords: Density forecasting; Weather risk management; Weather ensemble predictions; GARCH; Quantile regression; Quantile autoregression.

1. Introduction

Weather derivatives are a convenient way for businesses, such as energy companies, to protect themselves against weather risk. Since weather is not a tradable asset and there is no liquid market, the standard Black-Scholes style pricing is not applicable. Furthermore, liquidity will never be as good as in traditional price-hedging markets, as weather derivatives are location-specific (Campbell and Diebold, 2005). Consequently, in general, weather derivatives are priced using the expected discounted value approach, which relies heavily upon the quality of weather forecasts (Zeng, 2000; Davis, 2001).

Weather variable density forecasts are important for pricing weather derivatives because they can be used to forecast the density of the payoff (see Cao and Wei, 2000). The mean of the payoff density is the fair price of the derivative, and the distribution about the mean provides information regarding the uncertainty, which is needed for risk management purposes, such as value-at-risk (see Duffie and Pan, 1997). In this paper, we consider weather ensemble predictions, which are a relatively new type of weather forecast produced by large meteorological models of the earth's atmosphere. An ensemble prediction for a weather variable consists of different future scenarios, which are known as ensemble members. The distribution of the scenarios can be used as a forecast of the conditional probability density function of the weather variable.

The weather derivatives market is dominated by energy utilities (Dischel and Barriau, 2002). Due to the importance of temperature in modelling electricity and gas demand, about 90% of the weather derivatives currently traded are based on temperature (Weather Risk Management Association, 2002). In this study, we investigate the use of one to 10 day-ahead temperature ensemble predictions in forecasting the density of the payoff from a temperature put option. More specifically, we focus on forecasts of the mean and quantiles of the conditional density. The θ quantile of the conditional density of a variable y_t is the value, $Q_t(\theta)$, for which $P(y_t \leq Q_t(\theta)) = \theta$. We compare the ensemble-based density forecasts of the derivative payoff with

those based on a univariate time series model for temperature with variance modelled as a generalized autoregressive conditional heteroskedastic (GARCH) process. This paper is a follow up to a study that considered density forecasting for temperature variables (Taylor and Buizza, 2004). Although useful in establishing that there is strong potential for the use of ensemble predictions in temperature density forecasting, that study did not address the pricing of weather derivatives, which is the focus of the current paper.

In order to be consistent with the lead times for the weather ensemble predictions, the put option that we consider has a 10-day contract duration commencing in the period immediately following the forecast origin. Our analysis certainly has relevance for longer contracts because improved forecasting over the first 10 days implies an improvement over the full duration of a longer contract. Furthermore, 10 day-ahead prediction is precisely what is required for pricing longer contracts when one is just 10 days from the expiration date. We analyse daily air temperature recorded at the five locations used in the electricity demand forecasting models at National Grid Transco, which is the company responsible for electricity transmission in England and Wales. Since hedging electricity load is one of the main uses for weather derivatives, temperature recorded at these locations is an obvious candidate for underlying reference in derivative contracts (see Torr o *et al.*, 2003).

Section 2 introduces the temperature put option, which we use in this paper as a basis for investigating the different forecasting approaches. In Section 3, we review the literature on univariate models for temperature density forecasting and then present univariate models for our UK temperature data. Weather ensemble predictions are described in Section 4. Section 5 compares forecasts of the expected payoff from the different methods, and Section 6 compares quantile forecasting performance. The final section provides a summary and conclusions.

2. Temperature derivatives

2.1. A put option for UK temperature

Temperature derivatives are usually written with cumulative heating degree days (HDDs) or cooling degree days (CDDs) as the underlying weather index (Dischel, 2002). An HDD is a measure of demand due to heating and is defined for each day as the amount by which temperature for that day falls below a specified benchmark, usually taken to be 18°C. If the temperature is above the benchmark, the HDD for that day is zero. The value of 18°C is considered to be the approximate temperature at which users switch their heating on or off. A CDD is a measure of energy demand due to air conditioning and it is defined as the amount by which temperature exceeds a benchmark value. In the UK, since energy demand due to air conditioning is considerably lower than due to heating, contracts written for cumulative HDDs are far more common than for cumulative CDDs.

The payoff from a weather derivative swap contract is proportional to the weather index. In view of this, some authors have investigated derivative payoffs by analysing cumulative HDDs (e.g. Brody *et al.*, 2002). Dischel and Barrieu (2002) comment that although swaps are popular with speculators in the weather market, they have little appeal for end-users, such as energy providers. They explain that the simplest and most common hedging transaction for energy providers and many other end-users is a put or a call option. A cumulative HDD put enables the energy provider to protect against periods with lower than expected HDD accumulations. In this paper, we consider point forecasting and density forecasting of the payoff from the HDD put option with the following payoff, P_t :

$$P_t = \max(S_t - W_t, 0) \quad \text{where} \quad W_t = \sum_{i=0}^9 \max(18 - T_{t-i}, 0) \quad (1)$$

where T_t is temperature, W_t is the 10-day cumulative HDD weather index, and S_t is the strike price. For simplicity, we set the tick rate (payment per unit of weather index) to be one and we did not cap the payoff. As in the work of Cao and Wei (2000), we set the strike price, S_t ,

to be equal to the W_t seasonal average. We generated a different strike price for each day of the year by averaging the 10-day centred moving average values of W_t from the same day in each year of our estimation sample. Since the duration of the contract is just 10 days, we did not incorporate discounting. For the UK data in this study, there is on average only about three weeks in the year when the W_t is zero, and so for simplicity we considered forecasting for each day of the year.

2.2. Pricing temperature derivatives

Since the expected value of a non-linear function of random variables is not necessarily the same as the non-linear function of the expected values of the random variables, it would be a mistake to forecast the expected payoff simply by substituting temperature point forecasts into expression (1). Instead, the standard approach is to simulate different scenarios from a model for temperature, and, using expression (1), create multiple realisations for the payoff (e.g. Cao and Wei (2000) and Davis (2001)). The mean of the payoff realisations is then the forecast of the expected value of the payoff, and the distribution of the realisations is the payoff density forecast. This paper compares the payoff mean and density forecasts from two different approaches to generating the temperature scenarios: a univariate time series model for temperature and an atmospheric model.

An alternative to a univariate temperature model is to model directly cumulative HDDs or the payoff. Brix *et al.* (2002) advise against modelling the payoff directly because its distribution has discontinuities, for example at zero. Cao and Wei (2000) note that modelling cumulative HDDs directly has the disadvantage that it would be specific to the contract duration. A useful discussion of weather derivative pricing is provided by Alaton *et al.* (2002).

3. Univariate time series modelling of UK temperature

Our temperature series were recorded daily at midday from 1 January 1994 to 1 July 2000 and measured at a height of 2 meters at the following five locations in the UK: Birmingham, Bristol, Heathrow, Leeds and Manchester. We used the first five years of each of our daily UK temperature time series to identify and estimate AR-GARCH models. In later sections, we use the remaining 18 months for post-sample forecast comparison. Figure 1 shows a plot of the Heathrow series. As one would expect, there is strong within-year seasonality in the mean of the series, and a reasonable degree of variation about that seasonal pattern.

***** Figure 1 *****

We used the same AR-GARCH models considered in our previous study (see Taylor and Buizza, 2004). These models were proposed by Franses *et al.* (2001) and Campbell and Diebold (2005), and are presented in expression (2).

$$\begin{aligned}
 T_t &= s(\boldsymbol{\mu}, t) + \phi T_{t-1} + \varepsilon_t \\
 \varepsilon_t &= \sigma_t \eta_t \\
 \sigma_t^2 &= s(\boldsymbol{\omega}, t) + \alpha (\varepsilon_{t-1} - s(\boldsymbol{\gamma}, t))^2 + \beta \sigma_{t-1}^2
 \end{aligned} \tag{2}$$

where T_t is temperature, ε_t is an error term, σ_t is the conditional standard deviation (volatility), η_t is an i.i.d. error term, ϕ , α and β are scalar parameters, and $\boldsymbol{\mu}$, $\boldsymbol{\omega}$ and $\boldsymbol{\gamma}$ are vectors of parameters.

The three seasonality terms were defined as

$$\begin{aligned}
 s(\boldsymbol{\lambda}, t) &= \lambda_0 + \lambda_1 \sin\left(2\pi \frac{d(t)}{365}\right) + \lambda_2 \cos\left(2\pi \frac{d(t)}{365}\right) \\
 &\quad + \lambda_3 \sin\left(4\pi \frac{d(t)}{365}\right) + \lambda_4 \cos\left(4\pi \frac{d(t)}{365}\right)
 \end{aligned} \tag{3}$$

where $d(t)$ is a repeating step function that numbers the days from 1 to 365 within each year. We removed 29 February from each leap year in our sample in order to maintain 365 days in each year. We used the standard approach of maximum likelihood to estimate parameters with a Gaussian assumption for η_t . Table 1 presents our estimated model for each of our five temperature series. For more details regarding these models, see Taylor and Buizza (2004).

***** Table 1 *****

The AR-GARCH models enable predictions to be made for the mean and variance at a given forecast horizon. A density forecast can then be constructed using a Gaussian assumption or the empirical distribution of standardised residuals (see Granger *et al.*, 1989).

4. Atmospheric Modelling and weather ensemble predictions

The atmosphere is a complex dynamical system with many degrees of freedom (Buizza, 2003). In numerical weather prediction, the state of the atmosphere is described by the spatial distribution of wind, temperature, specific humidity, liquid water content and surface pressure. The mathematical differential equations used to predict the system's time evolution include Newton's law of motion and the laws of thermodynamics. These equations include parameterisation schemes that simulate the effects of physical processes such as radiative transfer, turbulent mixing, orographic forcing, and moist processes. Numerical weather prediction models predict the time evolution of the atmospheric state by solving the system equations numerically, using finite difference and/or spectral methods. These equations are integrated starting from the best estimate of the current state of the system, which is determined, typically, using observations from earth, ocean and satellite platforms taken during the preceding 12 hours.

The dynamical equations that describe the time evolution of the atmospheric flow are capable of identifying the relationship between present and future variables in a more accurate way than statistical methods. This explains why numerical weather prediction models are currently used as a preferred way to predict the weather on time scales of up to one month. The area where statistical methods are still competitive with dynamical models is seasonal prediction, i.e. predictions for time ranges of one to several months. The key reason is that limited computer resources do not yet allow meteorologists to integrate coupled ocean-atmospheric models that describe all the important phenomena. The interested reader is referred to Pearce (2002, pp 333) for a more thorough discussion of these and related issues.

Ensemble prediction aims to derive a more sophisticated estimate of the density function of a weather variable than that provided by the distribution of past values. Ensemble prediction systems generate multiple realisations of weather variables by using a range of different initial conditions in the numerical model of the atmosphere. The frequency distribution of the different realisations, which are known as ensemble members, provides an estimate of the density function. The initial conditions are not sampled as in a statistical simulation because this is not practical for the complex, high-dimensional weather model. Instead, they are designed to sample directions of maximum possible growth (measured in terms of energy) during the forecast horizon. This choice is based on the hypothesis that the components of the initial uncertainties with the fastest growth have the dominant effect on the forecast error. Thus, knowledge of these directions will help to estimate the ‘bounds’ of the probability density function of forecast states (Molteni *et al.*, 1996; Palmer *et al.*, 1993; Buizza *et al.*, 1998; Buizza 2003)

The number of ensemble members is limited by the necessity to produce forecasts in a reasonable amount of time with the available computer power. Ensemble predictions are generated using a lower resolution (horizontal grid spacing) than traditional single point forecasts. The predictions used in this study were supplied by the European Centre for Medium-range Weather Forecasts (ECMWF), who operate a 51-member system (Buizza *et al.*, 1998). The 51 consist of one forecast started from the unperturbed, best estimate of the atmosphere initial state plus 50 others generated by varying the initial conditions. Ensemble forecasts are produced every day for lead times from 12 hours ahead to 10 days ahead. In our work, we used ECMWF ensemble predictions for midday air temperature, recorded from 1 January 1997 to 1 July 2000 at the five UK locations specified in Section 3.2.

Taylor and Buizza (2002, 2003) consider the use of weather ensemble predictions in electricity demand forecasting. The ensemble members are used to produce scenarios for demand, which are then used as a basis for estimating the uncertainty in a demand forecast.

5. Empirical comparison of forecasts of the expected payoff

In this section, we use the period 1 January 1999 to 1 July 2000 to compare 10 methods for forecasting the expected payoff from the derivative in expression (1) for each of the five UK temperature series. We produced forecasts for each 10-day rolling window in the 18-month post-sample period. The forecast origin was the day immediately prior to the 10-day period.

5.1. Forecasting methods for the expected payoff

Methods E1 to E6 are univariate time series approaches. The first two of these involve simulation based on well-specified AR-GARCH models, while Methods E3 to E6 are naïve benchmark approaches. Methods E7 to E10 use predictions from an atmospheric model.

Method E1 - The AR-GARCH models in Table 1 were used to produce mean and variance forecasts for one to 10 days ahead. Using these forecasts and a Gaussian distribution, 10,000 values were simulated for temperature for each lead time. Substituting these into expression (1) delivered 10,000 payoff realisations. The mean of these was used as the forecast.

Method E2 - This method was identical to Method E1 except it used the empirical distribution of standardised AR-GARCH residuals from each lead time, instead of a Gaussian distribution.

Method E3 - The AR-GARCH temperature point forecasts were substituted into expression (1).

Method E4 - A random walk forecast was created by using the actual payoff recorded for the 10-day period that ended immediately prior to the start of the contract period to be predicted.

Method E5 - The mean of the observed payoffs for the five most recent 10-day periods prior to the start of the contract period to be predicted.

Method E6 - The mean of the observed payoff for the same 10-day period from each of the previous five years.

Method E7 - Using expression (1) and the 51 temperature ensemble members for lead times from one to 10 days, 51 realisations were generated for the derivative payoff. The mean of these was used as the forecast.

Method E8 - Preliminary analysis suggested that the distribution of the 51 ensemble members tends to underestimate the true uncertainty. To address this, we rescaled each ensemble member. We applied the following simple transformation to $\hat{T}_{t+k|t}^i$, the i th ensemble member for k -step-ahead prediction from forecast origin t :

$$\mu_{t+k|t}^{ENS} + \lambda_k \left(\hat{T}_{t+k|t}^i - \mu_{t+k|t}^{ENS} \right) \quad \text{where} \quad \lambda_k = \sqrt{\frac{\sum_j (T_{j+k} - \mu_{j+k|j}^{ENS})^2}{\sum_{j=1}^m (\sigma_{j+k|j}^{ENS})^2}}$$

$\mu_{t+k|t}^{ENS}$ and $\sigma_{t+k|t}^{ENS}$ are the mean and standard deviation, respectively, of the 51 ensemble members. The rescaling inflates the deviation of each ensemble member from its mean by a factor λ_k , which is the square root of the ratio of historic k -step-ahead forecast error variance to the average historic k -step-ahead ensemble variance. For the Heathrow data, λ_k ranged from $\lambda_1=3.35$ to $\lambda_{10}=1.15$. We elected to use a simple rescaling transformation in order not to lose the information in the ensemble predictions. The rescaled ensemble members were used to generate 51 payoff realisations as in Method E7. The mean of these was used as the forecast.

Method E9 - The temperature ensemble means, $\mu_{t+k|t}^{ENS}$, were substituted into expression (1). The ensemble mean is used here as temperature point forecast because it has been found to be a more accurate point forecast than the traditional prediction produced by running the model just once at high resolution using the best estimate of the initial state of the atmosphere (Leith, 1974; Molteni, *et al.* 1996).

Method E10 - The transformation in Method E8 rescales the spread of the ensemble members but it does not address the possibility of bias in the ensemble mean, which may transfer to the forecast for the expected payoff. Using only the estimation sample, actual payoffs were regressed on the forecasts for the expected payoff from Method E8 to deliver a set of debiased forecasts.

5.2. Results

Table 2 shows the mean absolute error (MAE) for the post-sample forecast errors from the 10 methods for each of the five series. The relative performance of the methods was similar when evaluated using MedAE and RMSE. To get some perspective on the MAE values in Table 2, it is worth noting that, in the post-sample period, the actual Heathrow payoffs ranged between zero and 36.1, with a mean of 7.7 and a median of 4.6. For each of the five series, we also calculated the MAE ranking for each method. The mean of the five rankings for each method are shown in the final column of Table 2 (low values are better).

***** Table 2 *****

It is reassuring to see that, for all five series, the naïve univariate methods, Methods E3 to E6, were outperformed by Methods E1 and E2, which are based on simulation from the AR-GARCH models. The results for Method E3 confirm that it would be a mistake simply to substitute temperature point forecasts into expression (1) to estimate the expected payoff.

Interestingly, the four methods that use predictions from the atmospheric model outperformed all six of the univariate methods. Indeed, the univariate methods were even outperformed by Method E9, which involves the naïve use of atmospheric model temperature point forecasts in the payoff formulae in expression (1). The improvement seen in moving from this method to Methods E7 and E8 shows that there is considerable benefit in using an approach based on the substitution of the 51 temperature ensemble members into the payoff formulae. Comparing the results for Methods E7, E8 and E10, we see that the simple rescaling of the ensemble members improved accuracy, and that this method was in turn improved by regression debiasing. Figure 2 shows observed payoffs for Heathrow with post-sample forecasts from this method, Method E10, and for Method E2, which was based on AR-GARCH modelling. The plot shows the ensemble-based forecasts varying more successfully with the observed data.

***** Figure 2 *****

6. Empirical comparison of quantile forecasts

In this section, we use the same 18-month post-sample period employed in Section 5 to evaluate seven methods for forecasting the quantiles of the payoff density. We focus on the following nine quantiles: 1%, 2.5%, 5%, 25%, 50%, 75%, 95%, 97.5% and 99%. Six of these are in the tails of the density, which are of great importance for risk management.

6.1. Quantile forecasting methods

The quantiles of the payoff distributions generated in Methods E1, E2, E7 and E8 of Section 5 are used in this section as quantile forecasts. We label the resultant quantile forecasting methods, Methods Q1 to Q4, respectively. The parameters in two of the remaining three methods are estimated using the quantile regression minimisation in expression (4), which was introduced by Koenker and Bassett (1978).

$$\min \left(\sum_{t | y_t \geq Q_t(\theta)} \theta |y_t - Q_t(\theta)| + \sum_{t | y_t < Q_t(\theta)} (1-\theta) |y_t - Q_t(\theta)| \right) \quad (4)$$

where $Q_t(\theta)$ is the model for the θ quantile of the dependent variable y_t .

Method Q5 - Ensemble-based payoff point forecast with quantile autoregression

An alternative to a pure univariate approach is to construct the payoff quantile forecast using an ensemble-based point forecast plus a univariate forecast of the quantile of the error in this point forecast. We used the payoff point forecast from Method E10 in Section 5, the debiased rescaled ensemble method. Given that the payoff density is very unlike Gaussian, we opted to forecast the error quantile using the nonparametric quantile autoregression approach devised by Engle and Manganelli (2004) for modelling the quantiles of financial returns:

$$\hat{Q}_{t+10|t}^e(\theta) = \hat{Q}_{t|t-10}^e(\theta) + \gamma(\theta) \left[\theta - I(e_{t|t-10} < \hat{Q}_{t|t-10}^e(\theta)) \right] \quad (5)$$

$\hat{Q}_{t+10|t}^e(\theta)$ is the quantile estimator for the error, $e_{t+10|t}$, resulting from the payoff point forecast for the contract period $t+1$ to $t+10$, conditional on information up to period t . $\gamma(\theta)$ is a parameter estimated separately for each quantile, θ , using the quantile regression minimisation in expression (4). $I()$ is an indicator function taking a value of one when the expression in the parentheses is true and zero otherwise. If the probability of the error falling below the θ error quantile estimator is θ , the expected value of the square parentheses in expression (5) is zero. The indicator function has the effect of reducing the next quantile estimate if, in the current period, the error is less than the estimated error quantile. If the error exceeds the error quantile estimate, the next estimate is increased. We used an extensive grid search to initialise the parameter, prior to numerical nonlinear optimisation.

Methods Q6 and Q7 - Combining quantile forecasts

When there is more than one appealing quantile forecast available, it can be beneficial to combine the individual forecasts. Using two different methods, we combined the forecasts from the substituted rescaled ensembles approach, Method Q4, and the quantile autoregression approach, Method Q5. *Method Q6* was the simple average of the forecasts from these two methods. *Method Q7* was the quantile regression combining approach proposed by Granger *et al.* (1989), with intercept omitted and weights restricted to sum to one.

6.2. Results

Unconditional coverage

The unconditional coverage is the percentage of observations falling below the quantile estimator. For the θ quantile estimator, the ideal percentage is θ . Expression (1) shows that the payoff is bounded below by zero and above by the strike price, which is the seasonal average of the weather index. Indeed, the probability of the payoff being zero is equal to the probability that

the index will be above its seasonal average. For many periods, this probability will be sizeable, with the consequence that, for many periods, the estimates of the lower quantiles (1%, 2.5%, 5% and 25%) and the actual payoff will both be zero. This creates a problem when calculating unconditional coverage. In view of this, we evaluated unconditional coverage only for those observations for which the actual payoff was not equal to the quantile estimate.

Table 3 shows the unconditional coverage results for the Heathrow data. An asterisk indicates that the percentage is significantly different from the ideal value at the 5% level. The test uses a Gaussian distribution and the standard error formula for a proportion. To summarise the unconditional coverage across the nine quantiles, we calculated chi-squared goodness-of-fit statistics for the total number of post-sample payoff values falling within the following 10 categories: below the 1% quantile estimator, between each successive pair of quantile estimators, and above the 99% quantile estimator. The chi-squared statistics are reported in the final column of Table 3. The worst value is for Method Q3, which uses the quantiles of the distribution of the ensemble-based payoff realisations. As anticipated, this method underestimated the uncertainty in the payoff. The results for Method Q4 show that the rescaling leads to considerable improvement. This method comfortably outperformed the two AR-GARCH methods. The only methods for which the chi-squared statistic is not significant at the 5% level (critical value is 16.9) are quantile autoregression and the simple average combination.

***** Tables 3 and 4 *****

To summarise across the five locations, we ranked the unconditional coverage percentages, according to how close they were to the ideal percentage, for each of the nine quantiles and five series, and then averaged the rankings across the five series to deliver the results in Table 4. The penultimate column in the table is the average of these ranks for each method. We also calculated the ranking of each method's chi-squared statistic for each of the five series, and then calculated the average of these five rankings. These average rankings are presented in the final column of Table 4. The rankings are reasonably consistent with those in

Table 3 for Heathrow. Simulation from the AR-GARCH model using the Gaussian assumption was more successful than the empirical distribution. The rescaling improved the ensemble-based approach, and both combinations with the quantile autoregression method led to further improvement. The method with the best rankings was the quantile autoregression approach.

Conditional coverage

The Engle and Manganelli (2004) test for conditional coverage evaluates the dynamic properties of the quantile estimator. It involves the joint test of whether the hit variable for the payoff, P_t , defined as $Hit_t \equiv I(P_t < \hat{Q}_{t|t-10}(\theta)) - \theta$, is distributed i.i.d. Bernoulli with probability θ , and is independent of the payoff quantile forecast, $\hat{Q}_{t|t-10}(\theta)$. The ideal is that Hit_t will have zero unconditional and conditional expectations. In our 10-step-ahead context, Hit_t should have no autocorrelation at lags of 10 or more. The test uses the following OLS regression:

$$Hit_t = \delta_0 + \delta_1 Hit_{t-10} + \delta_2 \hat{Q}_{t|t-10}(\theta) + u_t$$

Rewriting this as $Hit = X\delta + u$, the appropriate null hypothesis is that $\delta = \mathbf{0}$. Engle and Manganelli provide the following *dynamic quantile* (DQ) test statistic for this null hypothesis:

$$\frac{\hat{\delta}' X' X \hat{\delta}}{\theta(1-\theta)} \sim \chi^2(3)$$

In view of the existence of periods for which some of the quantile forecasts were identical to the actual payoff, when constructing the hit variable, we included only the periods for which the quantile forecasts were not identical to the actual payoff in that period and in the period 10 days previously. Table 5 shows the resulting DQ statistics for the Heathrow data (lower values are better). The empty entries in the table are due to the hit variable consisting entirely of zeros or entirely of ones, rendering the test's regression impossible.

***** Tables 5 and 6 *****

Table 5 shows that the two methods based on the AR-GARCH model performed poorly. The substituted ensemble approach, Method Q3, performed well for the quantiles in the upper half of the density, with none significant at the 5% level (critical value is 7.8). However, the results for this method are relatively poor for the lower half of the density. The other methods based on the atmospheric model also performed well for the upper half of the density.

To summarise conditional coverage across the five locations, we calculated the DQ statistic ranking of each method for each of the nine quantiles and five series, and then averaged the rankings across the five series to deliver the results in Table 6. The final column of the table shows the mean of the average rankings for each method. The best performing methods in Table 6 are the quantile autoregression approach and the two combining methods. The relative success of these methods is due, in part, to the quality of their unconditional coverage results, since the conditional coverage dynamic quantile statistic is influenced by the unconditional coverage.

Informational content

The DQ statistic is a measure of both the conditional and unconditional coverage of the estimator. The third measure that we use for post-sample evaluation is not affected by the level of unconditional coverage and, therefore, is a useful alternative measure of the dynamic properties of the quantile estimator. This measure, called $R^1(\theta)$, is the quantile regression analogue of the OLS regression R^2 (see Koenker and Machedo, 1999, and Taylor, 1999). $R^1(\theta)$ is recorded for the quantile regression performed using post-sample data with the quantile estimator as sole explanatory variable.

The Heathrow $R^1(\theta)$ results are presented in Table 7 (higher values are better). The most noticeable feature of the table is the sizeable difference between the values for the ensemble-based methods and the AR-GARCH methods. This difference is partly due to the superiority of the ensemble-based methods in forecasting the expected payoff. However, the fact that the quantile autoregression approach was comfortably outperformed by the other four methods

based on the atmospheric model shows that there is informational content in the distribution of the 51 ensemble-based payoff realisations that is not captured by the quantile autoregression.

***** Tables 7 and 8 *****

As a summary of $R^1(\theta)$ performance across the five locations, we calculated the ranking of each method for each of the nine quantiles and for each of the five series, and then averaged the rankings across the five series to give the results in Table 8. The results confirm our conclusions from Table 7 that Methods Q3, Q4, Q6 and Q7, which use the distribution of ensemble-based payoff realisations, outperform the methods that do not.

7. Summary and concluding comments

In this paper, we have investigated the use of ensemble predictions in forecasting the conditional expectation and conditional density of the payoff from a heating degree day put option. With regard to forecasting the expected value, post-sample results showed that the mean of ensemble-based payoff realisations comfortably outperformed the mean of payoff realisations produced by simulation from an AR-GARCH model.

We evaluated post-sample forecasts of the quantiles of the payoff density using three measures: unconditional coverage, the conditional coverage DQ statistic, and the informational content $R^1(\theta)$ measure. Using an ensemble-based forecast for the expected payoff with quantile autoregression produced excellent results in terms of the first two measures. However, this method was outperformed in terms of the third measure by the methods that use the distribution of ensemble-based payoff realisations. An attractive compromise is provided by the combining methods, which pool the information from the quantile autoregression method and the rescaled ensemble-based method to deliver forecasts that perform well across all three measures. Overall, we found that quantiles produced from the simulation of the AR-GARCH model did not match the quality of those from the ensemble-based methods. Therefore, our conclusion is that there is strong potential for the use of ensemble predictions in forecasting the payoff density.

Acknowledgements

We would like to acknowledge the helpful comments of two anonymous referees.

References

Alaton, P., Djehiche, B., & Stillberger, D. (2002). On modelling and pricing weather derivatives, *Applied Mathematical Finance*, 9, 1-20.

Brix, A., Jewson, S., & Ziehmann, C. (2002). Weather derivative modelling and valuation: a statistical perspective, in R.S. Dischel (Ed.), *Climate Risk and the Weather Market* (Chapter 8, pp. 127-150). London: Risk Books.

Brody, D.C., Syroka, J., & Zervos, M. (2002). Dynamical pricing of weather derivatives, *Quantitative Finance*, 2, 189-198.

Buizza, R. (2003). Weather prediction: Ensemble prediction, in J.R. Holton, J. Pyle, J.A. Curry (Eds.), *Encyclopaedia of Atmospheric Sciences*. London: Academic Press.

Buizza, R., Petroliagis, T., Palmer, T.N., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., & Wedi, N. (1998). Impact of model resolution and ensemble size on the performance of an ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society*, 124, 1935-1960.

Campbell, S., & Diebold, F.X. (2005). Weather forecasting for weather derivatives, *Journal of the American Statistical Association*, 100, 6-16.

Cao, M., & Wei, J. (2000). Pricing the weather, *Risk Magazine*, May, 67-70.

Davis, M. (2001). Pricing weather derivatives by marginal value, *Quantitative Finance*, 1, 1-4.

Dischel, R.S. (2002). Introduction to the weather market: dawn to mid-morning, in R.S. Dischel (Ed.), *Climate Risk and the Weather Market* (Chapter 1, pp. 3-24). London: Risk Books.

Dischel, R.S., & Barrieu, P. (2002). Financial weather contracts and their application to risk management, in R.S. Dischel (Ed.), *Climate Risk and the Weather Market* (Chapter 2, pp. 25-41). London: Risk Books.

Duffie, D., & Pan, J. (1997). An overview of value at risk, *Journal of Derivatives*, 4 Spring, 7-49.

Engle, R.F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles, *Journal of Business and Economic Statistics*, 22, 367-381.

Franses, P.H., Neele, J., & van Dijk, D. (2001). Modeling asymmetric volatility in weekly Dutch temperature data, *Environmental Modelling and Software*, 16, 131-137.

- Granger, C.W.J., White, H., & Kamstra, M. (1989). Interval forecasting: an analysis based upon ARCH-quantile estimators, *Journal of Econometrics*, 40, 87-96.
- Koenker, R.W. and Bassett, G.W. (1978). Regression quantiles, *Econometrica*, 46, 33-50.
- Koenker, R., & Machado, J.A.F. (1999). Goodness of fit and related inference processes for quantile regression, *Journal of the American Statistical Association*, 94, 1296-1310.
- Leith, C.E. (1974). Theoretical skill of Monte Carlo forecasts, *Monthly Weather Review*, 102, 409-418.
- Molteni F, Buizza R, Palmer T.N., & Petroliagis T. (1996). The new ECMWF ensemble prediction system: methodology and validation, *Quarterly Journal of the Royal Meteorological Society* 122: 73-119.
- Palmer, T.N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P. and Tribbia, J. (1993) Ensemble prediction. Proceedings of the ECMWF Seminar on *Validation of Models Over Europe: Vol. I*, ECMWF, Shinfield Park, Reading, RG2 9AX, UK.
- Pearce, R.P. (2002). *Meteorology at the Millennium*. Volume 83 in the International Geophysics Series. London: Academic Press
- Taylor, J.W. (1999). Evaluating volatility and interval forecasts, *Journal of Forecasting*, 18, 111-128.
- Taylor J.W., & Buizza, R. (2002). Neural network load forecasting with weather ensemble predictions, *IEEE Transactions on Power Systems* 17: 626-632.
- Taylor J.W., & Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting, *International Journal of Forecasting*, 19, 57-70, 2003.
- Taylor, J.W., & Buizza, R. (2004). Comparing Temperature Density Forecasts from GARCH and Atmospheric Models, *Journal of Forecasting*, 23, 337-355.
- Torró, H., Meneu, V., & Valor, E. (2003). Single factor stochastic models with seasonality applied to underlying weather derivatives variables, *Journal of Risk Finance*, Volume 4, Number 4, 6 - 17.
- Weather Risk Management Association. (2002). The weather risk management industry: survey findings for April 2001 to March 2002, Report by PriceWaterhouseCoopers.
- Zeng, L. (2000). Weather derivatives and weather insurance: concept application and analysis, *Bulletin of the American Meteorological Society*, 81, 2075-2982.

Table 1. Parameter estimates for the temperature AR-GARCH model in expression (2) with seasonality modelled using Fourier terms as in expression (3). Parentheses contain parameter t-statistics. Models estimated using daily data from 1994 to 1998, inclusive.

Parameters	Model				
	Birmingham	Bristol	Heathrow	Leeds	Manchester
Equation for Mean					
μ_0	3.37 (16.80)	3.40 (15.81)	3.61 (15.98)	3.31 (16.93)	3.49 (16.53)
μ_1	-0.75 (-8.99)	-0.72 (-9.12)	-0.75 (-9.10)	-0.78 (-9.16)	-0.81 (-9.37)
μ_2	-1.89 (-14.92)	-1.73 (-15.69)	-1.87 (-14.72)	-1.90 (-14.65)	-1.93 (-15.00)
μ_3	0.33 (4.49)	0.24 (3.76)	0.26 (3.69)	0.32 (4.16)	0.35 (4.65)
μ_4					
ϕ_1	0.71 (42.47)	0.72 (43.29)	0.72 (42.37)	0.71 (41.19)	0.70 (38.43)
Equation for Variance					
ω_0	0.85 (3.06)	0.49 (1.83)	1.40 (2.94)	0.70 (1.62)	1.32 (3.43)
ω_1					
ω_2	0.68 (2.96)			0.93 (2.16)	0.74 (3.10)
ω_3					
ω_4		-0.42 (-2.28)			
α	0.08 (4.30)	0.07 (4.21)	0.08 (3.29)	0.05 (2.33)	0.09 (4.49)
β	0.62 (9.88)	0.60 (8.80)	0.50 (4.27)	0.66 (8.09)	0.52 (5.83)
γ_0	-1.60 (-3.48)	-0.04 (-0.11)	-0.40 (-0.93)	-3.09 (-2.32)	-1.65 (-3.80)
γ_1					
γ_2	3.04 (4.29)	4.66 (5.00)	3.37 (3.36)	3.22 (2.17)	3.02 (4.45)
γ_3					
γ_4					
Diagnostics					
LB Q(7) for $\hat{\eta}_t$	8.17	15.98	10.23	5.85	7.69
LB Q(7) for $\hat{\eta}_t^2$	9.14	7.91	3.83	5.32	7.62
Adj R^2 (%)	86.5	88.0	87.7	86.1	85.6
SBC	4.41	4.16	4.32	4.23	4.43

Please note that this table has previously appeared in Taylor and Buizza (2004).

Table 2. Post-sample MAE ($^{\circ}\text{C}$) for forecasting the expected payoff.

	Birmingham MAE	Bristol MAE	Heathrow MAE	Leeds MAE	Manchester MAE	Mean rank
Univariate						
E1. Simulated AR-GARCH - Gaussian	6.8	5.8	6.2	7.7	7.7	6.4
E2. Simulated AR-GARCH - empirical	6.7	5.7	6.2	7.7	7.8	6.2
E3. Naïve use of AR-GARCH point forecasts	7.4	6.1	6.7	8.4	8.5	8.0
E4. Random walk	10.8	9.4	10.3	12.2	12.3	10.0
E5. Mean of 5 most recent periods	11.4	9.9	10.6	12.8	12.9	11.0
E6. Mean of same period from past 5 years	7.7	6.4	6.9	9.2	9.1	9.0
Atmospheric						
E7. Substituted ensembles	6.2	5.6	4.3	5.6	5.4	3.8
E8. Substituted rescaled ensembles	5.9	5.4	4.3	5.5	5.3	2.8
E9. Naïve use of temperature point forecasts	6.7	6.1	4.7	6.0	6.0	5.4
E10. Debiasing of Method E8	4.7	4.1	4.2	5.1	5.3	1.4

bold indicates best performing method for each quantile

Table 3. Post-sample unconditional coverage percentages for the Heathrow payoff quantiles. The chi-square statistic summarizes performance across the quantiles.

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%	ChiSq
Univariate										
Q1. Simulated AR-GARCH - Gaussian	0.3	0.8*	1.7*	11.9*	37.6*	74.5	94.4	97.0	98.0*	76.5*
Q2. Simulated AR-GARCH - empirical	0.3	0.8*	1.7*	11.9*	36.5*	72.4	94.1	96.9	98.0*	71.9*
Atmospheric										
Q3. Substituted ensembles	5.3*	7.2*	10.5*	33.2*	55.6*	78.0	94.9	96.9	97.9*	116.8*
Q4. Substituted rescaled ensembles	2.2*	3.7	5.7	25.5	51.3	78.6	95.9	98.4	98.7	17.5*
Q5. Quantile autoregression	1.1	1.8	3.7	23.6	47.2	72.2	96.2	97.8	99.3	9.9
Q6. Average of Q4 and Q5	0.9	2.0	4.4	23.9	52.5	76.2	96.5	98.4	99.5	6.5
Q7. Q.Reg. combination of Q4 and Q5	1.6	3.5	5.1	25.4	54.8*	80.1*	96.7	98.5	98.7	19.4*

* significant at 5% level

bold indicates best performing method for each quantile

Table 4. Mean rankings of the post-sample unconditional coverage percentages and chi-square statistic across the five locations.

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%	Row Mean	ChiSq
Univariate											
Q1. Simulated AR-GARCH - Gaussian	4.5	4.9	4.9	5.8	5.6	3.4	3.0	2.6	3.1	4.2	4.6
Q2. Simulated AR-GARCH - empirical	4.5	4.9	4.9	6.0	6.6	4.8	3.4	2.4	3.1	4.5	5.2
Atmospheric											
Q3. Substituted ensembles	7.0	7.0	7.0	5.8	4.6	4.8	5.0	5.0	5.8	5.8	7.0
Q4. Substituted rescaled ensembles	4.4	4.6	4.0	3.6	3.6	5.8	4.2	5.6	5.2	4.6	4.6
Q5. Quantile autoregression	2.1	1.9	2.1	1.6	2.0	2.8	3.4	1.8	3.0	2.3	1.2
Q6. Average of Q4 and Q5	2.3	1.3	2.0	2.8	3.1	3.2	4.1	5.0	3.5	3.0	2.0
Q7. Q.Reg. combination of Q4 and Q5	3.2	3.4	3.1	2.4	2.5	3.2	4.9	5.6	4.3	3.6	3.4

bold indicates best performing method for each quantile

Table 5. Post-sample dynamic quantile chi-square statistic for the Heathrow payoff quantiles.

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%
Univariate									
Q1. Simulated AR-GARCH - Gaussian	-	-	-	80.1*	48.4*	18.7*	18.4*	33.1*	82.2*
Q2. Simulated AR-GARCH - empirical	-	-	-	80.2*	58.2*	30.2*	22.0*	36.3*	81.4*
Atmospheric									
Q3. Substituted ensembles	487.0*	256.9*	208.4*	107.2*	50.9*	5.3	0.3	1.0	7.1
Q4. Substituted rescaled ensembles	133.8*	68.2*	86.6*	66.3*	29.6*	13.3*	1.4	2.0	0.6
Q5. Quantile autoregression	43.8*	26.2*	34.3*	40.7*	2.7	16.2*	4.4	1.3	1.0
Q6. Average of Q4 and Q5	43.5*	26.7*	54.1*	39.9*	5.5	2.8	2.9	1.6	1.1
Q7. Q.Reg. combination of Q4 and Q5	99.2*	49.1*	68.0*	53.7*	9.6*	8.6*	3.4	2.2	0.7

* significant at 5% level

bold indicates best performing method for each quantile

Table 6. Mean rankings of the post-sample dynamic quantile chi-square statistic across the five locations.

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%	Row Mean
Univariate										
Q1. Simulated AR-GARCH - Gaussian	6.4	6.5	6.5	3.4	5.0	5.2	4.8	5.2	5.6	5.4
Q2. Simulated AR-GARCH - empirical	6.4	6.5	6.5	4.4	6.4	6.2	5.6	4.4	5.0	5.7
Atmospheric										
Q3. Substituted ensembles	4.8	5.0	5.0	7.0	6.0	4.4	4.6	3.8	4.6	5.0
Q4. Substituted rescaled ensembles	3.8	3.0	3.6	5.2	4.4	4.8	4.2	4.4	4.0	4.2
Q5. Quantile autoregression	2.0	2.2	2.2	1.6	1.0	2.8	3.4	2.0	2.6	2.2
Q6. Average of Q4 and Q5	2.4	2.4	1.8	3.0	2.8	2.6	2.0	3.4	2.8	2.6
Q7. Q.Reg. combination of Q4 and Q5	2.2	2.4	2.4	3.4	2.4	2.0	3.4	4.8	3.4	2.9

bold indicates best performing method for each quantile

Table 7. Post-sample $R^1(\theta)$ percentages for the Heathrow payoff quantiles.

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%
Univariate									
Q1. Simulated AR-GARCH - Gaussian	0.0	69.6	0.0	-2.2	8.2	17.7	16.5	18.5	20.3
Q2. Simulated AR-GARCH - empirical	0.0	69.6	0.0	-2.2	7.7	17.3	16.3	18.4	19.6
Atmospheric									
Q3. Substituted ensembles	1.0	70.0	4.0	27.0	43.0	50.0	48.0	45.0	39.0
Q4. Substituted rescaled ensembles	4.0	71.0	5.0	27.0	43.0	50.0	48.0	46.0	39.0
Q5. Quantile autoregression	-0.2	69.6	0.0	18.7	40.4	43.1	24.2	22.1	22.2
Q6. Average of Q4 and Q5	0.0	69.6	0.2	23.9	42.2	47.9	38.8	38.3	33.9
Q7. Q.Reg. combination of Q4 and Q5	3.1	70.9	5.2	26.7	42.6	49.4	47.3	45.1	38.5

bold indicates best performing method for each quantile

Table 8. Mean rankings of post-sample $R^1(\theta)$ percentages across the five locations.

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%	Row Mean
Univariate										
Q1. Simulated AR-GARCH - Gaussian	3.6	4.2	5.3	6.2	6.0	6.0	6.6	6.6	6.2	5.6
Q2. Simulated AR-GARCH - empirical	3.6	4.2	5.3	6.8	7.0	7.0	6.4	6.4	6.8	5.9
Atmospheric										
Q3. Substituted ensembles	2.9	2.9	2.9	2.5	2.1	2.1	2.1	1.6	1.5	2.3
Q4. Substituted rescaled ensembles	1.6	1.4	1.2	1.1	1.1	1.1	1.5	3.2	2.9	1.7
Q5. Quantile autoregression	5.5	7.0	5.9	5.0	5.0	5.0	5.0	5.0	5.0	5.4
Q6. Average of Q4 and Q5	6.2	5.6	4.5	3.4	3.0	3.0	3.8	3.0	3.4	4.0
Q7. Q.Reg. combination of Q4 and Q5	4.6	2.7	2.9	3.0	3.8	3.8	2.6	2.2	2.2	3.1

bold indicates best performing method for each quantile

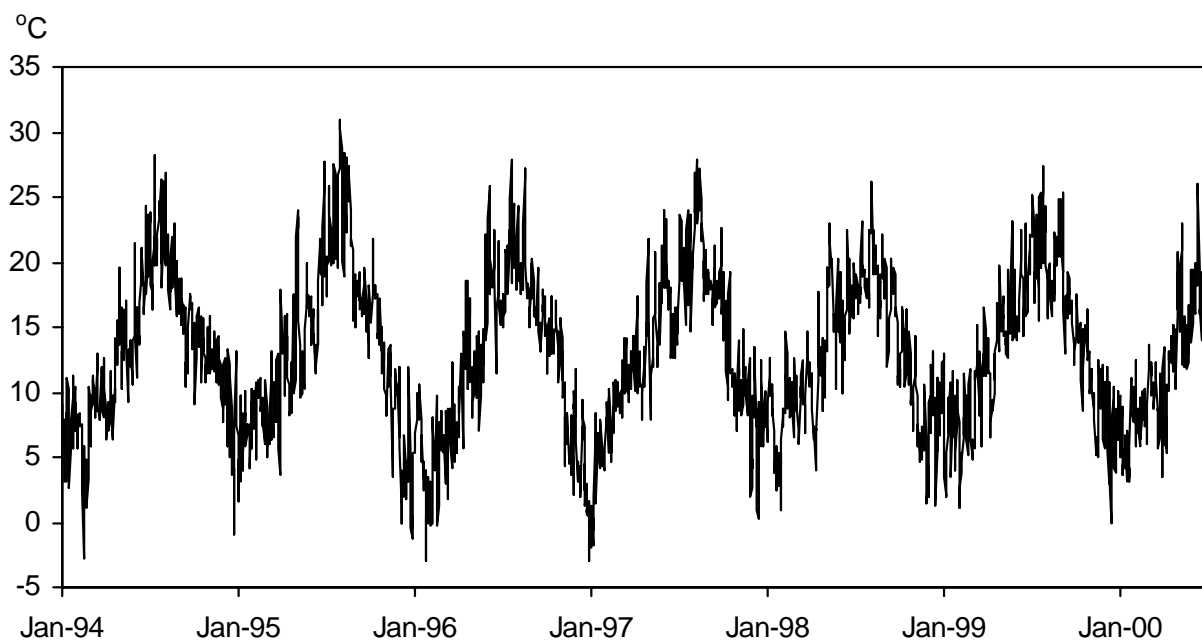


Fig. 1. Daily midday temperature observations at Heathrow.

Please note that this figure has previously appeared in Taylor and Buizza (2004).

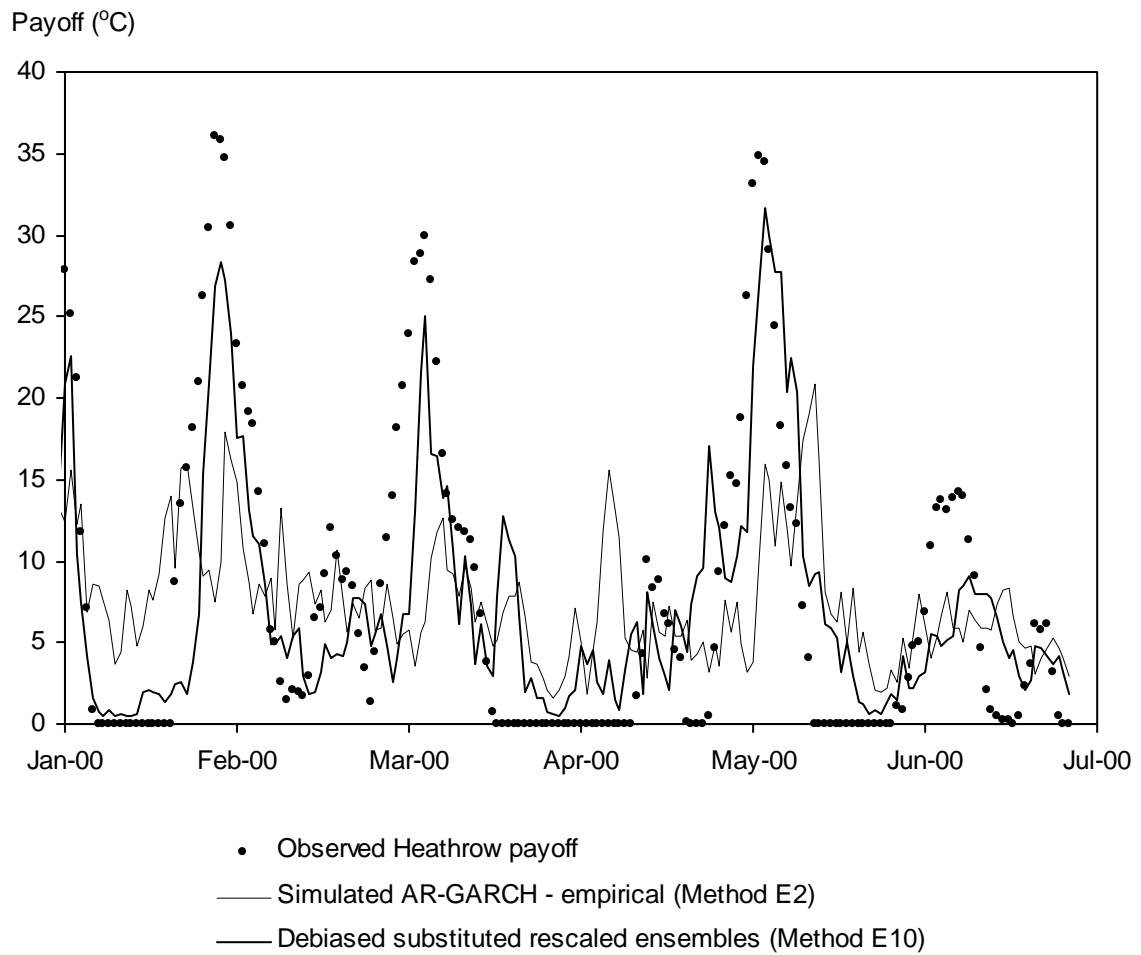


Fig. 2. Observed payoff values for Heathrow plotted with the forecasts for the expected payoff from two methods for the final 6 months of the post-sample evaluation period.