A Comparison of Temperature Density Forecasts from GARCH and Atmospheric Models

James W. Taylor

and

Roberto Buizza

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

**A Comparison of Temperature Density Forecasts
from GARCH and Atmospheric Models**

A COMPARISON OF TEMPERATURE DENSITY FORECASTS FROM GARCH AND
ATMOSPHERIC MODELS

Abstract

Density forecasts for weather variables are useful for the many industries exposed to weather
risk. Weather ensemble predictions are generated from atmospheric models and consist of
multiple future scenarios for a weather variable. The distribution of the scenarios can be used as
a density forecast, which is needed for pricing weather derivatives. We consider one to 10 day-
ahead density forecasts provided by temperature ensemble predictions. More specifically, we
evaluate forecasts of the mean and quantiles of the density. The mean of the ensemble scenarios
is the most accurate forecast for the mean of the density. We use quantile regression to debias
the quantiles of the distribution of the ensemble scenarios. The resultant quantile forecasts
compare favourably with those from a GARCH model. These results indicate the strong
potential for the use of ensemble prediction in temperature density forecasting.

INTRODUCTION

Weather has an important influence on many different industries, including agriculture, energy, retail and transportation. Due to the complexity of the earth's atmosphere, there is uncertainty in every weather forecast. To enable users of a forecast to plan for the different possible outcomes, it is important for the forecast to be accompanied by a measure of forecast uncertainty (Chatfield, 1993).

In this paper, we consider a new type of weather forecast, weather ensemble predictions, that are produced by large meteorological models of the earth's atmosphere. An ensemble prediction consists of multiple scenarios (51 in our specific case) for the future value of weather variables. The different scenarios are known as ensemble members. The ensemble prediction, therefore, conveys the degree of uncertainty in the weather variable. The distribution of the scenarios can be used as a forecast of the conditional density function of the weather variable. Density forecasts are particularly important in the area of weather derivatives because of their use in pricing the derivative (see Cao and Wei, 2000; Campbell and Diebold, 2002).

This study investigates the accuracy of the density forecasts based on temperature ensemble predictions for lead times from one to 10 days ahead. More specifically, we evaluate forecasts of the 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 quantile forecasts with those from a univariate time series model with variance modelled as a GARCH process. These models are widely used to forecast volatility in finance, but recently they have also been applied to temperature series. Campbell and Diebold (2002) estimate such models for US temperature data and compare the point forecasts produced from these models with those from meteorological atmospheric models. In this study, we compare both point forecasts and density forecasts from univariate time series models with those based on ensemble predictions from an atmospheric model. We analyse daily temperature at five UK locations.

In the next section, we review the literature on univariate time series models for temperature density forecasting before presenting our models for the temperature series considered in this paper. We then describe weather ensemble predictions. The next section compares point forecast accuracy of the different methods, and the section that follows compares quantile forecast performance. The final section provides a summary and conclusion.

UNIVARIATE MODELS FOR TEMPERATURE DENSITY FORECASTING

Previous Models for Temperature Time Series

Franses *et al.* (2001) estimate and evaluate a univariate model for weekly mean Dutch temperature data. Their preliminary analysis revealed four features of the time series: a yearly seasonal pattern in the mean; a yearly seasonal pattern in the volatility; large absolute deviations from the mean tend to cluster, as do small deviations; and the impact of temperatures lower than expected on conditional volatility is different from the impact of temperatures higher than expected, and this impact is seasonal. Since volatility clustering is also evident in high-frequency financial returns, Franses *et al.* consider generalized autoregressive conditional heteroskedastic (GARCH) models (Engle, 1982; Bollerslev, 1986), which have become widely used for modelling financial volatility. The simple GARCH(1,1) model is given by

$$\sigma_t^2 = \omega + \alpha \, \varepsilon_{t-1}^2 + \beta \, \sigma_{t-1}^2 \tag{1}$$

where σ_t is the conditional standard deviation (volatility), ε_t is a stochastic error term, and ω, α and β are parameters. Tol (1996) uses a GARCH(1,1) model for the volatility in daily mean Dutch temperature data, and an autoregressive (AR) model for the mean. He addresses the seasonality issue by estimating separate models for the summer and winter. By contrast, Franses *et al.* try to capture all of the four features that they had observed in their weekly data by estimating the following AR-GARCH model:

$$T_t = s(\pmb{\mu}, t) + \phi_1 T_{t-1} + \varepsilon_t$$
$$\varepsilon_t = \sigma_t \, \eta_t \qquad\qquad (2)$$
$$\sigma_t^2 = s(\pmb{\omega}, t) + \alpha \left(\varepsilon_{t-1} - s(\pmb{\gamma}, t) \right)^2 + \beta \, \sigma_{t-1}^2$$

where T_t is the temperature variable, η_t is an i.i.d. error term, and $\pmb{\mu}$, $\pmb{\omega}$ and $\pmb{\gamma}$ are vectors of parameters. The seasonality term, $s(\pmb{\mu}, t)$, appears in the equation for the mean along with a first order autoregressive term. Similar terms, $s(\pmb{\omega}, t)$ and $s(\pmb{\gamma}, t)$, are employed to model the seasonality in the volatility, and the asymmetric seasonal impact of temperatures lower and higher than expected on conditional volatility. Franses *et al.* model the seasonality as a quadratic function:

$$s(\pmb{\lambda}, t) = \lambda_0 + \lambda_1 \, w(t) + \lambda_2 \, w(t)^2$$

where $w(t)$ is a repeating step function that numbers the weeks from 1 to 52 within each year.

Campbell and Diebold (2002) estimate AR-ARCH time series models for average daily US temperature data. By contrast with the model of Franses *et al.*, Campbell and Diebold do not include either the lagged variance term, σ_t^2, nor the asymmetric seasonality term, $s(\pmb{\gamma}, t)$, in the variance model of expression (2), leaving a symmetric ARCH formulation. In addition, Campbell and Diebold use a low ordered Fourier series to model the seasonality, instead of a quadratic function. A second order Fourier modelling of seasonality has the form:

$$
\begin{aligned}
s(\pmb{\lambda}, t) = \lambda_0 &+ \lambda_1 \, \sin\!\left(2\,\pi\,\tfrac{d(t)}{365}\right) + \lambda_2 \, \cos\!\left(2\,\pi\,\tfrac{d(t)}{365}\right) \\
&+ \lambda_3 \, \sin\!\left(4\,\pi\,\tfrac{d(t)}{365}\right) + \lambda_4 \, \cos\!\left(4\,\pi\,\tfrac{d(t)}{365}\right)
\end{aligned}
\qquad (3)
$$

where $d(t)$ is a repeating step function that numbers the days from 1 to 365 within each year. Campbell and Diebold removed 29 February from each leap year in order to maintain 365 days in each year. Torró *et al.* (2001) fit an AR-GARCH model to daily Spanish temperature data. In order to try to model the seasonality in the variance, they use a GARCH(1,1) model, as in expression (1), multiplied by a power function of lagged temperature.

The lagged temperature term, T_{t-1}, in expression (2) enables the model to fit an annual seasonal cycle that is not constant from one year to the next. An appealing alternative approach would be to allow the parameters, $\pmb{\mu}$, of the seasonal term $s(\pmb{\mu}, t)$ to vary over time, as in the

dynamic harmonic regression (DHR) model of Young *et al*. (1999). Young *et al*. (1997) use the DHR approach to model the mean of a temperature time series. In order to produce density forecasts, it would be interesting to see the approach extended for the modelling of both the mean and variance. Hyndman and Wand (1997) find that the autocorrelation in a daily time series of Australian maximum temperatures varies throughout the year. In terms of AR-GARCH modelling, this also motivates the possible inclusion of time varying parameters.

Modelling Daily UK Temperature Time Series

We analyse daily air temperature, recorded 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. Weather data recorded at these locations is 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 Torró *et al*., 2001).

Figure 1 shows a plot of the Heathrow temperature 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. 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.

***** Figures 1 and 2 *****

To gain further insight into the seasonality, in Figure 2, we plot Heathrow temperature against the day of the year for the five-year estimation period. The plot indicates that the seasonality in the mean of the series does not appear to be quadratic, so an initial reaction is that Campbell and Diebold's Fourier modelling of seasonality is likely to be more effective than a

quadratic function of the type employed by Franses *et al*. There is slight evidence in Figure 2 of seasonality in the variance; it seems to be greatest in the winter months and least in the autumn. However, it is interesting to note that the seasonality in the variance is far less pronounced in our UK data than in the Dutch data of Franses *et al*. and the US data of Campbell and Diebold.

We estimated AR-GARCH models, as in expression (2), for our five daily UK temperature series using the standard approach of maximum likelihood to estimate parameters under the assumption that η_t was Gaussian. The estimation, and, indeed, all computational work in this paper, was performed using the statistical programming package, Gauss. Autoregressive terms of order greater than one and moving average terms of all orders were not significant for any of our five series. This was also the conclusion of Franses *et al*. for their Dutch data. We found that Campbell and Diebold's Fourier series modelling of seasonality gave better fit than quadratic modelling, which was used by Franses *et al*. Indeed, the quadratic terms, $d(t)$ and $d(t)^2$, were not significant when included alongside Fourier terms. We did not find significant Fourier terms of order more than two in any of the three seasonal features of the model in expression (2). We therefore used the seasonal function in expression (3) to represent seasonality. We followed Campbell and Diebold in removing 29 February from each leap year in our sample.

In Table I, we present our preferred model for each of our five temperature series. We selected models using the Schwarz Bayesian Criterion (SBC) to judge fit. For each model, the table presents each parameter with its t-statistic, adjusted R^2, SBC and Ljung-Box Q-statistic to test for autocorrelation in standardised residuals ($\hat{\eta}_t = \hat{\varepsilon}_t / \hat{\sigma}_t$) and squared standardised residuals. The only significant Q-statistic is for the residuals from the Bristol model (critical value is 12.59). This value is not significant at the 1% level (critical value is 16.81), and since we could not find a simple alternative model with better residuals, we decided to use this model. Table I shows that, for all but the Heathrow model, we found significant parameters within the

seasonal GARCH function, $s(\omega,t)$. Interestingly, with all five series, when fitting GARCH models, we found significant parameters in the asymmetric seasonal variance function, $s(\gamma,t)$.

***** Table I *****

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

WEATHER ENSEMBLE PREDICTIONS

The weather is a chaotic system. Small errors in the initial conditions of a forecast grow rapidly, and affect predictability. Furthermore, predictability is limited by model errors due to the approximate simulation of atmospheric processes in a numerical model. These two sources of uncertainty limit the accuracy of traditional single point forecasts, generated by running the model once with best estimates for the initial conditions.

The weather prediction problem can be described in terms of the time evolution of an appropriate probability density function in the atmosphere's phase space. An estimate of the density function provides forecasters with an objective way to gauge the uncertainty in single point predictions. Ensemble prediction aims to derive a more sophisticated estimate of the density function than that provided by the distribution of past atmospheric states. Ensemble prediction systems generate multiple realisations of numerical predictions 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 (Molteni *et al*., 1996; Palmer *et al*., 1993; Buizza *et al*., 1998).

The benefit of using ensemble predictions is illustrated in Figure 3. pdf_0, represents the initial uncertainties. From the best estimate of the initial state, a single point forecast is produced (bold solid curve). This point forecast fails to predict correctly the future state (dashed curve). The ensemble forecasts (thin solid curves), starting from perturbed initial conditions, can be used to estimate the probability of future states. In this example, the estimated probability density function, pdf_t, is bimodal. The figure shows that two of the perturbed forecasts almost correctly predicted the future state. Therefore, at time 0, the ensemble system would have given a non-zero probability of the future state.

***** Figure 3 *****

Since December 1992, both the US National Center for Environmental Predictions (NCEP, previously NMC) and the European Centre for Medium-range Weather Forecasts (ECMWF) have integrated their deterministic prediction with medium-range ensemble prediction (Palmer *et al*., 1993, Toth and Kalnay, 1993, Tracton and Kalnay, 1993). The number of ensemble members is limited by the necessity to produce forecasts in a reasonable amount of time with the available computer power. Traditional single point forecasts are produced using a high-resolution grid spacing of 40 km. In December 1996, after different system configurations had been considered, a 51-member system with a horizontal grid resolution of 120 km at mid-latitude was installed at ECMWF (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. Stochastic physics was introduced into the system in October 1998 (Buizza *et al*., 1999). This aims to simulate model uncertainties due to random model error. In November 2000, the resolution of the ECMWF ensemble system was further increased to a grid spacing of 80 km at mid-latitudes.

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

During the period spanned by the data used in this study, ensemble forecasts were produced every day for lead times from 12 hours ahead to 10 days ahead. The ensemble forecasts were archived every 12 hours, and are thus available for midday and midnight. The archived weather variables include both upper level variables (typically temperature, wind speed, humidity and vertical velocity at different heights) and surface variables (e.g. temperature, wind speed, precipitation, cloud cover). In our work, we used ECMWF ensemble predictions for midday air temperature, recorded from 1 January 1997 to 1 July 2000 and measured at a height of 2 meters at the five UK locations specified earlier.

EMPIRICAL COMPARISON OF POINT FORECASTS

Although the main aim of this paper is density forecasting, it is also interesting to consider the quality of the point forecasts produced by the different approaches to density estimation. The point forecast is, of course, the mean of the density, and so by evaluating its accuracy, we gain an understanding of the accuracy of the central location of the density forecast. We used the period from 1 January 1999 to 1 July 2000 for post-sample evaluation of forecasts for lead times from one to 10 days.

Forecasting Methods

Methods P1 to P4 are univariate time series approaches. The first uses the well-specified AR-GARCH models, while Methods P2 to P4 are naïve benchmark approaches. Methods P5 and P6 use predictions from an atmospheric model.

Method P1 - The AR-GARCH models in Table I were used to produce point forecasts.

Method P2 - Random walk forecasts were created, where the forecast for all lead times is the most recent period's observed value.

8

Method P3 - The average of the observed temperature on the corresponding day in each of the previous five years.

Method P4 - The average of the observed values on the five most recent days was used as the forecast for all lead times.

Method P5 - Traditional meteorological point forecasts generated by running the atmospheric model once at high resolution with best estimates for the initial conditions.

Method P6 - The mean of the 51 ensemble members. Perhaps surprisingly, this has been found to be a more accurate point forecast than the traditional high-resolution point forecast (Leith, 1974; Molteni *et al.*, 1996), indicating that the ensemble contains information not captured by the traditional forecast.

Results

We calculated the mean absolute error (MAE), root mean squared error (RMSE) and median absolute error (MedAE) for the post-sample forecast errors from the six methods for each of the 10 forecast horizons and for each of the five temperature series. We discuss only the MAE results because the relative performance of the methods was similar for all three measures.

In Figure 4, we present the MAE results for the Heathrow series. Reassuringly, the three more sophisticated methods, Methods P1, P5 and P6, outperform the three simple benchmark methods, Methods P2, P3 and P4. The one exception to this is that, beyond seven days ahead, the traditional high-resolution atmospheric model, Method P5, is beaten by the average of the corresponding day of the year from the previous five years, Method P4. Beyond seven days ahead, the traditional high-resolution atmospheric model is also outperformed by the AR-GARCH model, Method P1. These results are similar to those of Campbell and Diebold (2002), who found, for their US daily temperature series, that traditional atmospheric model forecasts outperform univariate time series model forecasts up to a horizon of about eight days. Interestingly, the method that outperforms all others, at all forecast horizons, is the mean of the

9

51 ensemble members, Method P6. Note that, up to about 9 days ahead, the MAE for this method is substantially lower than for the AR-GARCH model, with the difference being about 1°C for two and three days ahead. The relative rankings of the MAE results for the other four temperature series are similar to the results in Figure 4 for Heathrow, with the ensemble mean approach, Method P6, dominating for all horizons.

<div align="center">***** Figures 4 *****</div>

<div align="center">EMPIRICAL COMPARISON OF QUANTILE FORECASTS</div>

A probability density function can be described by its constituent quantiles. In this paper, we compare the ability of different methods to forecast the quantiles of the density. We focus on the following nine quantiles: 1%, 2.5%, 5%, 25%, 50%, 75%, 95%, 97.5% and 99%. We selected six quantiles in the tails of the density because this part of the distribution is of great importance from a risk management perspective. We compare post-sample quantile forecasts from five methods for lead times from one to 10 days for the same 18-month post-sample period considered in the previous section. Before introducing the methods, we briefly present quantile regression, which is used in three of the methods.

Quantile Regression

If the conditional θ quantile, $Q_t(\theta)$, of a variable y_t is a linear function of explanatory variables, we can write $Q_t(\theta) = x_t \beta(\theta)$, where x_t is a vector of explanatory variables and $\beta(\theta)$ is a vector of parameters dependent on θ. Koenker and Bassett (1978) showed that the *quantile regression* minimisation in (4) delivers parameters that asymptotically approach $\beta(\theta)$. Note that for computational convenience this minimisation can be formulated as a linear program.

$$\min_{\beta} \left(\sum_{t\,|\,y_t \geq x_t \beta} \theta \left| y_t - x_t \beta \right| + \sum_{t\,|\,y_t < x_t \beta} (1-\theta) \left| y_t - x_t \beta \right| \right) \tag{4}$$

Quantile Forecasting Methods

Method Q1 is a pure univariate approach. Methods Q2 and Q3 construct quantiles using the ensemble mean and a univariate model for the variation about the mean. Methods Q4 and Q5 base estimation on the quantiles of the distribution of ensemble members.

Method Q1 - AR-GARCH with Empirical Distribution

We used the AR-GARCH models in Table I to produce mean and variance forecasts. Quantiles were constructed separately using a Gaussian assumption and the empirical distribution of standardised residuals. The empirical distribution led to slightly more accurate quantile forecasts, and so for simplicity, in this paper we report only these results.

Method Q2 - Ensemble Mean with GARCH and Empirical Distribution

In the previous section, we found that the mean of the 51 ensemble members is a better point forecast than that provided by the univariate AR-GARCH models for all 10 lead times. In view of this, an alternative to the AR-GARCH models is to construct the density forecast using the ensemble mean as the estimate of the mean of the density with a univariate model for the uncertainty. The k-step-ahead conditional quantile estimator for the quantile of T_{t+k}, the temperature in period $t+k$, is then:

$$\hat{Q}_{t+k|t}(\theta) = \mu_{t+k|t}^{ENS} + \hat{Q}_{t+k|t}^{e}(\theta) \tag{5}$$

where $\mu_{t+k|t}^{ENS}$ is the mean of the k-step-ahead 51 ensemble members, $\hat{Q}_{t+k|t}^{e}(\theta)$ is the univariate estimator of the conditional quantile of the k-step-ahead forecast error, $e_{t+k|t} = T_{t+k} - \mu_{t+k|t}^{ENS}$. In estimating $\hat{Q}_{t+k|t}^{e}(\theta)$, the forecast errors for each lead time must be considered separately. Rather than laboriously specifying a different GARCH model for each lead time for each series, we opted to estimate a GARCH(1,1) model, as in expression (1), for them all. As with the AR-

GARCH model, an empirical distribution of standardised residuals led to slightly better quantile forecasts than using a Gaussian distribution.

Method Q3 - Ensemble Mean with Quantile Autoregression

We also estimated $\hat{Q}^e_{t+k|t}(\theta)$ in expression (5) using the following quantile autoregression approach devised by Engle and Manganelli (2002) for modeling the quantiles of financial returns:

$$\hat{Q}^e_{t+k|t}(\theta) = \hat{Q}^e_{t-1+k|t-1}(\theta) + \gamma_k(\theta)\left[\theta - I\left(e_{t|t-k} \leq \hat{Q}^e_{t|t-k}(\theta)\right)\right] \qquad (6)$$

I() is an indicator function taking a value of one when the expression in the parentheses is true and zero otherwise. The parameter, $\gamma_k(\theta)$, was estimated separately for each of the nine quantiles, θ, and 10 lead times, k, using the quantile regression minimisation in expression (4). The expected value of the expression within the square parentheses is zero if the probability of the error falling below the θ quantile estimator is θ. 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 quantile estimate, the next estimate is increased. The model focuses directly on the autoregressive nature of the quantiles. By contrast, GARCH approaches model autoregression in the variance, and then infer from this for the quantiles. We used an extensive grid search to initialise the parameters, prior to numerical nonlinear optimisation.

Method Q4 - Ensemble Quantiles Debiased Using Quantile Regression

Although the distribution of the 51 ensemble members can be used as a temperature density forecast, we found that this tends to underestimate substantially the true uncertainty. For example, the 99% 'ensemble quantile' considerably underestimated the true 99% quantile. In view of this, we used quantile regression to debias the ensemble quantiles with temperature as dependent variable and the ensemble quantile as regressor (see Granger, 1989). We used

ensemble predictions from 1 January 1997, the earliest date in our ensemble dataset, to 31 December 1998, the final date in our estimation sample. The form of the resultant estimator is:

$$\hat{Q}_{t+k|t}(\theta) = a_k(\theta) + b_k(\theta)\, Q_{t+k|t}^{ENS}(\theta)$$

where $Q_{t+k|t}^{ENS}(\theta)$ is the quantile of the k-step-ahead 51 ensemble members, and $a_k(\theta)$ and $b_k(\theta)$ are parameters. The debiasing is performed separately for each quantile, θ, and lead time, k.

An alternative to basing quantile estimation on the quantiles of the ensemble members is to use their standard deviation. This produced similar results to the ensemble quantiles.

Method Q5 - Ensemble Quantiles Debiased Using TVP Quantile Regression

The use of OLS regression to debias a point forecast was proposed by Theil (1971). In the context of judgmental point forecasting, Goodwin (1997) describes an approach that allows the debiasing regression parameters to vary over time if there is a changing relationship between actuals and point forecasts, such as when the quality of the forecasts are considered to have improved over time. In view of the developments in the ensemble generating system, such as the introduction of stochastic physics in October 1998, there is some appeal to debiasing the ensemble member quantiles using time varying parameters (TVP). We developed the following TVP quantile regression debiasing approach:

$$\hat{Q}_{t+k|t}(\theta) = a_{t+k|t}(\theta) + b_{t+k|t}(\theta)\, Q_{t+k|t}^{ENS}(\theta)$$

where

$$a_{t+k|t}(\theta) = a_{t-1+k|t-1}(\theta) + \alpha_k(\theta)\left[\theta - I\!\left(T_t \leq \hat{Q}_{t|t-k}(\theta)\right)\right]$$
$$b_{t+k|t}(\theta) = b_{t-1+k|t-1}(\theta) + \beta_k(\theta)\left[\theta - I\!\left(T_t \leq \hat{Q}_{t|t-k}(\theta)\right)\right]$$

where T_t is the temperature variable, and $\alpha_k(\theta)$ and $\beta_k(\theta)$ are parameters. The structure of the TVP parameters, $a_{t+k|t}(\theta)$ and $b_{t+k|t}(\theta)$, is based on the quantile autoregression models of Engle and Manganelli (2002). The effect of the indicator function is to reduce $a_{t+k|t}(\theta)$ and $b_{t+k|t}(\theta)$, if,

in the current period, the observed value for the temperature variable is less than the estimated quantile. Conversely, if temperature exceeds the quantile estimate, the parameters are increased.

Post-Sample Quantile Forecasting Results

Unconditional Coverage

The most fundamental requirement of a θ quantile estimator is that the percentage of observations falling below it is θ. This is termed *unconditional coverage* by Christoffersen (1998). Figure 5 compares the unconditional coverage of the five methods for estimation of the 5% quantile for the Heathrow data at the 10 different lead times for the post-sample period of 18 months. The dashed horizontal lines in Figure 5 are the bounds of the acceptance region for the test of whether the percentages are significantly different from 5% at the 5% level. The bounds are calculated using a Gaussian distribution and the standard error formula for a proportion.

***** Figure 5 and 6 *****

To summarise the unconditional coverage of the methods across the nine quantiles, we calculated chi-squared goodness of fit statistics for each of the 10 lead times and for each temperature series. For each method, at each lead time, we calculated the statistic for the total number of post-sample observations falling within the following 10 categories: below the 1% quantile estimator, between each successive pair of quantile estimators, and above the 99% quantile estimator. Figure 6 shows the resulting chi-squared statistics for the Heathrow data (lower values are better). The dashed horizontal line in the figure is the bound of the acceptance region for the 5% significance test. Overall, the results of Figure 6 are consistent with those shown in Figure 5 for the 5% quantile. The results show that the AR-GARCH approach, Method Q1, performs poorly beyond the early lead times. Performance for ensemble quantiles debiased using quantile regression, Method Q4, is particularly poor for the first three forecast horizons. However, it is interesting to see that Method Q5, which uses TVP quantile regression debiasing, offers substantial improvement. The best results are achieved with Method Q3, which uses the

14

ensemble mean with quantile autoregression. This method comfortably outperforms Method Q2, which uses the ensemble mean with a GARCH model for the variance. The relative performances of the methods, in terms of the unconditional coverage chi-squared statistic, for the other four locations are very similar to those for Heathrow.

Dynamic Quantile Test Statistic

Simply testing for unconditional coverage is insufficient, as it does not assess the dynamic properties of the quantile (Christoffersen, 1998). Engle and Manganelli (2002) test for *conditional coverage* by jointly testing whether the following hit variable is distributed i.i.d. Bernoulli with probability θ, and is independent of the value of the quantile estimator.

$$Hit_t \equiv I\left(y_t \le \hat{Q}_{t|t-k}(\theta)\right) - \theta$$

where $\hat{Q}_{t|t-k}(\theta)$ is the quantile estimator for the variable of interest, y_t. A similar hit variable was used in the quantile autoregression in expression (6). For an ideal quantile estimator, the hit variable has zero unconditional and conditional expectations. Engle and Manganelli consider only one period-ahead forecasting, for which the Hit_t variable should be serially uncorrelated. We can extend the test to the multi-period forecasting context by noting that the Hit_t variable should have no autocorrelation at lags of k or more, where k is the forecast horizon. The test can then proceed by running the following OLS regression:

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

Rewriting this in matrix form, we get:

$$Hit_t = X\delta + u_t \qquad u_t = \begin{cases} -\theta & with\ probability\ (1-\theta) \\ (1-\theta) & with\ probability\ \theta \end{cases}$$

The appropriate null hypothesis is that $\delta = \mathbf{0}$. Engle and Manganelli provide the following dynamic quantile test statistic for this null hypothesis:

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

Figure 7 shows the resulting dynamic quantile chi-squared statistics for estimation of the 5% quantile for the Heathrow data (lower values are better). The horizontal dashed line is the 5% critical value. The method constructed from the ensemble mean with quantile autoregression, Method Q3, performed well in terms of unconditional coverage for all lead times in Figure 5, but in Figure 7 the results for the dynamic quantile statistic are poor beyond five days-ahead. This shows that, although the estimator has acceptable unconditional coverage, it does not co-vary with the true 5% quantile at the longer lead times; it is not able to capture the dynamic behaviour of the true 5% quantile. The performance in Figure 7 is qualitatively similar to that in Figures 5 and 6 for Method Q1, which uses the AR-GARCH model, and also for Method Q4, which involves the ensemble quantile being debiased using quantile regression.

***** Figure 7 *****

To summarise overall performance, for each lead time, we calculated the ranking of each method, according to the dynamic quantile statistic, for each of the nine quantiles for each of the five series, and then calculated the average of these 45 rankings. The two methods that performed the best were Method Q3, which uses the ensemble mean with quantile autoregression, and Method Q5, which involves the ensemble quantiles being debiased using TVP quantile regression. The former was better for the early lead times, and the latter for the later lead times.

Informational Content

The third measure that we use for post-sample evaluation is the $R^1(\theta)$ measure, which was proposed by Koenker and Machedo (1999) as a measure of in-sample goodness-of-fit. $R^1(\theta)$ is the quantile regression analogue of the OLS regression R^2. Instead of using a sum of squares cost function, $R^1(\theta)$ uses the quantile regression cost function given in expression (4). In view of

the popularity of the OLS regression R^2 for evaluating the informational content of post-sample volatility forecasts, Taylor (1999) proposes the use of $R^1(\theta)$ to evaluate post-sample quantile forecasts. $R^1(\theta)$ is recorded for the quantile regression performed using post-sample data with the quantile estimator as sole regressor. The measure then assesses the degree to which the estimator co-varies with the true quantile, and, unlike the dynamic quantile statistic, $R^1(\theta)$ controls for unconditional coverage by first debiasing the estimator using quantile regression.

The $R^1(\theta)$ results for the 5% quantile for the Heathrow data are presented in Figure 8 (higher values are better). The most noticeable feature of the figure is the 20% to 30% difference between the $R^1(\theta)$ values for Methods Q4 and Q5, which use the ensemble quantiles, and Method Q1, which is based on the AR-GARCH model. This difference is largely due to the superiority of the ensemble-based methods in forecasting the mean of the series. However, it is important to note that in Figure 8, at all lead times, the methods based on ensemble quantiles, Methods Q4 and Q5, outperform those based on just the ensemble mean, Methods Q2 and Q3. This shows that there is informational content in the distribution of the 51 ensemble members that is not captured by the ensemble mean with univariate modelling of the uncertainty.

***** Figure 8 *****

To summarise $R^1(\theta)$ performance at each lead time, we calculated the ranking of each method for each of the nine quantiles for each of the five series, and then calculated the average of these 45 rankings. The relative performances of the methods were very similar to those shown in Figure 8 for the 5% Heathrow quantile.

SUMMARY AND CONCLUDING COMMENTS

Density forecasts provide an understanding of the uncertainty in weather variables, which is useful for the many industries exposed to weather risk. We have investigated the use of ensemble predictions in forecasting the density of temperature at five locations in the UK. We

first considered estimation of the mean of the density, or, in other words, point forecasting. Our results confirm that the mean of the 51 ensemble members is a better point forecast than the traditional high-resolution point forecasts from a meteorological atmospheric model. In addition, we found that the ensemble mean comfortably outperformed the point forecast from the AR-GARCH models for all lead times considered.

Although the distribution of the 51 ensemble members can be viewed as a temperature density forecast, this tends to underestimate substantially the true uncertainty. Some form of recalibration is, therefore, needed. Using a TVP quantile regression approach to debias the quantiles of the 51 ensemble members led to an estimator that performed consistently well across all three evaluation measures used. Overall, we found that quantiles produced from the AR-GARCH method did not match the quality of the ensemble-based methods. This should come as no great surprise given that the AR-GARCH models are based on far simpler information than the ensemble predictions. Therefore, our conclusion is that there is strong potential for the use of ensemble predictions in temperature density forecasting. We also considered combinations of quantile forecasts from the AR-GARCH and the ensemble-based methods (see Granger *et al*., 1989), but it did not offer improvement on the performance of the ensemble quantile debiased using TVP quantile regression. It would seem that this approach provides a good synthesis of univariate and ensemble information.

An area for further research is the analysis of other weather variables. From an energy perspective, wind speed is particularly interesting because it influences both demand and generation. Indeed, wind generation of electricity is currently receiving a lot of attention in Europe because of subsidies for renewable energy. Precipitation is important for hydroelectric power production, as well as for other industries, such as agriculture and transportation. Hyndman and Grumwald (2000) provide a promising time series model for precipitation.

REFERENCES

Bollerslev T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**: 307-327.

Buizza R, Petroliagis T, Palmer TN, 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.

Buizza R, Miller M, Palmer, TN. 1999. Stochastic Simulation of Model Uncertainties. *Quarterly Journal of the Royal Meteorological Society* **125**: 2887-2908.

Campbell S, Diebold FX. 2002. Weather forecasting for weather derivatives. Working Paper 02-046, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Cao M, Wei J. 2000. Pricing the Weather. *Risk Magazine* May: 67-70.

Chatfield C. 1993. Calculating Interval Forecasts. *Journal of Business and Economic Statistics* **11**: 121-135.

Christoffersen PF. 1998. Evaluating Interval Forecasts. *International Economic Review* **39**: 841-862.

Engle RF. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **50**: 987-1008.

Engle RF, Manganelli S. 2002. CAViaR: Conditional Autoregressive Value at Risk By Regression Quantiles. Department of Economics Discussion Paper, University of California, San Diego.

Franses PH, Neele J, van Dijk D. 2001. Modeling Asymmetric Volatility in Weekly Dutch Temperature Data. *Environmental Modelling and Software* **16**: 131-137.

Goodwin P. 1997. Adjusting Judgemental Extrapolations using Theil's Method and Discounted Weighted Regression. *Journal of Forecasting* **16**: 37-46.

Granger CWJ. 1989. Combining Forecasts - Twenty Years Later. *Journal of Forecasting* **8**: 167-173.

Granger CWJ, White H, Kamstra M. 1989. Interval Forecasting: An Analysis Based Upon ARCH-Quantile Estimators. *Journal of Econometrics* **40**: 87-96.

Hyndman RJ, Wand MP. 1997. Nonparametric Autocovariance Function Estimation. *Australian Journal of Statistics* **39**: 313-324.

Hyndman RJ, Grunwald GK. 2000. Generalized Additive Modelling of Mixed Distribution Markov Models with Application to Melbourne's Rainfall. *Australian and New Zealand Journal of Statistics* **42**: 145-158.

Koenker RW, Bassett GW. 1978. Regression Quantiles. *Econometrica* **46**: 33-50.

Koenker R, Machado JAF. 1999. Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association* **94**: 1296-1310.

Leith CE. 1974. Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review* **102**: 409-418.

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The New ECMWF Ensemble Prediction System: Methodology and Validation. *Quarterly Journal of the Royal Meteorological Society* **122**: 73-119.

Palmer TN, Molteni F, Mureau R, Buizza R, Chapelet P, 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.

Taylor JW. 1999. Evaluating Volatility and Interval Forecasts. *Journal of Forecasting* **18**: 111-128.

Taylor JW, Buizza R. 2002. Neural Network Load Forecasting with Weather Ensemble Predictions. *IEEE Transactions on Power Systems* **17**: 626-632.

Taylor JW, Buizza R. 2003. Using Weather Ensemble Predictions in Electricity Demand Forecasting. *International Journal of Forecasting* **19**, 57-70.

Theil H. 1971. Applied Economic Forecasting, North-Holland: Amsterdam.

Tol RSJ. 1996. Autoregressive Conditional Heteroscedasticity in Daily Temperature Measurements. *Environmetrics* **7**: 67-75.

Torró H, Meneu V, Valor E. 2001. Single Factor Stochastic Models with Seasonality Applied to Underlying Weather Derivatives Variables. Manuscript, University of Valencia.

Toth Z, Kalnay E. 1993. Ensemble Forecasting at NMC: The Generation of Perturbations. *Bulletin of the American Meteorological Society* **74**: 2317-2330.

Tracton MS, Kalnay E. 1993. Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects. *Weather and Forecasting* **8**: 379-398.

Young PC, Jakeman AJ, Post DA. 1997. Recent Advances in Data-Based Modelling and Analysis of Hydrological Systems. *Water Science and Technology* **36**: 99-116.

Young PC, Pedregal DJ, Tych W. 1999. Dynamic harmonic regression. *Journal of Forecasting* **18**: 369-394.

Table I. 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	3.40	3.61	3.31	3.49
	(16.80)	(15.81)	(15.98)	(16.93)	(16.53)
μ_1	-0.75	-0.72	-0.75	-0.78	-0.81
	(-8.99)	(-9.12)	(-9.10)	(-9.16)	(-9.37)
μ_2	-1.89	-1.73	-1.87	-1.90	-1.93
	(-14.92)	(-15.69)	(-14.72)	(-14.65)	(-15.00)
μ_3	0.33	0.24	0.26	0.32	0.35
	(4.49)	(3.76)	(3.69)	(4.16)	(4.65)
μ_4					
ϕ_1	0.71	0.72	0.72	0.71	0.70
	(42.47)	(43.29)	(42.37)	(41.19)	(38.43)
Equation for Variance					
ω_0	0.85	0.49	1.40	0.70	1.32
	(3.06)	(1.83)	(2.94)	(1.62)	(3.43)
ω_1					
ω_2	0.68			0.93	0.74
	(2.96)			(2.16)	(3.10)
ω_3					
ω_4		-0.42			
		(-2.28)			
α	0.08	0.07	0.08	0.05	0.09
	(4.30)	(4.21)	(3.29)	(2.33)	(4.49)
β	0.62	0.60	0.50	0.66	0.52
	(9.88)	(8.80)	(4.27)	(8.09)	(5.83)
γ_0	-1.60	-0.04	-0.40	-3.09	-1.65
	(-3.48)	(-0.11)	(-0.93)	(-2.32)	(-3.80)
γ_1					
γ_2	3.04	4.66	3.37	3.22	3.02
	(4.29)	(5.00)	(3.36)	(2.17)	(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
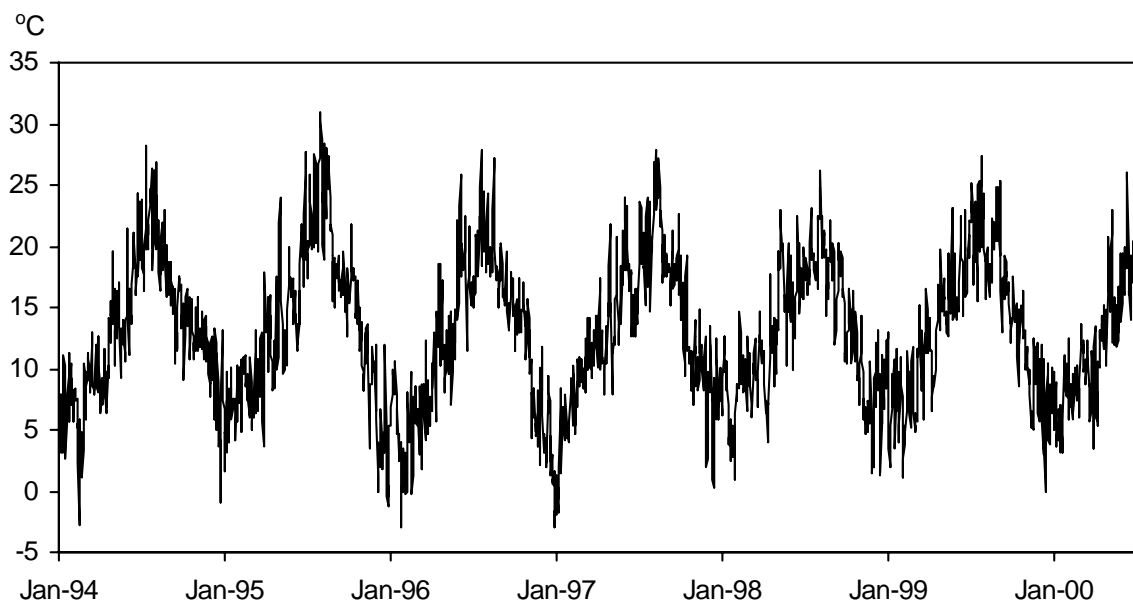
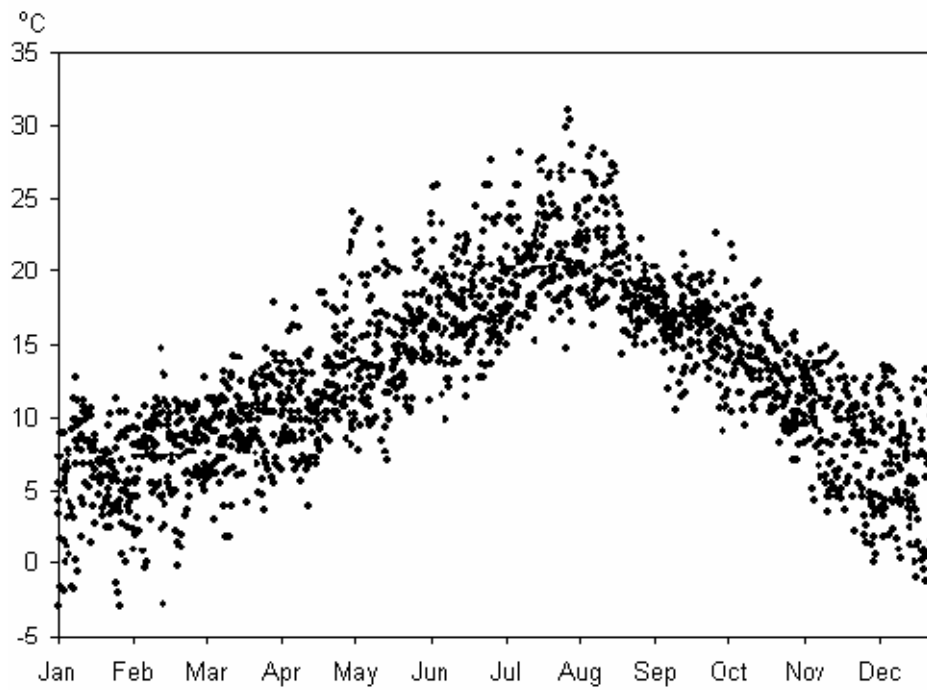Figure 1. Daily midday temperature observations at Heathrow

Figure 2. Daily midday temperature observations at Heathrow plotted
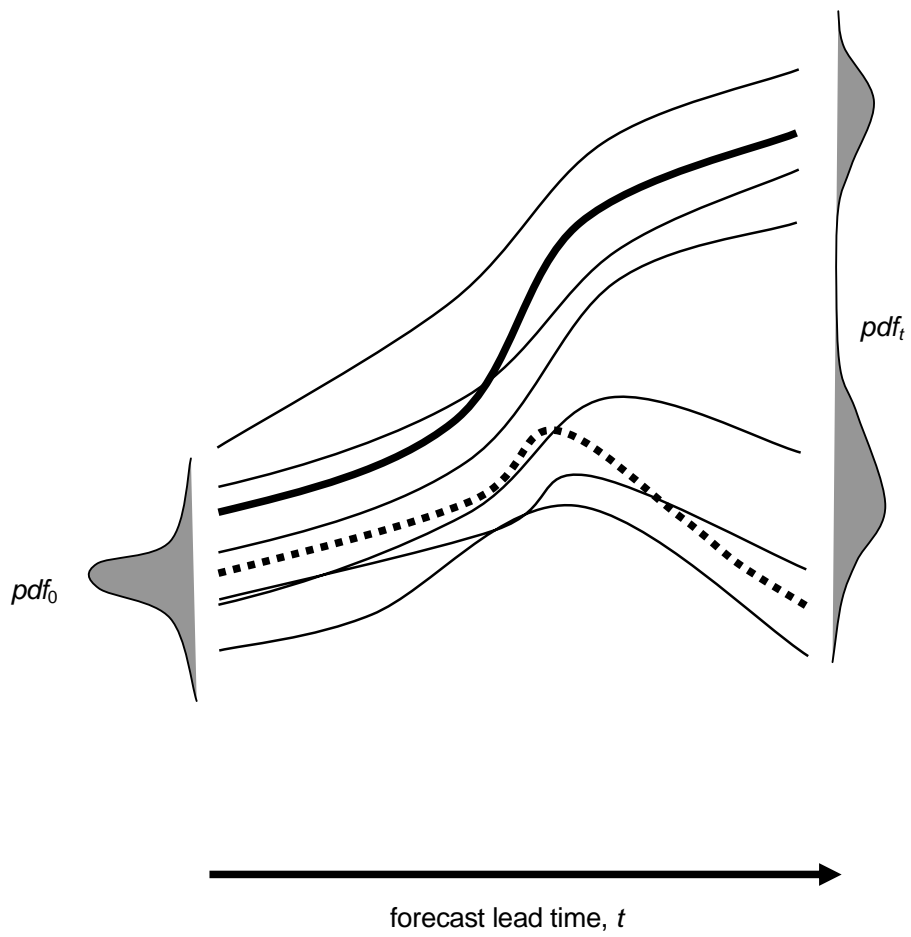against the day of the year for the estimation period 1994 to 1998

Figure 3. Schematic of ensemble prediction. Bold solid curve is the single point forecast. Dashed curve is the future state. Thin solid curves are the ensemble of perturbed forecasts.
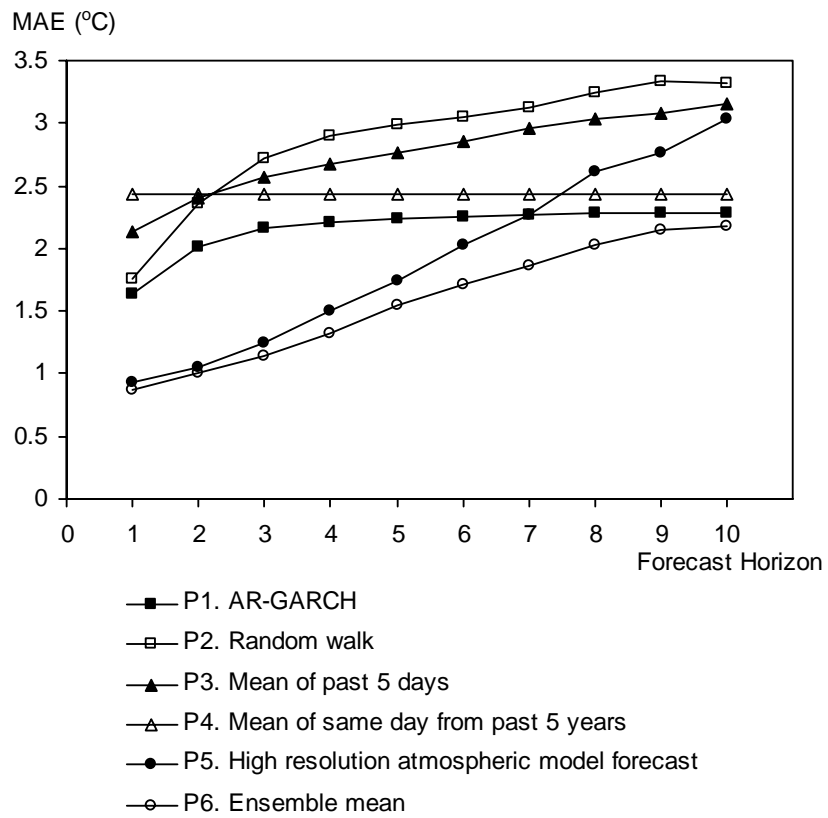
MAE (°C)

Forecast Horizon

—■— P1. AR-GARCH
—□— P2. Random walk
—▲— P3. Mean of past 5 days
—△— P4. Mean of same day from past 5 years
—●— P5. High resolution atmospheric model forecast
—○— P6. Ensemble mean

Figure 4. MAE for each of the different approaches to
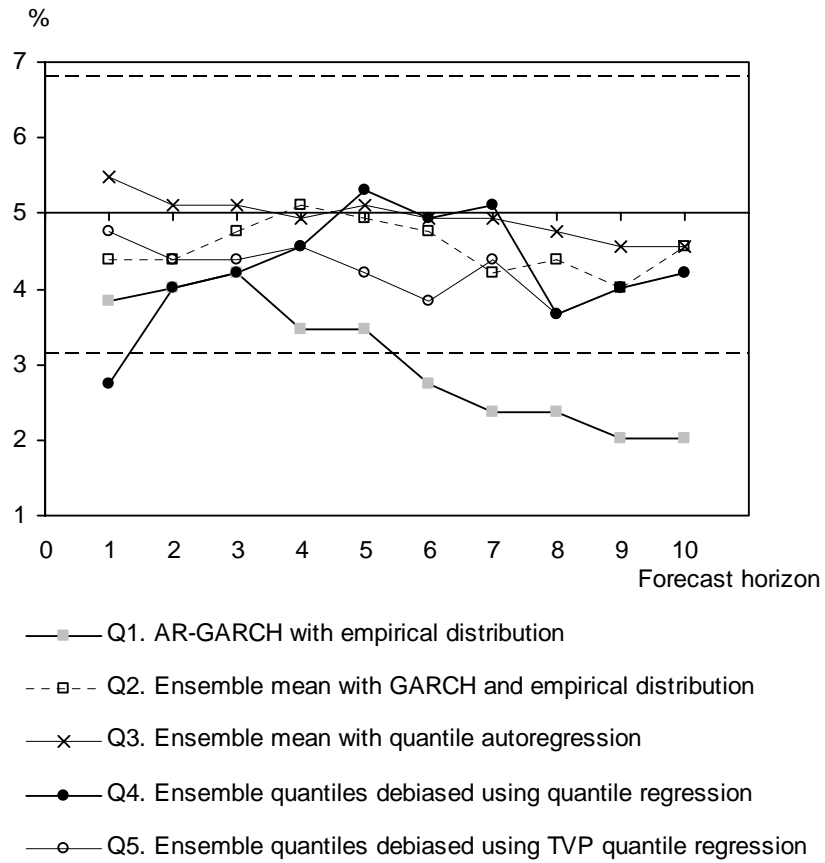point forecasting applied to the Heathrow temperature series.

Figure 5. Unconditional coverage percentage for the different approaches
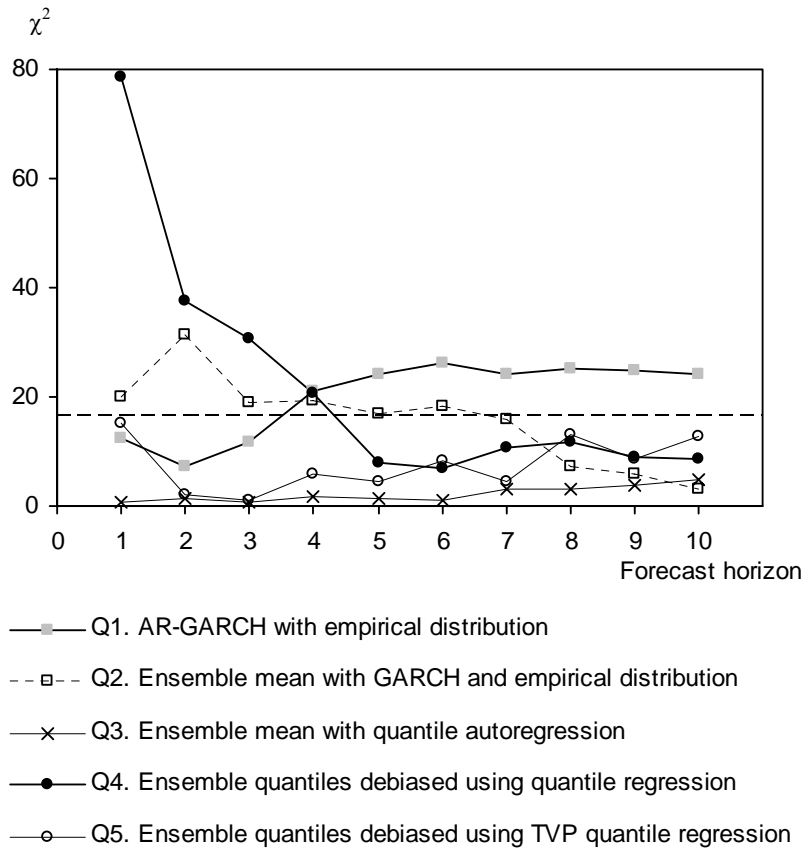to forecasting the 5% quantile of the Heathrow temperature data.

Figure 6. Unconditional coverage chi-square statistic for the different
approaches to forecasting the quantiles for the Heathrow temperature data.
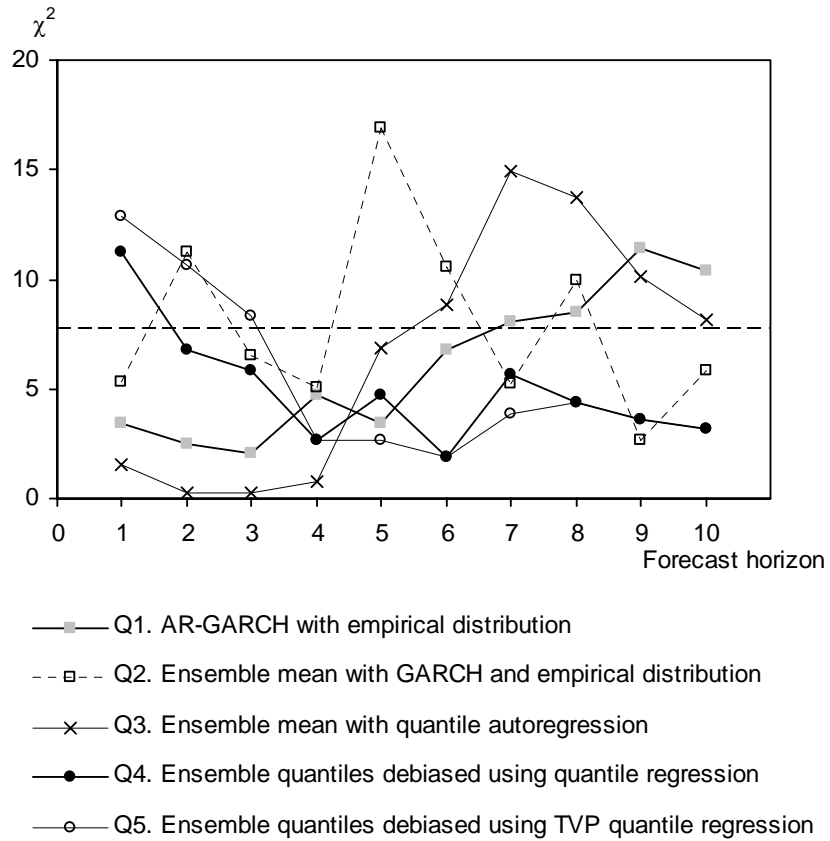The statistic summarises performance across all nine quantiles.

Figure 7. Dynamic quantile chi-square statistic for the different approaches
to forecasting the 5% quantile for the Heathrow temperature data.

%

100

80

60

40

20

0

0   1   2   3   4   5   6   7   8   9   10

Forecast horizon

⸺■⸺ Q1. AR-GARCH with empirical distribution

- - ⊡ - - Q2. Ensemble mean with GARCH and empirical distribution

⸺×⸺ Q3. Ensemble mean with quantile autoregression

⸺●⸺ Q4. Ensemble quantiles debiased using quantile regression

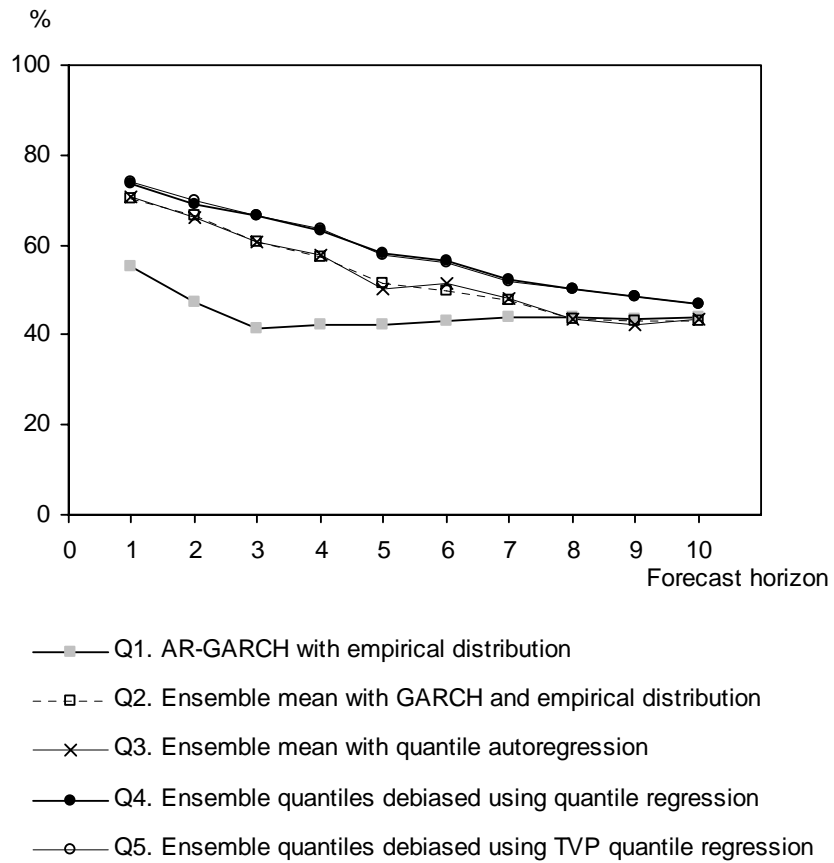⸺○⸺ Q5. Ensemble quantiles debiased using TVP quantile regression

Figure 8. $R^1(\theta)$ for the different approaches to forecasting
the 5% quantile for the Heathrow temperature data.