Probabilistic Forecasting of Daily Extreme Temperature using Bivariate Models of Daily Minimum and Maximum Time Series

31st January 2017

Xiaochun Meng and James W. Taylor

Saïd Business School, University of Oxford

Address for Correspondence:
Xiaochun Meng
Saïd Business School
University of Oxford
Park End Street Oxford
OX1 1HP, UK
Email: xiaochun.meng@sbs.ox.ac.uk

Abstract

Understanding changes in the frequency, severity, and seasonality of daily temperature extremes is important for public policy decisions regarding heat and cold waves. We model the daily minimum and maximum temperature using a bivariate VARMA-MGARCH model, with conditional dependency modelled using a dynamic copula. A useful by-product is the implicit modelling of the daily average and the diurnal temperature range, which has been used as an index of climate change. We model Spanish data recorded over a 65-year-period. To evaluate bivariate density estimates, we propose a new weighted energy score in order to focus on the extremes of the density.

Keywords: Climate; Daily Temperature; VARMA-MGARCH; Energy Score.

JEL: C22, C53, G10

1 Introduction

Extreme weather events can have negative impacts on our society and economy. For example, heat waves and cold snaps can lead to serious health problems and increased morbidity (Wang et al. 2013; Dupuis 2012, 2014). With climate change, the frequency and severity of extreme weather is increasing (Meehl and Tebaldi 2004). In this paper, we focus on the probabilistic forecasting of heat waves. Heat waves are loosely defined as periods of unusually hot weather. More precise definitions are typically based on the daily maximum temperature exceeding some threshold for one or possibly several successive days. However, heat waves are also sometimes defined as the simultaneous exceedances of the daily minimum and maximum temperature over some chosen thresholds (see, for example, Keellings and Waylen 2014). This has been the definition adopted by the UK Met Office in relation to some of their work for the UK Department of Public Health. Therefore, to support different definitions, in this paper, we aim to improve the estimation of both the marginal and joint distributions of the daily minimum and maximum temperature.

Probabilistic temperature forecasts can be obtained from a statistical time series model or a physical model, such as a Numerical Weather Prediction (NWP) system. By running an NWP model with multiple different initial conditions, an ensemble of predictions can be produced, and these can be used as the basis of probabilistic prediction. However, a procedure is required to convert the information in the ensemble into a probabilistic forecast (see, for example, Taylor and Buizza 2004). A further obstacle to their use is that their derivation is computationally intensive, which has implications for their cost, and geographical coverage. Ensemble predictions for the daily minimum and maximum are currently not available. By contrast, historical observations of daily minimum and maximum temperature are readily available for many locations, which allows statistical time series models to be produced. These models must be able to capture the relatively complex dynamics in daily temperature data. Indeed, a careful modelling of a long time series of temperature data provides an

opportunity to confirm, or perhaps correct, our understanding of how temperature extremes have been changing.

Common statistical approaches to estimating the conditional probability distributions of daily temperature involve univariate times series models. Typical examples are the use of autoregressive (AR) and generalised autoregressive conditional heteroskedasticity (GARCH) models, through either a single step or a multi-step approach (see, for example, Taylor and Buizza 2004; Campbell and Diebold 2005; Dupuis 2012, 2014; Erhardt et al. 2015). When such models are used for daily minimum or maximum data, a further modelling stage must be employed to produce the joint distribution, which is needed if the definition of a heat wave is based on the simultaneous exceedance of the minimum and maximum temperature above chosen thresholds. However, the obvious alternative is to model jointly the two temperature variables. This has the appeal of simplicity, efficiency, and allowing a richer model structure for each variable. Indeed, a strong argument for modelling the two variables jointly is that, even if an estimate of only the marginal distribution of one variable is required, a joint model can enable improved estimation because it makes better use of the information available. Another advantage of a joint model is that it can be used to produce density forecasts of the mean of the daily minimum and maximum, which is often studied under the nomenclature of *average daily temperature* (see, for example, Campbell and Diebold 2005), and also the density forecast of the diurnal temperature range (DTR), which has been used as an index of climate (see, for example, Qu et al. 2014). Indeed, the fuller information set used in the joint model may lead to more accurate forecasts of the mean and the DTR than two separate univariate models. Since multivariate models, such as the vector autoregressive moving average (VARMA) and the multivariate generalised autoregressive conditional heteroskedasticity (MGARCH) models, have been applied successfully in various applications (see, for example, Bauwens et al. 2006; Jeon and Taylor 2012), we consider their use for the new application of the simultaneous modelling of the joint distribution of the daily minimum and maximum temperature.

In this paper, we also propose a novel method for evaluating multivariate distributions.

In practice, a density forecast, which is accurate for the centre of the distribution, might have an unsatisfactory fit for the tails of the distribution. Indeed, in our study, the tails are of particular interest, and this should really be reflected in our evaluation of the forecasts. In this paper, we use several different evaluation measures. To evaluate the overall density forecast accuracy, we use the Continuously Ranked Probability Score (CRPS) and the energy score to evaluate to evaluate forecasts of marginal and joint densities, respectively (Gneiting and Raftery 2007). The use of the CRPS has become widespread, and interest in the energy score is increasing. To evaluate the accuracy of the tails of the density forecast, we use the weighted CRPS of Gneiting and Ranjan (2011), which allows us to put greater weight on the tails of the density. We contribute to this literature by introducing a new weighted form of the energy score, which enables us to focus more on the extremes of forecasts of joint densities. We also use the popular Brier score to evaluate probability forecasts of heat waves, defined as the simultaneous exceedance of the daily minimum and maximum temperature over some thresholds. In our empirical work, we use Spanish temperature data, which is particularly suitable for this study because Southern Europe has experienced an increased number of heat waves in recent years. The dataset consists of 65 years of observations of the daily minimum and maximum temperature. Our results show that a bivariate VARMA-MGARCH model is able to outperform univariate models in terms of both the marginal and joint distribution forecast accuracy.

The rest of the paper is structured as follows. Section 2 describes the data and its characteristics. Section 3 provides a literature review and a description of our proposed models. Section 4 briefly reviews the literature on the evaluation of density forecasts, and introduces the new weighted energy score. Section 5 presents empirical results, and Section 6 provides a brief summary and conclusion.

2 Data

We consider the historical data for the following four cities in Spain: Albacete, Seville, Cáceres and Madrid. Our dataset consists of daily minimum and maximum observations, measured in degrees Celsius, for the 65-year period from 1951 to 2015, inclusive. We obtained the data from the website of the European Climate Assessment and Dataset project (http://www.ecad.eu/). As has become standard in the literature, the observations for 29 February, occurring in each leap year, are removed from the series in order to have a constant 365 days in each year (see, for example, Campbell and Diebold 2005; Dupuis 2012). In our empirical work, we used the first 60 years of data to specify models, and then used rolling windows of the same length to estimate model parameters. We used the final 5 years of data to evaluate post-sample one day-ahead forecasts. Figure 1 is a plot of the daily minimum temperature at Seville for the first estimation period 1951 to 2010, and Figure 2 is the corresponding plot of the daily maximum. The plots provide some indication of an overall rising trend over the latter half of the 60-year series, which is consistent with the widely discussed rise in global temperatures since the 1970s. Naturally, the series possesses annual seasonal cyclicality, which accounts for the repeating periodic spikes in the time series plot. The seasonality is clear from Figure 3, which plots the daily minimum and maximum temperature observations for Seville against the day of the year for the 60-year period 1951 to 2010. Note that, at least in the minimum temperature observations, there is an annual cycle in both the mean and the variance, with the mean obviously at its highest in the summer months, while the variance is at its highest in the winter.

3 Time Series Models for Daily Temperature

In this section, we first provide a review of the literature that has considered the modelling of daily temperature time series. We then describe univariate and bivariate models that we implement for daily minimum and maximum temperature.

3.1 Literature Review

As we stated in Section 1, there is no existing literature considering the joint modelling of daily minimum and maximum temperature. There are, however, several studies that have modelled either the daily minimum, maximum or average using univariate models. It has been shown that daily series of the minimum, maximum and average, exhibit the following features: a trend in the mean and variance; a seasonal pattern in the mean and variance, where the pattern can be yearly or of different length; and both large and small absolute deviations from the mean tend to cluster (Taylor and Buizza 2004; Dupuis 2012, 2014). These characteristics are particularly suitable to be modeled by autoregressive models, where a plethora of literature can be found in financial econometrics. Tol (1996) and Taylor and Buizza (2004) use AR-GARCH models to estimate the distribution of daily average temperature. In their approaches, the modelling of the mean and variance involves the inclusion of seasonal terms based on quadratic functions of a counter for the day of the year. Campbell and Diebold (2005) also implement an AR-GARCH model for the average daily temperature, but their seasonal terms are based on Fourier terms. In fitting an ARMA model to daily maximum data, Wong (2015) uses a relatively complex model for the mean, but does not consider a model for the variance. Dupuis (2012; 2014) uses a multi-step approach in her univariate modelling of the daily minimum and maximum temperature; the series are preprocessed by an AR-GARCH model before an extreme value theory (EVT) approach is used to estimate the tails of the resultant residuals. Erhardt et al. (2015) uses a copula approach to estimate the joint distribution of the daily average temperature in different locations, where the marginal distributions are estimated by a multi-step AR model. Empirical results have provided support for the AR-GARCH models in comparison with a variety of simpler alternatives. In this paper, we extend the literature by considering ARMA structures for the mean, rather than just AR, and we implement a bivariate model for the minimum and maximum daily temperature.

3.2 ARMA-GARCH Models

The univariate model that we consider is the ARMA-GARCH model of expressions (1)-(5). In our empirical work, we applied the model to the daily minimum and maximum temperature, as well as the daily average temperature and the DTR.

$$
\begin{aligned}
T_t &= S_1(\boldsymbol{\mu}_1, t) + \psi_1 Trend_t + S_2(\boldsymbol{\mu}_2, t) Trend_t + y_t && (1)\\
y_t &= \sum_{i=1}^{m} \phi_i y_{t-i} + \sum_{i=1}^{n} \theta_i \epsilon_{t-i} + \epsilon_t && (2)\\
\epsilon_t &= h_t^{1/2} \eta_t && (3)\\
h_t &= S_3(\boldsymbol{\omega}_1, t) + \psi_2 Trend_t + S_4(\boldsymbol{\omega}_2, t) Trend_t \\
&\quad + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{o} \gamma_i \epsilon_{t-i}^2 I(\epsilon_{t-i} < 0) + \sum_{i=1}^{p} \beta_i h_{t-i} && (4)\\
S_i(\boldsymbol{\lambda}, t) &= \lambda_0 + \sum_{j=1}^{J_i} \lambda_{1,j} sin\left(2j\pi \frac{d(t)}{365}\right) + \lambda_{2,j} cos\left(2j\pi \frac{d(t)}{365}\right) && (5)
\end{aligned}
$$

where T_t is the temperature variable; $\boldsymbol{\mu}_i$ and $\boldsymbol{\omega}_i$ are vectors of parameters; ϕ_i, θ_i, ψ_i, α_i , γ_i, β_i , λ_0 and $\lambda_{i,j}$ are scalar parameters; $\boldsymbol{\lambda}$ is a vector with entries λ_0 and $\lambda_{i,j}$; $S_i(\boldsymbol{\lambda}, t)$ are the seasonal terms involving the sum of pairs of Fourier terms; $d(t)$ is a repeating step function that numbers the days of the year from 1 to 365 within each year; J_i denotes the number of pairs of Fourier terms; y_t is the stochastic part of T_t; ϵ_t is the error term in the ARMA process for y_t; h_t is the variance of ϵ_t; η_t is an i.i.d. distribution with mean 0 and variance 1; and m, n, q, o, and p are non-negative integers representing the orders of the ARMA and GARCH components. In Section 2, we discussed how our daily minimum and maximum temperature time series exhibited an apparent trend, which is obviously consistent with the literature on climate change. Unit root tests rejected the hypothesis of the trend being stochastic. We considered a variety of approaches to modelling a deterministic trend in the mean and variance of the series, including linear and quadratic functions of time, but these delivered relatively poor fit. Inspection of the time series reveals that they appear to rise steadily only from around the 1970s. This led us to consider a trend defined as being zero up until the start of a chosen year, and linear

thereafter. We chose the year by experimenting with different years, ranging from 1960 to 1990, in order to find the model with the best Schwarz Bayesian Criterion (SBC). We did this for the daily minimum and maximum series for all four locations, and found that the optimal starting point for the linear trend was close to 1974. In view of this, we defined the variable $Trend_t$ in expressions (1)-(5) as being a linear trend starting on 1 January 1974. In Figure 4, we show a trend of this type fitted to the daily minimum temperature time series for Seville. Although we feel this simplistic modelling of the trend is reasonable for our study, which has its emphasis on short-term probabilistic prediction, we acknowledge that there is potential for the incorporation of a more sophisticated approach, such as the semiparametric panel model used by Atak et al. (2011) for monthly data.

By contrast with most applications of autoregressive models to daily temperature, we include an ARMA process for the mean. However, a more novel feature of our model is that we have included, in the expressions for the mean and variance, an interaction term, which is the product of the trend and seasonality. This allows the model to accommodate a different trend for each day of the year, which would imply that climate change does not have a uniform affect on the different seasons of the year. Figure 5 shows how the resulting estimated trend differs across different days of the year. For clarity of presentation, the figure focuses on just the first day of each month of the year. The idea of allowing the trend to differ across the days of the year was motivated by the work of Proietti and Hillebrand (2016) in their analysis of monthly temperature data. In fitting a GARCH model to a series of weekly average temperature, Franses et al. (2001) include an asymmetric term to accommodate the effect that the impact of temperatures lower than expected on conditional volatility tends to be different from the impact of temperatures higher than expected. Taylor and Buizza (2004) and Dupuis (2014) also include this asymmetry in their GARCH modelling. We incorporate this by using a GJR-GARCH model for the variance in expression (4) (see Glosten et al. 1993).

In the proposed ARMA-GARCH method, e_t in expression (3) can be any i.i.d. distribution with zero mean and unit variance. We consider two different specifications in this study: the Gaussian distribution and the generalised asymmetric skew-t

distribution proposed by Zhu and Galbraith (2010). The latter can be viewed as a generalisation of the skew-t distribution proposed by Fernández and Steel (1998). In addition to a skewness parameter, the distribution allows the degrees of freedom to be different for each side of the mean. The Gaussian distribution, Student's t distribution, and the skew-t are special cases of this distribution. In addition to these distributional assumptions, we consider the EVT approach of McNeil and Frey (2000). This involves the peaks over threshold EVT approach being applied to the standardised residuals obtained from the ARMA-GARCH model fitted using a Gaussian assumption. In this approach, we set the threshold as the 90% quantile of the Gaussian distribution, and fitted a generalised Pareto distribution to the standardized residuals that exceed this threshold.

In addition to the annual cyclicality, we also experimented with the inclusion of Fourier terms to model a possible intraweek cycle, which has been observed in the DTR for a variety of locations (see, for example, Forster and Solomon 2003). However, these terms were not statistically significant in our models, and so we do not consider them further in this study.

The lag orders, the numbers of Fourier terms, and the specification of the trend term considered in the univariate ARMA-GARCH model were chosen using the SBC. For the mean, this led to the use of three pairs of Fourier terms, a trend variable that is zero prior to 1974 and linear thereafter, interaction terms involving this trend and three pairs of Fourier terms, and an ARMA(3,1) model. For the variance, we used two pairs of Fourier terms, the same trend variable as in the model for the mean, interaction terms involving this trend and two pairs of Fourier terms, and a GJR-GARCH(1,1) model. We then used the analogous specifications in the bivariate VARMA-MGARCH models.

Tables 1 and 2 present the parameters for the univariate ARMA-GARCH model with Gaussian assumption, estimated using the first 60-year rolling window of the daily minimum and maximum temperature data recorded at Seville. Most of the parameters are statistically significant at the 5% significance level. The insight provided by the model is consistent with previous studies. The coefficients for the trend terms in the ARMA

parts of both models are positive, which indicates that the levels of the maximal and minimal temperature are rising. The coefficient of the trend term in the ARMA part for the minimum temperature is larger than that for the maximum temperature, which suggests a decrease in the DTR. These findings are consistent with those in previous work (see, for example, Dupuis 2014; Qu et al. 2014). One term of each interaction pair is significant in both the ARMA and GARCH parts of the model with the GARCH part for the maximum temperature being the only exception, implying that the impact of the trend is not the same across the seasonal cycle. This was the finding of Proietti and Hillebrand (2016) for monthly data, while our work has confirmed the existence of the effect in the mean and variance of daily data. We revert from presenting the parameters for other models, because the insight was similar, and because, for the bivariate models, the number of parameters is relatively large.

3.3 VARMA-MGARCH Models

In this section, we present a bivariate model for the daily minimum and maximum temperature. We propose to use a vector autoregressive moving average (VARMA) to model the mean of a vector containing the two variables, and a multivariate autoregressive conditional heteroskedasticity (MGARCH) model for the covariance matrix. There exist many different choices of MGARCH specifications, such as the VEC model (Bollerslev et al. 1988), the dynamic conditional correlation model (Engle 2002), and the BEKK model (Engle and Kroner 1995). In this paper, we choose to use the VEC model, which has a very flexible and general specification. The implementation of VEC is usually difficult and computationally demanding in high-dimensional cases, but in this study, we are facing only a bivariate situation, which does not lead to overcomplicated numerical

implementation. The proposed VARMA-MGARCH model is described as follows:

$$\begin{aligned}
\boldsymbol{T}_t &= \boldsymbol{S}_1(\boldsymbol{\mu}_1, t) + \boldsymbol{\psi}_1 Trend_t + \boldsymbol{S}_2(\boldsymbol{\mu}_2, t) Trend_t + \boldsymbol{y}_t && (6)\\
\boldsymbol{y}_t &= \sum_{i=1}^{m} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \sum_{i=1}^{n} \boldsymbol{\theta}_i \boldsymbol{\epsilon}_{t-i} + \boldsymbol{\epsilon}_t && (7)\\
\boldsymbol{\epsilon}_t &= \boldsymbol{H}_t^{1/2} \boldsymbol{\eta}_t && (8)\\
vech(\boldsymbol{H}_t) &= \boldsymbol{S}_3(\boldsymbol{\omega}_1, t) + \boldsymbol{\psi}_2 Trend_t + \boldsymbol{S}_4(\boldsymbol{\omega}_2, t) Trend_t \\
&\quad + \sum_{i=1}^{q} \boldsymbol{\alpha}_i vech(\boldsymbol{\epsilon}_{t-i} \boldsymbol{\epsilon}_{t-i}^{'}) \\
&\quad + \sum_{i=1}^{o} \boldsymbol{\gamma}_{t-i} vech(\boldsymbol{\epsilon}_{t-i} \boldsymbol{\epsilon}_{t-i}^{'}) I\left(vech(\boldsymbol{\epsilon}_{t-i} \boldsymbol{\epsilon}_{t-i}^{'}) < \mathbf{0}\right) \\
&\quad + \sum_{i=1}^{p} \boldsymbol{\beta}_i vech(\boldsymbol{H}_{t-i}) && (9)\\
\boldsymbol{S}_i(\boldsymbol{\lambda}, t) &= \boldsymbol{\lambda}_0 + \sum_{j=1}^{J_i} \boldsymbol{\lambda}_{1,j} sin\left(2j\pi \frac{d(t)}{365}\right) + \boldsymbol{\lambda}_{2,j} cos\left(2j\pi \frac{d(t)}{365}\right) && (10)
\end{aligned}$$

where $\boldsymbol{T}_t = (T_1,\ T_2)$ is a vector of the two temperature variables; *vech* denotes the operator that stacks the lower triangular portion of a matrix as a column vector; $\boldsymbol{\mu}_i$, $\boldsymbol{\omega}_i$, $\boldsymbol{\psi}_i$, $\boldsymbol{\gamma}_i$, $\boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}_{i,j}$ are vectors of parameters; $\boldsymbol{\lambda}$ is a vector consisting of the concatenation of λ_0 and $\lambda_{i,j}$; $\boldsymbol{\phi}_i$, $\boldsymbol{\theta}_i$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}_i$ are parameter matrices; $\boldsymbol{S}_i(\boldsymbol{\lambda}, t)$ is a deterministic seasonal vector; $\boldsymbol{y}_t$ contains the stochastic parts of $\boldsymbol{T}_t$; J_i denotes the number of pairs of Fourier terms; $\boldsymbol{\epsilon}_t$ is the error term of $\boldsymbol{y}_t$; $\boldsymbol{H}_t$ is the covariance matrix of $\boldsymbol{\epsilon}_t$; $\boldsymbol{\eta}_t$ is a vector with entries that follow an i.i.d. distribution with zero mean vector and identity covariance matrix; m, n, q, o, and p are non-negative integers representing the order of lagged variables. We assume $\boldsymbol{\eta}_t$ is a standard multivariate Gaussian distribution, which is a common assumption in the MGARCH literature (Bauwens et al. 2006). $d(t)$ and $Trend_t$ are defined as in the univariate ARMA-GARCH model. In comparison with the ARMA-GARCH model of Section (3.2), the VARMA-MGARCH model has the same essential features, such as the trend, seasonal and interaction terms, as well as a model for the conditional covariance between the daily minimum and maximum temperatures.

3.4 VARMA-MGARCH with a Dynamic Copula

In a nutshell, the VARMA-MGARCH model is a multivariate generalisation of the univariate ARMA-GARCH model in Section 3.2. The choice of the distribution for $\boldsymbol{\eta}_t$ in expression (8) is, however, much more complicated than in the univariate case. Apart from the multivariate Gaussian distribution, other candidates that have been considered, in other applications, are the multivariate Student's t distribution and different versions of multivariate skew-t distribution (Azzalini and Capitanio 2003; Bauwens and Laurent 2005; Lee and Long 2009). However, a limitation of these Student's t and skew-t distributions is that the two marginal distributions are forced to have the same degrees of freedom. This is a limitation because we found that the marginal distributions of minimum and maximum temperature have very different degrees of freedom when we used the univariate forms of these distributions in the univariate ARMA-GARCH models of Section 3.2. To overcome this problem, and to enable a flexible modelling of the dependence structure between marginal distributions, we propose the use of a copula. By Sklar's theorem, a copula enables the decomposition of a joint distribution into the product of its marginal distributions and its dependence function, which is the copula. The joint distribution of a bivariate random variable (X, Y) can be decomposed as follows:

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y)) \tag{11}$$

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) c(F_X(x), F_Y(y)) \tag{12}$$

where C is the copula function; c is the copula density function; and expression (11) and expression (12) are the cumulative distribution function (CDF) and probability density function decompositions of (X, Y), respectively. Detailed reviews of copulae can be found in (Kolev et al. 2006; Patton 2012). We consider the Gaussian copula and the Student's t copula, which are standard in the literature (Patton 2006, 2012). The Gaussian copula has one parameter, which is a correlation parameter, and the Student's t copula has

two parameters, which are a correlation parameter and a parameter for the degrees of freedom. Note that the degrees of freedom of the copula relate only to the dependency structure between X and Y, and not to the marginal distributions of X or Y. The proposed copula-based VARMA-MGARCH model can be written as follows:

$$\begin{aligned}
\boldsymbol{T}_t &= \boldsymbol{S}_1(\boldsymbol{\mu}_1, t) + \boldsymbol{\psi}_1 Trend_t + \boldsymbol{S}_2(\boldsymbol{\mu}_2, t) Trend_t + \boldsymbol{y}_t && (13)\\
\boldsymbol{y}_t &= \sum_{i=1}^{m} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \sum_{i=1}^{n} \boldsymbol{\theta}_i \boldsymbol{\epsilon}_{t-i} + \boldsymbol{\epsilon}_t && (14)\\
\boldsymbol{\epsilon}_t &= (\epsilon_{1,t}, \epsilon_{2,t})' && (15)\\
F_{\boldsymbol{\epsilon}_t}(z_1, z_2) &= C_{R_t,\nu}(F_{\epsilon_{1,t}}(z_1), F_{\epsilon_{2,t}}(z_2)) && (16)\\
var(\epsilon_{j,t}) &= h_{j,t} && (17)\\
\boldsymbol{H}_t &= \begin{pmatrix} h_{1,t} & R_t\sqrt{(h_{1,t}h_{2,t})} \\ R_t\sqrt{(h_{1,t}h_{2,t})} & h_{2,t} \end{pmatrix} && (18)\\
vech(\boldsymbol{H}_t) &= \boldsymbol{S}_3(\boldsymbol{\omega}_1, t) + \boldsymbol{\psi}_2 Trend_t + \boldsymbol{S}_4(\boldsymbol{\omega}_2, t) Trend_t \\
&\quad + \sum_{i=1}^{q} \boldsymbol{\alpha}_i vech(\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-i}') \\
&\quad + \sum_{i=1}^{o} \boldsymbol{\gamma}_{t-i} vech(\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-i}') I\left(vech(\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-i}') < \mathbf{0}\right) \\
&\quad + \sum_{i=1}^{p} \boldsymbol{\beta}_i vech(\boldsymbol{H}_{t-i}) && (19)\\
\boldsymbol{S}_i(\boldsymbol{\lambda}, t) &= \boldsymbol{\lambda}_0 + \sum_{j=1}^{J_i} \boldsymbol{\lambda}_{1,j} sin\left(2j\pi\frac{d(t)}{365}\right) + \boldsymbol{\lambda}_{2,j} cos\left(2j\pi\frac{d(t)}{365}\right) && (20)
\end{aligned}$$

where $h_{j,t}$ is the variance of the error term $\epsilon_{j,t}$; C is the Gaussian or Student's t copula; R_t is the correlation parameter in the Gaussian or Student's t copula; ν is the degrees of freedom for the Student's t copula and is not needed for the Gaussian copula; $F_{\boldsymbol{\epsilon}_t}(\cdot,\cdot)$ is the distribution function of $\boldsymbol{\epsilon}_t$; $F_{\epsilon_{i,t}}(\cdot)$ is the distribution function of $\epsilon_{i,t}$; and the other notation is the same as in the expressions in the previous section. In other words, we replace the covariance matrix decomposition in expression (8) of Section 3.3 by the copula decomposition in expressions (15)-(16). Consequently, we do not use the VEC approach in the standard way to model the dynamics of the covariance matrix of $\boldsymbol{\epsilon}_t$, but instead we use the VEC approach to model the variances of $\epsilon_{j,t}$ and the copula's correlation

parameter simultaneously in expressions (17)-(18). Note that, as the copula's correlation parameter is modelled within the MGARCH VEC structure, the dependency between the error terms is being modelled using a dynamic copula. Other dynamic copulae are considered by Patton (2006; 2012). A feature of our work is that the variances of the marginal distributions and the copula correlation parameter are modelled simultaneously, while other studies have tended to estimate the copula parameter(s) after first modelling the individual marginal distributions.

4 Evaluating Density Forecasts

In this section, we first describe the CRPS and weighted CRPS, which are often used to evaluate univariate density forecasts. We then present the energy score, which is receiving increasing attention as a measure of the accuracy of multivariate density forecasts. We then introduce a new weighted form of the energy score.

4.1 Evaluating Univariate Density Forecasts Using the CRPS

A popular approach to evaluating density forecasts for univariate time series is to use the probability integral transform (PIT) (Diebold et al. 1998; Berkowitz 2001). If the forecasting method is correctly specified, then the PITs in the post-sample period should be independently and uniformly distributed between 0 and 1. A test for this is straightforward to implement, and is widely used (Diebold et al. 1998; Patton 2006; Dupuis 2012). However, it is not clear how to use it to compare the performance between different methods, and how to extend it for multivariate density forecast evaluation. These challenges can be addressed using scoring rules (Gneiting and Raftery 2007; Ziegel et al. 2014; Scheuerer and Hamill 2015). To evaluate the quality of an estimated distribution, a scoring rule assigns a numerical score based on the estimated distribution and on the event or value that materializes. A scoring rule is said to be *proper* if the expected score is always maximized for an observation drawn from the true data generating process (Gneiting and Raftery 2007). Therefore, if the scoring rule

is proper, a forecaster is encouraged to make careful and honest assessments (Garthwaite et al. 2005). Commonly used examples of proper scoring rules are the Brier score for evaluating the probability of a variable with binary or categorical outcomes, the quantile score for evaluating quantile estimates, and the CRPS for evaluating continuous probability distribution estimates (Gneiting and Raftery 2007). For each of these proper scoring rules, after it has been calculated for each post-sample period, the resulting scores are averaged to produce an overall measure of accuracy. Note that throughout Section 4, in order to simplify the notation, we do not include a time subscript on any terms in any of the expressions. Before discussing the CRPS, let us start with the definition of the Brier score:

$$BS(p) = (p - o)^2$$

where p is the predicted probability, o is the actual outcome of the event, which is 0 if the event occurs, and 1 otherwise.

The CRPS can be written, as shown in the following expressions:

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - I(y \geq x))^2 dx \tag{21}$$

$$= \int_{0}^{1} QS_\alpha(F^{-1}(\alpha), y) d\alpha \tag{22}$$

$$= -\frac{1}{2} E_P|X - X'| + E_P|X - y| \tag{23}$$

where $F(\cdot)$ is the CDF of the predictive distribution; X and X' are random variables with distribution F; y is the actual observation; $I(\cdot)$ is the indicator function; and $QS_\alpha(q, y) = 2(I(y < q) - \alpha)(q - y)$ is the quantile score (Gneiting and Raftery 2007). Expression (21) is referred to as the threshold version of the CRPS. This expression shows that CRPS is equal to the integral of the Brier score for the binary event $I(y \geq x)$ at each possible threshold value x. Expression (22) is the quantile version of the CRPS, and it shows the CRPS can be written as the integral of the quantile score for all possible values of the

probability level α. Expression (23) is a potentially useful way to calculate the CRPS in practice (Gneiting and Raftery 2007). There is another equivalent expression for the CRPS in terms of characteristics functions (Gneiting and Raftery 2007; Székely and Rizzo 2013):

$$CRPS(F, y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|\phi_X(r) - e^{i\langle r, y\rangle}|}{r^2} dr \tag{24}$$

where $\phi_F(\cdot)$ is the characteristic function of F. The above expression is of theoretical importance. We will use this expression to help us define a new weighted energy score in Section 4.4. We will term this expression the *characteristic function based CRPS*.

4.2 Evaluating Univariate Density Forecasts Using the Weighted CRPS

In this paper, we are interested in probabilistic prediction of heat waves, thus it is crucial that our methods deliver reliable estimates of the upper tail of the temperature distributions. It has been proposed that weighting schemes can be used in the context of proper scoring rules to emphasise regions of interest (Diks et al. 2011; Gneiting and Ranjan 2011). The weighted CRPS is an example of this (Gneiting and Ranjan 2011). The weighted CRPS can take either of the following two forms:

$$WCRPS_u(F, y) = \int_{-\infty}^{\infty} (F(x) - I(y \leq x))^2 u(x) dx \tag{25}$$

$$WCRPS_v(F, y) = \int_{0}^{1} QS_\alpha(F^{-1}(\alpha), x) v(\alpha) d\alpha \tag{26}$$

where $u(\cdot)$ is a non-negative weight function on the real line, and $v(\cdot)$ is a non-negative weight function on the unit interval. Gneiting and Ranjan (2011) refer to expression (25) as the threshold-weighted CRPS, and expression (26) as the quantile-weighted CRPS. The threshold-weighted CRPS allows higher weights to be assigned to particularly important regions of the real line. The quantile-weighted CRPS enables higher weights to be assigned to regions of the probability level of particular interest. In this paper, we wish to put

a greater emphasis on the upper tail of the density, and so we implement the quantile-weighted CRPS with $v(\alpha) = \alpha^2$, which is the weight suggested by Gneiting and Ranjan (2011) for emphasising the upper tail of a density. For the threshold-weighted CRPS, Gneiting and Ranjan (2011) suggest that the weight function $u(x)$ is set as a Gaussian CDF with mean and variance subjectively chosen. In Section 5, we report the results for an extreme form of this, where we define the weight as $u(x) = I(x > 35)$ for the daily maximum temperature, and $u(x) = I(x > 17.5)$ for the daily minimum.

The weighted CRPS can also be expressed as a weighted integral based on expression (24):

$$WCRPS_w(F, y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|\phi_X(r) - e^{i\langle r,y\rangle}|}{r^2} w(r)dr$$

where $w(r)$ is a non-negative weight function. The above expression has not be proposed and studied so far. We will refer to this expression as the *characteristic function weighted CRPS* .

4.3 Evaluating Multivariate Density Forecasts Using the Energy Score

Gneiting and Raftery (2007) introduce the energy score, as a generalisation of the CRPS, to evaluate multivariate distributions. The definition of the energy score is as follows:

$$ES_\beta(\boldsymbol{P}, \boldsymbol{y}) = -\frac{1}{2} E_P \|\boldsymbol{X} - \boldsymbol{X}'\|^\beta + E_P \|\boldsymbol{X} - \boldsymbol{y}\|^\beta \tag{27}$$

where $\boldsymbol{P}$ is the estimated multivariate distribution, $\boldsymbol{y}$ is the actual observation; $\|\cdot\|$ is the Euclidean norm; $\boldsymbol{X}$ and $\boldsymbol{X}'$ are two independent copies of a random vector drawn from the distribution $\boldsymbol{P}$; and β is a parameter to be decided by the user from the interval $(0, 2)$. When the observation vector is one-dimensional, and we set $\beta = 1$, expression (27) reduces to the form of the CRPS in expression (23). The energy score is based on the work

of Székely and Rizzo (2013) on generalised energy distance. We refer to expression (27) as the *generalised energy score*, and refer to expression (27) when $\beta = 1$ simply as the *energy score*.

In our empirical work, we used the energy score to evaluate forecasts of bivariate distributions. However, just as Gneiting and Ranjan (2011) were motivated to develop a weighted CRPS, we see strong appeal in using a weighted form of the energy score, in order to put greater emphasis on the regions of particular interest in the multivariate distribution. In the remainder of this subsection, we discuss the generalised energy score in greater depth as preparation for presenting a weighted energy score in Section 4.4.

Gneiting and Raftery (2007) present an equivalent expression for the generalised energy score in terms of characteristics functions:

$$ES_\beta(\boldsymbol{P}, \boldsymbol{y}) = \frac{\beta 2^{\beta-2}\Gamma(\frac{d}{2}+\frac{\beta}{2})}{\pi^{\frac{d}{2}}\Gamma(1-\frac{\beta}{2})} \int\limits_{R^d} \frac{|\phi_{\boldsymbol{P}}(\boldsymbol{x}) - e^{i\langle \boldsymbol{x}, \boldsymbol{y}\rangle}|}{\|\boldsymbol{x}\|^{d+\beta}} d\boldsymbol{x} \tag{28}$$

where $\phi_{\boldsymbol{P}}(\cdot)$ is the characteristic function of $\boldsymbol{P}$, $\langle \cdot, \cdot \rangle$ is the inner product, and $i^2 = -1$. The integral in expression (28) can also be written in spherical coordinates:

$$ES_\beta(\boldsymbol{P}, \boldsymbol{y}) = \frac{\beta 2^{\beta-2}\Gamma(\frac{d}{2}+\frac{\beta}{2})}{\pi^{\frac{d}{2}}\Gamma(1-\frac{\beta}{2})} \int\limits_{R^d} \frac{|\phi_{\boldsymbol{P}}(\boldsymbol{x}) - e^{i\langle \boldsymbol{x}, \boldsymbol{y}\rangle}|}{\|\boldsymbol{x}\|^{d+\beta}} d\boldsymbol{x} \tag{29}$$

$$= \frac{\beta 2^{\beta-2}\Gamma(\frac{d}{2}+\frac{\beta}{2})}{\pi^{\frac{d}{2}}\Gamma(1-\frac{\beta}{2})} \int\limits_{S_{d-1}} \int\limits_{0}^{\infty} \frac{|\phi_{\boldsymbol{P}}(r\boldsymbol{a}) - e^{i\langle r\boldsymbol{a}, \boldsymbol{y}\rangle}|}{r^{d+\beta}} r^{d-1} dr d\mu(\boldsymbol{a}) \tag{30}$$

$$= \frac{\beta 2^{\beta-2}\Gamma(\frac{d}{2}+\frac{\beta}{2})}{2\pi^{\frac{d}{2}}\Gamma(1-\frac{\beta}{2})} \int\limits_{S_{d-1}} \int\limits_{-\infty}^{\infty} \frac{|\phi_{\boldsymbol{P}}(r\boldsymbol{a}) - e^{i\langle r\boldsymbol{a}, \boldsymbol{y}\rangle}|}{r^{d+\beta}} r^{d-1} dr d\mu(\boldsymbol{a}) \tag{31}$$

$$= \frac{\beta 2^{\beta-2}\Gamma(\frac{d}{2}+\frac{\beta}{2})}{2\pi^{\frac{d}{2}}\Gamma(1-\frac{\beta}{2})} \int\limits_{S_{d-1}} \int\limits_{-\infty}^{\infty} \frac{|\phi_{\boldsymbol{P}}(r\boldsymbol{a}) - e^{i\langle r\boldsymbol{a}, \boldsymbol{y}\rangle}|}{r^{d+\beta}} dr d\mu(\boldsymbol{a}) \tag{32}$$

$$= \frac{\Gamma(\frac{d}{2}+\frac{\beta}{2})}{2\pi^{\frac{d-1}{2}}\Gamma(\frac{1}{2}+\frac{\beta}{2})}$$

$$\int\limits_{S_{d-1}} \int\limits_{-\infty}^{\infty} \frac{\beta 2^{\beta-2}\Gamma(\frac{1}{2}+\frac{\beta}{2})}{\pi^{\frac{1}{2}}\Gamma(1-\frac{\beta}{2})} \frac{|\phi_{\boldsymbol{P}}(r\boldsymbol{a}) - e^{ir\langle \boldsymbol{a}, \boldsymbol{y}\rangle}|}{r^{d+\beta}} dr d\mu(\boldsymbol{a}) \tag{33}$$

$$= \frac{\Gamma(\frac{d}{2}+\frac{\beta}{2})}{2\pi^{\frac{d-1}{2}}\Gamma(\frac{1}{2}+\frac{\beta}{2})}$$
$$\int_{S_{d-1}}\int_{-\infty}^{\infty} \frac{\beta 2^{\beta-2}\Gamma(\frac{1}{2}+\frac{\beta}{2})}{\pi^{\frac{1}{2}}\Gamma(1-\frac{\beta}{2})} \frac{|\phi_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}(r) - e^{ir\langle \boldsymbol{a},\boldsymbol{y}\rangle}|}{r^{d+\beta}} dr d\mu(\boldsymbol{a}) \tag{34}$$

where $\Gamma(\cdot)$ denotes the Gamma function; d is the dimension of $\boldsymbol{y}$; $\phi_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}(r)$ denotes the characteristic function of the random variable $\langle \boldsymbol{a}, \boldsymbol{X}\rangle$, where $\boldsymbol{X}$ has distribution $\boldsymbol{P}$; S_{d-1} is the unit $(d-1)$ dimensional sphere; $\boldsymbol{a}$ is any vector on S_{d-1}; and $\mu(\boldsymbol{a})$ is the uniform distribution on S_{d-1}. Expression (30) comes from the fact that any $\boldsymbol{x} \in R^d$ can be expressed in the form of $r\boldsymbol{a}$, where $\boldsymbol{a} \in S_{d-1}$ and r is a non-negative scalar. Expression (31) comes from the fact that $r\boldsymbol{a} = -r(-\boldsymbol{a})$, and as $\boldsymbol{a}$ runs through S_{d-1}, so does $-a$, therefore, expression (30) is equal to the same expression with the integration with respect to r having lower limit of $-\infty$ and upper limit of 0. Expression (34) uses the fact that $\phi_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}(r) = E(\exp(i\langle \boldsymbol{a}, \boldsymbol{X}\rangle)) = E(\exp(i\boldsymbol{a}'\boldsymbol{X})) = \phi_{\boldsymbol{P}}(r\boldsymbol{a})$. Notice that the inner integral in expression (34) is just the one-dimensional generalised energy score. Expression (34) provides intuition for the generalised energy score: given realization $\boldsymbol{y}$, a probabilistic distribution is optimal in terms of the generalised energy score if and only if it is optimal in terms of the generalised energy score for every one dimensional realization $\langle \boldsymbol{a}, \boldsymbol{X}\rangle$.

In our work with multivariate densities in this paper, we consider the bivariate case of $d = 2$, and we set $\beta = 1$, as this has become standard in the literature that has used the energy score. With these values of d and β, by expression (24), expression (34) can be written as:

$$ES_1(\boldsymbol{P}, \boldsymbol{y}) = \frac{1}{4}\int_{S_1} CRPS(P_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}, \langle \boldsymbol{a}, \boldsymbol{y}\rangle) d\mu(\boldsymbol{a}) \tag{35}$$

where $P_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}$ is the distribution function of $\langle \boldsymbol{a}, \boldsymbol{X}\rangle$. This expression shows that, for $d = 2$, the energy score is essentially equal to the CRPS integrated over the unit circle.

4.4 Evaluating Multivariate Density Forecasts Using the Energy Score

Let us now propose a definition for a weighted energy score, building on the form of the generalised energy score in expression (34):

Definition 1. We define the weighted generalised energy score for the predictive distribution $\boldsymbol{P}$ and the realization $\boldsymbol{y}$ as:

$$WES_{\beta}(\boldsymbol{P}, \boldsymbol{y}) = \frac{\Gamma(\frac{d}{2} + \frac{\beta}{2})}{2\pi^{\frac{d-1}{2}}\Gamma(\frac{1}{2} + \frac{\beta}{2})} \int_{S_{d-1}} \int_{-\infty}^{\infty} \frac{\beta 2^{\beta-2}\Gamma(\frac{1}{2} + \frac{\beta}{2})}{\pi^{\frac{1}{2}}\Gamma(1 - \frac{\beta}{2})} \frac{|\phi_{\langle \boldsymbol{a}, \boldsymbol{X} \rangle}(r) - e^{ir\langle \boldsymbol{a}, \boldsymbol{y} \rangle}|}{r^{\beta+1}} w(\boldsymbol{a}, r) dr d\mu(\boldsymbol{a}) \tag{36}$$

where $w(\boldsymbol{a}, r)$ is a non-negative weight function. The weight function $w(\boldsymbol{a}, r)$ can be different for different values of $\boldsymbol{a}$. The fact that the generalised energy score is a proper scoring rule, and the weight function is non-negative, immediately implies that the proposed weighted generalised energy score is a proper scoring rule.

For our empirical work with bivariate distributions, we set $d = 2$ and $\beta = 1$ in expression (36), which leads to the following weighted energy score:

$$WES_{1}(\boldsymbol{P}, \boldsymbol{y}) = \frac{1}{4} \int_{S_1} \int_{-\infty}^{\infty} \frac{|\phi_{\langle \boldsymbol{a}, \boldsymbol{X} \rangle}(r) - e^{ir\langle \boldsymbol{a}, \boldsymbol{y} \rangle}|}{2\pi r^2} w(\boldsymbol{a}, r) dr d\mu(\boldsymbol{a}) \tag{37}$$

The inner integral is nothing but the characteristic function form of weighted CRPS, thus, we will refer to expression as the *characteristic function weighted energy score*.

Recall that we have two different weighting methods for the CRPS, the quantile-weighted CRPS and the threshold-weighted CRPS. Replacing the inner integral in expression (37) by the quantile-weighted CRPS or the threshold-weighted

CRPS, we obtain the following alternative expressions for the weighted energy score:

$$WES_1(\boldsymbol{P}, \boldsymbol{y}) = \frac{1}{4}\int_{S_1}\int_{-\infty}^{\infty}(F_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}(r) - I(\langle \boldsymbol{a}, \boldsymbol{y}\rangle \leq r))^2 u(\boldsymbol{a}, r)drd\mu(\boldsymbol{a}) \tag{38}$$

$$WES_1(\boldsymbol{P}, \boldsymbol{y}) = \frac{1}{4}\int_{S_1}\int_{0}^{1} QS_\alpha(F^{-1}_{\langle \boldsymbol{a},\boldsymbol{X}\rangle}(\alpha), \langle \boldsymbol{a}, \boldsymbol{y}\rangle)v(\boldsymbol{a}, \alpha)d\alpha d\mu(\boldsymbol{a}) \tag{39}$$

Expression (38) and expression (39) are the threshold-weighted energy score and the quantile-weighted energy score, respectively. Expression (37)-(39) show that, for $d = 2$, the weighted energy score is essentially equal to the weighted CRPS integrated over the unit circle. Therefore, the weight function discussed in Section 4.2 can be extended here to produce weighted energy scores.

In our empirical work, for the threshold-weight function, we consider the use of a binary variable, which is defined as $u(\boldsymbol{a}, r) = I(\boldsymbol{a}r \in A)$, where $A = \{x, y | x > 17.5 \; and \; y > 35\}$. This weight function has a physical meaning: the bivariate random variable $X = (X_1, X_2)$ only receives weight 1 if $X_1 > 17.5$ and and $X_2 > 35$ and is ignored for other regions. As X_1 and X_2 are the minimum and maximum temperature in our study, the region A is exactly the region where heat waves occur. For the quantile-weight function, we consider the universal weight $v(\boldsymbol{a}, \alpha) = \alpha^2$ for every $\boldsymbol{a}$ on the unit circle. This weight function emphasises the extreme quantiles for each univariate random variable $\langle \boldsymbol{a}, \boldsymbol{X}\rangle$.

5 Empirical Study

5.1 Models

We implemented univariate ARMA-GARCH models for the daily minimum, maximum and average temperature, and the DTR. We used four different distributional assumptions: Gaussian distribution, generalised asymmetric skew-t distribution,

Gaussian distribution with EVT for the tails, and generalised asymmetric skew-t distribution with EVT for the tails. We implemented bivariate VARMA-MGARCH models for the daily minimum and maximum temperature. For these models, we considered five different distributional assumptions: Gaussian marginal distributions with no dependence structure; generalised asymmetric skew-t marginal distributions with no dependence structure; a bivariate Gaussian distribution; generalised asymmetric skew-t marginal distributions with a Gaussian copula; and generalised asymmetric skew-t marginal distributions with a Student's t copula. Having produced forecasts for the joint distribution of the daily minimum and maximum temperature, we were able to produce corresponding forecasts of the marginal distributions of the daily average temperature and the DTR through simulation. Our density forecast evaluation in Sections 5.2 and 5.3 focuses on 1 day-ahead prediction, and in Section 5.4, we consider exceedance probability forecasts for a horizon of up to three days.

5.2 Evaluating Density Forecasts Using the CRPS and Energy Score

We calculated the unweighted and weighted forms of the CRPS and energy score, which we described in Section 4. As the values of these scores do not have an intuitive interpretation, we calculated Theil's U statistics using the univariate ARMA-GARCH model with Gaussian distribution as the benchmark. We averaged the resulting ratios over the four Spanish locations, and then subtracted this from one to deliver a measure that reflects the percentage by which each method is better than the benchmark method. This measure is known in the probabilistic forecast evaluation literature as a skill score. Higher values indicate superior accuracy, and positive values imply greater accuracy than the benchmark method. In our tables of results, and in the remainder of this section, we use Tmin, Tmax and Tavg as abbreviations for the daily minimum, maximum and average temperature, respectively.

Table 3 presents the skill scores for the unweighted CRPS and energy score of Sections

4.1 and 4.3. Each of the first four columns of values presents the skill scores for one of the four temperature variables. For example, the first column of values shows the skill scores resulting from density forecasts of Tmin. Each row in the table can be viewed as representing the forecasts produced by a pair of the temperature variables. For example, the first four rows were produced by the univariate models of Tmin and Tmax. For these rows, the Tmin forecasts were produced using just the model for Tmin; the Tmax forecasts were produced using just the model for Tmax; and the Tavg and DTR forecasts were produced using both the model for Tmin and Tmax.

The first four rows of Table 3 show that using EVT did not lead to a noticeable improvement in forecast accuracy for the univariate models applied to the daily minimum and maximum. For this reason, and to save space, we omit from the tables the results of using EVT with the models applied to the daily mean and DTR. The results for the univariate models show that models that used the generalised asymmetric skew-t were generally more accurate than models employing a Gaussian assumption. It is interesting to note from Table 3 that generating forecasts for Tmin from the univariate models for Tavg and DTR led to greater accuracy than simply using the univariate model for Tmin. The table does not show a similar finding for Tmax. This reflects the view of Dupuis (2014) that Tmin is more challenging to model than Tmax.

Turning to the bivariate VARMA-MGARCH models, we see from Table 3 that each of these produced better CRPS results than the univariate models for almost all four temperature variables. Overall the best performing bivariate models are the two that used the generalised asymmetric skew-t distribution with dynamic Gaussian or Student's t copula. It is noteworthy that these bivariate models for Tmin and Tmax delivered clearly better accuracy for DTR than a univariate model for DTR, and that, in terms of forecasting Tavg, these bivariate models even slightly outperformed a univariate model for Tavg.

The final column of Table 3 presents the energy score for forecasts of the joint distribution of Tmin and Tmax. Note that this distribution can be used to generate a forecast of the joint distribution of Tavg and DTR, and so the energy score obtained for

these two joint distributions will be identical. The energy scores in Table 3 show that the bivariate modelling led to improved forecast accuracy, and that the best results were produced using the generalised asymmetric skew-t distribution with dynamic Gaussian or Student's t copula.

5.3 Evaluating Density Forecasts Using the Weighted CRPS and Weighted Energy Score

Tables 4 and 5 present the skill scores corresponding to the quantile-weighted and threshold-weighted CRPS and energy score described in Sections 4.2 and 4.4. The quantile-weighted results of Table 4 are similar in magnitude to the unweighted skill score results of Table 3, suggesting that perhaps the quantile-weighting scheme proposed by Gneiting and Ranjan (2011) does not, at least for our case, put sufficient emphasis on the tail of the distribution. The threshold-weighted results of Table 5 are more noticeably different from the corresponding unweighted scores in Table 3. However, the ranking of methods remains broadly the same, with the bivariate models performing well in terms of forecasting the marginal distributions of Tmin and Tmax, as well as the joint distribution of these two variables.

5.4 Evaluating Exceedance Probability Forecasts Using the Brier Score

Finally, to simulate the probabilistic forecasting of heat waves, we evaluated the performance of the models in terms of forecasting the probability of the event that the Tmin and Tmax exceed specified thresholds for 3 consecutive days. We chose the thresholds as 17.5 and 35 for Tmin and Tmax, respectively. Table 6 reports the Brier skill scores for three different events. The first column corresponds to the event that Tmin exceeds 17.5 on each of the next three days; the second column reports the results for the event that Tmax exceeds 35 on each of the next three days; and the third column corresponds to the event that Tmin exceeds 17.5 and Tmax exceeds 35 on the

next three days. Note that this was the only part of our analysis that evaluated prediction beyond 1 step-ahead. It is reassuring to see from Table 6 that the ranking of methods is very similar to the ranking that we saw in Tables 3 to 5.

6 Summary and Conclusion

In this paper, we have considered univariate and bivariate methods for density forecasting of the daily minimum and maximum temperature. We implemented univariate ARMA-GARCH models with different distributional assumptions, and with the use of EVT. We considered bivariate VARMA-MGARCH models with a variety of different distributional and dependency specifications. These included a bivariate normal distribution, with and without an assumption of independence, and a bivariate distribution composed of skew-t marginal distributions and a dynamic copula to capture the conditional dependency. We evaluated the goodness-of-fit of the models by evaluating short-term density forecast accuracy for each day in the years 2011 to 2015. We evaluated predictions of the marginal and joint distribution for the minimum and maximum temperature, as well as for average temperature and DTR.

We introduced a weighted energy score, which is a proper scoring rule that enables the user to put greater weight on parts of the multivariate distribution of interest. In our study, we used a weighting scheme that emphasised the tails of the bivariate temperature distributions. We also evaluated the accuracy of probabilistic forecasts of heat waves, defined as the minimum and maximum simultaneously exceeding chosen threshold. Our empirical analysis shows that the bivariate VARMA-MGARCH methods with dynamic dependence structure noticeably outperform the other methods. From a methodology perspective, it was interesting to find that the models with more complex structure did produce better empirical results. We note that the most successful model was a VARMA-MGARCH model with 94 parameters. Although this is a reasonably large number of parameters for a model of this type, estimation benefited from the relatively long time series consisting of 21,900 daily observations.

The models revealed some interesting results: the daily minimum and maximum temperature have predictive power for each other, a dynamic dependence structure is essential for modelling the joint distribution of the minimum and maximum temperatures, and this joint distribution produces competitive density forecasts for the marginal distributions for the daily average temperature and the DTR. Our models also showed increasing trends in the minimum and maximum temperatures, a decreasing trend in the DTR, and different parts of the seasonal cycle changing at different rates, which are findings that are consistent with those of other researchers.

References

Atak, A., Linton, O., and Xiao, Z. (2011). A semiparametric panel model for unbalanced data with application to climate change in the united kingdom. *Journal of Econometrics*, 164(1):92–115.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389.

Bauwens, L. and Laurent, S. (2005). A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *Journal of Business and Economic Statistics*, 23(3):346–354.

Bauwens, L., Laurent, S., and Rombouts, J. V. (2006). Multivariate garch models: a survey. *Journal of Applied Econometrics*, 21(1):79–109.

Berkowitz, J. (2001). Testing the accuracy of density forecasts. *Journal of Business and Economic Statistics*.

Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–131.

Campbell, S. D. and Diebold, F. X. (2005). Weather forecasting for weather derivatives. *Journal of the American Statistical Association*, 100(469):6–16.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, pages 863–883.

Diks, C., Panchenko, V., and Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230.

Dupuis, D. J. (2012). Modeling waves of extreme temperature: the changing tails of four cities. *Journal of the American Statistical Association*, 107(497):24–39.

Dupuis, D. J. (2014). A model for nighttime minimum temperatures. *Journal of Climate*, 27(19):7207–7229.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics*, 20(3):339–350.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized arch. *Econometric Theory*, 11(01):122–150.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2015). R-vine models for spatial time series with an application to daily mean temperature. *Biometrics*, 71(2):323–332.

Fernández, C. and Steel, M. F. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Forster, P. M. d. F. and Solomon, S. (2003). Observations of a 'weekend effect' in diurnal temperature range. *Proceedings of the National Academy of Sciences*, 100(20):11225–11230.

Franses, P. H., Neele, J., and van Dijk, D. (2001). Modeling asymmetric volatility in weekly dutch temperature data. *Environmental Modelling and Software*, 16(2):131–137.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29(3):411–422.

Jeon, J. and Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79.

Keellings, D. and Waylen, P. (2014). Increased risk of heat waves in Florida: Characterizing changes in bivariate heat wave risk using extreme value analysis. *Applied Geography*, 46:90–97.

Kolev, N., Anjos, U. d., and Mendes, B. V. d. M. (2006). Copulas: a review and recent developments. *Stochastic Models*, 22(4):617–660.

Lee, T.-H. and Long, X. (2009). Copula-based multivariate garch model with uncorrelated dependent errors. *Journal of Econometrics*, 150(2):207–218.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3):271–300.

Meehl, G. A. and Tebaldi, C. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, 305(5686):994–997.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.

Proietti, T. and Hillebrand, E. (2016). Seasonal changes in central england temperatures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. Forthcoming.

Qu, M., Wan, J., and Hao, X. (2014). Analysis of diurnal air temperature range change in the continental united states. *Weather and Climate Extremes*, 4:86–95.

Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Taylor, J. W. and Buizza, R. (2004). A comparison of temperature density forecasts from garch and atmospheric models. *Journal of Forecasting*, 23(5):337–355.

Tol, R. S. (1996). Autoregressive conditional heteroscedasticity in daily temperature measurements. *Environmetrics*, 7(1):67–75.

Wang, M.-z., Zheng, S., He, S.-l., Li, B., Teng, H.-j., Wang, S.-g., Yin, L., Shang, K.-z., and Li, T.-s. (2013). The association between diurnal temperature range and emergency room admissions for cardiovascular, respiratory, digestive and genitourinary disease among the elderly: a time series study. *Science of The Total Environment*, 456:370–375.

Wong, T. (2015). Statistical analysis of heat waves in the state of victoria in australia. *Australian and New Zealand Journal of Statistics*, 57(4):463–480.

Zhu, D. and Galbraith, J. W. (2010). A generalized asymmetric student-t distribution with application to financial econometrics. *Journal of Econometrics*, 157(2):297–305.

Ziegel, J. F., Gneiting, T., et al. (2014). Copula calibration. *Electronic Journal of Statistics*, 8(2):2619–2638.

Figure 1: Seville daily minimum temperature time series (in degrees Celsius) for the first 60-year estimation sample.

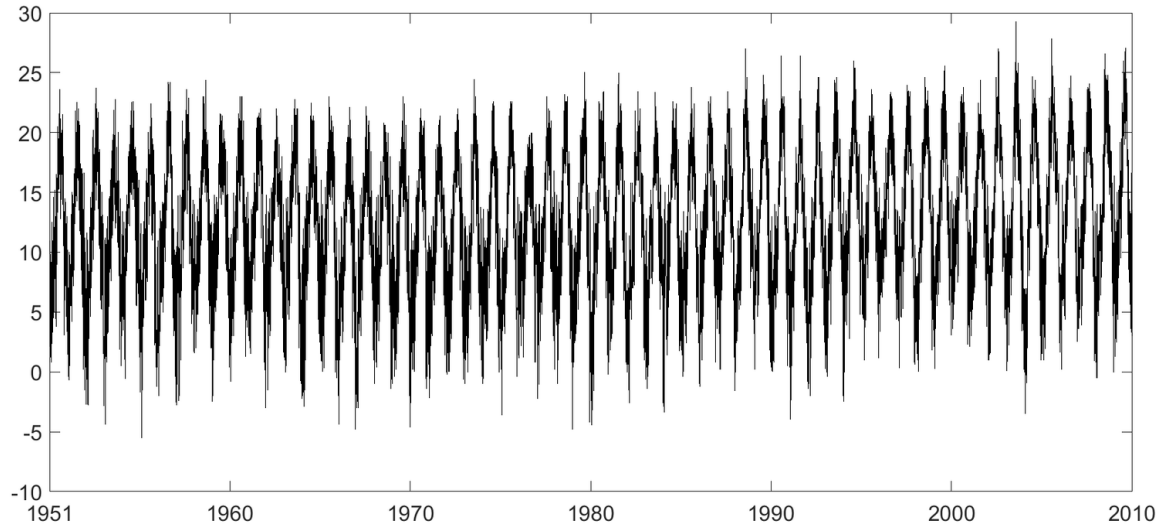

Figure 2: Seville daily maximum temperature time series (in degrees Celsius) for the first 60-year estimation sample.

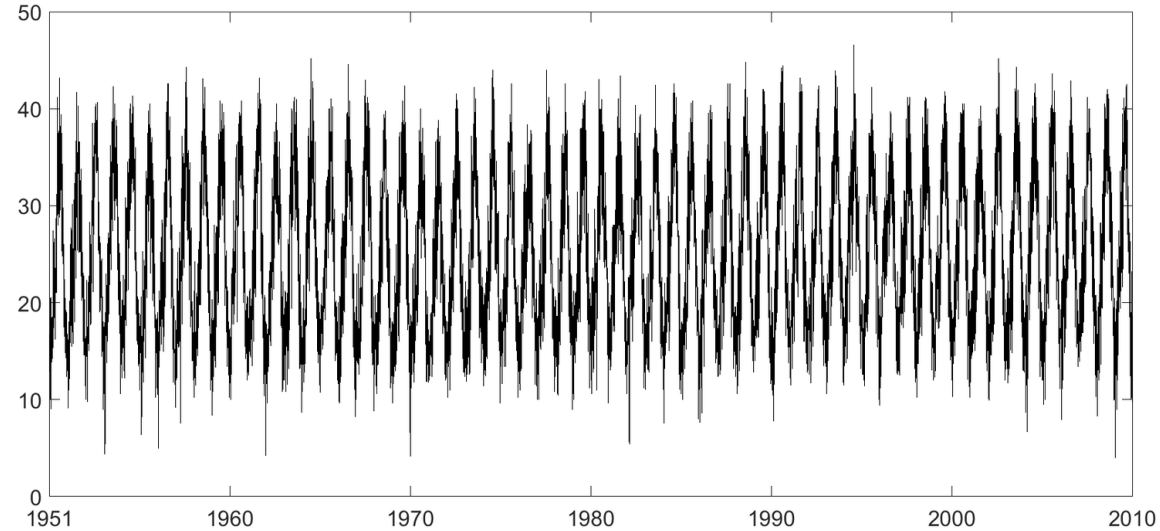

Figure 3: Seville daily minimum and maximum temperature (in degrees Celsius) plotted against the day of the year for the first 60-year estimation sample.

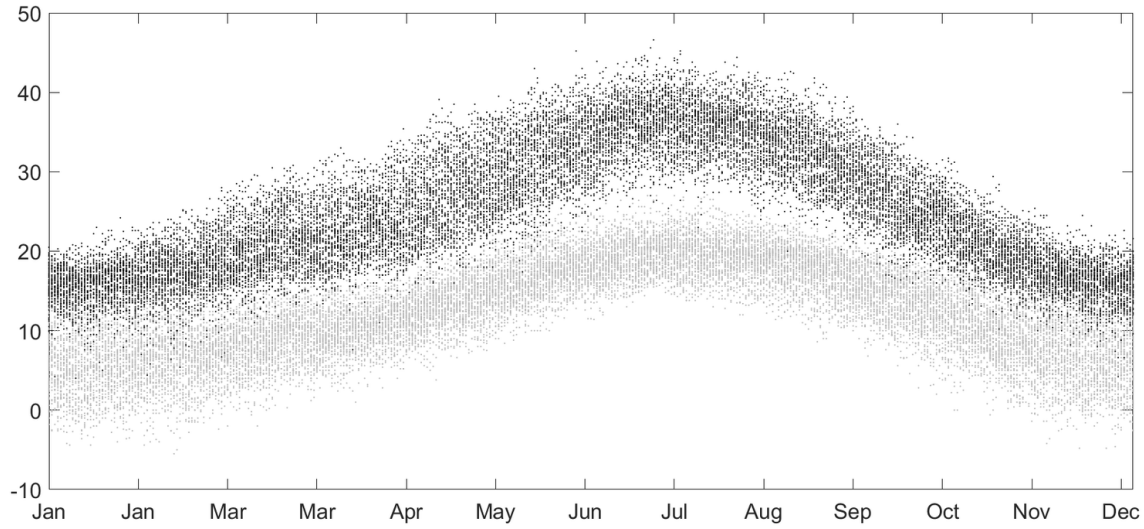

Figure 4: Seville daily minimum temperature (in degrees Celsius) for the first 60-year estimation sample, with a linear deterministic trend starting from the beginning of 1974.

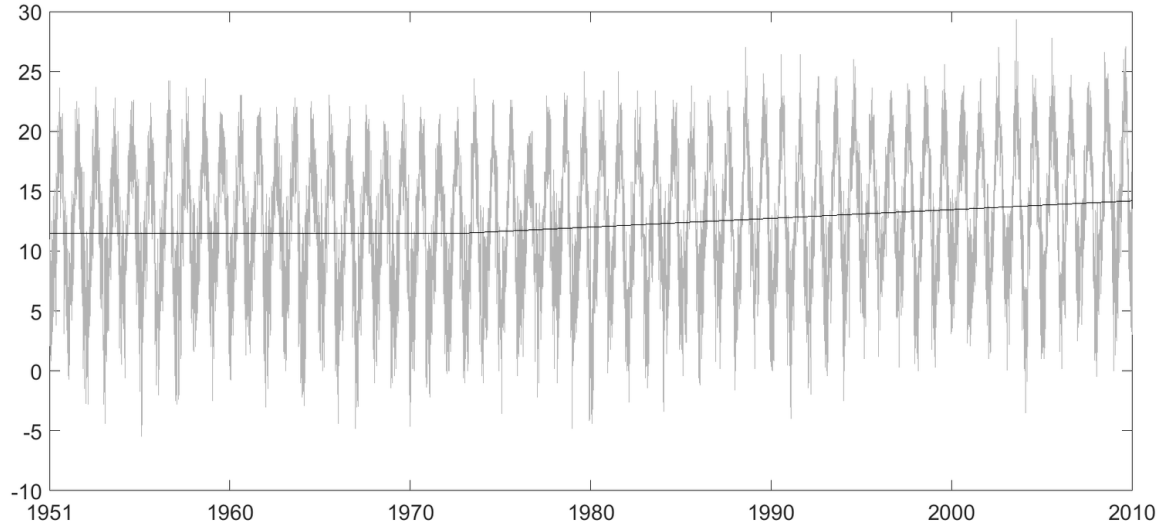

Figure 5: For the first day of each month, Seville daily minimum temperature (in degrees Celsius) plotted for the first 60-year estimation sample, along with the trend estimated by the univariate ARMA-GARCH model with Gaussian assumption.

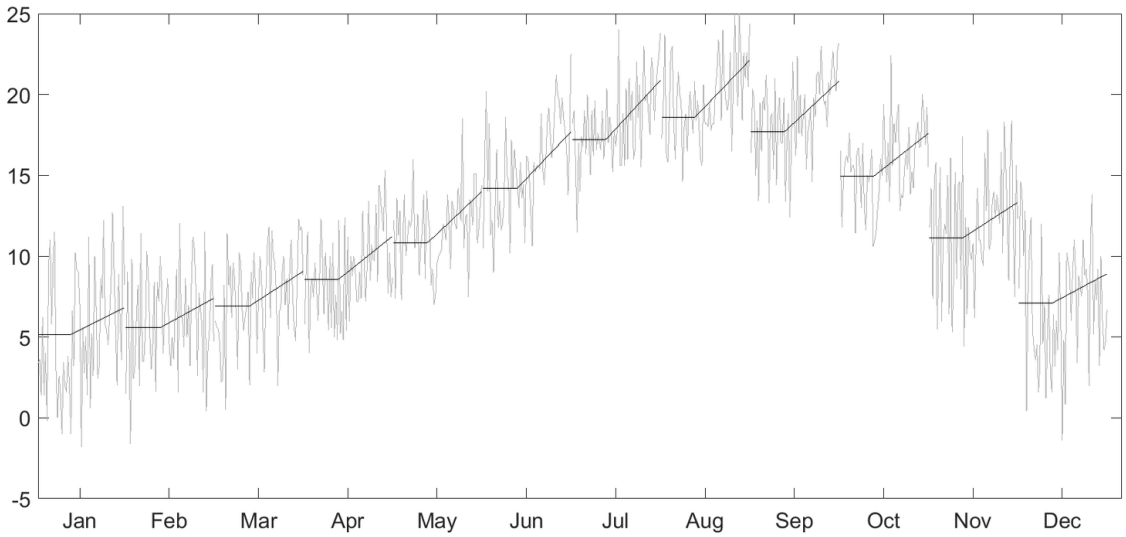

Table 1: Parameter estimates and p-values for the mean component of the univariate ARMA-GARCH model with Gaussian assumption, derived using the first 60 years of daily minimum (Tmin) and maximum (Tmax) temperature observations for Seville.

	Tmin	Tmax
$(\mu_1)_0$	114.963(0.00**)	245.444(0.00**)
$(\mu_1)_{1,1}$	-31.358(0.00**)	-40.320(0.00**)
$(\mu_1)_{2,1}$	8.108(0.00**)	17.312(0.00**)
$(\mu_1)_{1,2}$	0.578(0.41)	-2.189(0.01**)
$(\mu_1)_{2,2}$	-58.162(0.00**)	-87.862(0.00**)
$(\mu_1)_{1,3}$	-2.767(0.00**)	-2.020(0.03*)
$(\mu_1)_{2,3}$	-2.295(0.00**)	-2.362(0.00**)
ψ_1	0.002(0.00**)	0.001(0.00**)
$(\mu_2)_{1,1}$	0.000(0.92)	0.001(0.00**)
$(\mu_2)_{1,2}$	-0.001(0.00**)	0.000(0.20)
ϕ_1	1.485(0.00**)	1.711(0.00**)
ϕ_2	-0.513(0.00**)	-0.797(0.00**)
ϕ_3	-0.003(0.70)	0.073(0.00**)
θ_1	-0.863(0.00**)	-0.932(0.00**)

Note: Significance of the p-values at 5% and 1% levels is indicated by * and **, respectively.

Table 2: Parameter estimates and p-values for the variance component of the univariate ARMA-GARCH model with Gaussian assumption, derived using the first 60 years of daily minimum (Tmin) and maximum (Tmax) temperature observations for Seville.

	Tmin	Tmax
$(\omega_1)_0$	100.744(0.00**)	233.680(0.00**)
$(\omega_1)_{1,1}$	4.896(0.04*)	22.044(0.00**)
$(\omega_1)_{2,1}$	0.651(0.62)	-7.919(0.02*)
$(\omega_1)_{1,2}$	51.948(0.00**)	-65.497(0.00**)
$(\omega_1)_{2,2}$	8.636(0.00**)	-8.237(0.01*)
ψ_2	-0.003(0.00**)	0.001(0.04*)
$(\omega_2)_{1,1}$	0.000(0.32)	0.000(0.69)
$(\omega_2)_{2,1}$	-0.002(0.00**)	0.001(0.10)
α_1	0.028(0.00**)	0.124(0.00**))
γ_1	0.038(0.00**)	-0.068(0.00**)
β_1	0.753(0.00**)	0.447(0.00**)

Note: Significance of the p-values at 5% and 1% levels is indicated by * and **, respectively.

Table 3: CRPS skill scores for forecasts of the marginal distribution of daily minimum (Tmin), maximum (Tmax) and average (Tavg) temperature, and DTR. Energy skill scores for forecasts of the joint distribution of Tmin and Tmax. Skill scores averaged across the four locations.

	CRPS				Energy Score
	Tmin	Tmax	Tavg	DTR	
Univariate ARMA-GARCH for Tmin and Tmax					
Gaussian	0.0	0.0	0.0	0.0	0.0
Gen asym skew-t	-0.1	0.7	-0.1	0.5	0.3
Gaussian with EVT	0.0	0.0	0.1	0.0	0.0
Gen asym skew-t with EVT	-0.1	0.7	-0.1	0.5	0.3
Univariate ARMA-GARCH for Tavg and DTR					
Gaussian	5.0	-2.4	2.6	1.0	1.4
Gen asym skew-t	5.3	-2.3	2.7	1.3	1.6
Bivariate VARMA-MGARCH for Tmin and Tmax					
Gaussian with no dependency	7.1	0.0	1.6	3.6	2.9
Gen asym skew-t with no dependency	7.2	0.7	2.2	3.9	3.4
Bivariate Gaussian	6.9	0.1	2.5	3.6	3.2
Gen asym skew-t with Gaussian copula	7.2	0.8	2.9	4.0	3.6
Gen asym skew-t with Student's t copula	7.2	0.8	2.9	4.0	3.6

Table 4: Quantile-weighted CRPS skill scores for forecasts of the marginal distribution of daily minimum (Tmin), maximum (Tmax) and average (Tavg) temperature, and DTR. Quantile-weighted energy skill scores for forecasts of the joint distribution of Tmin and Tmax. Skill scores averaged across the four locations.

	Weighted CRPS		Weighted Energy Score
	Tmin	Tmax	
Univariate ARMA-GARCH for Tmin and Tmax			
Gaussian	0.0	0.0	0.0
Gen asym skew-t	0.0	0.7	0.3
Gaussian with EVT	0.0	0.1	0.0
Gen asym skew-t with EVT	0.0	0.7	0.3
Univariate ARMA-GARCH for Tavg and DTR			
Gaussian	4.3	-2.4	1.3
Gen asym skew-t	4.5	-2.2	1.6
Bivariate VARMA-MGARCH for Tmin and Tmax			
Gaussian with no dependency	6.6	0.0	2.9
Gen asym skew-t with no dependency	6.8	0.8	3.4
Bivariate Gaussian	6.6	0.1	3.2
Gen asym skew-t with Gaussian copula	6.7	0.8	3.5
Gen asym skew-t with Student's t copula	6.8	0.9	3.7

Table 5: Thershold-weighted CRPS skill scores for forecasts of the marginal distribution of daily minimum (Tmin), maximum (Tmax) and average (Tavg) temperature, and DTR. Threshold-weighted energy skill scores for forecasts of the joint distribution of Tmin and Tmax. Skill scores averaged across the four locations.

	Weighted CRPS		Weighted Energy Score
	Tmin	Tmax	
Univariate ARMA-GARCH for Tmin and Tmax			
Gaussian	0.0	0.0	0.0
Gen asym skew-t	-0.2	1.9	0.8
Gaussian with EVT	0.0	0.3	0.0
Gen asym skew-t with EVT	-0.2	1.9	0.8
Univariate ARMA-GARCH for Tavg and DTR			
Gaussian	6.6	-0.8	2.5
Gen asym skew-t	6.9	-0.3	3.1
Bivariate VARMA-MGARCH for Tmin and Tmax			
Gaussian with no dependency	13.4	0.0	8.2
Gen asym skew-t with no dependency	13.6	1.9	8.8
Bivariate Gaussian	13.2	0.2	8.1
Gen asym skew-t with Gaussian copula	14.0	1.7	9.0
Gen asym skew-t with Student's t copula	13.7	1.7	8.9

Table 6: Brier skill scores for probability forecasts of exceedance over all of the next 3 days. Skill scores averaged across the four locations.

	Tmin > 17.5	Tmax > 35	Tmin>17.5 and Tmax>35
Univariate ARMA-GARCH for Tmin and Tmax			
Gaussian	0.0	0.0	0.0
Gen asym skew-t	-0.1	2.6	2.0
Gaussian with EVT	0.0	0.1	0.0
Gen asym skew-t with EVT	0.0	2.5	2.0
Univariate ARMA-GARCH for Tavg and DTR			
Gaussian	6.6	-1.4	0.9
Gen asym skew-t	8.1	-0.9	1.9
Bivariate VARMA-MGARCH for Tmin and Tmax			
Gaussian with no dependency	8.7	0.2	2.9
Gen asym skew-t with no dependency	8.5	2.8	5.0
Bivariate Gaussian	8.2	0.9	4.4
Gen asym skew-t with Gaussian copula	9.1	2.9	5.8
Gen asym skew-t with Student's t copula	9.1	2.9	5.7