

Comparing Probabilistic Forecasts of the Daily Minimum and Maximum Temperature

12th April 2022

Xiaochun Meng^{*} and James W. Taylor[‡]

^{*} University of Sussex Business School, University of Sussex, UK

[‡] Saïd Business School, University of Oxford, UK

International Journal of Forecasting, 2022, 38(1), 267-281.

Abstract

Understanding changes in the frequency, severity and seasonality of daily temperature extremes is important for public policy decisions regarding heat waves and cold snaps. A heat wave is sometimes defined in terms of both the daily minimum and maximum temperature, which necessitates the generation of forecasts of their joint distribution. In this paper, we develop time series models with the aim of providing insight, and producing forecasts of the joint distribution that can challenge the accuracy of forecasts based on ensemble predictions from a Numerical Weather Prediction model. We use ensemble model output statistics to recalibrate the raw ensemble predictions for the marginal distributions, with ensemble copula coupling used to capture the dependency between the marginal distributions. In terms of time series modelling, we consider a bivariate VARMA-MGARCH model. We use daily Spanish data recorded over a 65-year period, and find that, for the 5-year out-of-sample period, the recalibrated ensemble predictions outperform the time series models in terms of forecast accuracy.

Keywords: Probabilistic forecasting; Weather ensemble predictions; VARMA-MGARCH.

1 Introduction

Extreme weather events can have negative impacts on society and the economy. For example, heat waves and cold snaps can lead to serious health problems and increased morbidity (Wang et al. 2013; Dupuis 2012, 2014). The exposure of a variety of businesses to weather extremes has prompted the development of weather derivatives (Alexandridis and Zapranis 2012; Alexandridis et al. 2017; Campbell and Diebold 2005). With climate change, the frequency and severity of extreme weather is increasing (Fildes and Kourentzes 2011; Meehl and Tebaldi 2004). Forecasts of extreme weather events are of great importance, as they enable warnings to be made to the public, and allow services, such as hospitals, to prepare. Although accurate point forecasting is an aim, the inevitable uncertainty in the weather means that probabilistic forecasting is appropriate. In this paper, we consider the probabilistic forecasting of heat waves. The definition of a heat wave is sometimes based on the daily maximum temperature exceeding a specified threshold on consecutive days, as it has been found that people’s health is sensitive to the exposure to high temperatures for this duration (Pattenden et al. 2003; Le Tertre et al. 2006; Hajat et al. 2006; Perkins and Alexander 2013). However, heat waves are also sometimes defined as the simultaneous exceedances of the daily minimum and maximum temperature over specified thresholds (see, for example, Keellings and Waylen 2014). In view of this, we aim to improve the prediction 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 of the earth’s atmosphere, such as a Numerical Weather Prediction (NWP) system (see, for example, Bauer et al. 2015; Baran and Lerch 2018). The NWP model uses nonlinear partial differential equations to describe the physical dynamics of the atmosphere, where complex weather variables such as wind, pressure and temperature are taken into consideration (see, for example, Bauer et al. 2015). To account for the uncertainty existing in the weather variables, ensemble predictions are produced by considering multiple different initial conditions and model physics in the NWP models. Ensemble predictions are typically subject to forecast bias and dispersion errors. To address this, post-processing methods have been proposed (see, for example, Gneiting 2014; Baran and Lerch 2018). Among the existing post-processing approaches, the ensemble model output statistics (EMOS) and ensemble copula coupling (ECC) methods have been widely studied and applied. The former post-processes univariate weather variables, while the latter aims to capture the dependency

between univariate weather variables.

Statistical time series models, such as the well-established autoregressive moving average (ARMA) and generalised autoregressive conditional heteroskedasticity (GARCH) models, provide an alternative because historical observations of daily minimum and maximum temperature are readily available for many locations (see, for example, Taylor and Buizza 2004; Campbell and Diebold 2005; Dupuis 2012). 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 improve, our understanding of how temperature extremes have been changing. With this in mind, the first of our two main aims in this paper is to develop suitable time series models for both the daily minimum and maximum temperature. Our second aim is to investigate whether such carefully chosen time series models are able to compete with the accuracy of forecasts based on ensemble predictions from an NWP model. To achieve this, we propose a novel vector ARMA (VARMA) and multivariate (GARCH) model, to model the joint distribution of the daily minimum and maximum temperature. Our empirical analysis uses Spanish data, which is of particular interest because Southern Europe has experienced severe heat waves in recent years (see, for example, Kew et al. 2019).

The rest of the paper is structured as follows. Section 2 introduces our dataset of daily time series of Spanish temperature observations. In Section 3, we describe our dataset of NWP ensemble predictions, and discuss the steps required to obtain probabilistic forecasts from the raw ensemble data. Univariate and bivariate candidate time series models are described in Sections 4 and 5, respectively. Section 6 presents an empirical comparison of forecast accuracy. Section 7 summarises the paper.

2 Daily Temperature Time Series

We considered historical data for the following four cities in Spain: Albacete, Seville, Cáceres and Madrid. For all four cities, heat waves have serious consequences, because each city has a sizeable population. We intentionally chose locations that were not close, as this would have led to similar temperatures and forecasting results. Our dataset consisted of daily minimum and maximum observations, measured in degrees Celsius ($^{\circ}\text{C}$), for the 65-year period from 1951 to 2015, inclusive. For each day, the values recorded were the minimum and maximum

temperature occurring on that day. In this paper, we used the common notation TN, TX, and TG as abbreviations for the daily miNimum, maXimum and averaGe temperature, respectively, where TG is defined as the mean of TN and TX. 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 on temperature time series modelling, the observations for 29 February, occurring in each leap year, were removed from the series in order to have a constant 365 days in each year (see, for example, Campbell and Diebold 2005; Dupuis 2012). For each of our time series models, we used the first 60 years of data to specify the models, which involved determining the orders of the ARMA and GARCH parts of the model, as well as the types of functions to use for the trend and seasonality. We used the final 5 years of data to evaluate out-of-sample forecasts. To produce forecasts from each day in the out-of-sample period, we used a rolling window approach, where we used the model with parameters estimated using the previous 60 years of data. For example, to obtain a forecast for the first day in the out-of-sample period, 1 January 2015, we used the data between 1 January 1951 and 31 December 2014 to estimate the model parameters.

Figure 1 is a plot of TN at Seville for the first estimation period 1951 to 2010, and Figure 2 is the corresponding plot of TX. The trends in the plots will be discussed in Section 4.2. 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 possess annual seasonality, which accounts for the repeating periodic spikes. The seasonality is clear from Figure 3, which plots TN and TX observations for Seville against the day of the year for the 60-year period 1951 to 2010. Note that, at least in the TN 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.

Figure 1: Seville daily minimum temperature (TN) time series (in °C) for the entire 65-year sample, with a linear trend starting from the beginning of 1974.

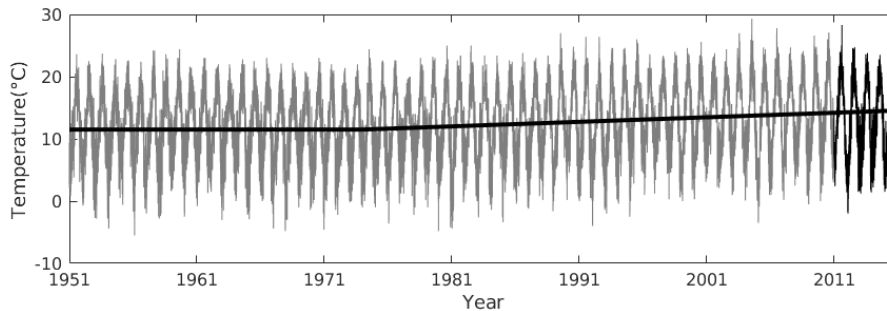


Figure 2: Seville daily maximum temperature (TX) time series (in °C) for the entire 65-year sample, with a linear trend starting from the beginning of 1974.

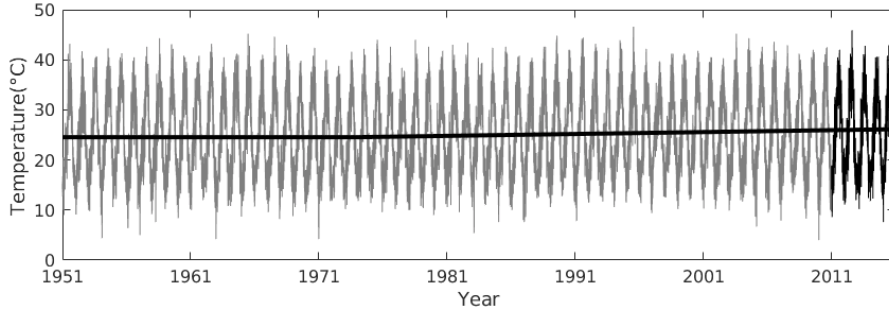
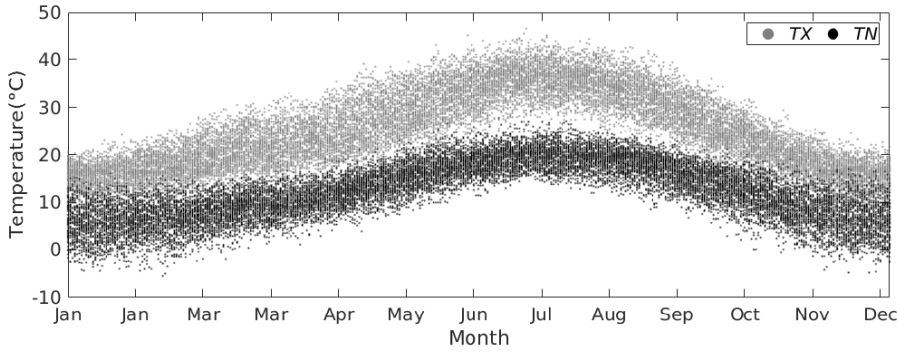


Figure 3: Seville daily minimum (TN) and maximum temperature (TX) (in °C) plotted against the day of the year for the first 60-year estimation sample.



3 Weather Ensemble Predictions

We obtained the ensemble predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF). Each ensemble consisted of 51 scenarios for the future value of temperature (see, for example, Molteni et al. 1996; ECMWF Directorate 2012). The ECMWF NWP model has a grid resolution of 0.125° in both latitude and longitude directions. We used linear interpolation to convert ensemble predictions at the grid points into ensemble predictions for the locations of the four Spanish cities in our study. For our out-of-sample period (2011-2015), ECMWF provided ensemble predictions at lead times every 3 hours up to 6 days ahead then at lead times every 6 hours up to 15 days ahead, with midnight and midday as forecast origins. With the forecast origin chosen as midnight, each of the 51 temperature ensemble scenarios consists of a forecast function extending from midnight, with a single forecast at each of the 3-hourly intervals up to 6 days ahead and then at each of the 6-hourly intervals up to 15 days ahead. For each day, we converted each of these 51 scenarios into a single ensemble member for TN and TX as follows. For each scenario, the forecast for the TN

and TX were computed as the minimum and maximum values of the ensemble predictions within a certain day. For example, the 1 day-ahead forecast for the TN and TX were computed as the minimum and maximum values of the ensemble predictions at lead times 3, 6, ..., and 24 hours ahead; and the 15 day-ahead forecast for the TN and TX were computed as the minimum and maximum values of the ensemble predictions at lead times 342, 348, 354 and 360 hours ahead. In this way, we produced 51 ensemble members for TN and TX for each day under consideration.

As ensemble predictions are typically subject to forecast bias and dispersion errors, we used the EMOS-ECC approach to recalibrate the ensemble using the historical temperature values. This is a two-step approach. In the first step, EMOS fits a Gaussian distribution to the marginal distribution of the temperature variable, where the mean and variance of the Gaussian distribution are linear transformations of the mean and standard deviation of the 51 raw ensemble members (see, for example, Gneiting et al. 2005; Baran and Lerch 2018). EMOS can be represented as follows,

$$T_{EMOS,t} \sim N(a_{t,0} + a_{t,1}\mu_{E,t}, b_{t,0} + b_{t,1}\sigma_{E,t}^2) \quad (1)$$

where $T_{EMOS,t}$ denotes the estimated temperature variable; $a_{t,i}$ and $b_{t,i}$ are constant parameters, and $\mu_{E,t}$ and $\sigma_{E,t}^2$ denote the mean and variance of the 51 raw ensemble members. We set $b_{t,0}, b_{t,1} \geq 0$ to ensure the non-negativity of the post-processed variances. We estimated the parameters $a_{t,i}$ and $b_{t,i}$ using maximum likelihood, where the conditions $b_{t,0}, b_{t,1} \geq 0$ were ensured by applying a constrained optimisation algorithm. The parameters for the EMOS approach are fitted using a rolling window estimation period. It has been found in the literature that the choice of the length for the parameter estimation period can have a noticeable impact on the forecast performance, particularly for longer lead times (see, for example, Gneiting et al. 2005; Gneiting 2014; Hemri et al. 2014; Feldmann et al. 2019; Baran et al. 2020). Therefore, we considered multiple lengths for the estimation period. Following Gneiting et al. (2005), we considered the most recent 40 days as the estimation period and following Gneiting (2014), we considered the most recent 100 and 200 days as the estimation period. Because the longest forecast horizon that we consider in this paper is relatively long (15 days), we also considered the most recent 300 days as the estimation period. To not lose any out-of-sample data for evaluation, the NWP ensemble predictions in 2010 were also used in the estimation period.

Once we had recalibrated the marginal distributions of TN and TX, we used ECC to capture the dependency between the marginal distributions (Scheffzik et al. 2013) in the second step. More specifically, we first obtained a discrete sample of size 51 from each recalibrated marginal distribution up to 15 days-ahead using the following expression,

$$T_{m,t} = F_{EMOS,t}^{-1} \left(\frac{m}{52} \right) \quad (2)$$

where $m = 1, 2, \dots, 51$, and $F_{EMOS,t}$ denotes the Gaussian distribution in expression (1). We then formed pairs of TN and TX values by arranging the 51 values for TN and TX, so that the samples had the same rank order as the 51 pairs of TN and TX values in the raw ensemble. In other words, the samples of TN and TX were arranged to have the same Spearman's rank correlation as the raw ensemble. Note that for consistency, in the empirical study in Section 6, we evaluate the marginal distributions via the 51 values for TN and TX obtained from expression (2).

4 Univariate ARMA-GARCH Models

4.1 Model Specification

There are several studies that have modelled either TN, TX or TG using univariate models. It has been shown that daily series of these variables exhibit the following features: a trend in the mean and variance; a seasonal pattern in the mean and variance; 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 for autoregressive modelling of the mean and variance.

Tol (1996) and Taylor and Buizza (2004) use AR-GARCH models to estimate the distribution of TG, and Taylor and Buizza (2006) use AR-GARCH models to estimate the distribution of temperature series recorded daily at midday. They model the mean and variance using 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 TG, but their seasonal modelling is based on Fourier terms. In fitting an ARMA model to TX 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 TN and TX; the series are preprocessed by an AR-GARCH model before an extreme value theory 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 TG at different locations, with marginal distributions estimated by a multi-step AR model.

Empirical results have provided support for the AR-GARCH models in comparison with a variety of simpler time series models. In this paper, we extend the literature on univariate modelling by considering ARMA structures for the mean, rather than just AR, and by including novel trend features. The univariate model that we consider is the ARMA-GARCH model of expressions (3)-(7). In our empirical work, we applied the model separately to TN and TX. Our model has the following expression:

$$T_t = S_1(\boldsymbol{\mu}_1, t) + \psi_1 Trend_t + S_2(\boldsymbol{\mu}_2, t) Trend_t + y_t \quad (3)$$

$$y_t = \sum_{i=1}^m \phi_i y_{t-i} + \sum_{i=1}^n \theta_i \epsilon_{t-i} + \epsilon_t \quad (4)$$

$$\epsilon_t = h_t^{1/2} \eta_t \quad (5)$$

$$h_t = S_3(\boldsymbol{\omega}_1, t) + \psi_2 Trend_t + S_4(\boldsymbol{\omega}_2, t) Trend_t + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^o \gamma_i \epsilon_{t-i}^2 \mathbb{1}(\epsilon_{t-i} < 0) + \sum_{i=1}^p \beta_i h_{t-i} \quad (6)$$

$$S_i(\boldsymbol{\lambda}, t) = \lambda_0 + \sum_{j=1}^{J_i} \lambda_{1,j} \sin\left(2j\pi \frac{d(t)}{365}\right) + \lambda_{2,j} \cos\left(2j\pi \frac{d(t)}{365}\right) \quad (7)$$

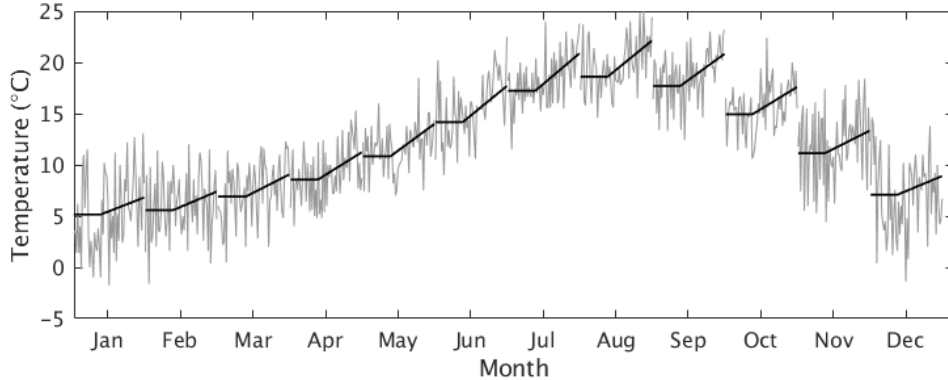
T_t is the temperature variable; $\boldsymbol{\mu}_i$ and $\boldsymbol{\omega}_i$ are vectors of parameters; ψ_i , ϕ_i , θ_i , α_i , γ_i , β_i , λ_0 and $\lambda_{i,j}$ are scalar parameters; $\boldsymbol{\lambda}$ is a vector with entries λ_0 and $\lambda_{i,j}$; $S_i(\boldsymbol{\lambda}, t)$ are the 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 of a particular periodicity; 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; $\mathbb{1}(\cdot)$ is the indicator function; and m , n , q , o , and p are the orders of the ARMA and GARCH components. Following Franses et al. (2001), Taylor and Buizza (2004) and Dupuis (2014), we included an asymmetric GARCH term, $\sum_{i=1}^o \gamma_i \epsilon_{t-i}^2 \mathbb{1}(\epsilon_{t-i} < 0)$, in expression (6) 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. In the proposed formulation, η_t is assumed to be an i.i.d. Gaussian distribution with zero mean and unit variance. We also experimented with the assumptions of Student-t and skew-t distributions, but this did not lead to improved forecast accuracy.

4.2 Model Discussion

In Section 2, we described how our TN and TX time series exhibited an apparent trend, which is consistent with the literature on climate change. Unit root tests rejected the hypothesis of the trend being stochastic (see, for example, Cavaliere and Taylor 2009). 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. Visual inspection of the time series suggests a steady rise 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 starting year using the Schwarz Bayesian Criterion (SBC). We did this for the TN and TX 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 (3)-(7) as being a linear trend starting on 1 January 1974. In Figure 1 and Figure 2, we show the trends of this type fitted to the TN and TX 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.

A 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 effect on the different seasons of the year. Figure 4 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. Allowing the trend to differ across the days of the year was motivated by the work of Proietti and Hillebrand (2017) with monthly temperature data.

Figure 4: For the first day of each month, Seville daily minimum temperature (TN) (in °C) plotted for the first 60-year estimation sample, along with the trend estimated by the univariate ARMA-GARCH model with Gaussian assumption.



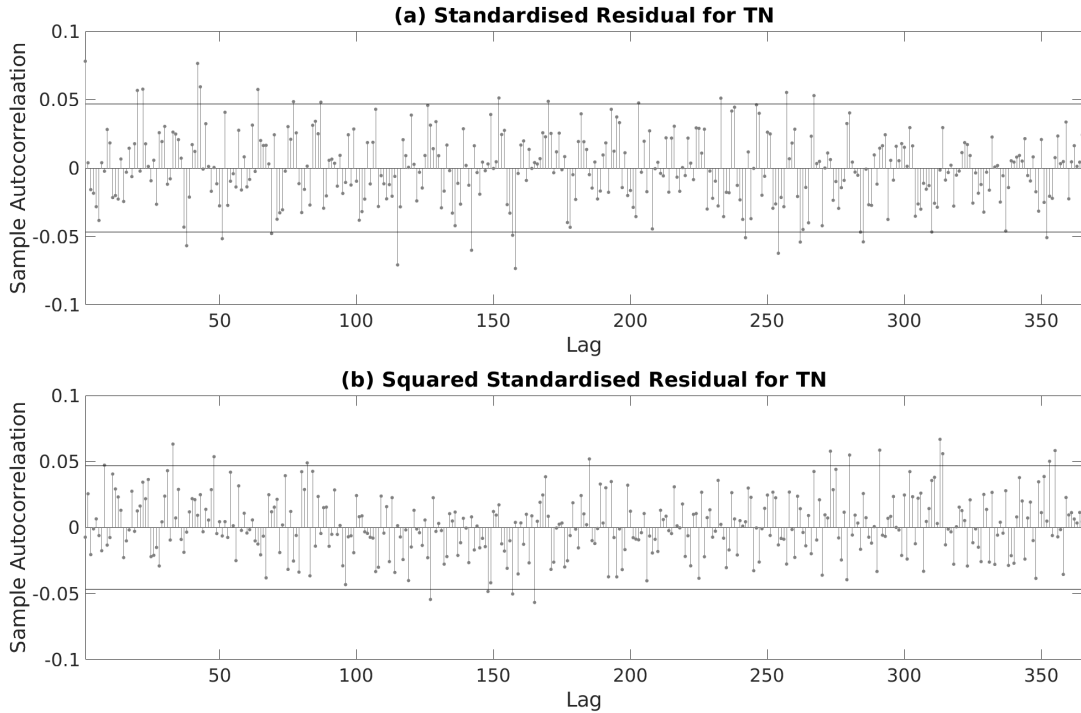
The lag orders, the numbers of Fourier terms, and the specification of the trend term were chosen using the SBC. For the mean, this led to the use of three pairs of Fourier terms, interaction terms involving the trend and three pairs of Fourier terms, and an ARMA(3,1) model. For the variance, this led to 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 an asymmetric GARCH term.

In the appendix, Table 9 presents the parameters for the univariate ARMA-GARCH model with Gaussian assumption, estimated using the first 60-year rolling window of TN and TX recorded at Seville. The insight provided by the model is consistent with previous studies. The coefficients ψ_1 for the trend terms in the ARMA parts of both models are positive, which indicates that the levels of TN and TX are rising. The coefficient ψ_1 of the trend term in the ARMA part for TN is larger than that for TX, which suggests a decrease in the diurnal temperature range (DTR), defined as the difference between TN and TX. These findings are consistent with those in previous work (see, for example, Dupuis 2014; Qu et al. 2014). One term of each interaction pair $\{\mu_2, \omega_2\}$ is significant in both the ARMA and GARCH parts of the model, implying that the impact of the trend is not the same across the seasonal cycle. This was the finding of Proietti and Hillebrand (2017) for monthly data; our work has confirmed the existence of the effect in the mean and variance of daily data. In the rest of the paper, 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.

Figure 5 shows the out-of-sample autocorrelation plots of the standardised residuals and squared standardised residuals for TN, where the residuals were obtained using the ARMA-

GARCH model with Gaussian distribution. It can be seen that the residuals and squared residuals only have little autocorrelation structure left. The autocorrelation plots for TX were similar to that for TN, hence we omit them in the paper.

Figure 5: Out-of-sample autocorrelation plots for (a) the standardised residuals for TN, and (b) the squared standardised residuals for TN. The plots in (a) and (b) were produced using the ARMA-GARCH model with Gaussian distribution, estimated using the first 60 years of data for Seville.



5 Bivariate VARMA-MGARCH Model

In this section, we present our proposed bivariate model for TN and TX. Although multivariate time series models, such as the vector autoregressive moving average (VARMA) and multivariate GARCH (MGARCH) models, have been applied successfully in various applications (see, for example, Carroll et al. 2017; Taylor and Jeon 2018), they have not been used to model daily maximum and minimum temperature. In this paper, we contribute to the literature on the modelling and probabilistic forecasting of heat waves by evaluating bivariate and univariate time series models of the daily minimum and maximum temperature.

We use a VARMA model for the mean of a vector containing the two variables, TN and TX, and a MGARCH model for the covariance matrix. There exist different 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 use the VEC model,

which has a very flexible and general specification. In comparison with the univariate model of Section 4, our bivariate model has the same essential features, with the addition of a model for the conditional covariance. Having said that, it is important to note that the lag structure is far richer in the bivariate model, as the two temperature variables are also modelled in terms of lags of each other. The proposed model is expressed as follows:

$$\mathbf{T}_t = \mathbf{S}_1(\boldsymbol{\mu}_1, t) + \psi_1 Trend_t + \mathbf{S}_2(\boldsymbol{\mu}_2, t)Trend_t + \mathbf{y}_t \quad (8)$$

$$\mathbf{y}_t = \sum_{i=1}^m \phi_i \mathbf{y}_{t-i} + \sum_{i=1}^n \theta_i \boldsymbol{\epsilon}_{t-i} + \boldsymbol{\epsilon}_t \quad (9)$$

$$\boldsymbol{\epsilon}_t = \mathbf{H}_t^{1/2} \boldsymbol{\eta}_t \quad (10)$$

$$\begin{aligned} vech(\mathbf{H}_t) &= \mathbf{S}_3(\boldsymbol{\omega}_1, t) + \psi_2 Trend_t + \mathbf{S}_4(\boldsymbol{\omega}_2, t)Trend_t \\ &\quad + \sum_{i=1}^q \boldsymbol{\alpha}_i vech(\boldsymbol{\epsilon}_{t-i} \boldsymbol{\epsilon}_{t-i}^T) \\ &\quad + \sum_{i=1}^o (\boldsymbol{\gamma}_i vech(\boldsymbol{\epsilon}_{t-i} \boldsymbol{\epsilon}_{t-i}^T)) \odot \begin{pmatrix} \mathbb{1}(\epsilon_{1,t} < 0) \\ \mathbb{1}(\epsilon_{2,t} < 0) \\ \mathbb{1}(\epsilon_{1,t} \epsilon_{2,t} < 0) \end{pmatrix} \\ &\quad + \sum_{i=1}^p \boldsymbol{\beta}_i vech(\mathbf{H}_{t-i}) \end{aligned} \quad (11)$$

$$\mathbf{S}_i(\boldsymbol{\lambda}, t) = \boldsymbol{\lambda}_0 + \sum_{j=1}^{J_i} \boldsymbol{\lambda}_{1,j} \sin\left(2j\pi \frac{d(t)}{365}\right) + \boldsymbol{\lambda}_{2,j} \cos\left(2j\pi \frac{d(t)}{365}\right) \quad (12)$$

where $\mathbf{T}_t = (TN_t, TX_t)$; $vech$ denotes the operator that stacks the lower triangular portion of a matrix as a column vector; $(\cdot)^T$ denotes matrix transpose; $\boldsymbol{\mu}_i$, $\boldsymbol{\omega}_i$, ψ_i , $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_{i,j}$ are vectors of parameters; $\boldsymbol{\lambda}$ is a vector consisting of the concatenation of $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_{i,j}$; ϕ_i , θ_i , $\boldsymbol{\alpha}_i$, $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}_i$ are parameter matrices; $\mathbf{S}_i(\boldsymbol{\lambda}, t)$ is a deterministic seasonal vector; \mathbf{y}_t contains the stochastic parts of \mathbf{T}_t ; J_i denotes the number of pairs of Fourier terms of a particular periodicity; $\boldsymbol{\epsilon}_t$ is the error term of \mathbf{y}_t ; \mathbf{H}_t is the covariance matrix of $\boldsymbol{\epsilon}_t$; and m , n , q , o and p are the VARMA and MGARCH orders. 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 Section 4. For practicality, in specifying the bivariate model, we used the analogous terms that we had selected for the univariate model. For example, we used the same orders of ARMA and GARCH terms.

6 Empirical Forecasting Study

6.1 Forecasting Models

We considered the raw ensemble prediction as well as the recalibrated weather ensemble predictions by implementing the EMOS-ECC, as described in Section 3. We recalibrated the raw weather ensemble predictions using the most recent 40, 100, 200 or 300 days as the estimation period. The EMOS-ECC method was implemented in Matlab.

We implemented the univariate ARMA-GARCH model of Section 4 for TN and TX. To obtain the bivariate distribution of TN and TX, we took the simplistic approach of assuming independence between the univariate distributions. We then implemented the Gaussian bivariate VARMA-MGARCH model for TN and TX, which was described in Section 5. In the VEC approach that we used within the VARMA-MGARCH models, we ensured the positive definiteness of the matrix \mathbf{H}_t by adding a large penalty term to the likelihood in the case of a violation occurring. For the optimisation algorithm, we set the initial parameter values to be those estimated using the univariate Gaussian models for TN and TX, with the remaining parameters set to zero. We then passed this initial parameter vector to the interior point optimisation algorithm in Matlab. It is worth pointing out that, to achieve better estimation efficiency, we estimated all the parameters simultaneously, rather than utilising a multi-step approach. The multi-step sequential regression approaches were merely used to obtain initial parameter vectors for the numerical optimisation algorithm. The univariate and bivariate models were both implemented in Matlab.

Producing multi-day-ahead forecasts from the time series models is not trivial. The ARMA-GARCH and VARMA-MGARCH models have closed form 1 day-ahead distributional forecasts, but such analytic expressions do not exist for multi-step-ahead prediction. To address this, we used simulation from each model with 1500 sample paths simulated over the forecast horizon of multiple days. The marginal and joint distributions were constructed from these sample paths. We make two remarks here. Firstly, in practice, TX should obviously be no less than TN. We did not address this issue explicitly in our models because such temperature crossing occurred very infrequently due to the distributions of TN and TX being reasonably far apart throughout the year. In our simulation study, the frequency of crossing occurred in fewer than 1% of the simulated sample paths, and so we took the pragmatic approach of setting TN and TX to be their average when crossing occurred.

6.2 Evaluating Forecasts of Marginal and Joint Distributions

In this section, we consider the evaluation of forecasts of univariate distributions of daily TN and TX, and the joint distributions for daily TN and TX up to 15 days ahead for the out-of-sample period 2011-2015. To evaluate univariate and multivariate distributional forecasts, we used the continuous ranked probability score (CRPS) and energy score, respectively. The CRPS is widely used to evaluate univariate distributional forecasts (Gneiting and Raftery 2007).

The CRPS is expressed as follows:

$$\text{CRPS}(F_t, y_t) = -\frac{1}{2}\mathbb{E}|X_t - X'_t| + \mathbb{E}|X_t - y_t| \quad (13)$$

where y_t is the actual observation; X_t and X'_t are random variables with CDF F_t ; and $\mathbb{E}(\cdot)$ is the expectation. The score is averaged over the out-of-sample period to produce an overall measure of accuracy.

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

$$\text{ES}(F_t, \mathbf{y}_t) = -\frac{1}{2}\mathbb{E}\|\mathbf{X}_t - \mathbf{X}'_t\| + \mathbb{E}\|\mathbf{X}_t - \mathbf{y}_t\| \quad (14)$$

where \mathbf{y}_t is the actual observation; $\|\cdot\|$ is the Euclidean norm; \mathbf{X}_t and \mathbf{X}'_t are two independent copies of a random vector with distribution F_t .

The CRPS and energy score are said to be *proper* because their expectation is minimised by the distributional forecast that is equal to the true distribution. This property is important, as it encourages honest assessments (Garthwaite et al. 2005). In this paper, the CRPS and the energy score are both negatively oriented, that is the lower the better. For simplicity and clarity, we present the skill scores instead of the raw values for the CRPS and energy score. For each score S and each method F , the skill score is calculated as $100 \times \left(1 - \frac{\bar{S}_F}{\bar{S}_{ref}}\right)$, where \bar{S}_F and \bar{S}_{ref} are the score values averaged over the out-of-sample period and four locations corresponding to forecast F and the reference forecast, respectively. Positive values imply greater accuracy than the reference method, and higher values indicate superior accuracy. The reference method was chosen as the univariate ARMA-GARCH models for TN and TX.

Table 1 presents the CRPS skill scores, computed for the whole 5-year out-of-sample period, for the CRPS for the univariate distributions of TN and TX. In these tables, Uni-TN&TX

denotes the univariate ARMA-GARCH models for TN and TX; Bi-TN&TX denotes the bivariate VARMA-MGARCH model for the joint distribution of TN and TX; RAW NWP denotes the raw NWP predictions; and NWP-40, NWP-100, NWP-200 and NWP-300 denote the recalibrated NWP ensemble predictions via EMOS-ECC with the length of the estimation period being 40, 100, 200, and 300 days, respectively. Each row in the tables can be viewed as representing the forecasts produced by a pair of the temperature variables. For example, the first row in each table was produced by the univariate models of TN and TX, where the TN forecasts were produced using just the model for TN; and the TX forecasts were produced using just the model for TX.

Table 1 provides three main insights. Firstly, the time series models were beaten by the recalibrated NWP ensemble predictions. Secondly, comparing just the time series models, the results show that, at least for TN, the bivariate model was better than the univariate models for the shorter lead time, but the bivariate model was generally outperformed for the longer lead times. Thirdly, we find that the recalibrated NWP outperformed the raw NWP predictions, and longer estimation periods for the EMOS-ECC generally led to better performance for longer lead times.

Table 1: CRPS skill scores for the univariate distributions of TN and TX for lead times up to 15 days for the full 5-year out-of-sample period. Skill scores are computed relative to Uni-TX&TN.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
TN															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	7.0	4.7	2.1	0.7	0.3	-0.1	-0.3	-0.4	-0.4	-0.5	-0.3	-0.4	-0.4	-0.5	-0.6
RAW NWP	5.7	24.4	28.8	29.2	27.5	23.4	18.8	13.0	7.6	3.9	-4.2	-6.9	-8.8	-10.0	-11.8
NWP-40	37.1	45.5	46.1	44.1	40.1	33.6	25.6	17.6	9.0	1.2	-5.0	-10.5	-14.9	-17.9	-20.8
NWP-100	36.1	45.3	46.1	44.5	41.1	35.2	27.9	20.6	13.4	8.1	2.5	-1.8	-5.4	-8.0	-9.7
NWP-200	35.4	45.1	46.3	44.9	41.6	36.0	29.4	22.3	15.9	10.8	6.8	3.1	0.6	-1.6	-4.0
NWP-300	34.4	44.2	45.5	44.3	41.2	35.6	28.7	21.8	15.7	10.7	7.1	3.6	1.0	-0.8	-2.8
TX															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	0.1	0.0	-0.1	0.0	-0.0	-0.0	-0.1	-0.0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
RAW NWP	8.3	28.2	34.2	35.4	33.2	29.5	-8.2	-10.2	-12.1	-14.7	-22.1	-23.9	-25.8	-27.5	-29.1
NWP-40	54.4	60.5	59.2	55.7	50.0	42.6	32.7	24.0	15.1	6.2	-2.4	-9.1	-14.3	-18.6	-25.6
NWP-100	53.4	60.2	59.6	56.6	51.5	44.6	35.1	27.4	19.9	13.5	7.8	3.1	-1.4	-4.2	-6.7
NWP-200	51.9	59.4	59.2	56.8	51.8	44.9	36.7	29.6	22.8	17.0	11.8	8.1	4.8	2.2	-0.0
NWP-300	51.7	59.4	59.5	57.1	52.2	45.4	37.2	30.3	24.1	18.8	14.2	10.8	7.9	5.5	3.3

Note: Large skill scores are better. The best method of each column in each block is indicated in bold.

For the two temperature variables, Tables 2-3 provide a breakdown of the CRPS results for the seasons of the year, with each table containing a block of results for each of the four seasons: spring (March-May), summer (June-August), autumn (September-November), and winter (December-February). Table 4 presents the skill scores for the energy score for the joint distributions of TN and TX. These tables confirm the superiority of the NWP approaches,

with the time series models becoming relatively more competitive for longer lead times.

Table 2: CRPS skill scores for the univariate distribution of TN for out-of-sample forecasts for lead times up to 15 days for each season of the year. Skill scores are computed relative to Uni-TX&TN.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spring (March-May)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	9.6	8.0	4.4	2.3	1.3	0.4	-0.1	-0.4	-0.5	-0.9	-0.8	-0.8	-0.9	-0.8	-0.9
RAW NWP	4.2	22.6	27.4	28.2	27.0	24.0	20.1	14.5	8.1	3.9	-6.0	-8.0	-10.9	-14.2	-17.8
NWP-40	34.9	44.4	45.9	44.7	41.8	37.3	29.9	21.3	12.0	5.0	-2.7	-8.6	-12.6	-17.8	-23.7
NWP-100	33.7	44.1	45.9	45.3	43.2	39.0	31.5	24.1	15.7	10.6	1.9	-3.7	-10.5	-16.5	-17.7
NWP-200	33.3	44.2	46.5	46.0	44.0	40.0	34.1	27.1	20.5	15.4	11.8	8.6	5.5	1.6	-1.6
NWP-300	33.4	44.7	47.1	46.8	45.3	41.2	34.9	28.4	21.8	16.8	13.6	11.0	8.2	4.7	1.8
Summer (June-August)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	17.2	8.2	3.0	0.2	-0.2	-0.6	-1.0	-1.0	-0.9	-0.4	-0.1	-0.1	0.1	-0.3	-0.5
RAW NWP	15.9	30.4	32.8	32.6	30.3	27.1	21.0	12.7	5.8	1.9	-2.7	-5.9	-6.2	-6.9	-9.4
NWP-40	41.9	47.2	47.0	44.9	40.4	35.6	29.1	19.9	11.1	2.9	-1.2	-9.3	-11.9	-14.3	-19.0
NWP-100	41.6	47.6	47.6	45.3	41.4	37.1	30.4	21.3	13.9	8.4	5.4	1.0	-0.9	-1.8	-3.4
NWP-200	40.5	46.7	46.7	44.3	40.5	36.0	29.7	20.8	14.2	8.6	6.0	1.9	-0.2	-2.3	-5.5
NWP-300	39.9	46.0	46.1	43.9	40.2	35.7	29.6	20.9	14.6	9.9	8.0	4.1	1.9	0.3	-2.4
Autumn (September-November)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	3.6	3.8	1.6	0.5	0.0	-0.2	0.2	0.3	0.3	0.1	0.2	0.1	-0.2	-0.0	-0.1
RAW NWP	7.2	28.9	33.3	32.3	29.0	23.7	19.6	14.2	9.9	6.7	-1.5	-4.4	-6.7	-8.7	-10.1
NWP-40	34.3	43.6	43.5	39.8	34.2	27.3	18.9	9.6	1.3	-6.7	-11.4	-15.8	-25.1	-32.6	-35.2
NWP-100	33.2	43.6	43.9	40.7	35.9	29.1	22.0	14.1	7.4	3.3	-0.3	-4.0	-6.7	-10.5	-13.2
NWP-200	33.3	44.3	45.1	42.4	38.0	31.8	25.6	18.6	12.4	7.8	4.7	1.1	-1.2	-3.5	-5.2
NWP-300	31.4	42.3	43.4	41.3	36.6	30.5	24.6	18.5	13.0	8.5	6.2	3.5	1.7	0.4	-0.4
Winter (December-February)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	-1.4	-0.7	-0.5	-0.2	0.0	0.1	-0.3	-0.6	-0.7	-0.7	-0.6	-0.7	-0.6	-0.7	-0.8
RAW NWP	-2.9	17.0	22.6	24.7	24.4	19.5	15.2	10.9	6.5	2.9	-6.3	-8.8	-10.6	-9.4	-9.4
NWP-40	37.5	46.8	47.9	46.8	43.3	34.0	24.6	19.2	11.0	3.1	-4.7	-8.4	-10.4	-7.2	-6.2
NWP-100	36.2	45.9	47.2	46.5	43.4	35.4	27.5	22.2	16.0	9.8	3.3	-0.1	-2.9	-2.4	-3.7
NWP-200	35.0	45.3	46.7	46.3	43.2	35.9	28.0	22.2	15.8	10.6	4.6	0.8	-1.8	-2.3	-3.9
NWP-300	33.2	44.0	45.5	45.0	42.0	34.5	25.8	19.2	13.1	7.3	1.2	-3.7	-7.2	-7.9	-9.8

Note: Large skill scores are better. The best method of each column in each block is indicated in bold.

Table 3: CRPS skill scores for the univariate distribution of TX temperature for out-of-sample forecasts for lead times up to 15 days for each season of the year. Skill scores are computed relative to Uni-TX&TN.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spring (March-May)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	0.1	-0.0	-0.1	-0.0	-0.2	-0.2	-0.4	-0.3	-0.5	-0.4	-0.4	-0.4	-0.5	-0.5	-0.4
RAW NWP	9.3	29.0	36.1	39.0	37.4	32.7	4.9	3.8	1.4	-1.2	-7.3	-10.8	-14.3	-18.7	-21.6
NWP-40	59.7	65.4	65.2	62.2	57.4	49.0	38.7	30.3	21.5	13.8	6.9	0.8	-4.6	-10.1	-18.8
NWP-100	58.9	65.1	65.3	63.1	58.3	50.5	41.1	33.4	24.9	18.2	11.9	6.6	-0.6	-5.5	-8.7
NWP-200	57.0	64.1	64.8	62.9	58.3	50.9	42.4	35.2	27.5	22.2	17.0	12.9	8.6	4.9	2.6
NWP-300	56.2	63.3	64.0	62.3	57.9	50.6	42.9	36.2	29.3	24.5	19.9	15.8	11.7	7.6	4.6
Summer (June-August)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	0.1	0.1	0.1	0.2	0.1	-0.1	0.0	0.1	0.0	0.0	0.0	0.3	0.3	0.2	0.1
RAW NWP	12.1	39.3	46.8	46.4	43.2	39.7	17.5	13.3	8.3	3.9	-1.3	-3.3	-4.0	-5.2	-6.9
NWP-40	59.2	63.8	60.9	55.3	48.0	41.0	31.1	21.0	9.4	1.0	-10.2	-17.2	-20.5	-23.8	-31.5
NWP-100	57.6	63.3	60.9	55.3	48.1	40.8	32.3	23.1	15.0	8.9	2.4	-1.7	-3.1	-4.9	-7.0
NWP-200	55.7	62.0	59.8	54.9	48.0	40.3	32.7	24.3	16.8	10.3	4.0	-0.4	-2.6	-5.2	-6.9
NWP-300	55.9	62.5	60.6	55.8	48.7	40.7	32.3	23.9	16.6	10.9	5.3	1.9	0.3	-1.6	-3.6
Autumn (September-November)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	-0.0	-0.2	-0.4	-0.3	-0.3	-0.2	-0.2	-0.1	-0.1	-0.2	-0.1	-0.1	-0.0	-0.0	0.0
RAW NWP	11.1	29.3	33.4	34.0	31.6	27.1	-16.6	-18.1	-19.2	-20.7	-29.5	-31.3	-33.1	-34.7	-36.6
NWP-40	50.5	57.4	56.1	53.0	46.5	38.5	28.6	19.0	9.5	-2.0	-11.6	-21.6	-30.2	-36.0	-41.5
NWP-100	49.8	57.6	57.0	54.5	49.6	42.7	33.0	25.0	17.0	10.2	4.7	-0.6	-4.9	-7.7	-10.8
NWP-200	47.1	56.3	56.7	55.1	50.4	43.4	35.8	28.4	21.3	15.4	10.1	6.4	2.9	0.4	-2.6
NWP-300	46.8	56.5	57.4	55.9	51.2	45.1	36.7	29.8	24.0	19.1	14.6	11.5	9.1	7.7	5.7
Winter (December-February)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	0.2	0.2	0.2	0.2	0.4	0.6	0.6	0.4	0.4	0.2	0.2	0.1	0.2	0.2	0.1
RAW NWP	-0.6	11.3	15.7	16.9	15.5	14.3	-51.5	-52.6	-50.6	-51.9	-62.0	-60.9	-61.1	-58.9	-57.3
NWP-40	45.7	52.3	51.2	49.1	44.9	39.0	29.7	23.0	17.9	9.3	1.9	-1.0	-4.7	-6.9	-12.0
NWP-100	44.9	51.8	52.0	50.5	46.9	41.7	31.4	25.7	20.9	14.9	11.0	7.2	3.1	2.4	1.3
NWP-200	45.5	52.6	52.8	51.4	47.8	42.7	33.4	28.2	23.9	18.0	14.3	12.3	9.5	8.3	6.4
NWP-300	45.9	52.9	53.3	51.9	48.3	42.9	34.3	29.0	24.4	18.6	14.6	12.4	9.6	8.1	6.1

Note: Large skill scores are better. The best method of each column in each block is indicated in bold.

Table 4: Energy score skill scores for the joint distribution of TN and TX for lead times up to 15 days for each season and for the full 5-year out-of-sample period. Skill scores are computed relative to Uni-TX&TN.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spring (March-May)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	3.7	3.2	2.0	1.4	1.1	0.9	0.8	0.7	0.5	0.5	0.5	0.4	0.7	0.7	0.5
RAW NWP	8.5	27.4	33.6	35.4	33.8	29.5	10.2	7.2	3.5	0.5	-5.9	-8.8	-11.7	-15.6	-18.8
NWP-40	48.4	56.4	57.3	55.2	51.1	44.2	35.1	26.6	17.6	10.3	3.6	-2.4	-6.7	-11.6	-19.4
NWP-100	47.5	56.2	57.4	56.0	52.2	45.7	37.2	29.5	21.2	15.0	8.2	2.9	-3.4	-8.6	-11.2
NWP-200	46.4	55.8	57.4	56.2	52.5	46.4	39.0	31.9	24.7	19.4	15.0	11.3	7.7	4.1	1.3
NWP-300	45.9	55.5	57.1	56.1	52.8	46.6	39.7	33.1	26.4	21.6	17.7	14.2	10.9	7.2	4.3
Summer (June-August)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	7.4	4.4	2.7	1.9	1.8	1.6	1.7	1.6	1.5	1.5	1.5	1.6	2.0	1.7	1.4
RAW NWP	15.0	36.4	41.4	40.9	38.0	34.8	18.2	12.4	6.7	2.6	-1.6	-4.1	-4.3	-5.4	-7.6
NWP-40	51.3	56.8	54.9	50.6	44.5	38.6	30.4	21.0	10.8	2.8	-6.0	-13.1	-15.3	-17.9	-25.0
NWP-100	50.2	56.6	55.1	50.7	44.8	38.9	31.5	22.6	15.0	9.2	4.1	0.1	-1.1	-2.8	-4.9
NWP-200	48.6	55.4	54.1	50.1	44.6	38.3	31.6	23.2	16.2	10.3	5.4	1.2	-0.6	-2.9	-5.2
NWP-300	48.4	55.4	54.4	50.6	45.0	38.5	31.4	23.2	16.5	11.3	7.1	3.7	2.2	0.5	-1.9
Autumn (September-November)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	1.7	1.9	1.0	0.6	0.4	0.4	0.6	0.6	0.5	0.4	0.4	0.4	0.6	0.6	0.5
RAW NWP	9.3	29.3	33.6	33.4	30.5	25.8	-0.8	-4.1	-6.3	-8.3	-16.0	-18.3	-19.9	-21.6	-23.3
NWP-40	42.1	50.7	50.1	46.7	40.7	33.3	24.0	14.5	6.0	-3.6	-10.9	-18.6	-27.6	-34.3	-38.5
NWP-100	41.3	50.8	50.8	48.0	43.3	36.7	28.3	20.2	12.8	7.3	2.6	-1.9	-5.5	-8.8	-11.7
NWP-200	40.0	50.5	51.2	49.1	44.7	38.3	31.5	24.2	17.6	12.4	8.0	4.5	1.6	-0.8	-3.3
NWP-300	39.0	49.7	50.8	49.0	44.5	38.6	31.4	24.8	19.2	14.6	11.0	8.0	5.9	4.5	2.9
Winter (December-February)															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	-0.2	0.3	0.2	0.5	0.5	0.7	0.6	0.3	0.2	-0.1	-0.1	-0.2	0.0	-0.1	-0.2
RAW NWP	-0.1	15.5	20.3	22.0	20.8	17.8	-15.8	-18.5	-20.1	-22.6	-31.5	-32.5	-33.6	-32.2	-31.8
NWP-40	40.7	48.6	48.7	47.2	43.3	35.8	26.3	20.4	13.7	5.3	-2.3	-5.5	-8.1	-7.7	-10.0
NWP-100	39.6	47.9	48.7	47.7	44.3	37.8	28.7	23.2	17.6	11.4	6.0	2.5	-0.7	-0.7	-2.0
NWP-200	39.2	47.8	48.7	47.9	44.6	38.5	29.9	24.3	19.0	13.4	8.5	5.6	3.1	2.5	0.7
NWP-300	38.3	47.3	48.3	47.5	44.2	38.0	29.2	23.2	17.9	12.1	7.0	3.5	0.6	-0.2	-2.2
Whole out-of-sample period															
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	3.2	2.5	1.5	1.1	1.0	0.9	0.9	0.8	0.7	0.6	0.6	0.5	0.8	0.7	0.5
RAW NWP	8.2	27.3	32.5	33.2	31.1	27.2	3.5	-0.2	-3.6	-6.5	-13.3	-15.5	-17.1	-18.6	-20.4
NWP-40	45.8	53.3	53.0	50.3	45.3	38.4	29.4	21.0	12.4	4.1	-3.4	-9.4	-14.0	-17.5	-23.0
NWP-100	44.8	53.1	53.3	50.9	46.5	40.2	31.8	24.3	17.0	11.0	5.4	1.0	-2.8	-5.4	-7.7
NWP-200	43.7	52.6	53.1	51.2	47.0	40.8	33.4	26.3	19.7	14.3	9.6	6.0	3.3	0.9	-1.4
NWP-300	43.1	52.2	52.9	51.1	47.0	40.8	33.4	26.6	20.5	15.4	11.2	7.8	5.3	3.3	1.0

Note: Large skill scores are better. The best method of each column in each block is indicated in bold.

6.3 Model Confidence Set

In Section 6.2, we presented the skill scores for the CRPS and energy score. In this section, we consider statistical significance in the differences between the scores of a pair of methods, by reporting model confidence set results (Hansen et al. 2011). The model confidence set method is a sequential algorithm to select the best-performing models out of a pool of candidate models, where in each step, the Diebold and Mariano test is used to test for statistical significance between the scores of a pair of methods (Diebold and Mariano 2002).¹

The output of the model confidence set method is a set of models that contains the best-

¹To implement the model confidence set approach to evaluation, we used the MATLAB code provided by Kevin Sheppard in his MFE Toolbox, which is freely available on https://www.kevinsheppard.com/MFE_Toolbox.

performing model with a given confidence level. Following Hansen et al. (2011), we considered the 75% and 90% confidence levels. The results for these confidence levels were similar, hence, to save space, we only report the results for the 90% confidence level.

In the appendix, we report two tables of the MCS results, Table 10 and Table 11: the former presents the number of locations for which each model lies inside the model confidence set for the 90% confidence levels for the 5-year out-of-sample period, for the CRPS for the univariate distributions of TN and TX, and the latter 11 presents the model confidence set results for the energy score for the joint distributions of TN and TX. A large number is preferred, where the best and worst values are 4 and 0 respectively.

The results in these tables are consistent with the results in Section 6.2. The recalibrated NWP approaches with the length of the estimation period being 200 and 300 performed the best. Although the time series models were beaten by the NWP approaches, they became relatively more competitive at the longer lead times such as 14 and 15 days.

6.4 Evaluating Heat Wave Forecasts

We evaluated forecasts of heat waves defined as the simultaneous exceedances of TN and TX over specified thresholds for two consecutive days. For example, the 1 day-ahead forecast for a simultaneous exceedance is based on the 1 and 2 day-ahead predictions for the joint distribution of TN and TX. Hence, we can produce such forecasts up to 14 days-ahead.

There is no universal choice for the thresholds, as they are generally location specific. We considered the thresholds 15°C and 30°C for TN and TX, respectively. These thresholds have been used by the UK Met Office. As Spanish temperatures are generally hotter than in the UK, we also considered the thresholds 17.5°C and 35°C. To give an idea of the extremity of these temperatures, for each city, Table 5 presents the proportion of the in-sample and out-of-sample summer days (June, July, and August) for which TN and TX exceeded each threshold individually and simultaneously, and Table 6 presents the proportion of the in-sample and out-of-sample summer days (June, July, and August) for which TN and TX exceeded each threshold for two consecutive days, individually and simultaneously.

Table 5: Percentage of summer observations exceeding the thresholds for each of the four Spanish locations.

	TN>15°C	TX>30°C	TN>15°C & TX>30°C	TN>17.5°C	TX>35°C	TN>17.5°C & TX>35°C
Period 1951-2010						
Albacete	53	64	46	21	15	8
Seville	88	82	77	64	46	39
Cáceres	76	66	63	53	23	22
Madrid	75	50	50	49	8	8
Period 2011-2015						
Albacete	74	77	66	44	26	22
Seville	96	90	88	83	58	55
Cáceres	79	75	70	54	34	31
Madrid	87	72	71	66	21	21

Table 6: Percentage of summer observations exceeding the thresholds for two consecutive days for each of the four Spanish locations.

	TN>15°C	TX>30°C	TN>15°C & TX>30°C	TN>17.5°C	TX>35°C	TN>17.5°C & TX>35°C
Period 1951-2010						
Albacete	42	56	36	12	9	4
Seville	84	77	71	55	36	30
Cáceres	69	58	54	42	16	15
Madrid	69	41	41	40	4	4
Period 2011-2015						
Albacete	67	70	58	32	19	15
Seville	95	85	84	76	48	43
Cáceres	71	68	62	43	23	20
Madrid	83	65	65	52	13	13

To evaluate forecasts of the probability of the event that TN and TX exceeded the specified thresholds for two consecutive days, we calculated the widely used *Brier score* as follows:

$$BS(p_t, o_t) = (p_t - o_t)^2$$

where p_t is the predicted probability, o_t is the actual outcome of the event, which is 1 if the event occurs, and 0 otherwise. In a similar way to the calculation of skill scores in Section 6.2, we computed the skill score for the Brier score of each model as $100 \times \left(1 - \frac{\overline{BS}_F}{\overline{BS}_{ref}}\right)$.

Table 7 presents the skill scores and, in the appendix, Table 12 presents the corresponding model confidence set results for the thresholds 15°C and 30°C. In these tables, the first vertical block corresponds to the event that TN exceeds 15°C on each of the next two days; the second vertical block reports the results for the event that TX exceeds 30°C on each of the next two days; and the third vertical block corresponds to the event that TN exceeds 15°C and TX exceeds 30°C on each of the next two days. Similarly, Table 8 presents the Brier skill scores

and, in the appendix, Table 13 presents the corresponding model confidence set results for the more extreme thresholds of 17.5°C and 35°C.

In Tables 7-8 and 12-13, the rankings of methods are similar to that for the CRPS and energy score in the summer months in Tables 2-4 and 10-11 discussed in Section 6.2. Overall, the raw and recalibrated NWP dominated the time series models, and the time series methods became relatively more competitive at the longest of the lead times such as 13 and 14 days.

Table 7: Brier skill scores for out-of-sample probability forecasts of heat waves for the less extreme pair of thresholds. Skill scores averaged across the four locations.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TN>15°C														
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	12.4	6.2	2.2	0.6	0.2	-0.7	-1.0	-0.9	-0.9	-0.6	-0.7	-0.5	-0.6	-0.9
RAW NWP	16.6	25.8	27.4	26.9	24.5	22.1	15.6	11.0	5.8	-1.3	-7.3	-9.8	-10.5	-10.4
NWP-40	44.5	46.0	43.9	37.1	31.2	26.0	17.5	6.9	-3.7	-10.9	-16.9	-24.4	-27.2	-31.5
NWP-100	44.4	46.6	44.4	39.8	34.5	30.2	22.1	14.2	6.7	1.9	-1.6	-5.6	-8.9	-13.6
NWP-200	42.6	44.7	42.9	38.7	33.9	29.9	21.9	14.6	7.9	3.3	-0.1	-2.9	-5.0	-7.8
NWP-300	41.7	44.7	43.4	39.9	35.2	31.0	23.6	16.8	10.7	6.6	3.4	0.9	-0.7	-2.4
TX>30°C														
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	-0.2	-0.0	0.0	0.2	0.1	-0.1	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4
RAW NWP	24.7	38.1	42.2	40.0	36.7	21.7	9.2	5.2	3.3	-0.1	-2.6	-4.2	-6.0	-7.9
NWP-40	58.9	61.2	58.2	52.0	44.7	36.2	25.9	14.2	6.0	-3.7	-11.4	-18.7	-22.9	-28.1
NWP-100	59.3	62.2	59.0	52.7	45.2	37.9	28.0	19.2	14.9	9.0	5.8	2.7	0.3	-3.2
NWP-200	58.7	61.9	59.4	53.6	46.3	39.5	30.0	21.7	16.9	11.5	8.0	4.8	2.1	-0.4
NWP-300	57.0	60.8	58.3	52.7	45.8	38.9	30.2	22.3	18.0	12.6	9.2	6.5	3.3	0.6
TN>15°C & TX>30°C														
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	6.3	4.8	3.8	3.3	2.8	2.6	2.8	3.1	3.3	3.1	3.1	3.8	4.1	3.5
RAW NWP	35.3	47.4	50.2	45.6	41.3	30.5	20.5	16.7	14.4	9.9	6.0	5.1	2.8	0.1
NWP-40	54.2	56.3	54.7	46.8	40.4	32.9	24.8	16.7	9.1	-0.6	-7.7	-14.6	-17.9	-22.2
NWP-100	54.3	57.4	55.3	48.2	41.8	35.0	26.4	19.5	15.2	9.6	6.1	3.4	1.0	-2.8
NWP-200	53.0	56.4	54.5	47.7	41.9	35.7	27.7	21.2	16.8	11.9	8.6	5.9	3.2	0.1
NWP-300	53.3	56.8	55.0	48.1	42.4	36.5	28.6	22.4	18.7	13.8	10.7	8.4	5.3	2.2

Note: Large skill scores are better. The best method of each column in each block is indicated in bold.

Table 8: Brier skill scores for out-of-sample probability forecasts of heat waves for the more extreme pair of thresholds. Skill scores averaged across the four locations.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TN>17.5°C														
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	13.0	6.7	1.8	0.6	0.1	-0.1	-0.1	-0.2	-0.1	0.1	0.3	0.3	0.5	0.7
RAW NWP	12.8	26.0	28.3	27.9	24.7	20.6	14.9	8.9	4.3	1.2	-2.9	-4.9	-5.8	-9.1
NWP-40	39.6	45.1	44.7	41.5	35.7	30.1	22.7	14.1	4.9	-1.9	-8.4	-17.2	-22.2	-25.3
NWP-100	38.6	45.1	45.4	42.4	37.2	32.5	24.2	16.5	8.8	5.6	2.5	-2.8	-6.1	-10.0
NWP-200	35.2	42.3	42.7	40.1	34.8	30.4	22.3	15.3	9.1	5.7	3.2	-1.3	-4.0	-8.4
NWP-300	34.6	41.2	41.5	39.1	34.3	29.7	22.1	15.6	10.0	7.9	6.3	2.9	0.2	-3.5
TX>35°C														
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	0.1	0.3	0.5	0.4	0.3	0.3	0.4	0.4	0.3	0.6	0.5	0.4	0.5	0.3
RAW NWP	31.5	44.1	45.4	42.8	39.5	29.6	22.0	18.3	12.2	9.1	5.9	5.1	3.6	0.6
NWP-40	60.7	61.1	55.8	49.2	40.1	32.6	23.5	13.2	3.9	-3.4	-11.5	-18.4	-20.5	-29.6
NWP-100	58.1	59.0	54.7	48.4	39.3	33.3	25.7	18.5	11.5	6.1	-0.2	-2.9	-5.8	-10.0
NWP-200	57.0	57.8	54.2	48.5	40.0	34.9	27.8	21.0	13.8	8.4	2.6	-0.2	-3.3	-8.2
NWP-300	57.9	58.5	54.8	49.0	39.9	34.6	27.0	20.1	13.2	8.5	3.9	1.9	-0.2	-3.7
TN>17.5°C & TX>35°C														
Uni-TN&TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bivariate-TN&TX	6.1	4.0	3.0	2.4	2.4	2.5	2.6	3.0	2.9	2.5	2.5	2.5	2.5	2.1
RAW NWP	42.8	51.9	51.0	48.4	44.8	37.4	30.5	26.5	19.2	14.9	11.2	10.2	7.8	4.5
NWP-40	59.9	61.6	56.6	50.3	42.3	35.7	26.3	16.8	8.6	2.5	-4.7	-11.0	-14.2	-20.1
NWP-100	56.6	58.3	54.0	48.5	41.0	35.9	28.3	21.5	14.6	8.7	2.6	-0.8	-3.5	-7.4
NWP-200	54.8	56.5	52.6	47.8	40.9	36.8	29.8	23.1	16.3	10.6	5.2	1.9	-1.5	-6.9
NWP-300	55.3	56.6	52.7	48.0	40.4	36.0	28.9	22.1	15.7	10.7	6.5	4.2	1.6	-2.4

Note: Large values are better, with the best possible value being 4. The best method of each column in each block is indicated in bold.

7 Summary

In this paper, we have considered distributional forecasting of the daily minimum and maximum temperature. We had two main aims. Firstly, to develop a suitable time series model that would provide insight into the evolution of these two temperature variables, as well as provide accurate distributional forecasts. Secondly, we wanted to establish whether such carefully chosen time series models can compete with the accuracy of distributional forecasts based on ensemble predictions from an NWP model. We implemented a univariate ARMA-GARCH model and a bivariate VARMA-MGARCH model. For the NWP ensemble predictions, we recalibrated the raw predictions using the EMOS-ECC methods.

We evaluated forecasts of the marginal and joint distribution for the minimum and maximum temperature. We also evaluated the accuracy of probabilistic forecasts of heat waves, defined as the minimum and maximum simultaneously exceeding chosen thresholds. Overall, we found that recalibrated NWP ensemble predictions with the 200 or 300-day estimation period, comfortably performed the best. The time series models were generally dominated by the raw and recalibrated NWP predictions. For the NWP ensemble predictions, our finding that the recalibration delivers better performance for longer estimation periods,

namely, 200 or 300 days, is consistent with the literature. For the time series models, the performance for longer lead times is more competitive than for short lead times. We also found that the proposed bivariate time series model outperformed the univariate time series models for short lead times, but the advantage did not maintain for long lead times. The time series models revealed some interesting results: there are increasing trends in the minimum and maximum temperatures, there is a decreasing trend in the DTR, and different parts of the seasonal cycle experience different rates of climate change, which are findings that are consistent with those of other researchers.

In terms of future work, we are interested in considering the use of a series of intraday temperature data to forecast TN and TX, although a concern with this is the availability of a long history of intraday temperature observations. Other interesting future research directions include the use of spatial analysis or perhaps the incorporation of covariates in the time series models, such as the global mean temperature, which could be used to capture the trend.

Acknowledgement. The authors are grateful to the European Centre for Medium-Range Weather Forecasts for providing the weather ensemble predictions used in this study. The authors are also very grateful to the three reviewers for providing comments that helped greatly to improve the paper.

Appendix

Table 9: Parameter estimates and standard errors for the mean component of the univariate ARMA-GARCH model with Gaussian assumption, derived using the first 60 years of daily minimum (TN) and maximum (TX) temperature observations for Seville.

	TN	TX
ARMA		
$(\mu_1)_0$	115(0.86)	245(1.08)
$(\mu_1)_{1,1}$	-31.4(1.12)	-40.3(1.43)
$(\mu_1)_{2,1}$	8.11(0.773)	17.3(0.958)
$(\mu_1)_{1,2}$	0.578(0.695)	-2.19(0.8)
$(\mu_1)_{2,2}$	-58.2(1.19)	-87.9(1.39)
$(\mu_1)_{1,3}$	-2.77(0.774)	-2.02(0.94)
$(\mu_1)_{2,3}$	-2.29(0.705)	-2.36(0.792)
ψ_1	0.002(0.000124)	0.00104(0.000175)
$(\mu_2)_{1,1}$	0.0000171(0.000163)	0.000816(0.000233)
$(\mu_2)_{1,2}$	-0.000739(0.000171)	-0.000296(0.00023)
ϕ_1	1.48(0.0218)	1.71(0.0159)
ϕ_2	-0.513(0.0162)	-0.797(0.0164)
ϕ_3	-0.00311(0.00803)	0.073(0.00737)
θ_1	-0.863(0.0208)	-0.932(0.0142)
GARCH		
$(\omega_1)_0$	99.03(15.5)	236(22.3)
$(\omega_1)_{1,1}$	4.74(2.34)	22.2(5.05)
$(\omega_1)_{2,1}$	0.6203(1.31)	-7.99(3.50)
$(\omega_1)_{1,2}$	51.09(8.15)	-65.9(7.53)
$(\omega_1)_{2,2}$	8.50(1.80)	-8.22(3.39)
ψ_2	-0.00297(0.000499)	0.00121(0.000583)
$(\omega_2)_{1,1}$	-0.000238(0.000242)	0.000299(0.000750)
$(\omega_2)_{1,2}$	-0.00208(0.000408)	0.00124(0.000770)
α_1	0.0275(0.00606)	0.124(0.0117)
γ_1	0.0385(0.00870)	-0.0683(0.0119)
β_1	0.757(0.0344)	0.444(0.0475)

Table 10: Model confidence set results for the CRPS of the univariate distributions of TN and TX for lead times up to 15 days for the full 5-year out-of-sample period. The values are the number of locations for which each model lies inside the model confidence set for the 90% confidence levels.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
TN															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	2	2	4	4	4
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	0	2	2	4	4	4
RAW NWP	0	0	0	0	0	0	1	2	3	3	3	3	3	2	2
NWP-40	4	3	3	3	2	2	2	2	1	1	1	1	0	0	0
NWP-100	3	2	3	3	3	2	2	2	1	1	1	1	2	1	0
NWP-200	2	3	3	3	3	4	4	4	4	4	4	4	3	3	2
NWP-300	2	3	3	4	4	4	4	4	4	4	4	4	4	3	3
TX															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
RAW NWP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NWP-40	4	4	4	1	0	0	0	0	0	0	0	0	0	0	0
NWP-100	2	4	3	2	2	3	1	0	0	0	0	0	0	0	0
NWP-200	2	3	3	3	3	3	3	3	2	0	0	0	0	0	2
NWP-300	1	3	3	4	4	4	4	4	4	4	4	4	4	4	4

Note: Large values are better, with the best possible value being 4. The best method of each column in each block is indicated in bold.

Table 11: Model confidence set results for the energy score of the joint distribution of TN and TX for lead times up to 15 days for each season and for the full 5-year out-of-sample period. The values are the number of locations for which each model lies inside the model confidence set for the 90% confidence levels.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spring (March-May)															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
RAW NWP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NWP-40	4	4	4	2	1	0	0	0	0	0	0	0	0	0	0
NWP-100	3	4	4	4	3	3	2	1	0	1	0	0	0	0	0
NWP-200	1	3	4	4	3	3	3	3	2	2	0	0	1	4	4
NWP-300	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4
Summer (June-August)															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	1	2	1	2	2
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	1	2	4	4	4	3
RAW NWP	0	0	0	0	0	3	0	0	0	3	2	2	2	1	0
NWP-40	4	4	4	3	3	4	4	3	1	2	1	1	1	1	0
NWP-100	2	4	4	3	3	4	4	3	3	3	3	3	2	2	1
NWP-200	1	2	3	3	3	3	3	3	3	3	3	2	2	2	1
NWP-300	1	3	4	4	4	4	4	4	4	4	4	4	4	3	2
Autumn (September-November)															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2
RAW NWP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NWP-40	4	4	3	3	0	0	0	0	0	0	0	0	0	0	0
NWP-100	3	4	4	4	4	3	1	0	0	0	0	0	0	0	0
NWP-200	3	3	4	4	4	3	4	3	3	3	3	3	3	3	2
NWP-300	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4
Winter (December-February)															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	1	2	3	3	4
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	0	1	1	4	4	4
RAW NWP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NWP-40	4	4	3	3	3	2	1	2	1	1	0	1	1	1	3
NWP-100	3	4	3	3	3	2	1	2	3	3	2	2	1	2	4
NWP-200	2	3	3	3	3	4	4	4	4	4	4	4	4	4	4
NWP-300	3	4	4	4	4	4	4	3	4	4	3	3	2	3	4
Whole out-of-sample period															
Uni-TN&TX	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3
Bivariate-TN&TX	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
RAW NWP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NWP-40	4	4	4	2	1	0	0	0	0	0	0	0	0	0	0
NWP-100	0	4	4	3	3	2	0	0	0	0	0	0	0	0	0
NWP-200	0	3	3	3	3	3	3	3	3	3	2	2	1	3	3
NWP-300	0	2	3	4	4	4	4	4	4	4	4	4	4	4	4

Note: Large values are better, with the best possible value being 4. The best method of each column in each block is indicated in bold.

Table 12: Model confidence set results for out-of-sample probability forecasts of heat waves for the less extreme pair of thresholds. The values are the number of locations for which each model lies inside the model confidence set for the 90% confidence levels.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TN>15°C														
Uni-TN&TX	0	0	0	0	1	1	1	2	4	4	4	4	4	4
Bivariate-TN&TX	1	0	0	0	0	0	2	2	4	4	4	4	4	4
RAW NWP	1	1	1	1	1	2	3	3	4	3	2	2	3	3
NWP-40	4	3	3	3	3	4	4	4	4	2	2	0	1	1
NWP-100	4	4	4	4	4	4	4	4	4	4	4	4	4	4
NWP-200	3	3	3	4	4	4	4	4	4	4	4	4	4	4
NWP-300	3	3	3	4	4	4	4	4	4	4	4	4	4	4
TX>30°C														
Uni-TN&TX	0	0	0	0	0	0	0	1	1	2	3	4	4	4
Bivariate-TN&TX	0	0	0	0	0	0	0	1	1	2	3	4	4	4
RAW NWP	0	0	1	1	2	2	2	2	2	2	2	2	2	2
NWP-40	4	4	4	4	4	4	4	4	3	3	1	1	1	2
NWP-100	4	4	4	3	4	4	4	4	4	4	4	4	4	4
NWP-200	4	3	4	4	4	4	4	4	4	4	4	4	4	4
NWP-300	4	3	3	4	4	4	4	4	4	4	4	4	4	4
TN>15°C & TX>30°C														
Uni-TN&TX	0	0	0	0	0	0	0	0	1	2	2	3	4	4
Bivariate-TN&TX	0	0	0	0	0	0	0	1	2	2	2	3	4	4
RAW NWP	1	3	3	3	4	4	3	3	3	4	4	4	4	4
NWP-40	4	4	4	4	4	4	3	2	2	3	2	2	2	1
NWP-100	4	4	4	4	4	4	3	3	4	3	3	3	3	3
NWP-200	3	3	4	4	4	4	3	3	3	3	3	3	4	3
NWP-300	3	4	4	4	4	4	4	4	4	3	3	3	4	3

Note: Large values are better, with the best possible value being 4. The best method of each column in each block is indicated in bold.

Table 13: Model confidence set results for out-of-sample probability forecasts of heat waves for the more extreme pair of thresholds. The values are the number of locations for which each model lies inside the model confidence set for the 90% confidence levels.

Lead time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TN>17.5°C														
Uni-TN&TX	0	0	0	0	0	0	2	2	2	2	3	4	3	4
Bivariate-TN&TX	2	0	0	0	0	0	2	2	2	3	3	4	4	4
RAW NWP	0	0	1	2	3	3	3	3	3	3	2	3	3	3
NWP-40	4	4	3	3	3	3	3	4	3	3	2	2	1	2
NWP-100	4	4	4	4	4	4	4	4	4	4	4	4	4	4
NWP-200	4	4	4	4	4	4	4	4	4	4	4	4	3	4
NWP-300	4	3	3	4	4	4	4	4	4	4	4	4	4	4
TX>35°C														
Uni-TN&TX	0	0	0	0	0	1	2	2	3	3	4	4	4	4
Bivariate-TN&TX	0	0	0	0	0	1	2	2	3	3	4	4	4	4
RAW NWP	1	1	3	3	4	4	4	4	4	4	4	4	4	4
NWP-40	4	4	4	4	4	4	4	4	3	2	2	2	1	0
NWP-100	3	3	4	4	4	4	4	4	4	4	4	4	4	4
NWP-200	3	3	4	4	4	4	4	4	4	4	4	4	4	4
NWP-300	4	4	4	4	4	4	4	4	4	4	4	4	4	4
TN>17.5°C & TX>35°C														
Uni-TN&TX	0	0	0	0	0	2	2	2	2	3	4	4	4	4
Bivariate-TN&TX	0	0	0	0	0	1	2	2	2	3	4	4	4	4
RAW NWP	1	3	3	4	4	4	4	4	4	4	4	4	4	4
NWP-40	4	4	4	4	4	4	4	3	3	3	2	2	2	2
NWP-100	2	2	3	4	4	4	4	4	4	4	4	4	4	4
NWP-200	2	2	3	4	4	4	4	4	4	4	4	4	4	4
NWP-300	3	3	3	4	4	4	4	4	4	4	4	4	4	4

Note: Large values are better. The best method of each column in each block is indicated in bold.

References

- Alexandridis, A. and Zapranis, A. D. (2012). *Weather derivatives: modeling and pricing weather-related risk*. Springer Science & Business Media.
- Alexandridis, A. K., Kampouridis, M., and Cramer, S. (2017). A comparison of wavelet networks and genetic programming in the context of temperature derivatives. *International Journal of Forecasting*, 33(1):21–47.
- 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.
- Baran, S., Baran, Á., Pappenberger, F., and Ben Bouallègue, Z. (2020). Statistical post-processing of heat index ensemble forecasts: is there a royal road? *Quarterly Journal of the Royal Meteorological Society*, 146(732):3416–3434.
- Baran, S. and Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34(3):477–496.
- Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.
- Bauwens, L., Laurent, S., and Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, 21(1):79–109.
- 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.
- Carroll, R., Conlon, T., Cotter, J., and Salvador, E. (2017). Asset allocation with correlation: A composite trade-off. *European Journal of Operational Research*, 262(3):1164–1180.
- Cavaliere, G. and Taylor, A. R. (2009). Heteroskedastic time series with a unit root. *Econometric Theory*, pages 1228–1276.
- 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.
- ECMWF Directorate (2012). Describing ECMWF’s forecasts and forecasting system. *ECMWF Newsletter*, 133:11–13.
- 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.
- Feldmann, K., Richardson, D. S., and Gneiting, T. (2019). Grid-versus station-based postprocessing of ensemble temperature forecasts. *Geophysical Research Letters*, 46(13):7744–7751.
- Fildes, R. and Kourentzes, N. (2011). Validation and forecasting accuracy in models of climate change. *International Journal of Forecasting*, 27(4):968–995.
- 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.
- Gneiting, T. (2014). *Calibration of medium-range weather forecasts*. ECMWF Technical Memorandum No.719.
- 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., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.
- Hajat, S., Armstrong, B., Baccini, M., Biggeri, A., Bisanti, L., Russo, A., Paldy, A., Menne, B., and Kosatsky, T. (2006). Impact of high temperatures on mortality: Is there an added heat wave effect? *Epidemiology*, 17(6):632–638.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

- Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41(24):9197–9205.
- 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.
- Kew, S. f., Philip, S. Y., Jan van Oldenborgh, G., van der Schrier, G., Otto, F. E., and Vautard, R. (2019). The exceptional summer heat wave in southern europe 2017. *Bulletin of the American Meteorological Society*, 100(1):S49–S53.
- Le Tertre, A., Lefranc, A., Eilstein, D., Declercq, C., Medina, S., Blanchard, M., Chardon, B., Fabre, P., Filleul, L., Jusot, J.-F., et al. (2006). Impact of the 2003 heatwave on all-cause mortality in 9 French cities. *Epidemiology*, 17(1):75–79.
- 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.
- Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119.
- Pattenden, S., Nikiforov, B., and Armstrong, B. (2003). Mortality and temperature in Sofia and London. *Journal of Epidemiology & Community Health*, 57(8):628–633.
- Perkins, S. and Alexander, L. (2013). On the measurement of heat waves. *Journal of Climate*, 26(13):4500–4517.
- Proietti, T. and Hillebrand, E. (2017). Seasonal changes in central England temperatures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):769–791.
- 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.
- Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., et al. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640.
- 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.
- Taylor, J. W. and Buizza, R. (2006). Density forecasting for weather derivative pricing. *International Journal of Forecasting*, 22(1):29–42.
- Taylor, J. W. and Jeon, J. (2018). Probabilistic forecasting of wave height for offshore wind turbine maintenance. *European Journal of Operational Research*, 267(3):877–890.
- 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.