2022-03-26

A Bayesian Non-linear State Space Copula Model for Air Pollution in Beijing

Kreuzer, A

10.1111/rssc.12548
Journal of the Royal Statistical Society Series C: Applied Statistics
Royal Statistical Society

A Bayesian Non-linear State Space Copula Model for Air Pollution in Beijing

Alexander Kreuzer
Technische Universität München, München, Germany.
E-mail: a.kreuzer@tum.de
Luciana Dalla Valle
University of Plymouth, Plymouth, England.
Claudia Czado
Munich Data Science Institute, Technische Universität München, München, Germany.

Abstract. Air pollution is a serious issue that currently affects many industrial cities in the world and can cause severe illness to the population. In particular, it has been proven that extreme high levels of airborne contaminants have dangerous short-term effects on human health, in terms of increased hospital admissions for cardiovascular and respiratory diseases and increased mortality risk. For these reasons, an accurate estimation of airborne pollutant concentrations is crucial. In this paper, we propose a flexible novel approach to model hourly measurements of fine particulate matter and meteorological data collected in Beijing in 2014. We show that the standard state space model, based on Gaussian assumptions, does not correctly capture the time dynamics of the observations. Therefore, we propose a non-linear non-Gaussian state space model where both the observation and the state equations are defined by copula specifications, and we perform Bayesian inference using the Hamiltonian Monte Carlo method. The proposed copula state space approach is very flexible, since it allows us to separately model the marginal distributions and to accommodate a wide variety of dependence structures in the data dynamics. We show that the proposed approach allows us not only to accurately estimate particulate matter measurements, but also to capture unusual high levels of air pollution, which were not detected by measured effects.

Keywords: Air Pollution; Bayes; Hamiltonian Monte Carlo; State Space Models.

1. Introduction

Over recent decades, rapid economic development and urbanization lead to severe and chronic air pollution in China, which is currently listed as one of the most polluted countries in the world. Airborne pollutants contribute not only to the contamination of the air, but also of food and water, making inhalation and ingestion the major routes of pollutant exposure, in addition to dermal contact, to a minor extent (Kampa and Castanas, 2008). Exposure to ambient air pollution has been associated with a variety of adverse health effects, ranging from cardiovascular and respiratory illnesses, such as stroke and ischemic heart disease, to cancer and even death. Human health effects

include birth defects, serious developmental delays in children, and reduced activity of the immune system, leading to a number of diseases (Liang et al., 2015). It has been shown that air pollution increases mortality and morbidity and shortens life expectancy (World Health Organization, 2013), with heavy consequences in terms of health care and economy (Song et al., 2017). Outdoor PM2.5 has been established as one of the best metrics of air pollution-related risk to public health, since it is considered to be the fraction of air pollution that is most reliably associated with human disease (Liu et al., 2017). In particular, PM2.5 is known to be a better predictor for acute and chronic health effects than other types of particulate matter pollutants (Matus et al., 2012). PM2.5 consists of fine particulate matter with aerodynamic diameters of less than 2.5 micrometers (μm). PM2.5 is a portion of air pollution that is made up of extremely small particles and liquid droplets containing acids, organic chemicals, metals, and soil or dust particles, that are able to travel deeply into the respiratory tract, reaching the lungs. Sources of PM2.5 include combustion in mechanical and industrial processes, vehicle emissions, and tobacco smoke. It has been estimated that in China, ambient PM2.5 was the first-ranking mortality risk factor in 2015 and exposure to this pollutant caused 1.1 million deaths in that year (Cohen et al., 2017).

Fine particulate matter is a key driver of global health and therefore it is vital to accurately model and estimate the exposure to PM2.5, especially in areas of severe and persistent air pollution such as China and its biggest cities like Beijing. A precise estimation of air pollution is crucial for a realistic appraisal of the risks that airborne contaminants pose and for the design and implementation of effective environmental and public health policies to control and limit those risks (Shaddick et al., 2018).

Several contributions in the literature focus on modelling the observed concentrations of ambient air pollution via spatio-temporal models based on Gaussian assumptions. For example, Sahu et al. (2006) modelled fine atmospheric particulate matter data collected in the US using a Bayesian hierarchical spatio-temporal approach. Sahu and Mardia (2005) used a spatio-temporal process based on a Bayesian kriged Kalman filtering to model atmospheric particulate matter in New York City. Calder (2008) adopted a Bayesian dynamic process convolution approach to provide space-time interpolations of PM2.5 and PM10 concentrations readings taken across the state of Ohio.

Another stream of research in the atmospheric science literature is devoted to forecasting pollutant concentrations using deep learning models (Ayturan et al., 2018). Feng et al. (2015) used PM2.5 and meteorological variables as input nodes to a multi-layer perceptron (MLP) back-propagation artificial neural network (ANN). The authors used air mass trajectory analysis to identify air corridors to different air pollution monitoring stations in China, and they applied wavelets decompositions to pollutant predictors to improve the ANN forecast accuracy. Li et al. (2016) proposed a spatio-temporal deep learning (STDL)-based air quality prediction method using data from 12 air quality monitoring stations in Beijing. Other authors adopted recurrent neural networks (RNNs) with long short-term memory (LSTM) for forecasting air pollution concentrations. For example, Bui et al. (2018) applied LSTM to predict urban air quality in South Korea. Liu et al. (2020) combined LSTM-based wind-sensitive attention mechanisms with an LSTM neural network for predicting PM2.5 concentrations. Other applications of deep models include the paper by Rangapuram et al. (2018), which presents a forecasting

method that parametrizes a particular linear Gaussian state space model using an RNN. For multivariate time series forecasting, Salinas et al. (2019) combine an RNN-based time series model with a Gaussian copula process, relying on a low-rank approximation of the covariance structure.

In this paper, we propose a novel flexible non-linear non-Gaussian state space model based on copulas, that includes a dynamic latent smoothing effect. As opposed to traditional methods, our approach, in addition to model the observed concentrations of air pollution, allows us to obtain an estimate of underlying non-measured factors and to identify time-points where the latent states have a considerable impact on the response, which are critical to assess pollution-related health risks. These points correspond to unusual high levels of air pollution, which cannot be accommodated for simply by the model including covariate effects such as weather conditions and seasonal patterns. These can have dangerous effects on human health. Extreme air pollution levels need to be carefully monitored, since it is proven that acute exposures increase the rate of cardiovascular, respiratory and mortality events (Anderson et al., 2012). Recent studies in various countries confirm the severity of short- and long-term effects of the exposure to increased levels of airborne contaminants on human health, including respiratory diseases, decreased lung functions, recurrent health care utilization, reduced life expectancy and increased mortality. Vulnerable people, such as infants and elderly, are particularly susceptible to extreme air pollution levels. In particular, children who are exposed to an excess level of PM2.5 are under a significantly high risk of hospitalization for respiratory symptoms, asthma medication use, and reduced lung function, while PM2.5 pollution is linked to an increased risk of hospital admission for heart failure among the elderly. In addition, air pollution has a substantial economic impact, since it multiplies the world wide healthcare burden (Anderson et al., 2012; Kan et al., 2012; Kim et al., 2015). We will show that our methodology performs better than traditional approaches in accurately modelling unusual high levels of air pollutants, allowing us to better assess the effects of human exposure to airborne contaminants.

1.1. Linear Gaussian state space models

State space models are dynamic statistical analysis techniques, which assume that the state of a system at time t can only be observed indirectly through observed time series data (Durbin and Koopman, 2000). State space models contain two classes of variables, the unobserved state variables, which describe the development over time of the underlying system, and the observed variables (Durbin and Koopman, 2002). Let's consider a specific univariate linear Gaussian state space model, which is a first order unobserved component model, with continuous states and discrete time points $t = 1, \ldots, T$

$$Z_t = \rho_t^O W_t + \sigma_t^O \eta_t^O \tag{1}$$
$$W_t = \rho_t^L W_{t-1} + \sigma_t^L \eta_t^L. \tag{2}$$

Here, $(Z_t)_{t=1,\ldots,T}$ is a random vector corresponding to the observations, $(W_t)_{t=1,\ldots,T}$ is an unobserved state vector and η_t^O and η_t^L are independent disturbances, with $\eta_t^O \sim \mathrm{N}(0,1)$ and $\eta_t^L \sim \mathrm{N}(0,1)$ for $t = 1, \ldots, T$. Further, it holds that $\rho_t^O \in (-1,1)$, $\rho_t^L \in$

$(-1, 1)$, $\sigma_t^O \in (0, \infty)$ and $\sigma_t^L \in (0, \infty)$. It is also assumed that $W_0 \sim \mathrm{N}(\mu_0^L, (\sigma_0^L)^2)$ is independent of ρ_t^L and ρ_t^L for all t, where μ_0^L and σ_0^L are generally known. Equation (1) is commonly referred to as the observation equation and it describes how the observed series depends on the unobserved state variables W_t and on the disturbances η_t^O. Equation (2) is referred to as the state equation and it describes how these state variables evolve over time (Van den Brakel and Roels, 2010).

Typically, Kalman filter recursions are used for determining the optimal estimates of the state vector W_t given information available at time t (Durbin and Koopman, 2012). Other methods, such as Empirical Bayes, were proposed by Koopman and Mesters (2017) to efficiently estimate dynamic factor models defined by latent stochastic processes, adopting a shrinkage-based approach. Ippoliti et al. (2012) used a linear Gaussian state space model to produce predictions of airborne pollutants in Italy and in Mexico.

1.2. Beijing ambient air pollution data

In this paper, we aim at accurately estimating the concentration of airborne particulate matter using a flexible state space model. We consider a data set of hourly PM2.5 readings (μg/m^3) and meteorological measurements, such as dew point (DEWP, degrees Celsius), temperature (TEMP, degrees Celsius), pressure (PRES, hPa), wind direction (CBWD, taking values: northwest (NW), northeast (NE), southeast (SE) and calm and variable (CV)), cumulated wind speed (IWS, m/s) and precipitations (PREC), collected in 2014 in Beijing. The data set refers to a single location, with PM2.5 measurements collected at the US Embassy in Beijing, and meteorological data collected at the Beijing Capital International Airport. The data set used in this paper is part of a larger data set collected in Beijing during a 5-year time period, from January 1st, 2010 to December 31st, 2014, for a total of 43,824 observations. The data are available at `https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data` (Liang et al., 2015). We split the data into 12 monthly sub-sets, since this allows us to adjust the model over time periods. In order to consider the effects of meteorological conditions on airborne contaminants concentrations, we assume a generalized additive model (GAM) (Hastie and Tibshirani, 1986). However, the choice of using a GAM is arbitrary, since it might be replaced by any regression-type model able to remove the covariate effects. More precisely, we suppose that, for each month, the relationship between the logarithm of PM2.5 concentrations Y_t and covariates $\boldsymbol{x}_t$ for each hourly data point $t = 1, \ldots, T$ (where T is the total number of monthly observations) is described by a GAM, such that

$$Y_t = f(\boldsymbol{x}_t) + \sigma\varepsilon_t, \tag{3}$$

where $\boldsymbol{x}_t$ contains the meteorological covariates and seasonal covariates capturing within-day and -week patterns. Further, $f(\cdot)$ is a smooth function of the covariates, expressing the mean of the GAM, and $\varepsilon_t \overset{\text{iid}}{\sim} N(0, 1)$. As suggested by many authors in the air pollution literature, we are going to investigate the presence of structures in the residual processes (Laird and Ware, 1982; Burnett and Krewski, 1994; Lee et al., 2014). For estimation we make use of the two step approach which is commonly used for copula models: we first estimate the GAM, fix the GAM parameters at point estimates, and then estimate the copula model to capture temporal effects. (Joe and Xu, 1996). For

$t = 1, \ldots, T$, we define the standardized errors Z_t as

$$Z_t = \frac{Y_t - f(\boldsymbol{x}_t)}{\sigma}. \tag{4}$$

This step allows us to account for weather and seasonal patterns. High values of Z_t are of interest to detect unusual high levels of pollution so far not accounted for. Using the estimates $\hat{f}$ and $\hat{\sigma}$ of the GAM, we obtain approximately standard normal data $\hat{z}_t$ from the (4). Empirical autocorrelation functions of $(\hat{z}_t)_{t=1,\ldots,T}$ for each month show dependence among succeeding observations (supplement, Figure 1). Thus the independence assumption for the errors ε_t of the GAM in (3) seems to be inappropriate. We also estimated a GAM with heteroscedastic errors for each of the 12 months, allowing for a time varying variance. However, the results showed only very minor differences with a GAM model with homoscedastic errors. We employ a state space model, as specified in (1) and (2), to allow for time effects in the GAM. Here ρ_t^O and ρ_t^L will be estimated from the data. Further, we assume that they do not depend on time, i.e. we set $\rho_t^O = \rho_O$ and $\rho_t^L = \rho_L$. In our data application we split the data into monthly periods to make this assumption more plausible.

We now consider a state space model for Z_t, which is standardized by a GAM. Under our assumptions we have $\sigma_t^O = \sqrt{Var(Z_t|W_t)} = \sqrt{1-\rho_O^2}$ and $\sigma_t^L = \sqrt{Var(W_t|W_{t-1})} = \sqrt{1-\rho_L^2}$ with $\rho_O, \rho_L \in (-1, 1)$. For the initial conditions we assume $\mu_0^L = 0$ and $\sigma_0^L = 1$. With these assumptions the state space model in (1) and (2) becomes

$$\begin{aligned} Z_t &= \rho_O W_t + \sqrt{1-\rho_O^2}\eta_t^O \\ W_t &= \rho_L W_{t-1} + \sqrt{1-\rho_L^2}\eta_t^L \end{aligned} \tag{5}$$

with $\eta_t^O, \eta_t^L \overset{\text{iid}}{\sim} N(0,1)$ and $W_0 \sim N(0,1)$. Note that representation (5) induces the following bivariate normal distributions

$$\begin{pmatrix} Z_t \\ W_t \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_O \\ \rho_O & 1 \end{pmatrix}\right) \quad \text{and} \quad \begin{pmatrix} W_t \\ W_{t-1} \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_L \\ \rho_L & 1 \end{pmatrix}\right).$$

In order to assess the suitability of the linear Gaussian state space model to the Beijing air pollution data, we display in Figure 1 the bivariate normalized contour plots of the pairs $(\hat{z}_t, \hat{z}_{t-1})_{t=2,\ldots,T}$ for each month, to visualize the dependence structure between two successive time points in the series. Using (5) we see that Z_t can be written as a linear function of Z_{t-1} and independent normally distributed disturbances. Since Z_1 is normally distributed, it follows that (Z_t, Z_{t-1}) are jointly normal. In particular, we have

$$Z_t \sim N(0,1) \qquad \text{and} \qquad Cov(Z_t, Z_{t-1}) = \rho_O^2 \rho_L \qquad \forall t \geq 1. \tag{6}$$

However, Figure 1 reveals that the normalized contour plots of the Beijing monthly data deviate from the elliptical shape of a Gaussian dependence structure (which, to aid comparisons, is depicted in the top left panel of Figure 3 in the supplement). For

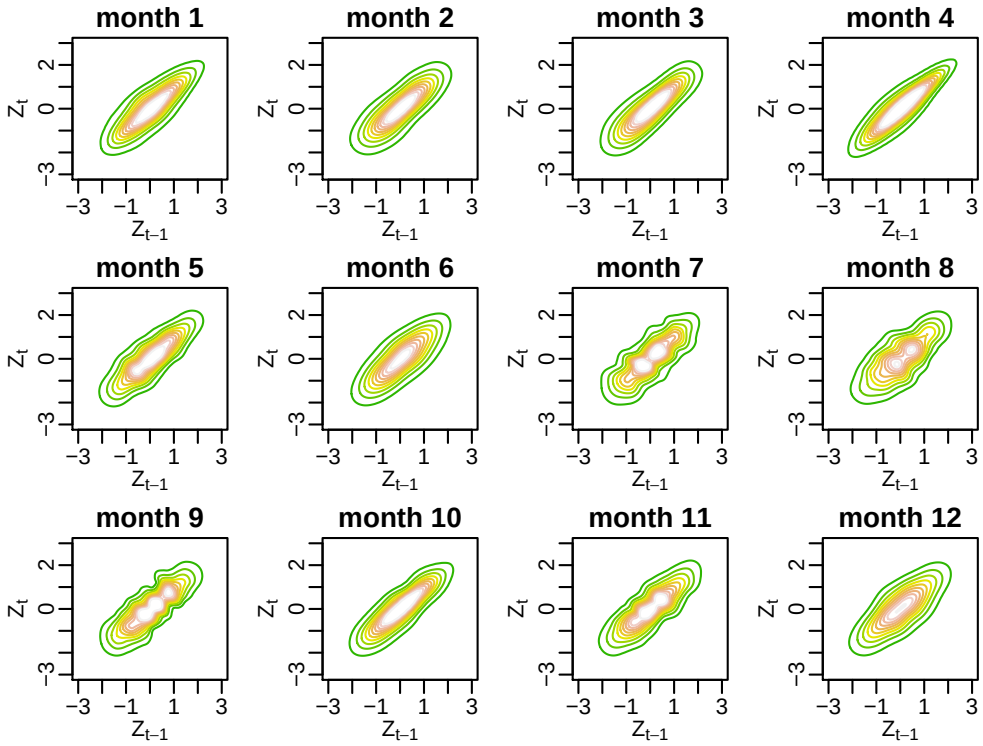

Figure 1. Normalized contour plots of pairs $(\hat{z}_t, \hat{z}_{t-1})_{t=2,\ldots,T}$ ignoring serial dependence for each of the 12 Beijing air pollution monthly data sets.

example, the normalized contour plots for January and October (months 1 and 10) show tail dependence and/or asymmetry in the tails, which cannot be modeled with a Gaussian distribution. This suggests that the linear Gaussian state space model is too restrictive for the Beijing air pollution data and a more flexible approach needs to be adopted.

1.3. Our proposal

In the literature, extensions of the linear Gaussian state space model, relaxing the assumptions of linearity and normality, have been studied, for example, by Johns and Shumway (2005). They adopted a non-linear and non-Gaussian state space formulation to model airborne particulate matter, yet relying on the Normal distribution to describe the errors in the state and observation equations. Chen et al. (2012) implemented a non-linear state space model to predict the global burden of infectious diseases using the extended Kalman filter approach. Non-linear state and observation equations of this model were derived from differential equations, however the authors still used Gaussian noise terms in the observation and state equations.

We propose a very flexible Bayesian non-linear and non-Gaussian state space model, where both the observation and the state equations are described by copulas. Copu-

las are flexible mathematical tools, which allow us to model separately the marginals from the dependence structure, and the use of different copula families are suitable to accommodate various types of dependences. More formally, a d-dimensional copula is a multivariate distribution function on the d-dimensional hyper cube $[0,1]^d$ with uniformly distributed marginals. A thorough overview about copulas is provided in Joe (2014) and Nelsen (2007). We point out that our approach is different from the one introduced by Smith and Maneesoonthorn (2018), who proposed the construction of copulas through the inversion of nonlinear state space models. First, we find an equivalent formulation of the Gaussian state space model in (5) in terms of copulas. The representation is given by

$$\begin{aligned}(U_t\,,\,V_t) &\sim \mathbb{C}^N_{U,V}(\,\cdot\,,\,\cdot;\,\tau_O)\\ (V_t\,,\,V_{t-1}) &\sim \mathbb{C}^N_{V_2,V_1}(\,\cdot\,,\,\cdot;\,\tau_L),\end{aligned} \tag{7}$$

where

$$U_t = \Phi\left(Z_t\right), V_t = \Phi\left(W_t\right), \tag{8}$$

with Φ denoting the standard normal cumulative distribution function. The variables U_t and V_t are marginally uniformly distributed on $(0,1)$ and Z_t and W_t are standard normal. Here the Gaussian copulas $\mathbb{C}^N_{U,V}$ and $\mathbb{C}^N_{V_2,V_1}$ are parametrized by Kendall's τ, obtained as $\tau_O = \frac{2}{\pi}\arcsin(\rho_O)$ and $\tau_L = \frac{2}{\pi}\arcsin(\rho_L)$. Corresponding approximately uniform pseudo-copula data, that can be used for estimating the model in (7), are obtained as

$$\hat{u}_t = \Phi\left(\hat{z}_t\right). \tag{9}$$

By reformulating the state space representation in (5) in terms of copulas in (7), it is straightforward to see how we can generalize the Gaussian linear state space model by replacing the Gaussian copulas in (7) with arbitrary bivariate copulas. Typical restrictions of the Gaussian copula, such as symmetric tails, can be circumvented. For example, a Gumbel copula would allow for asymmetric tails. Koopman et al. (2016) incorporated the symmetric-tailed Gaussian and Student t copulas in non-linear non-Gaussian state space models; however, asymmetric tail dependence could not be captured, since the authors ignored non-symmetric copula families and restricted their attention solely to autoregressive state equations. The proposed copula-based state space model allows us to specify various dependence structures to model the relationships between the observations and the underlying states, and to describe the states evolution over time. We will show that our methodology is able to accurately model the levels of PM2.5 in Beijing.

The remainder of the paper is organized as follows. Section 2 introduces a copula-based state space model, Section 3 illustrates Bayesian inference for the proposed approach and Section 4 is devoted to the application of the copula state space model to the Beijing pollution data. Concluding remarks are given in Section 5.

2. The copula state space model

The copula state space model extends the linear Gaussian state space approach, allowing copula specifications in place of normal distributions as in the observation equation as

well as in the state equation. In particular, we assume that the dynamic behaviour of the residuals $Z_t := \Phi^{-1}(U_t)$ for the GAM model introduced in equation (3), with $Z_t \sim N(0,1)$ and $U_t \sim \mathrm{U}(0,1)$ defined as in (8), depends on the latent variable $W_t := \Phi^{-1}(V_t)$, with $W_t \sim N(0,1)$ and $V_t \sim \mathrm{U}(0,1)$, according to a bivariate copula distribution given in the observation equation. The evolution of the latent variable W_t over time is also described by a bivariate copula distribution, which defines the state equation of the model. Our approach is more flexible than traditional copula-based time series models, since it captures covariate effects using a semiparametric regression, and then constructs a state space model on the standardized residuals, with copulas modelling the observation as well as the state equation. Other approaches use copulas to model the observed time-series directly as Markov processes, however they do not consider latent variables to capture the series temporal dynamics. The Markov process approach has been first followed by Chen and Fan (2006) and then extended to financial time series models by Patton (2009). In Smith et al. (2010) higher order Markov dependence is allowed and modelled in a D-vine copula framework. The copula distributions defining the observation and state equations of the proposed state space approach do not necessarily belong to the same family, allowing maximum flexibility in the specification of the model. However, we restrict our model to bivariate copula families with a single parameter. This gives still a flexible class of copula families, including e.g. Gaussian, Gumbel, Clayton or Frank copulas. The Student t copula can also be included if we fix the degrees of freedom parameter. An overview of different bivariate copula families can be found in Joe (2014), Chapter 4. Further, we are able to express the copula dependence parameters in the observation and state equations in terms of Kendall's τ. This is convenient for comparison of the dependence strength, since the parameter space of distinct copula families may be different. More formally, we assume that the joint distributions for the uniformly transformed variables U_t and V_t, with $t = 1, \ldots, T$, are described by copulas, similarly to the (7); however, the Gaussian copula in the observation equation is replaced by $\mathbb{C}^{O}_{U,V}(\cdot\,,\cdot;\,\tau_O)$ and the Gaussian copula in the states equation is replaced by $\mathbb{C}^{L}_{V_2,V_1}(\cdot\,,\cdot;\,\tau_L)$, where $\tau_O = g(\theta_O)$ is the Kendall's τ of the copula of the observations and $\tau_L = g(\theta_L)$ is the Kendall's τ of the copula of the states (latent variables), respectively. The function g is an appropriate one-to-one transformation function, and θ_O and θ_L are the parameters of the bivariate copulas $\mathbb{C}^{O}_{U,V}$ and $\mathbb{C}^{L}_{V_2,V_1}$, respectively. For the specification of g for some one-parameter copula families see Joe (2014), Chapter 4.

The copula state space model is defined on the uniform scale as follows

$$U_t \,|\, V_t = v_t \;\sim\; \mathbb{C}^{O}_{U|V}(\cdot \,|\, v_t;\, \tau_O) \tag{10}$$

$$V_t \,|\, V_{t-1} = v_{t-1} \;\sim\; \mathbb{C}^{L}_{V_2|V_1}(\cdot \,|\, v_{t-1};\, \tau_L) \tag{11}$$

where (10) is the observation equation and (11) is the state equation. We assume, as in the linear Gaussian state space model, that U_t is independent of U_{t-1} given the latent state V_t. The copula state space model introduced in equations (10) and (11) can be visualized as in Figure 2.

We now derive the joint distributions of $(Z_t\,, W_t) \sim F_{Z_t,W_t}$ and $(W_t\,, W_{t-1}) \sim F_{W_t,W_{t-1}}$. By Sklar's theorem (Sklar, 1959), we have that

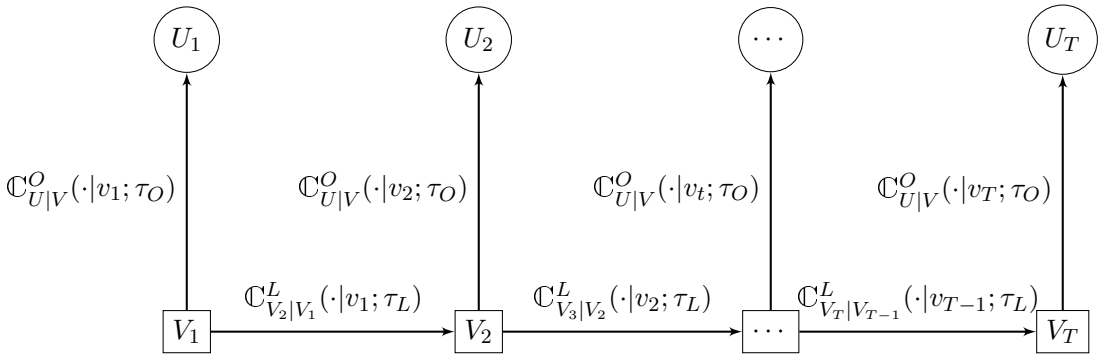

Figure 2. Graphical visualization of the copula state space model.

$$F_{Z_t,W_t}(z_t,w_t) = \mathbb{C}^O_{U,V}(\,\Phi(z_t)\,,\,\Phi(w_t);\,\tau_O) = \mathbb{C}^O_{U,V}(\,u_t\,,\,v_t;\,\tau_O).$$

Hence,

$$\begin{aligned} F_{Z_t|W_t=w_t}(z_t|w_t) &= \frac{\partial}{\partial v_t}\mathbb{C}^O_{U,V}(\,\Phi(z_t)\,,\,v_t;\,\tau_O)\Big|_{v_t=\Phi(w_t)} = \mathbb{C}^O_{U|V}(\,u_t\,|\,v_t;\,\tau_O)\Big|_{u_t=\Phi(z_t),v_t=\Phi(w_t)} \\ &= \mathbb{C}^O_{U|V}(\,\Phi(z_t)\,|\,\Phi(w_t);\,\tau_O). \end{aligned}$$

Similarly,

$$F_{W_t|W_{t-1}=w_{t-1}}(w_t|w_{t-1}) = \mathbb{C}^L_{V_2|V_1}(\,\Phi(w_t)\,|\,\Phi(w_{t-1});\,\tau_L).$$

Therefore, the model can also be expressed on the normalized scale as follows

$$Z_t\,|\,W_t = w_t \;\sim\; \mathbb{C}^O_{U|V}(\,\Phi(z_t)\,|\,\Phi(w_t);\,\tau_O) \tag{12}$$

$$W_t\,|\,W_{t-1} = w_{t-1} \;\sim\; \mathbb{C}^L_{V_2|V_1}(\,\Phi(w_t)\,|\,\Phi(w_{t-1});\,\tau_L), \tag{13}$$

where (12) is the observation equation and (13) is the state equation. Contour plots of (Z_t, Z_{t-1}) of this model for different choices of bivariate copulas are shown in Figure 3, illustrating different shapes that the model can deal with.

The copula state space model has the advantage of allowing flexibility in the specification of the observation and state equations, and thus is able to accommodate a wide variety of dependence structures in the air pollution data dynamics. In the standard GAM the errors are assumed to be independent. Our methodology allows us to account for autoregressive effects in the error through the underlying latent variable σW_t, as defined on the original scale of the GAM residuals, or via the proxy V_t, on the uniform scale. These latent variables can be interpreted as non-measured nonlinear autoregressive effects. As we will see in Section 4.3, our model's flexibility allows us to detect extreme air pollution levels, where the response is more susceptible to the effect of the

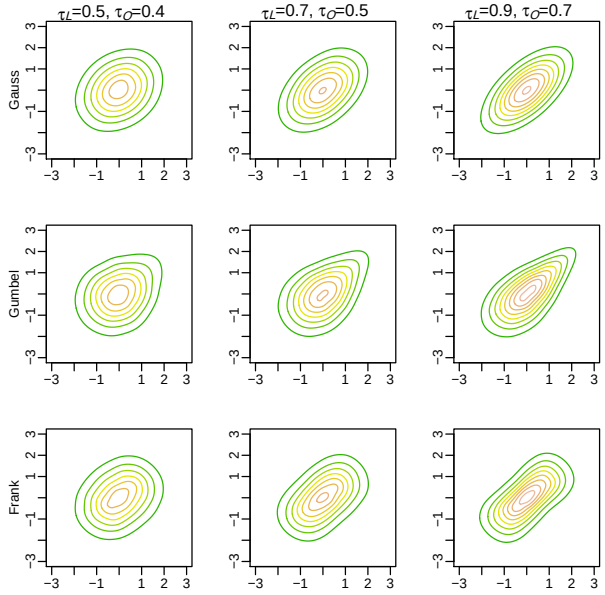

Figure 3. Normalized contour plots for (Z_t, Z_{t-1}) of the copula state space model for different bivariate copulas. In the state and observation equation we choose the same copula family.

underlying latent variable. Capturing unusual air contaminant levels is very important, since human exposure to pollution spikes have a substantial impact on general health, causing severe cardiovascular and respiratory illness, and increasing mortality.

First we study the normalized copula state space model (12) and (13), when the copulas $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$ are Gaussian copulas with parameters ρ_O and ρ_L. It is straight forward to see that we can identify (12) and (13) as the Gaussian state space model (1) and (2) with time constant parameters given by $\rho^O_t = \rho_O, \sigma^O_t = \sqrt{1-\rho^2_O}$ and $\rho^L_t = \rho_L, \sigma^L_t = \sqrt{1-\rho^2_L}$, respectively. In particular the associated state equation is a stationary Gaussian AR(1) process, whose variance and covariance expressions can be used to determine that the joint distribution of the observation vector $(Z_1, \ldots, Z_T)$ for any integer value $T > 0$ is jointly normal with zero mean vector, marginal unit variances and autocorrelations given by

$$\text{cor}(Z_t, Z_{t+s}) = \rho^2_O \rho^s_L \tag{14}$$

for $t = 1, ...T$ and $s + t \in \{1, ..., T\}$. For $s = 1$ we recover the result stated in (6).

Now we consider the general copula state space case described by $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$. The corresponding joint density of the observations $\boldsymbol{u}_T = (u_1, \ldots, u_T)$ and states $\boldsymbol{v}_T = (v_1, \ldots, v_T)$ on the copula scale is given by

$$c(\boldsymbol{u}_T, \boldsymbol{v}_T) = \prod_{t=1}^{T} c^O_{U,V}(u_t, v_t; \tau_O) \prod_{t=1}^{T-1} c^L_{V_2,V_1}(v_t, v_{t+1}; \tau_L). \tag{15}$$

Integration over the latent states will give the joint distribution of the observations $\boldsymbol{u}_T$. Since this would require a T dimensional integration it is numerically intractable. However the bivariate density of (U_1, U_2) is tractable. In particular we have

$$c(u_1, u_2) = \int_0^1 [\int_0^1 c^O_{U,V}(u_1, v_1; \tau_O) c^L_{V_2,V_1}(v_1, v_2; \tau_L) dv_1] c^O_{U,V}(u_2, v_2; \tau_O) dv_2. \tag{16}$$

The expression in the square bracket is a bivariate copula density for (U_1, V_2) (denoted by c_{U_1,V_2}) associated with a three dimensional C-vine on the nodes $\{U_1, V_1, V_2\}$ with pair copula $\mathbb{C}^O_{U,V}$ for (U_1, V_1), $\mathbb{C}^L_{V_2,V_1}$ for (V_1, V_2) and the independence copula for the conditional copula of (U_1, V_2) given V_1. For an elementary introduction to vine copulas see for example Czado (2019), where the three dimensional case is covered in Section 4.1. Using this bivariate copula density c_{U_1,V_2} as the pair copula density for (U_1, V_2) in a three dimensional C-vine (U_1, V_2, U_2), the pair copula density $c^O_{U,V}$ for (V_2, U_2) together with the independence copula for the conditional copula for (U_1, U_2) given V_2, then we can identify $c(u_1, u_2)$ in (16) as the bivariate copula density arising from the so specified three dimensional C-vine for (U_1, V_2, U_2). Using the same approach we can show that the marginal density of $(U_{1+\ell}, U_{2+\ell})$ for $\ell \geq 1$ coincides with the marginal density of (U_1, U_2), i.e. $c(u_{1+\ell}, u_{2+\ell} = c(u_1, u_2)$ holds for all $\ell = 1, \ldots, T$.

Using recursively three dimensional C-vines with a conditional independence pair copula we can also determine the density of (U_1, U_3). In particular we first use the three dimensional C-vine (V_2, V_3, U_3), then (V_1, V_2, U_3) and finally (U_1, V_1, U_3). The pair copula associated with (U_3, V_2) in the second C-vine (V_1, V_2, U_3) is set to the bivariate marginal copula using integration from the first C-vine (V_2, V_3, U_3). Finally the pair copula for (V_1, U_3) in the third C-vine (U_1, V_1, U_3) is set to by the integrated bivariate marginal copula of the second C-vine. This recursive procedure can then be extended to determine the density of (U_1, U_k) for arbitrary $k = 2, \ldots, T$. It is also easy to see that $c(u_{1+\ell}, u_{k+\ell}) = c(u_1, u_k)$ holds for all $\ell = 1, \ldots, T$ and $k = 1, \ldots, T - \ell$. In this sense the copula state space model is stationary.

If we use the bivariate Farlie-Gumbel-Morgenstein copula for $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$ in the copula state space model, we can perform the required integration analytically. In particular consider the Farlie-Gumbel-Morgen copula density given as $c_{FGM}(u_1, u_2; \theta) = \theta(2u_1 - 1)(2U_2 - 1) + 1$ with parameter $\theta \in [-1, 1]$. Then we have

$$\int_0^1 c_{FGM}(u_1, t; \theta_1) c_{FGM}(t, u_2; \theta_2) dt = c_{FGM}(u_1, u_2; \frac{\theta_1 \theta_2}{3})$$

as shown for example in Stoica (2013). However this copula is not interesting in practice, because it allows only for low dependence. In particular the associated Kendall's τ is $2\theta/9$. For the general case it follows that the Kendall's τ associated with $c(u_1, u_t)$ is given by

$$\tau^{FGM}_{1,t} = \frac{2\theta_L \theta_O^t}{3^{t+2}}. \tag{17}$$

2.1. Identifiability constraints

We notice some identifiability issues related to the model. In particular, if $\tau_O = 1$, the observed and latent variables are equivalent and hence the state equation becomes unnecessary. In addition, if $\tau_L = 0$, then the latent variables $(V_t)_{t=1,\dots,T}$ at different time points become uncorrelated. Therefore, we need to set identifiability constraints for the copula state space model by establishing a relationship between τ_O and τ_L. In order to do that, we notice that the dependence between two successive time points U_{t-1} and U_t is determined by both τ_L and τ_O. The form of the correlation between $Z_{t-1} = \Phi^{-1}(U_{t-1})$ and $Z_t = \Phi^{-1}(U_t)$ can be derived exactly when $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$ are both Gaussian copulas. Since in the Gaussian case the parameter of the observation equation copula is the correlation coefficient ρ_O and the parameter of the state equation copula is the correlation coefficient ρ_L, then the correlation between Z_{t-1} and Z_t is $\mathrm{cor}(Z_t, Z_{t+1}) = \rho_O^2\rho_L$ and more generally (14) holds. For larger values of s and larger values of ρ_L we see that $\mathrm{cor}(Z_t, Z_{t+s})$ is close to zero, so we observe an approximate nonidentifiability in that case. Similar remarks can be made in the case of the Farlie-Gumbel-Morgenstein copula based model using Equation (17). The higher the value of ρ_L the smoother the latent states are. Higher smoothness of the latent states induces a lower prediction uncertainty for the latent states. To guarantee a certain degree of smoothness, we need to set ρ_L greater than some specific value and therefore impose $\rho_O \leq \rho_L$ in our approach. In particular, we assume the identifiability constraint in the Gaussian case

$$\rho_O = \rho_L^c \quad \text{for some suitable value} \quad c \geq 1.$$

In this case, the correlation between Z_{t-1} and Z_t becomes $\mathrm{cor}(Z_{t-1}, Z_t) = \rho_L^{2c+1}$. Transforming the correlation coefficients into Kendall's τ, in the Gaussian case, we obtain the following relationships

$$\tau_O = \frac{2}{\pi}\arcsin(\rho_L^c) \qquad \text{and} \qquad \tau_L = \frac{2}{\pi}\arcsin(\rho_L),$$

hence, τ_O is a function of τ_L and c.

Figure 4 visualizes the relationship between the parameter τ_O (on the y-axis) plotted against τ_L (on the x-axis) in the Gaussian case for different values of $c = 1, 3, 6, 10$. Considering that the strength of dependence between U_{t-1} and U_t is increasing in τ_L and in τ_O, Figure 4 shows that the higher the value of c the higher τ_L needs to be to achieve a fixed strength of dependence between U_{t-1} and U_t. Therefore, for higher values of c we expect to obtain a smoother behaviour of the latent states $(V_t)_{t=1,\dots,T}$. We propose to use a similar relationship between τ_L, τ_O and c, not only in the Gaussian case, but also for arbitrary bivariate copula families. Therefore, in general, we impose the following identifiability constraint on the copula parameter for all bivariate copula families with a single parameter identified uniquely by Kendall's τ as follows

$$\sin\left(\frac{\pi}{2}\tau_O\right) = \left(\sin\left(\frac{\pi}{2}\tau_L\right)\right)^c \qquad \text{for some suitable value } c \geq 1. \tag{18}$$

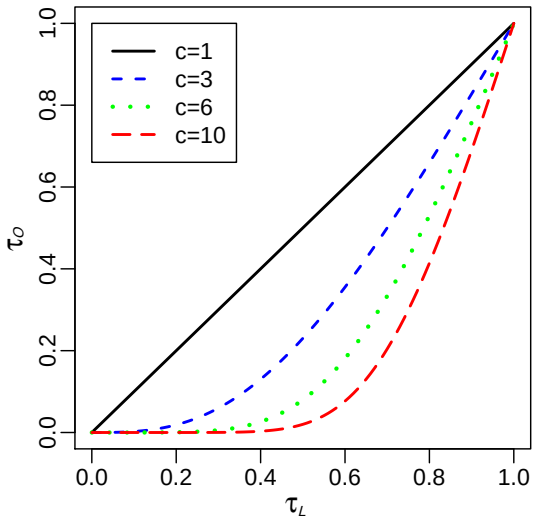

Figure 4. Graphical representation of the relationship between the parameter τ_O plotted against τ_L in the Gaussian case for different values of $c = 1, 3, 6, 10$.

3. Bayesian analysis of the copula state space model

3.1. Hamiltonian Monte Carlo

The copula state space model is a highly non-linear and non-Gaussian model, which provides great flexibility by allowing for different bivariate copulas. The downside of this flexibility is that inference for this model is not straight forward, e.g. it is not possible to implement a Gibbs sampler, where we can directly sample from the corresponding full conditionals. For inference for the copula state space model we rely on the No-U-Turn sampler of Hoffman and Gelman (2014) implemented within the STAN framework (Carpenter et al., 2016). The No-U-Turn sampler extends Hamiltonian Monte Carlo (HMC) and adaptively selects tuning parameters. HMC can be considered as a Metropolis Hastings algorithm, where new states are efficiently obtained by using information on the gradient of the log posterior density. The gradient is obtained through automatic differentiation (Carpenter et al., 2015) in STAN. The HMC sampler has shown good performance in several other cases (Hajian, 2007; Pakman and Paninski, 2014; Hartmann and Ehlers, 2017). We provide a short introduction to HMC in the supplement (Section D) and refer to Neal et al. (2011) or Betancourt (2017) for more details.

An alternative Bayesian approach for jointly estimating parameters and states in non-linear non-Gaussian state space models is presented by Barra et al. (2017), who designed flexible proposal densities for the independent Metropolis-Hasting and the importance sampling algorithms.

3.2. Posterior inference

As prior distribution for τ_L we use a uniform prior on (0,1), which is a non-informative prior restricted to positive dependence, since we do not expect negative dependence in

our application. With this prior choice we obtain a fully specified Bayesian model with posterior density

$$\pi(\tau_L, v_1, \ldots, v_T | \hat{u}_1, \ldots, \hat{u}_T) = \prod_{t=1}^{T} c_{U,V}(\hat{u}_t, v_t; \tau_O) \prod_{t=2}^{T} c_{V_2,V_1}(v_t, v_{t-1}; \tau_L),$$

where τ_O is a function of τ_L as given in (18). Note that for the Bayesian approach the latent variables of the state equation are considered as parameters. We run the No-U-Turn sampler to sample from this posterior density. For a chosen c we obtain a posterior sample for τ_L

$$\tau_L^r(c), \quad r = 1, \ldots, R$$

and, similarly, for τ_O, using the relationship in (18),

$$\tau_O^r(c), \quad r = 1, \ldots, R$$

where R is the total number of HMC iterations. Additionally, posterior samples for the latent variables V_t, for $t = 1, \ldots, T$, are denoted by

$$v_t^r(c), \quad t = 1, \ldots, T \quad \text{and} \quad r = 1, \ldots, R.$$

3.3. Predictive simulation

An advantage of the Bayesian approach is that our model already specifies the predictive distribution, which is the distribution of the response for new data points conditional on observed data points. From this distribution uncertainty is easy to be quantified through credible intervals.

We consider the set $\{\tau_L^r(c),\ v_t^r(c),\ r = 1, \ldots, R,\ t = 1, \ldots, T\}$, a posterior sample of the model parameters. Simulations for a new value at time $t \in \{1, \ldots, T\}$ on the copula scale can be obtained by

- simulate $u_t^r(c)$ from $\mathbb{C}_{U|V}^{O}\left(\cdot | v_t^r(c); \tau_O^r(c)\right)$.

We refer to the corresponding distribution as the in-sample predictive distribution on the copula scale. The out-of-sample predictive distribution refers to new values at time $t > T$. Simulated values from the one-day-ahead predictive distribution of U_{T+1} given U_T, can be obtained as follows

- simulate $v_{T+1}^r(c)$ from $\mathbb{C}_{V_2|V_1}^{L}\left(\cdot | v_T^r(c); \tau_L^r(c)\right)$,
- simulate $u_{T+1}^r(c)$ from $\mathbb{C}_{U|V}^{O}\left(\cdot | v_{T+1}^r(c); \tau_O^r(c)\right)$.

In general, simulations from the i-days-ahead out-of-sample predictive distribution on the copula scale can be obtained recursively through:

- simulate $v_{T+i}^r(c) \sim \mathbb{C}_{V_2|V_1}^L \left(\cdot|v_{T+i-1}^r(c); \tau_L^r(c)\right)$,
- simulate $u_{T+i}^r(c) \sim \mathbb{C}_{U|V}^O \left(\cdot|v_{T+i}^r(c); \tau_O^r(c)\right)$.

Based on a simulation of the (in-sample or out-of-sample) predictive distribution on the copula scale $u_t^r(c)$, we further define

$$\varepsilon_t^r(c) := \Phi^{-1}\left(u_t^r(c)\right)$$

as a sample of the predictive distribution of the error of the GAM model specified in (3). In particular we estimate $E(Y_t)$ by $\hat{f}(\boldsymbol{x}_t)$ with estimated error variance $\hat{\sigma}^2$. So,

$$y_t^r(c) := \hat{f}(\boldsymbol{x}_t) + \hat{\sigma}\varepsilon_t^r(c)$$

gives a sample of the predictive distribution of the response. Note that to obtain this predictive sample we ignore the uncertainty in the marginal distribution.

4. Data analysis

Recall the hourly data set discussed in Section 1.2 divided into 12 sub data sets, one data set for each month.

4.1. Marginal models

For each of the 12 data sets we fit a GAM using the R package `mgcv` of Wood and Wood (2015), where the response is the logarithm of PM2.5 and the covariates are DEWP, TEMP, PRES, IWS, PREC and CBWD, as described in Section 1.2. We define an additional covariate PREC_ind, which indicates if there is precipitation, i.e. PREC_ind $= \mathbb{1}_{\text{PREC}>0}$. We also use the hour denoted by H and the weekday denoted by D as covariates. Liang et al. (2015) showed that the wind direction not only has influence on the response itself, but might also influence the relationship between the other covariates DEWP, TEMP, PRES, IWS, PREC and the response. Therefore we allow for different smooth terms corresponding to different wind directions. More precisely, we create four indicator variables corresponding to the four wind directions $\mathbb{1}_{\text{CBWD=CV}}$, $\mathbb{1}_{\text{CBWD=NE}}$, $\mathbb{1}_{\text{CBWD=NW}}$ and $\mathbb{1}_{\text{CBWD=SE}}$. Then, we replicate the part of the model matrix corresponding to a covariate x four times and multiply each of the four parts with one of the indicator variables $\mathbb{1}_{\text{CBWD=CV}}$, $\mathbb{1}_{\text{CBWD=NE}}$, $\mathbb{1}_{\text{CBWD=NW}}$ and $\mathbb{1}_{\text{CBWD=SE}}$. So, we obtain four smooth terms for each of the covariates DEWP, TEMP, PRES and IWS. We do not allow for these interactions with the covariate PREC since this variable has only few values not equal to zero. For variable selection the approach of Marra and Wood (2011) is used, which allows terms to be penalized to zero.

Plots of the different estimated smooth components are shown in the supplement (Figure 2) for the January data set. These plots indicate the covariate effects on PM2.5. For example, with northwestern winds (NW), PM2.5 is lower for higher temperatures. Further, we draw the same conclusion as Liang et al. (2015), that different smooth terms are necessary for different wind directions. For example, with northeastern winds (NE), we do not see any influence of the covariate PRES on PM2.5, but with northwestern winds (NW), we observe a non-linear relationship between PRES and PM2.5.

4.2. Model selection of monthly copula family and value of c based on the Watanabe Akaike Information Criterion (in-sample)

We now consider model selection for the copula state space model. This includes the selection of the copula families and the selection of the value of c. We fit models with different copula families and different values of c and select the model which minimizes the Watanabe Akaike Information Criterion (WAIC) introduced by Watanabe (2010). For our model AIC and BIC would require to integrate out all the latent variables. Therefore we stick to the WAIC which is easy to evaluate for such Bayesian models with latent variables. We define by $\ell_t^r := c(\hat{u}_t, v_t^r; \tau_O^r(c))$ the likelihood contribution of iteration r at time t. Following Vehtari et al. (2017) the WAIC can then be estimated by $\text{WAIC} = -2\sum_{t=1}^{T}\left[\ln\left(\hat{\text{E}}((\ell_t^r)_{r=1,\dots,R})\right) - \widehat{\text{Var}}\left((\ln(\ell_t^r))_{r=1,\dots,R}\right)\right]$, where $\hat{\text{E}}$ denotes the sample mean and $\widehat{\text{Var}}$ the sample variance.

We have one GAM specification for each month and obtain, for each month, approximately Uniform(0,1) pseudo-copula data $\hat{u}_t$ by the probability integral transform $\hat{u}_t = \Phi\left(\frac{y_t - \hat{f}(\boldsymbol{x}_t)}{\hat{\sigma}}\right)$ for $t = 1, \dots, T$ as in (9). Here $\hat{f}$ and $\hat{\sigma}$ are the estimates of the GAM and T denotes the number of observations in the corresponding monthly data set. To simplify notation we avoid indexing the models by month.

In the following we study several models that can be divided into three model classes.

- **Gaussian state space model** $\mathscr{M}_N$: $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$ are both Gaussian copulas.
- **Copula based state space model** $\mathscr{M}_C$: $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$ are from the same bivariate copula family.
- **GAM model with independent errors** $\mathscr{M}_I$: $\mathbb{C}^O_{U,V}$ and $\mathbb{C}^L_{V_2,V_1}$ are both independence copulas. This corresponds to a standard GAM model with independent errors.

For each of the 12 monthly data sets on the copula scale, the three model classes are fitted. To estimate model parameters we run the No-U-Turn sampler with 2 chains, where each chain contains 2000 iterations. The first 500 iterations are discarded for burnin. Preliminary analysis showed that this burnin choice is sufficient. We fit the independence model $\mathscr{M}_I$, the Gaussian model $\mathscr{M}_N$ for every value of $c = 1, 3, 6, 10$ and several latent copula models for the class $\mathscr{M}_C$. The different state space copula models correspond to all combinations of the values of $c = 1, 3, 6, 10$ and the following bivariate parametric copula families: Student t (df=3), Student t (df=6), Gumbel, Clayton and Frank. This set includes copula families that are appropriate for the observed contour plots in Figure 1. So for one specific monthly data set a model is specified by the value of c and the copula family.

As an example we have a closer look at the model for January with Student t copulas with 6 degrees of freedom and $c = 1$. Figure 5 shows the traceplots of the dependence parameter τ_L and the latent state at time point 100 (V_{100}) for the first chain. The traceplots suggest that the chains have converged. The chain for τ_L converges to values far away from zero, thus showing dependence. Figure 6 illustrates the effect of the different values of c on the posterior mode estimates of the latent states $\hat{v}_t$. As expected, we observe that the size of the oscillations decreases as the value of c increases.

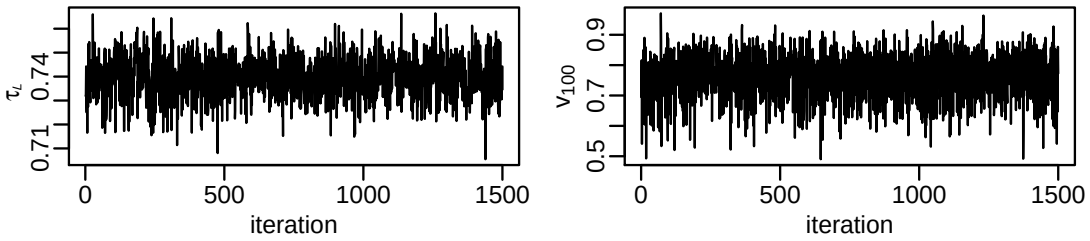

Figure 5. Traceplots of 1500 posterior draws after a burnin of 500 iterations of τ_L (left) and V_{100} (right) of the first chain of the HMC sampler for the model with Student t copulas with 6 degrees of freedom and $c = 1$ using the data set for January.

Table 1 shows the best model in $\mathscr{M}_C$, characterized by the value of c and the copula family, and the best model in $\mathscr{M}_N$, characterized by the value of c. In addition Table 1 shows the WAIC of the best model within the model classes $\mathscr{M}_C$, $\mathscr{M}_N$ and $\mathscr{M}_I$. We see that for $\mathscr{M}_N$ and $\mathscr{M}_C$ the value of c of the best model is always equal to 1 thus allowing for higher oscillations in the posterior of the latent states. The best model according to the WAIC is provided by the copula based model class $\mathscr{M}_C$ for every month, since this model is always associated to the smallest WAIC.

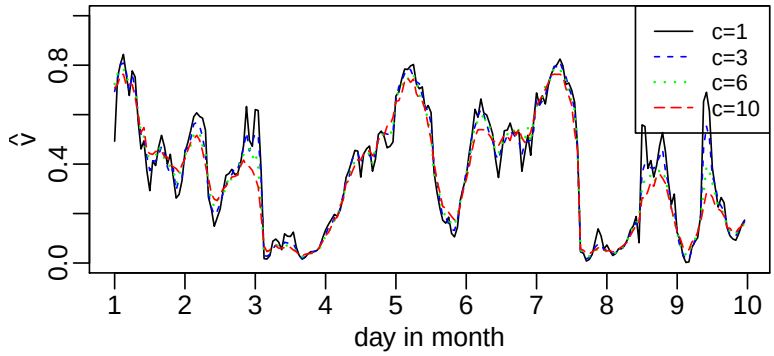

Figure 6. Estimated hourly posterior mode of the latent state $\hat{v}_t$ at time t plotted against t for the first 9 days of January for models with Student t copulas with 6 degrees of freedom and different values of c ($c = 1, 3, 6, 10$). The posterior mode estimates are obtained as modes of univariate kernel density estimates and are based on 3000 iterations from two chains.

4.3. Analysis of fitted models

In the previous section we selected the best copula state space models according to the lowest WAIC. This gave the copula family choice and the value of c for $\mathscr{M}_C$ and the value of c for $\mathscr{M}_N$. Figure 7 shows the estimated posterior densities for the dependence parameter τ_L for these models. We observe that most of the mass of the posterior density

Table 1. Family of the best model in $\mathscr{M}_C$, value of c of the best model in $\mathscr{M}_C$ and the best model in $\mathscr{M}_N$ and the WAIC of the best model within each class $\mathscr{M}_C$, $\mathscr{M}_N$ and $\mathscr{M}_I$. The best model is selected with respect to the WAIC.

	family	c		$WAIC$		
month	$\mathscr{M}_C$	$\mathscr{M}_C$	$\mathscr{M}_N$	$\mathscr{M}_C$	$\mathscr{M}_N$	$\mathscr{M}_I$
1	t(6)	1	1	-926	-887	0
2	Frank	1	1	-755	-702	0
3	Frank	1	1	-1000	-898	0
4	t(3)	1	1	-1200	-1103	0
5	t(6)	1	1	-982	-945	0
6	t(3)	1	1	-672	-604	0
7	t(3)	1	1	-808	-722	0
8	t(3)	1	1	-680	-653	0
9	t(6)	1	1	-972	-873	0
10	Gumbel	1	1	-1130	-1102	0
11	t(6)	1	1	-910	-900	0
12	t(6)	1	1	-765	-758	0

concentrates between 0.6 and 0.8 for all monthly models. This range for τ_L coincides with positive dependence between two succeeding time points. We also see that the Kendall's τ values of the $\mathscr{M}_C$ model class are slightly higher than those of the $\mathscr{M}_N$ model class for all months.

The copula based state space model was fitted to the standardized residuals of the GAM $\hat{z}_t$ as defined in (4). To further evaluate our model, we simulate from the predictive distribution of the error for each $t \in \{1, \ldots, T\}$, as explained in Section 3.3, and compare it to the standardized residuals of the corresponding GAM model. Figure 8 shows that the copula state space model is able to recover the dynamics of the standardized residuals.

If we ignored the latent effect, the distribution of the error would be standard normal. Simulating from the predictive distribution of the error can be considered as taking the latent effect into account. Therefore a concentration of the predictive distribution that is far away from zero indicates time points where the latent variable has higher effects. These are time points where the level of the response is unusually high or low for the corresponding specification of the covariates.

We see from Figure 8 that on January 18th, the estimated mode of the predictive density of the error is high. On this day unusual high pollution was recorded in Beijing where PM2.5 reached around 500 micrograms per cubic meter (μg/m^3), skyrocketing to more than 20 times the level considered unhealthy by the World Health Organization (See `www.takepart.com/article/2014/01/18/beijing-china-air-pollution-billboard`). The copula based state space model with a Student t copula has a high peak on that day and is able to capture this unusual behaviour.

The ability to model unusual high peaks of airborne contaminants is fundamental to accurately assess the effect of exposure to human health. Indeed, many studies in the literature show that increased levels of air pollutants may have a dramatic effect on human health. In particular, for a 10-μg/m^3 increase in PM2.5, hospital admissions for

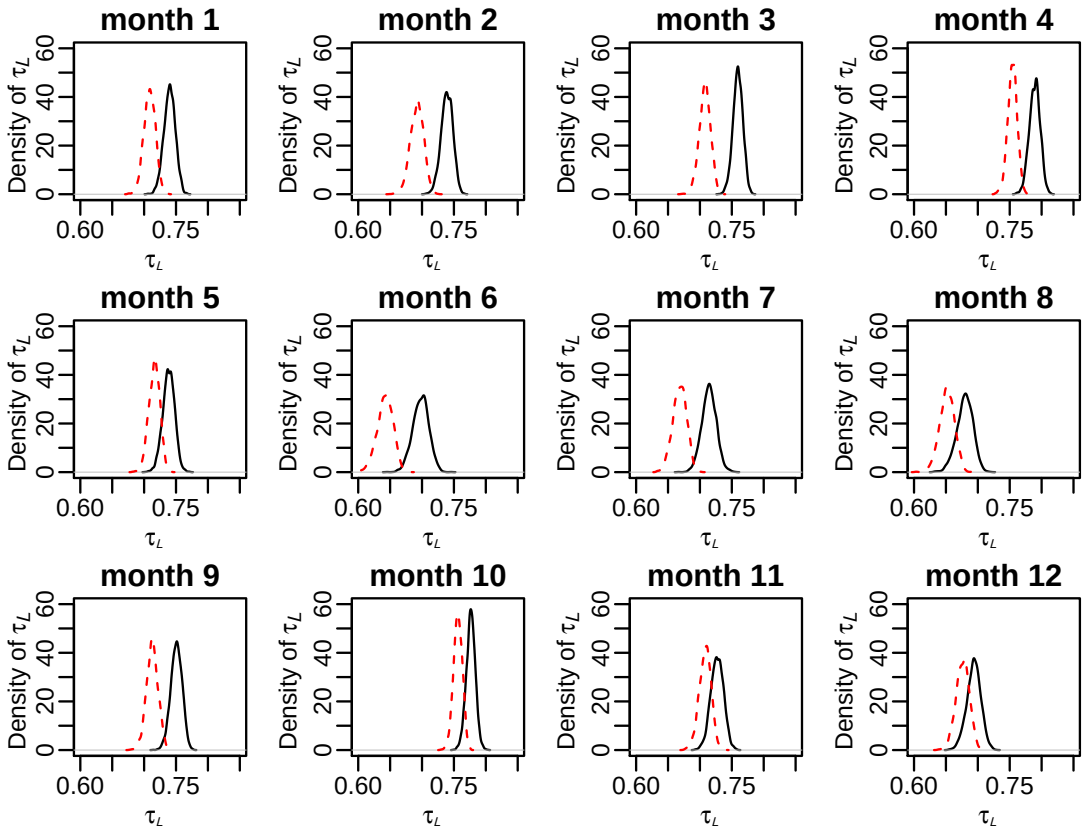

Figure 7. Estimated posterior density of the dependence parameter τ_L for the best model in $\mathscr{M}_C$ (black) and $\mathscr{M}_N$ (red, dashed) according to the WAIC for all 12 data sets.

ischemic cardiac events and heart failures may increase by 4.5% and 3.6%, respectively; respiratory and pneumonia hospitalizations may increase by 17% and 6.5%; respiratory and lung cancer mortality may increase by 2.2% and 8%. In addition, exposure to PM2.5 is estimated to reduce the life expectancy of the population by about 8.6 months on average (Anderson et al., 2012). The economic impact of PM2.5 pollution is also relevant, since fine particulate matter-related illness can ultimately lead to financial and non-financial welfare losses of not only patients and their families but also a significant portion of gross domestic product (GDP). Indeed, it was estimated that in 2009 China suffered a health-related economic loss of 2.1% of its GDP, corresponding to 106.5 billion US dollars (Kim et al., 2015). Therefore, the consequences on citizens' health and economy of an extremely high value of PM2.5, such as the one experienced in Beijing on the 18th January 2014, may be very severe and extensive.

The proposed Bayesian non-linear non-Gaussian state space model allows us to capture unusual extreme air pollution events appropriately and could provide accurate information to stakeholders such as doctors and policy makers to better evaluate the consequences of pollution on citizens.

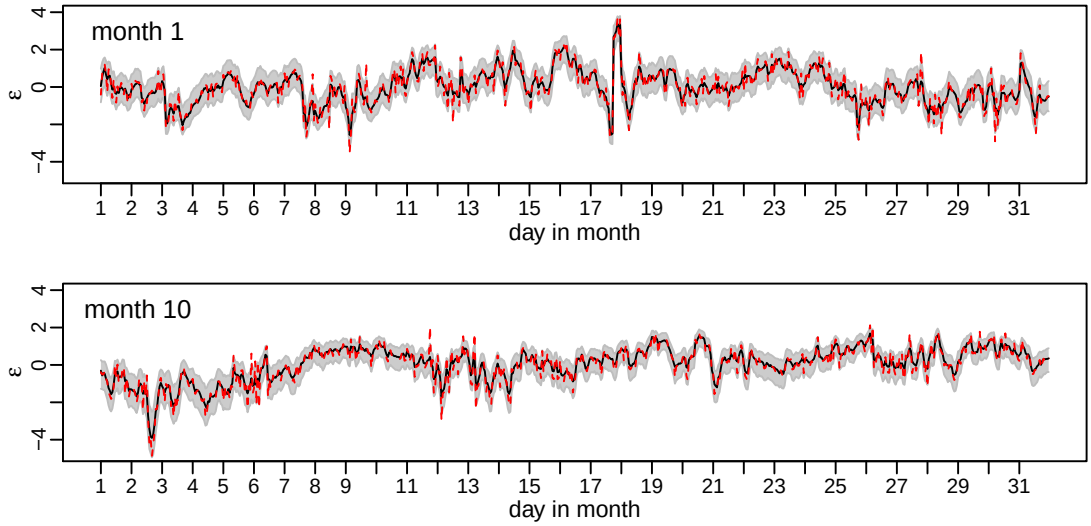

Figure 8. Estimated mode of the predictive density of the error ϵ_t plotted against t for every data point in January (top row) and October (bottom row) using the best models in $\mathscr{M}_C$ as selected by WAIC. A 90% credible region, constructed from the 5% and 95% empirical quantiles of simulations from the predictive distribution of the error, is added in grey. Further, the standardized residual of the GAM $\hat{z}_t$ is added in red (dashed).

4.4. Out-of-sample predictions

Short term predictions of PM2.5 levels can be used to alert citizens of high pollution periods which are dangerous to health. In this section we construct predictions several hours up to two days ahead. More precisely, we consider the best copula state space model for March and use it to predict the first 48 hours of April. We choose March, since it is the month for which the non-elliptical Frank copula was selected.

We first simulate from the out-of-sample predictive distribution of the error as explained in Section 3.3. Figure 9 shows predictive densities for different time-steps ahead for this model, more precisely the estimated forecast density of ϵ_{T+t} for $t = 1, 12, 24, 48$ hours based on 3000 HMC iterations from two chains. As we see from Figure 9, we obtain non-Gaussian forecast densities. Further, the densities are more disperse for a longer time period ahead, reflecting the fact that uncertainty increases if we predict a longer time period ahead.

To obtain predictions for the PM2.5 levels the simulations for the error needs to be combined with the mean prediction of the GAM, according to our model $Y_t = f(\boldsymbol{x}_t) + \sigma\varepsilon_t$.

To obtain the predicted mean of the GAM the covariate values are required. Ideally, we would have to use in our model the predicted values of the covariates published by Chinese meteorological authorities, as in Feng et al. (2015), for example. Unfortunately, except for the weekday D and the hour H, future covariate levels are not known. Therefore, as the best proxy for an unknown covariate vector with hour H=h, we use the covariate specifications of the last observed time point with the same hour H=h. We denote this covariate vector by $\boldsymbol{x}_t^l$ and obtain predictive simulations of the response at

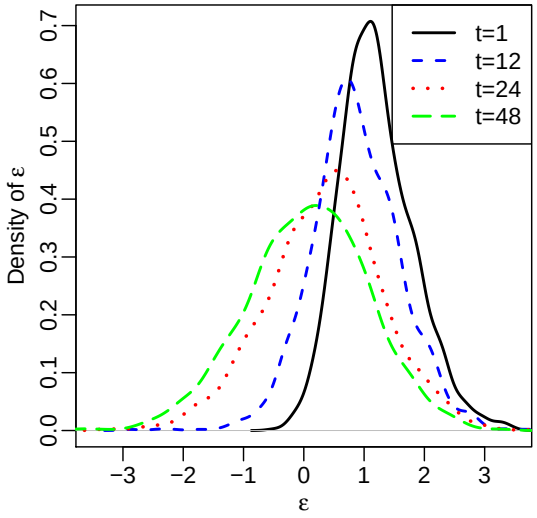

Figure 9. Estimated predictive density of ϵ_{T+t} using the best copula state space model for March for different time steps (hours) ahead ($t = 1, 12, 24, 48$). The estimated predictive density is the kernel density estimate of simulations from the corresponding predictive distribution.

time $t > T$ as follows

$$y_t^r = \hat{f}(\boldsymbol{x}_t^l) + \hat{\sigma}\varepsilon_t^r, \tag{19}$$

for $r = 1, \ldots, R$. These predictive simulations are visualized in Figure 10. We see that the observed values are most of the time within the 90% credible interval.

We compared our out-of-sample predictions with those obtained by applying a deep model to our data. We implemented an LSTM RNN model using the same data setting adopted for the proposed non-linear non-Gaussian state space model. In particular, we used the March data as training set, the first 48 hours of April as test set, with covariates set to the last observed time point with the same hour. We included the PM2.5 variable, together with the meteorological covariates in the input layer. All variables were normalized in order to give similar impact of all inputs. The model was built using `keras` and `tensorflow` in R. For each predicted value, we computed 5% and 95% bootstrap prediction intervals. Figure 10 shows predictions and credible regions obtained from the copula state space approach and from the deep model. The Figure indicates that the copula state space model provides better forecasts. For further comparison, we made use of the mean squared error (MSE) to evaluate point forecasts and of the interval score (IS) (Gneiting and Raftery, 2007) to evaluate the accuracy of the credible regions. The MSE of the copula state space model is 0.73, while the MSE of the deep model is 1.01. The IS of the copula state space model is 4.92, while the IS of the deep model is 9.82. Since a lower IS indicates a more accurate credible region, both evaluation metrics are in favor of the copula state space approach.

Therefore, the out-of-sample results show that RNN produce less accurate probabilistic forecasts than a well-formulated statistical time series model. This is consistent with

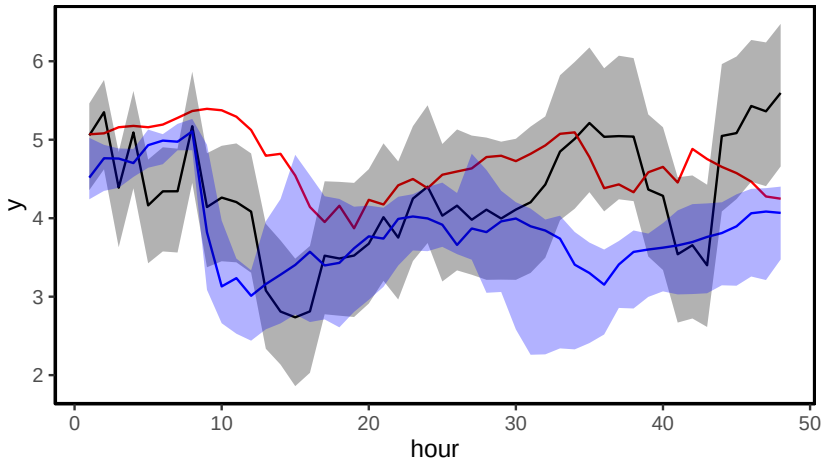

Figure 10. The black line shows the estimated mode of the predictive density of the response t hours ahead plotted against t. The simulations of the predictive distribution of the response 1 up to 48 hours ahead are obtained according to (19) based on the best copula state space model for March. A 90% credible region, constructed from the 5% and 95% empirical quantiles of simulations from the predictive distribution of the response, is added in grey. Further, predictions and a corresponding 90% confidence region obtained from the deep model are added in blue and the observed response values are shown in red.

findings by a number of other authors, such as, for example, Klein et al. (2020) in their paper about electricity price forecasting.

In addition, the simulations for the error may be combined with mean predictions obtained from the GAM with different covariate specifications. Since the covariates several hours ahead are random, different scenarios as specified by different covariate levels are possible and should be taken into account. Here, we first consider two cases where the temperature at each time point in $\boldsymbol{x}_t^l$ is increased and decreased by 1 degree. Second, we also investigate more extreme scenarios for $\boldsymbol{x}_t^l$ where we decrease and increase the temperature at each time point by 4 degrees and in addition change the wind direction at each time point to the same value. The value for the wind direction CBWD is set to either CV or SE. This yields four different scenarios. The mode estimates of the resulting predictive densities are visualized in Figure 11. It is not surprising that the first case where we only change the temperature by 1 degree results in less changes in the mode estimates compared to the more extreme case. There are many more scenarios that can be analysed in a similar fashion. In particular, relevant scenarios suggested by experts could be analysed. A conservative warning system could alert citizens if at least one of the scenarios results in dangerous air pollution levels.

5. Summary and Outlook

The starting point of this paper was the question of how to capture not only non-linear effects of meteorological variables on pollution measures such as airborne particulate

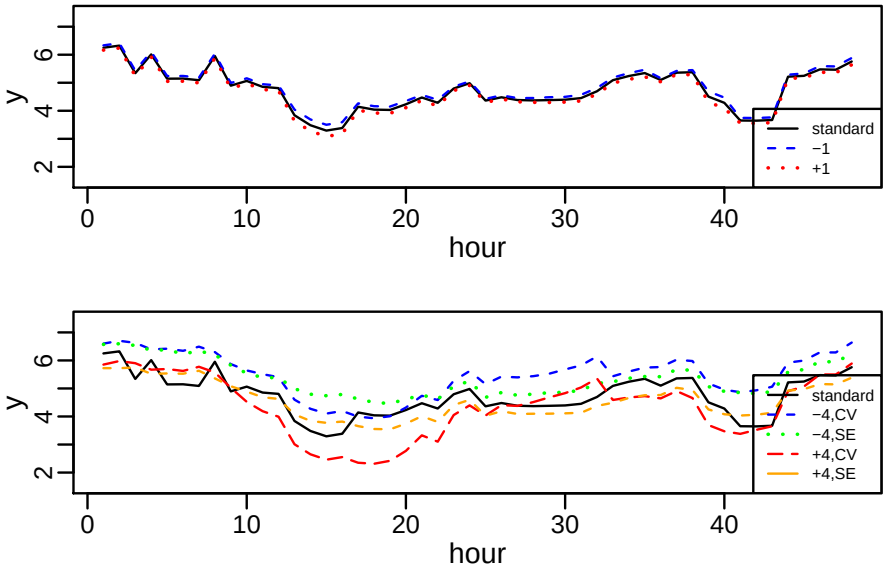

Figure 11. We show the estimated mode of the predictive density of the response t hours ahead plotted against t for different specifications of the covariates. The simulations of the corresponding predictive distribution of the response 1 up to 48 hours ahead are obtained according to (19) based on the best copula state space model for March (black line). In the top row, we consider additionally predictive distributions where the temperature of $\boldsymbol{x}_t^l$ is changed by ± 1 degree. In the bottom row, we consider additionally predictive distributions where the temperature of $\boldsymbol{x}_t^l$ is changed by ± 4 degree and the covariate CBWD is set equal to SE or CV.

matter, but also to allow for further time dynamics of the observations not covered by the meteorological variables. For this we investigated hourly data of ambient air pollution in Beijing and illustrated that the lag-one time dynamics is not a Gaussian one, thus ruling out standard linear state space models.

To deal with this non-Gaussian dependence we proposed a novel non-linear state space model based on a copula formulation for univariate observation and state equations. The observation and state variables are coupled using two bivariate copulas. Since the copula approach allows for separate modelling of the margins and dependence, the observation variables are allowed to follow any time invariant statistical model. In the application we utilized a GAM to allow for non-linear effects of covariates. Once the marginal distribution of the response variables is specified, they can be transformed to the uniform scale using the probability integral transform. The resulting value on the uniform scale at time t, U_t, is then coupled with a [0,1] valued state variable for time t using a bivariate copula. Therefore, the observation equation of the copula based state space formulation is given by the conditional distribution of U_t given the value of the state variable at time t. The time dynamics of the state variables is then similarly modeled as the conditional distribution of the state variable at time t given the state variable at time $t-1$, where these two state variables are jointly modeled by a bivariate copula. We first show that,

in the case of a bivariate Gaussian copula, standard linear state space models result. Since many different parametric bivariate copulas exist, the flexibility of the copula-based state space model is evident and thus a significant extension of linear Gaussian state space models is possible.

Of course, such an extension has its price. In our case this means we cannot follow a standard estimation approach such as the Kalman filter for linear state space models. Therefore we propose and develop a Bayesian approach based on HMC. Further we deal with some identifiability issues of the copula state space model, which we solve by restricting the strength of the dependence among the lag-one state space variables to be at least as high as the one of the observation variable U_t and the state variable at time t.

The state variables can be interpreted as a way to capture non-measured effects and thus are very appropriate for the data set analyzed in this paper. It allowed us to identify unusual high levels of pollution, which were not captured by the measured variables. We also present, with appropriate normalized bivariate contour plots, explorative tools to detect non-Gaussian dependence structures.

The proposed approach can be used to accurately model extreme air pollution events and can assist stakeholders in the evaluation of the health consequences of exposure. The analysis of high temporal resolution particulate matter data allows us to immediately detect quick upsurges of airborne contaminants and anticipate lower temporal resolution health effects. Nevertheless, we point out that the applicability of copula state space approach is not restricted to air pollution, but could be adopted in numerous settings.

The approach first proposed here allows a wide range of extensions, such as adding covariates for the dependence parameter of the bivariate copulas as well as extending to multivariate response data and adopting wavelets expansions for marginal models. Here the use of vine copulas can be envisioned wherever higher-dimensional than bivariate copulas are needed. Another route of extension would be to model the bivariate copulas completely nonparameteric. In this case the identifiability issues have to be reworked.

Acknowledgements

The second author was supported by a Global Challenges for Women in Math Science Entrepreneurial Programme grant for the project "Bayesian Analysis of State Space Factor Copula Models" provided by the Technical University of Munich. The third author is supported by the German Research Foundation (DFG grant CZ 86/6-1). Computations were performed on a Linux cluster supported by DFG grant INST 95/919-1 FUGG.

References

Anderson, J. O., J. G. Thundiyil, and A. Stolbach (2012). Clearing the air: a review of the effects of particulate matter air pollution on human health. *Journal of Medical Toxicology 8*(2), 166–175.

Ayturan, Y. A., Z. C. Ayturan, and H. O. Altun (2018). Air pollution modelling with deep learning: a review. *International Journal of Environmental Pollution and Environmental Modelling 1*(3), 58–62.

Barra, I., L. Hoogerheide, S. J. Koopman, and A. Lucas (2017). Joint Bayesian Analysis of Parameters and States in Nonlinear non-Gaussian State Space Models. *Journal of Applied Econometrics 32*(5), 1003–1026.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bui, T.-C., V.-D. Le, and S.-K. Cha (2018). A deep learning approach for forecasting air pollution in south korea using lstm. *arXiv preprint arXiv:1804.07891*.

Burnett, R. and D. Krewski (1994). Air pollution effects on hospital admission rates: a random effects modeling approach. *Canadian Journal of Statistics 22*(4), 441–458.

Calder, C. A. (2008). A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics: The official journal of the International Environmetrics Society 19*(1), 39–48.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2016). Stan: A probabilistic programming language. *Journal of Statistical Software 20*.

Carpenter, B., M. D. Hoffman, M. Brubaker, D. Lee, P. Li, and M. Betancourt (2015). The stan math library: Reverse-mode automatic differentiation in C++. *arXiv preprint arXiv:1509.07164*.

Chen, S., J. Fricks, and M. J. Ferrari (2012). Tracking measles infection through nonlinear state space models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(1), 117–134.

Chen, X. and Y. Fan (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics 130*(2), 307–335.

Cohen, A. J., M. Brauer, R. Burnett, H. R. Anderson, J. Frostad, K. Estep, K. Balakrishnan, B. Brunekreef, L. Dandona, R. Dandona, et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet 389*(10082), 1907–1918.

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*.

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(1), 3–56.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*(3), 603–616.

Durbin, J. and S. J. Koopman (2012). *Time series analysis by state space methods*, Volume 38. Oxford University Press.

Feng, X., Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang (2015). Artificial neural networks forecasting of pm2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment 107*, 118–128.

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

Hajian, A. (2007). Efficient cosmological parameter estimation with Hamiltonian Monte Carlo technique. *Physical Review D 75*(8), 083525.

Hartmann, M. and R. S. Ehlers (2017). Bayesian inference for generalized extreme value distributions via Hamiltonian Monte Carlo. *Communications in Statistics-Simulation and Computation*, 1–18.

Hastie, T. and R. Tibshirani (1986). Generalized Additive Models. *Statistical Science 1*(3), 297–318.

Hoffman, M. D. and A. Gelman (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Ippoliti, L., P. Valentini, and D. Gamerman (2012). Space–time modelling of coupled spatiotemporal environmental variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(2), 175–200.

Joe, H. (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

Joe, H. and J. J. Xu (1996). The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia.

Johns, C. J. and R. H. Shumway (2005). A non-linear and non-Gaussian state-space model for censored air pollution data. *Environmetrics: The official journal of the International Environmetrics Society 16*(2), 167–180.

Kampa, M. and E. Castanas (2008). Human health effects of air pollution. *Environmental pollution 151*(2), 362–367.

Kan, H., R. Chen, and S. Tong (2012). Ambient air pollution, climate change, and population health in China. *Environment international 42*, 10–19.

Kim, K.-H., E. Kabir, and S. Kabir (2015). A review on the human health impact of airborne particulate matter. *Environment international 74*, 136–143.

Klein, N., M. S. Smith, and D. J. Nott (2020). Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. *arXiv preprint arXiv:2010.01844*.

Koopman, S. J., A. Lucas, and M. Scharth (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics 98*(1), 97–110.

Koopman, S. J. and G. Mesters (2017). Empirical Bayes Methods for Dynamic Factor Models. *Review of Economics and Statistics 99*(3), 486–498.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

Lee, D., A. Rushworth, and S. K. Sahu (2014). A bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics 70*(2), 419–429.

Li, X., L. Peng, Y. Hu, J. Shao, and T. Chi (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research 23*(22), 22408–22417.

Liang, X., T. Zou, B. Guo, S. Li, H. Zhang, S. Zhang, H. Huang, and S. X. Chen (2015). Assessing Beijing's PM2. 5 pollution: severity, weather impact, APEC and winter heating. *Proc. R. Soc. A 471*(2182), 20150257.

Liu, D.-R., S.-J. Lee, Y. Huang, and C.-J. Chiu (2020). Air pollution forecasting based on attention-based lstm neural network and ensemble learning. *Expert Systems 37*(3), e12511.

Liu, M., Y. Huang, Z. Ma, Z. Jin, X. Liu, H. Wang, Y. Liu, J. Wang, M. Jantunen, J. Bi, et al. (2017). Spatial and temporal trends in the mortality burden of air pollution in China: 2004–2012. *Environment international 98*, 75–81.

Marra, G. and S. N. Wood (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis 55*(7), 2372–2387.

Matus, K., K.-M. Nam, N. E. Selin, L. N. Lamsal, J. M. Reilly, and S. Paltsev (2012). Health damages from air pollution in China. *Global environmental change 22*(1), 55–66.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Pakman, A. and L. Paninski (2014). Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics 23*(2), 518–542.

Patton, A. J. (2009). Copula–based models for financial time series. In *Handbook of financial time series*, pp. 767–785. Springer.

Rangapuram, S. S., M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski (2018). Deep state space models for time series forecasting. *Advances in neural information processing systems 31*, 7785–7794.

Sahu, S. K., A. E. Gelfand, and D. M. Holland (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics 11*(1), 61.

Sahu, S. K. and K. V. Mardia (2005). A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(1), 223–244.

Salinas, D., M. Bohlke-Schneider, L. Callot, R. Medico, and J. Gasthaus (2019). High-dimensional multivariate forecasting with low-rank gaussian copula processes. In *Advances in Neural Information Processing Systems*, pp. 6827–6837.

Shaddick, G., M. L. Thomas, A. Green, M. Brauer, A. van Donkelaar, R. Burnett, H. H. Chang, A. Cohen, R. Van Dingenen, C. Dora, et al. (2018). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 67*(1), 231–253.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris 8*, 229–231.

Smith, M., A. Min, C. Almeida, and C. Czado (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association 105*(492), 1467–1479.

Smith, M. S. and W. Maneesoonthorn (2018). Inversion copulas from nonlinear state space models with an application to inflation forecasting. *International Journal of Forecasting 34*(3), 389–407.

Song, C., L. Wu, Y. Xie, J. He, X. Chen, T. Wang, Y. Lin, T. Jin, A. Wang, Y. Liu, et al. (2017). Air pollution in China: status and spatiotemporal variations. *Environmental pollution 227*, 334–347.

Stoica, E. (2013). A stability property of Farlie-Gumbel-Morgenstern distributions. *Unpublished manuscript:HAL Id: hal-00861234*.

Van den Brakel, J. and J. Roels (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *The Annals of Applied Statistics*, 1105–1138.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing 27*(5), 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*(Dec), 3571–3594.

Wood, S. and M. S. Wood (2015). Package 'mgcv'. *R package version 1*, 29.

World Health Organization (2013). Review of evidence on health aspects of air pollution–REVIHAAP Project. *World Health Organization, Copenhagen, Denmark*.