Biometrics 71, 323–332 DOI: 10.1111/biom.12279
June 2015

R-vine Models for Spatial Time Series with an Application to Daily Mean Temperature

Tobias Michael Erhardt*, Claudia Czado and Ulf Schepsmeier

Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany

**email:* tobias.erhardt@tum.de

Summary:

We introduce an extension of R-vine copula models to allow for spatial dependencies and model based prediction at unobserved locations. The proposed spatial R-vine model combines the flexibility of vine copulas with the classical geostatistical idea of modeling spatial dependencies using the distances between the variable locations. In particular the model is able to capture non-Gaussian spatial dependencies. To develop and illustrate our approach we consider daily mean temperature data observed at 54 monitoring stations in Germany. We identify relationships between the vine copula parameters and the station distances and exploit these in order to reduce the huge number of parameters needed to parametrize a 54-dimensional R-vine model fitted to the data. The new distance based model parametrization results in a distinct reduction in the number of parameters and makes parameter estimation and prediction at unobserved locations feasible. The prediction capabilities are validated using adequate scoring techniques, showing a better performance of the spatial R-vine copula model compared to a Gaussian spatial model.

Key words: Daily mean temperature; Marginal model; Spatial R-vine model; Spatial statistics; Vine copulas.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Understanding the earth's climate system is of vital interest to every aspect of human life. Recently the class of vine copulas has captured attention as a flexible class to model high dimensional dependencies (see Czado, 2010; Czado, Brechmann, and Gruber, 2013; Kurowicka and Cooke, 2006; Kurowicka and Joe, 2011, and reference therein). We present a new vine copula based approach for the spatial modeling of climatic time series. Utilization of available spatial information will lead to a distinct reduction in the number of parameters needed to parametrize the high dimensional (spatial) regular vine (R-vine) copula model. Model selection, estimation and a prediction method at arbitrary locations will be developed.

Many different approaches to spatio-temporal (dependency) modeling can be found in the literature. We refer the reader to the comprehensive monograph of Cressie and Wikle (2011) and references therein. Multivariate Gaussian distributions are customary used for dependency modeling. However, they are not appropriate to model all data, since they require symmetry and do not allow for extreme dependency. Therefore, we apply vine copula models, which are designed to overcome these limitations. Copulas are d-dimensional distribution functions on $[0,1]^d$ with uniform margins. They can be understood as a tie between a multivariate distribution function F and its marginals $(F_1, \ldots, F_d)$ and capture all dependency information (see Sklar, 1959). In particular, we have $F(\boldsymbol{y}) = C\left(F_1(y^1), \ldots, F_d(y^d)\right)$, where $\boldsymbol{y} = (y^1, \ldots, y^d)'$ is the realization of a random vector $\boldsymbol{Y} \in \mathbb{R}^d$. Vine copulas are constructions of d-dimensional copulas built on bivariate copulas only. They are well understood and easy to compute (see Aas et al., 2009; Brechmann and Schepsmeier, 2013; Dißmann et al., 2013). A short introduction to R-vines will be given in Section 2.

We develop our approach for daily mean temperature time series collected over the period 01/01/2010-12/31/2012 by the German Meteorological Service (Deutscher Wetterdienst) (Section 3). The common modeling of all marginal distributions is discussed in Section 3. It

captures seasonality effects and temporal dependencies of the time series. Spatially varying parameters allow to approximate these effects and dependencies at unobserved locations.

The main contribution is the development of a new vine copula based spatial dependency model introduced in Section 4. It relies on a reparametrization of an R-vine copula model, which exploits the relationship between the model parameters and the available spatial information. Different model specifications based on distances and elevation differences were considered in Erhardt (2013), the most promising one is highlighted here. Parameter estimation methods are followed by model based prediction at unobserved locations. A geostatistical model is developed in Section 5 and used for comparison. The resulting model evaluation is conducted in Section 6. A validation data set of 19 additional locations allows to calculate adequate scores, based on which the quality of the predictions can be compared. The paper closes with a discussion section.

2. Regular vine copula models

Vine copulas provide an easy and flexible way to model multivariate distributions having different marginal distributions and allow for complex (non-Gaussian) dependencies. The R package `VineCopula` (see Schepsmeier et al., 2014) eases their application in practice.

Vine copulas in general were introduced by Bedford and Cooke (2001, 2002) and trace back to ideas of Joe (1996). They are build using a cascade of $d(d-1)/2$ bivariate copulas, called *pair copulas*. This cascade is identified by a set of nested trees called a regular vine tree sequence or short *regular vine* (R-vine). The R-vine tree sequence $\mathscr{V} = (\mathcal{T}_1, \dots, \mathcal{T}_{d-1})$ satisfies the following conditions (see Bedford and Cooke, 2001):

(1) $\mathcal{T}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ is a tree with vertices $\mathcal{V}_1 = \{1, \dots, d\}$ and edge set $\mathcal{E}_1 \subset \mathcal{V}_1 \times \mathcal{V}_1$.

(2) $\mathcal{T}_l = (\mathcal{V}_l, \mathcal{E}_l)$ is a tree with $\mathcal{V}_l = \mathcal{E}_{l-1}$ and $\mathcal{E}_l \subset \mathcal{V}_l \times \mathcal{V}_l$, for all $l = 2, \dots, d-1$.

(3) For all vertex pairs in $\mathcal{V}_l$ connected by an edge $e \in \mathcal{E}_l$, $l = 2, \ldots, d-1$, the corresponding edges in $\mathcal{E}_{l-1}$ have to share a common vertex (*proximity condition*).

Aas et al. (2009) were the first to develop statistical inference for non-Gaussian pair copulas.

The notation for vine edges will follow Czado (2010). An edge $e \in \mathcal{E}_l$, $l = 1, \ldots, d-1$, will be denoted by $i(e), j(e); \mathcal{D}_e$, where $i(e) < j(e)$ make up the *conditioned set* $\mathcal{C}_e = \{i(e), j(e)\}$ and $\mathcal{D}_e$ is called *conditioning set*. An example in five dimensions is given in Web Figure 1. It depicts the four nested trees of an R-vine tree sequence $\mathscr{V} = (\mathcal{T}_1, \ldots, \mathcal{T}_4)$.

Next we introduce the link of an R-vine tree sequence $\mathscr{V}$ to the multivariate copula distribution of some random vector $\boldsymbol{U} = (U^1, \ldots, U^d) \in [0,1]^d$ with $U^1, \ldots, U^d \sim \mathcal{U}(0,1)$. We define the set $\mathcal{B} \coloneqq \left\{C_{i(e),j(e);\mathcal{D}_e} : e \in \mathcal{E}_l, l = 1, \ldots, d-1\right\}$ of bivariate copulas $C_{i(e),j(e);\mathcal{D}_e}$ corresponding to the R-vine edges $e \in \mathcal{E}_l$, $l = 1, \ldots, d-1$. The copula families of these (parametric) pair-copulas are denoted by $b_{i(e),j(e);\mathcal{D}_e}$. For an overview of frequently used bivariate copula families we refer to Brechmann and Schepsmeier (2013). Further we define $\boldsymbol{u}^{\mathcal{I}} \coloneqq \left\{u^k : k \in \mathcal{I}\right\}$ for arbitrary index sets $\mathcal{I} \subseteq \{1, \ldots, d\}$. This allows to express the vine copula density of $\boldsymbol{U}$ associated with the R-vine tree sequence $\mathscr{V}$ as

$$c_{1,\ldots,d}(\boldsymbol{u}) = \prod_{l=1}^{d-1} \prod_{e \in \mathcal{E}_l} c_{i(e),j(e);\mathcal{D}_e} \left\{C_{i(e)|\mathcal{D}_e}(u^{i(e)} \mid \boldsymbol{u}^{\mathcal{D}_e}), C_{j(e)|\mathcal{D}_e}(u^{j(e)} \mid \boldsymbol{u}^{\mathcal{D}_e})\right\}, \tag{1}$$

where $c_{i(e),j(e);\mathcal{D}_e}\{\cdot,\cdot\}$ are the densities corresponding to the bivariate copulas $C_{i(e),j(e);\mathcal{D}_e} \in \mathcal{B}$. For a derivation of (1) see Bedford and Cooke (2001). To evaluate such a density we need to calculate the so called *transformed variables* $C_{i(e)|\mathcal{D}_e}(u^{i(e)} \mid \boldsymbol{u}^{\mathcal{D}_e})$ and $C_{j(e)|\mathcal{D}_e}(u^{j(e)} \mid \boldsymbol{u}^{\mathcal{D}_e})$. Here $C_{i(e)|\mathcal{D}_e}$ and $C_{j(e)|\mathcal{D}_e}$ are conditional distributions obtained from $C_{i(e),j(e);\mathcal{D}_e}$. The calculation is performed recursively according to Joe (1996) using the formula

$$C_{k|\mathcal{J}}(u^k \mid \boldsymbol{u}^{\mathcal{J}}) = \frac{\partial\, C_{kl;\mathcal{J}_{-l}}\left\{C_{k|\mathcal{J}_{-l}}(u^k \mid \boldsymbol{u}^{\mathcal{J}_{-l}}), C_{l|\mathcal{J}_{-l}}(u^l \mid \boldsymbol{u}^{\mathcal{J}_{-l}})\right\}}{\partial C_{l|\mathcal{J}_{-l}}(u^l \mid \boldsymbol{u}^{\mathcal{J}_{-l}})}, \tag{2}$$

where $k, l \in \{1, \ldots, d\}$, $k \neq l$, $\{l\} \subset \mathcal{J} \subset \{1, \ldots, d\} \backslash \{k\}$ and $\mathcal{J}_{-l} \coloneqq \mathcal{J} \backslash \{l\}$. We implicitly made the *simplifying assumption*, that the copulas in $\mathcal{B}$ do not depend on the conditioning

value $\boldsymbol{u}^{\mathcal{D}_e}$ other than through its arguments given in (1). Moreover, in Section 4 we will use *truncated* R-vines (Brechmann, Czado, and Aas, 2012). Truncation after level $k < d-1$ means that independence copulas are selected for all edges $e \in \mathcal{E}_l$, $k < l < d$.

In a spatio-temporal setting the data y_t^s, $s = 1, \ldots, d$, $t = 1, \ldots, N$, is not restricted to the unit hypercube $[0,1]^d$ and does not necessarily have uniformly distributed margins. Therefore the data has to be transformed to so called *copula data* $u_t^s \sim \mathcal{U}(0,1)$, $s = 1, \ldots, d$, $t = 1, \ldots, N$, before vine copula models can be applied. We consider a regression model $Y_t^s = g(t, \boldsymbol{x}^s; \boldsymbol{\beta}) + \varepsilon_t^s$, $\varepsilon_t^s \sim F^s$, with spatial covariates $\boldsymbol{x}^s$, to adjust for spatial as well as seasonality effects and temporal dependencies. The resulting residuals $\widehat{\varepsilon}_t^s := y_t^s - g(t, \boldsymbol{x}^s; \widehat{\boldsymbol{\beta}})$, $t = 1, \ldots, N$, are approximately independent for each location $s = 1, \ldots, d$. We transform these residuals by their respective parametric marginal distribution functions F^s, i.e. we calculate $u_t^s := F^s(\widehat{\varepsilon}_t^s)$. This transformation is called a probability integral transform. We prefer to use parametric probability integral transformations (see Joe and Xu, 1996) over empirical rank transformations (proposed for example by Genest, Ghoudi, and Rivest, 1995), since we are interested in predictions on the original scale using the proposed marginal models. Referring to Kim, Silvapulle, and Silvapulle (2007) we emphasize, that the choice of a suitable marginal model is important, as a gross misspecification of the distributional shape of the marginal distribution can distort the joint modeling using copulas. To ensure that the chosen model is adequate we advise to check if the transformed data is uniform.

3. A Marginal Model for Daily Mean Temperatures

The data set consists of daily mean temperatures in °C collected over the period 2010 to 2012 by the German Meteorological Service (Deutscher Wetterdienst) for 73 selected observation stations across Germany. We split the data into a training ($s = 1, \ldots, 54$) and a validation data set ($s = 55, \ldots, 73$). Hence we build our models on $d = 54$ times $N = 1096$ observations y_t^s of mean temperatures, which are considered as realizations of random variables Y_t^s ($t =$

$1, \dots, N$, $s = 1, \dots, d$). Lists with details about the location s (longitude (x^s_{lo}), latitude (x^s_{la}) and elevation (x^s_{el})) and the names of all 73 stations are given in Web Table 1. Their locations in Germany are illustrated in Figure 1.

[Figure 1 about here.]

For vine copula based models we need to transform our data to copula data. For this, we use the marginal model of Erhardt (2013, Chapter 3), which is a tailor-made model for the marginal mean temperatures at arbitrary locations in Germany. To ensure homoscedasticity, i.e. $\text{Var}(\varepsilon^s_t) = \sigma^2 > 0$, $t = 1, \dots, N$, $s = 1, \dots, d$, the model considers appropriately weighted observations $\widetilde{Y}^s_t := Y^s_t / \sqrt{\widehat{w}_t}$. Raw weights $\widetilde{w}_t$, $t = 1, \dots, N$, obtained as the sample variances $\widetilde{w}_t := \frac{1}{d-1} \sum_{s=1}^{d} (y^s_t - \overline{y}_t)^2$, where $\overline{y}_t := \frac{1}{d} \sum_{s=1}^{d} y^s_t$, $t = 1, \dots, N$, are smoothed using least-squares. This results in the *smoothed weights* $\widehat{w}_t := \exp\{q(t; \widehat{\boldsymbol{\alpha}})\}$. Here q is chosen to be a polynomial in t of degree nine with estimated parameter vector $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^{10}$.

3.1 *Model Components*

For details on the following model components we refer to Erhardt (2013, Chapter 3).

Annual seasonality. Yearly temperature fluctuations can be captured by sine curves of the form $\lambda \sin(\omega t + \delta)$, parametrized by λ (amplitude), ω (angular frequency) and δ (phase shift). A substitution of these parameters, inspired by Simmons (1990), leads to the linear model component $\beta_{\text{s}} \sin(\omega t) + \beta_{\text{c}} \cos(\omega t)$, with ω set to $2\pi/365.25$, due to the annual context.

Autoregression. Temporal dependence is eliminated using an autoregression component $\sum_{j=1}^{q} \gamma_j Y_{t-j}$ in the marginal model. Investigations show that the choice of $q = 3$ lagged responses as additional covariates is appropriate.

Skew-t distributed errors. Detailed investigations showed that skew-t distributed errors $\varepsilon_1, \dots, \varepsilon_N \overset{\text{i.i.d.}}{\sim} \text{skew-}t\,(\xi, \omega, \alpha, \nu)$ appropriately capture the observed skewness and heavy tails. The parametrization of Azzalini and Capitanio (2003) is utilized. In particular the probability

density function of the errors is given by

$$f_{\text{skew-}t}\left(x;\xi,\omega,\alpha,\nu\right)=\frac{2}{\omega}t_{\nu}(\widetilde{x})T_{\nu+1}\left\{\alpha\widetilde{x}\left(\frac{\nu+1}{\nu+\widetilde{x}^{2}}\right)^{\frac{1}{2}}\right\}, \tag{3}$$

where $\widetilde{x}\coloneqq(x-\xi)/\omega$. Here t_ν is the density and $T_{\nu+1}$ the cumulative distribution function of a univariate Student-t distribution with ν and $\nu+1$ degrees of freedom, respectively. The parameters ξ, ω and α can be interpreted as location, scale and shape parameter, respectively.

Aggregated parameters. The parameters of the previously described model components are replaced by polynomial structures to account for spatial variation in the temperatures depending on longitude (x^s_{lo}), latitude (x^s_{la}) and elevation (x^s_{el}). We call them aggregated or spatially varying parameters.

3.2 *The Marginal Model*

The marginal model for the daily mean temperatures is given as

$$\widetilde{Y}^s_t=\mu^s_t+\varepsilon^s_t,\quad \varepsilon^s_t\sim\text{skew-}t\left(\xi^s,\omega^s,\alpha^s,\nu^s\right),\quad t=1,\ldots,N,\ s=1,\ldots,d, \tag{4}$$

with mean function $\mu^s_t\coloneqq g\left(t,\widetilde{Y}^s_{t-1},\widetilde{Y}^s_{t-2},\widetilde{Y}^s_{t-3},x^s_{\text{el}},x^s_{\text{lo}},x^s_{\text{la}};\boldsymbol{\beta}\right)\coloneqq\beta^s_0+\beta^s_{\text{s}}\sin\left(2\pi t/365.25\right)+\beta^s_{\text{c}}\cos\left(2\pi t/365.25\right)+\gamma^s_1\widetilde{Y}^s_{t-1}+\gamma^s_2\widetilde{Y}^s_{t-2}+\gamma^s_3\widetilde{Y}^s_{t-3}$. The spatially varying parameters are divided into the *aggregated intercept and seasonality parameters*

$$\begin{aligned}
\beta^s_0&\coloneqq\beta_{00}+\beta_{011}x^s_{\text{el}}+\beta_{031}x^s_{\text{la}},\\
\beta^s_{\text{s}}&\coloneqq\beta_{\text{s}0}+\sum_{j=1}^{4}\beta_{\text{s}1j}\left(x^s_{\text{el}}\right)^j+\beta_{\text{s}21}x^s_{\text{lo}}+\sum_{l=1}^{6}\beta_{\text{s}3l}\left(x^s_{\text{la}}\right)^l,\\
\beta^s_{\text{c}}&\coloneqq\beta_{\text{c}0}+\sum_{j=1}^{6}\beta_{\text{c}1j}\left(x^s_{\text{el}}\right)^j+\sum_{k=1}^{2}\beta_{\text{c}2k}\left(x^s_{\text{lo}}\right)^k+\beta_{\text{c}31}x^s_{\text{la}},
\end{aligned}$$

the *aggregated autoregression parameters*

$$\begin{aligned}
\gamma^s_1&\coloneqq\gamma_{10}+\gamma_{111}x^s_{\text{el}}+\sum_{k=1}^{2}\gamma_{12k}\left(x^s_{\text{lo}}\right)^k+\sum_{l=1}^{6}\gamma_{13l}\left(x^s_{\text{la}}\right)^l,\\
\gamma^s_2&\coloneqq\gamma_{20}+\gamma_{211}x^s_{\text{el}}+\sum_{k=1}^{2}\gamma_{22k}\left(x^s_{\text{lo}}\right)^k+\sum_{l=1}^{6}\gamma_{23l}\left(x^s_{\text{la}}\right)^l,
\end{aligned}$$

$$\gamma_3^s \coloneqq \gamma_{30} + \sum_{k=1}^{4} \gamma_{32k} \left(x_{\text{lo}}^s\right)^k + \sum_{l=1}^{7} \gamma_{33l} \left(x_{\text{la}}^s\right)^l,$$

and the *aggregated skew-t parameters*

$$\begin{aligned}
\xi^s &\coloneqq \xi_0 + \xi_{11} x_{\text{el}}^s + \sum_{k=1}^{2} \xi_{2k} \left(x_{\text{lo}}^s\right)^k + \xi_{31} x_{\text{la}}^s, \\
\omega^s &\coloneqq \exp\left\{\omega_0 + \sum_{j=1}^{3} \omega_{1j} \left(x_{\text{el}}^s\right)^j + \omega_{21} x_{\text{lo}}^s + \sum_{l=1}^{6} \omega_{3l} \left(x_{\text{la}}^s\right)^l\right\}, \\
\alpha^s &\coloneqq \alpha_0 + \sum_{j=1}^{4} \alpha_{1j} \left(x_{\text{el}}^s\right)^j + \sum_{k=1}^{2} \alpha_{2k} \left(x_{\text{lo}}^s\right)^k + \alpha_{31} x_{\text{la}}^s, \\
\nu^s &\coloneqq \exp\left\{\nu_0 + \sum_{j=1}^{2} \nu_{1j} \left(x_{\text{el}}^s\right)^j + \sum_{k=1}^{2} \nu_{2k} \left(x_{\text{lo}}^s\right)^k + \sum_{l=1}^{4} \nu_{3l} \left(x_{\text{la}}^s\right)^l\right\},
\end{aligned}$$

parametrized by $\boldsymbol{\beta} \coloneqq \left(\boldsymbol{\beta}_0', \boldsymbol{\beta}_{\text{s}}', \boldsymbol{\beta}_{\text{c}}', \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2', \boldsymbol{\gamma}_3'\right)' \in \mathbb{R}^{57}$ and $\boldsymbol{\eta} \coloneqq \left(\boldsymbol{\xi}', \boldsymbol{\omega}', \boldsymbol{\alpha}', \boldsymbol{\nu}'\right)' \in \mathbb{R}^{33}$.

3.3 *Marginal Model Parameter Estimation*

Parameter estimation follows a two step approach. First, the parameters $\boldsymbol{\beta}$ are estimated using least-squares and the raw residuals $\widehat{\varepsilon}_t^s \coloneqq \widetilde{y}_t^s - \widehat{\mu}_t^s$, $t = 4, \ldots, N$, $s = 1, \ldots, d$, are calculated. They cannot be computed for $t = 1, 2, 3$, due to the autoregression of $\widetilde{y}_t^s$ on the three previous points in time. Maximization of the pseudo-likelihood $\mathcal{L}_{\text{skew-}t}\left(\boldsymbol{\eta} \,|\, \widehat{\boldsymbol{\varepsilon}}^1, \ldots, \widehat{\boldsymbol{\varepsilon}}^d\right) = \prod_{s=1}^{d} \prod_{t=4}^{N} f_{\text{skew-}t}\left(\widehat{\varepsilon}_t^s; \xi^s, \omega^s, \alpha^s, \nu^s\right)$ in a second step leads to estimates of the skew-t parameters $\boldsymbol{\eta}$. This results in the vector $\widehat{\boldsymbol{\theta}} \coloneqq \left(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\eta}}'\right)'$ of parameter estimates for Model (4).

3.4 *Transformation to Copula Data*

Finally we use the fitted Model (4) to transform our data to copula data, i.e. we transform our original time series $y_1^s, \ldots, y_N^s$, to $u_4^s, \ldots, u_N^s \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$ for all $s = 1, \ldots, d$. Since for all $s = 1, \ldots, d$ we modeled the errors $\varepsilon_1^s, \ldots, \varepsilon_N^s$ as i.i.d. skew-t distributed with spatially varying parameters ξ^s, ω^s, α^s and ν^s, the desired copula data is obtained as $u_t^s \coloneqq F_{\text{skew-}t}\left(\widehat{\varepsilon}_t^s \,|\, \widehat{\xi}^s, \widehat{\omega}^s, \widehat{\alpha}^s, \widehat{\nu}^s\right), t = 4, \ldots, N$, $s = 1, \ldots, d$, where $F_{\text{skew-}t}\left(\cdot \,|\, \xi, \omega, \alpha, \nu\right)$ is the cumulative distribution function corresponding to (3).

4. A Spatial R-vine Model for Daily Mean Temperatures

As climatic data such as temperature is measured at a large number of spatial locations, we face a high dimensional problem. With rising dimensionality ordinary R-vine copula models become computationally infeasible since the number of parameters increases quadratically. Exploitation of spatial information in our new approach of a *spatial R-vine copula model (SV)* allows to reduce the number of parameters significantly.

4.1 *Preliminary Analyses*

To develop a spatial R-vine model we consider the copula data $\boldsymbol{u}^1, \ldots, \boldsymbol{u}^d$ where $\boldsymbol{u}^s = (u_1^s, \ldots, u_N^s)'$ and $u_1^s, \ldots, u_N^s \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$ for all $s = 1, \ldots, d$, i.e. we have copula data series of length N for d different observation stations. From the quantities elevation, longitude and latitude we calculate the *distance* $D_{i,j}$ and the *elevation difference* $E_{i,j}$ between each pair of observation stations (i,j) with $1 \leqslant i < j \leqslant d$.

We allow for one- and two-parametric pair-copula families, whose first and second copula parameters are called $\theta_{i,j;\mathcal{D}_e}$ and $\nu_{i,j;\mathcal{D}_e}$. The corresponding Kendall's τ's are denoted by $\tau_{i,j}$ respectively $\tau_{i,j;\mathcal{D}_e}$, depending on whether they are calculated directly from the data or based on transformed variables in the trees $\mathcal{T}_2, \mathcal{T}_3, \ldots, \mathcal{T}_{d-1}$ of the R-vine.

Returning to the mean temperature data set ($d = 54$) we further investigate the spatial dependencies of the given variables. We are interested in identifying a relationship between the dependence strength on the one hand and distance and elevation difference of station pairs on the other hand. For all $d(d-1)/2 = 1431$ possible station pairs (i,j), $1 \leqslant i < j \leqslant 54$, the empirical Kendall's τ values $\widehat{\tau}_{i,j}$ are estimated, to quantify the dependence of these pairs. Since they are restricted to $(-1,1)$ we apply the Fisher z-transform

$$g_z(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right), \; r \in (-1,1), \tag{5}$$

first introduced by Fisher (1915), to transform to $(-\infty, \infty)$. The left panel of Figure 2

illustrates the Fisher z-transformed estimated Kendall's τ's against the logarithmized distances $\ln(D_{i,j})$. A distinct linear relationship can be observed. The right panel gives the corresponding plot against the logarithmized elevation differences $\ln(E_{i,j})$. The observed linear relationship is not as strong as for the distances.

[Figure 2 about here.]

The tree-wise analysis of an R-vine model fitted to the mean temperature data gives deeper insight into the relationship of the R-vine copula parameters and the available spatial information. We allow for bivariate Gaussian (Φ), Student-t (t), Clayton (C), Gumbel (G) and Frank (F) copulas as pair-copulas. Additionally rotated versions of the Clayton and Gumbel copula are utilized to capture possible negative and asymmetric dependencies. The copula families are selected separately for each bivariate building block according to the Akaike information criterion. For more details on these copula families, copula rotation and copula selection we refer to Brechmann and Schepsmeier (2013). The R-vine tree structure is selected by tree-wise selection of maximum spanning trees, where Kendall's τ's are used as edge weights (see Dißmann et al., 2013). Application of a bivariate asymptotic independence test (Genest and Favre, 2007) in the copula family selection procedure yields a share of independence copulas of more than 50% in all trees $\mathcal{T}_l$ with $l \geqslant 10$. Thus, a truncation after tree $\mathcal{T}_{10}$ results in a significant reduction in the number of model parameters.

Subsequently we add a superscript $l \leqslant 10$ to emphasize the corresponding tree number. Table 1 summarizes the structure of the R-vine which we will investigate in more detail. The copula family which occurs most often is the bivariate Student-t copula. It is the only two-parametric copula family under consideration. The number of other copula families increases with the tree number. In tree $\mathcal{T}_{10}$ the Gumbel family dominates. Further we observe from Table 1, that the strong dependencies are already captured in tree $\mathcal{T}_1$ and that the associations in higher trees vary mostly between -0.2 and 0.3, i.e. negative dependencies

occur as well. Figure 3 shows the logarithmized estimated degrees of freedom parameters $\ln\left\{\widehat{\nu}^l_{i(e),j(e);\mathcal{D}_e}\right\}$ of the Student-t copulas plotted against the respective tree number l. We discover a quadratic trend (dashed gray line) with regard to the tree number. This finding will be used to model the second copula parameters $\nu^l_{i(e),j(e);\mathcal{D}_e}$ jointly for all trees $\mathcal{T}_l$, $l = 1, \ldots, 10$.

[Table 1 about here.]

[Figure 3 about here.]

It remains to study the relationships between the first copula parameters $\theta^l_{i(e),j(e);\mathcal{D}_e}$ and the corresponding distances $D_{i(e),j(e)}$ and elevation differences $E_{i(e),j(e)}$, distinguishing which tree $\mathcal{T}_l$, $l \leqslant 10$, the edge e belongs to. Depending on the copula family $b_{i(e),j(e);\mathcal{D}_e}$ there exists a known relationship

$$\tau^l_{i(e),j(e);\mathcal{D}_e} = g_\tau\left\{\theta^l_{i(e),j(e);\mathcal{D}_e}; b_{i(e),j(e);\mathcal{D}_e}\right\}, \tag{6}$$

between the copula parameter $\theta^l_{i(e),j(e);\mathcal{D}_e}$ and the Kendall's τ $\tau^l_{i(e),j(e);\mathcal{D}_e}$. Hence, we study separately for each tree relationships between the Fisher z-transformed Kendall's τ's and the distances and elevation differences. A similar modeling approach was already followed by Gräler and Pebesma (2011). For the purpose of the tree-wise analysis we define the average distances and elevations $\overline{D_{i(e),\mathcal{D}_e}} := \frac{1}{l-1}\sum_{k\in\mathcal{D}_e} D_{i(e),k}$, $\overline{D_{j(e),\mathcal{D}_e}} := \frac{1}{l-1}\sum_{k\in\mathcal{D}_e} D_{j(e),k}$, $\overline{E_{i(e),\mathcal{D}_e}} := \frac{1}{l-1}\sum_{k\in\mathcal{D}_e} E_{i(e),k}$ and $\overline{E_{j(e),\mathcal{D}_e}} := \frac{1}{l-1}\sum_{k\in\mathcal{D}_e} E_{j(e),k}$, for all edges $e \in \mathcal{E}_l$ of trees $\mathcal{T}_l$ with $l > 1$, where the conditioning set $\mathcal{D}_e$ is non-empty, and consider them as potential predictors in our models. For details on the tree-wise analysis we refer the reader to Chapter 5 of Erhardt (2013). It shows that in general the distance based predictors capture more dependence information than the elevation based ones and that the direct unconditioned distances $D_{i(e),j(e)}$ are suited best to model the corresponding Kendall's τ.

4.2 *Model Formulation and Selection*

Our preliminary analyses suggest first copula parameter model specifications of the form

$$\theta^l_{i(e),j(e);\mathcal{D}_e} \coloneqq g_\tau^{-1}\left[g_z^{-1}\left\{h_l(e|\boldsymbol{\beta}_l)\right\}; b_{i(e),j(e);\mathcal{D}_e}\right], \quad e \in \mathcal{E}_l, \quad l = 1,\dots,10. \tag{7}$$

The inclusion of different combinations of the available spatial predictors $D_{i(e),j(e)}$, $E_{i(e),j(e)}$, $\overline{D_{i(e),\mathcal{D}_e}}$, $\overline{E_{i(e),\mathcal{D}_e}}$, $\overline{D_{j(e),\mathcal{D}_e}}$ and $\overline{E_{j(e),\mathcal{D}_e}}$ into the model is controlled by the model function $h_l\left(e|\boldsymbol{\beta}_l\right)$, $e \in \mathcal{E}_l$, $l = 1,\dots,10$, which is linear in the logarithmized predictors. A tree-wise comparison of different model specifications in Tables 5.4-5.7 and Figure 5.3.1 in Erhardt (2013) led to the selection of a model, which includes all available distance based predictors. The investigations showed, that an inclusion of the elevation based predictors does not lead to a significant improvement in terms of explanatory power. The model function $h_l(e|\boldsymbol{\beta}_l)$ of the distance model specification is defined tree-wise. For the first tree $\mathcal{T}_1$ it is defined as

$$h_1\left(e|\boldsymbol{\beta}_1\right) \coloneqq \beta_{1,0} + \beta_{1,1}\ln\left(D_{i(e),j(e)}\right), \quad e \in \mathcal{E}_1, \tag{8}$$

with $\boldsymbol{\beta}_1 = (\beta_{1,0}, \beta_{1,1})' \in \mathbb{R}^2$. For all trees $\mathcal{T}_l$, $l = 2,\dots,10$, the model function is given as

$$h_l\left(e|\boldsymbol{\beta}_l\right) \coloneqq \beta_{l,0} + \beta_{l,1}\ln\left(D_{i(e),j(e)}\right) + \beta_{l,2}\ln\left(\overline{D_{i(e),\mathcal{D}_e}}\right) + \beta_{l,3}\ln\left(\overline{D_{j(e),\mathcal{D}_e}}\right), \quad e \in \mathcal{E}_l, \tag{9}$$

with parameters $\boldsymbol{\beta}_l \coloneqq (\beta_{l,0}, \beta_{l,1}, \beta_{l,2}, \beta_{l,3})' \in \mathbb{R}^4$. We summarize the parameters for all trees as $\boldsymbol{\beta}^{\text{SV}}_{\text{dist}} \coloneqq (\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_{10}')' \in \mathbb{R}^{38}$. Moreover, Figure 3 suggests a quadratic model specification

$$\nu^l_{i(e),j(e);\mathcal{D}_e} \coloneqq \exp\left(\beta^\nu_0 + \beta^\nu_1 l + \beta^\nu_2 l^2\right), \quad e \in \mathcal{E}_l,\ l = 1,\dots,10, \tag{10}$$

for the second copula parameters, where $\boldsymbol{\beta}^{\text{SV}}_\nu \coloneqq (\beta^\nu_0, \beta^\nu_1, \beta^\nu_2)' \in \mathbb{R}^3$.

4.3 *Model Fit*

Maximum-likelihood estimation. For parameter estimation, we now specify the likelihood corresponding to the selected model. Since the parameters $\theta^l_{i(e),j(e);\mathcal{D}_e}$ may change their sign during the numerical maximization of the log-likelihood, the corresponding copula family might have to change with respect to rotation. Using the model specifications (10) for the

degrees of freedom $\nu^l_{i(e),j(e);\mathcal{D}_e}$ and (7) for $\theta^l_{i(e),j(e);\mathcal{D}_e}$, the usual R-vine likelihood changes to

$$\mathcal{L}_{\text{SV}}\left(\boldsymbol{\beta}^{\text{SV}}_{\text{dist}}, \boldsymbol{\beta}^{\text{SV}}_{\nu} \mid \boldsymbol{u}^1, \ldots, \boldsymbol{u}^d\right) = \prod_{t=1}^{N} \prod_{l=1}^{10} \prod_{e \in \mathcal{E}_l} c_{i(e),j(e);\mathcal{D}_e}\left\{\widetilde{u}^{i(e)}_t, \widetilde{u}^{j(e)}_t; \theta^l_{i(e),j(e);\mathcal{D}_e}, \nu^l_{i(e),j(e);\mathcal{D}_e}\right\},$$

where the transformed variables are defined as $\widetilde{u}^{i(e)}_t := C_{i(e)|\mathcal{D}_e}\left\{u^{i(e)}_t \mid \boldsymbol{u}^{\mathcal{D}_e}_t\right\}$ and $\widetilde{u}^{j(e)}_t := C_{j(e)|\mathcal{D}_e}\left\{u^{j(e)}_t \mid \boldsymbol{u}^{\mathcal{D}_e}_t\right\}$ with $\boldsymbol{u}^{\mathcal{D}_e}_t := \{u^s_t : s \in \mathcal{D}_e\}$.

Sequential estimation. Since in higher dimensions full maximum likelihood estimation becomes computationally demanding, we suggest a sequential estimation approach (cp. Aas et al., 2009). Tree-wise maximization of $\prod_{e \in \mathcal{E}_l} c_{i(e),j(e);\mathcal{D}_e}\left\{\widetilde{u}^{i(e)}_t, \widetilde{u}^{j(e)}_t; \theta^l_{i(e),j(e);\mathcal{D}_e}, \nu^l_{i(e),j(e);\mathcal{D}_e}\right\}$ by looping through the trees $\mathcal{T}_l$, $l = 1, \ldots, 10$, yields the sequential (seq) parameter estimates $\widehat{\boldsymbol{\beta}}^{\text{SV}}_{\text{seq}} = \left(\widehat{\boldsymbol{\beta}}^{\text{SV}}_{\text{dist}}, \widehat{\boldsymbol{\beta}}^{\text{SV}}_{\nu}\right)' \in \mathbb{R}^{41}$. All results in the following sections are based on sequential parameter estimates. For the selection of suitable starting values for the optimization we refer to Subsection 5.3.2 of Erhardt (2013).

Results. We provide an illustration of the dependencies modeled by the spatial R-vine model. Web Figure 2 shows all 54 observation stations and all edges that occur in the ten trees of the fitted (spatial) R-vine model. The magnitude of association between station pairs is indicated through edge width and edge color. The thicker and darker the edges are, the higher is the corresponding estimated association. The Student-t copula degrees of freedom resulting from our (sequential) estimation are visualized in Figure 3 (dashed gray line). We conclude from the plot, that our model shows strong tail dependencies in the first trees, which get weaker with increasing tree number.

4.4 *Prediction*

Now we address the model based prediction of mean temperatures at new locations. We illustrate the introduced methodology using the validation data introduced in Section 3.

Methodology. Predictions based on the spatial R-vine model will be on the copula data level. Thus a back transformation to the original level of mean temperatures is needed,

which is based on the marginal models presented in Section 3. For technical details we refer to Web Appendix A. To predict mean temperatures at a new location s for an arbitrary time point t with corresponding copula data u_t^s, we need to specify the conditional distribution $C_{s|1,\dots,d}(u_t^s|u_t^1,\dots,u_t^d)$ of the variable u_t^s conditioned on $u_t^1,\dots,u_t^d$. Here $u_t^1,\dots,u_t^d$ is the copula data at time t given by the training data set. The model specifies the joint distribution of $u_t^1,\dots,u_t^d$, as an R-vine distribution. Therefore, access to the conditional distribution function $C_{s|1,\dots,d}(u_t^s|u_t^1,\dots,u_t^d)$ can be achieved by extending the underlying spatial R-vine by one further vertex s.

Since the structure of the modeled R-vine should be preserved, we add the new variable as a leaf to the first R-vine tree. We estimate the Kendall's τ's $\tau_{i(e_1),j(e_1)}$ for all d edges $e_1 = \{i(e_1), j(e_1)\} = \{r, s\}$, $r = 1,\dots,d$, which may be added, by $\widehat{\tau}_{i(e_1),j(e_1);\mathcal{D}_{e_1}} \coloneqq \widehat{\tau}_{i(e_1),j(e_1)} = \widehat{\tau}_{r,s} = g_z^{-1}\left\{h_1\left(e_1 = \{r,s\}|\widehat{\boldsymbol{\beta}}_1\right)\right\}$. Here the conditioning set $\mathcal{D}_{e_1}$ is the empty set, h_1 is the model function defined in (8) and g_z is given by (5). The edge e_1^* which yields the biggest Kendall's τ estimate is selected to extend the first R-vine tree. A corresponding copula family $b_{i(e_1^*),j(e_1^*)}$ has to be selected. We select the family which occurs most often in the original R-vine, however other selection criteria might be chosen. Then the corresponding first copula parameter $\widehat{\theta}^1_{i(e_1^*),j(e_1^*)} = \widehat{\theta}^1_{i(e_1^*),j(e_1^*);\mathcal{D}_{e_1^*}}$ is estimated by

$$\widehat{\theta}^1_{i(e_1^*),j(e_1^*);\mathcal{D}_{e_1^*}} = g_\tau^{-1}\left\{\widehat{\tau}_{i(e_1^*),j(e_1^*);\mathcal{D}_{e_1^*}}; b_{i(e_1^*),j(e_1^*)}\right\} \tag{11}$$

using (6). If needed the second copula parameter $\widehat{\nu}^1_{i(e_1^*),j(e_1^*);\mathcal{D}_{e_1^*}} = \widehat{\nu}^1_{i(e_1^*),j(e_1^*)}$ is estimated by

$$\widehat{\nu}^1_{i(e_1^*),j(e_1^*);\mathcal{D}_{e_1^*}} = h_\nu\left(e_1^*, 1|\widehat{\boldsymbol{\beta}}_\nu^{\mathrm{SV}}\right), \tag{12}$$

where the function $h_\nu\left(e, l|\boldsymbol{\beta}_\nu^{\mathrm{SV}}\right)$, which depends on the respective edge e and tree number l and is parametrized by $\boldsymbol{\beta}_\nu^{\mathrm{SV}}$, represents the model specification for the second copula parameters (see Equation (10)).

The remainder of the R-vine is extended tree-wise starting from tree $\mathcal{T}_2$. For each tree $\mathcal{T}_l$ we

have to ensure that the proximity condition is fulfilled after a new edge e_l has been added. For all edges e_l with $j(e_l) = s$ and $\mathcal{D}_{e_l} = \mathcal{D}_{e^*_{l-1}} \cup i(e^*_{l-1})$ which fulfill the proximity condition, we estimate the corresponding Kendall's τ's using (9) by $\widehat{\tau}_{i(e_l),j(e_l);\mathcal{D}_{e_l}} = g_z^{-1}\left\{h_l\left(e_l|\widehat{\boldsymbol{\beta}}_l\right)\right\}$. Again, the edge e_l^* with the biggest Kendall's τ estimate is selected and included into the R-vine and a copula family $b_{i(e_l^*),j(e_l^*);\mathcal{D}_{e_l^*}}$ has to be selected. The corresponding parameters $\theta^l_{i(e_l^*),j(e_l^*);\mathcal{D}_{e_l^*}}$ and $\nu^l_{i(e_l^*),j(e_l^*);\mathcal{D}_{e_l^*}}$ are estimated in analogy to (11) and (12), respectively. For trees exceeding the truncation level $k < d$, arbitrary edges which fulfill the proximity condition can be chosen. The copulas corresponding to these edges are selected to be independence copulas. Thus, no parameters have to be specified for these copulas.

The above procedure yields an R-vine copula specification corresponding to the variables $u_t^s, u_t^1, \ldots, u_t^d$ with distribution function $C(u_t^s, u_t^1, \ldots, u_t^d)$. Using the recursion (2), we can calculate $C_{s|1,\ldots,d}(u_t^s|u_t^1, \ldots, u_t^d)$ iteratively. Thus, we are able to simulate from the predictive distribution $C_{s|1,\ldots,d}(u_t^s|u_t^1, \ldots, u_t^d)$ using the probability integral transform. We simulate $v \sim \mathcal{U}(0,1)$ and use $\breve{u}_t^s := C^{-1}_{s|1,\ldots,d}(v|u_t^1, \ldots, u_t^d)$ as a simulated copula data point at location s and time t. After a back transformation of the copula data $\breve{u}_t^s$ to the level of the originally modeled data $\breve{y}_t^s$, point predictions $\widehat{y}_t^s$ can be calculated as the mean of the simulated $\breve{y}_t^s$.

Omitting all arguments, the prediction density $c_{s|1,\ldots,d}$ corresponding to $C_{s|1,\ldots,d}$ can be obtained by decomposing its numerator $c_{s,1,\ldots,d}$ and denominator $c_{1,\ldots,d}$ according to Equation (1) into products of pair-copulas. Since the R-vine copula specification corresponding to $c_{s,1,\ldots,d}$ differs from the one corresponding to $c_{1,\ldots,d}$ only in terms of the additional edges $e_1^*, \ldots, e_{d-1}^*$, it holds $j(e_l^*) = s$ by construction and we truncate at level $k < d$, we obtain $c_{s|1,\ldots,d} = \prod_{l=1}^{k} c_{i(e_l^*),s;\mathcal{D}_{e_l^*}}\left\{C_{i(e_l^*)|\mathcal{D}_{e_l^*}}, C_{s|\mathcal{D}_{e_l^*}}; \widehat{\theta}^l_{i(e_l^*),s;\mathcal{D}_{e_l^*}}, \widehat{\nu}^l_{i(e_l^*),s;\mathcal{D}_{e_l^*}}\right\}$.

For the mean temperature data we perform the above calculations based on the distance model specification (7) and on the model specification (10) for the second copula parameters. Due to our previous investigations on the structure of the R-vine underlying the spatial R-

vine model (see Table 1) we select a Student-t copula for every edge which is added to the truncated R-vine. The resulting predictions of the 19 mean temperature time series of the validation data set are based on 1000 simulations of each time series.

Results. We select the stations *Grambek* (67) and *Arkona* (56) as representatives for a detailed analysis. Their predictions are compared in Web Figure 3. For comparison we plotted the observed values in black and the prediction in gray. Moreover, the corresponding 95% prediction intervals are indicated by the light gray area around the point predictions. Whereas the predictions for *Grambek* are very close to the observed values and the prediction intervals are very narrow, we observe noticeable deviations for *Arkona*. There seems to be more uncertainty in the predictions for *Arkona*, which is reflected in the comparatively broad prediction intervals. This seems to be due to the special location of *Arkona* on an island in the Baltic Sea, where the temperatures might be exposed to several factors which are not included in our model. Web Figure 4 highlights the prediction errors for *Grambek* (67) and *Arkona* (56). For *Arkona* we observe systematic deviations from zero, which points to a misspecification of the seasonality parameters. A possible reason is that the latitude of *Arkona* lies outside the latitude range of our training data set. This first analysis of predictions from our spatial R-vine model illustrates the prediction capabilities and limitations of our model. We see a good performance, as long as we predict within the range of the training data. However, our marginal model is not able to capture the temperature trends of stations which lie outside of this range.

5. A Spatial Gaussian Model for Daily Mean Temperatures

For comparison we introduce a *spatial Gaussian model* (SG). As before let $\widetilde{Y}_t^s$ be a real valued random variable, which represents the (weighted) mean temperature at a location s and a time point t and define $\widetilde{\boldsymbol{Y}}_t \coloneqq (\widetilde{Y}_t^1, \dots, \widetilde{Y}_t^d)' \in \mathbb{R}^d$ for all $t = 1, \dots, N$. Then we specify

a spatial Gaussian model by

$$\widetilde{\boldsymbol{Y}}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d\left\{\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{SG}})\right\}, \quad t = 1, \ldots, N,$$

where $\boldsymbol{\mu}_t \coloneqq (\mu_t^1, \ldots, \mu_t^d)' \in \mathbb{R}^d$ is a vector of means for all $t = 1, \ldots, N$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) \in \mathbb{R}^{d \times d}$ is a positive definite covariance matrix depending on a parameter vector $\boldsymbol{\theta}^{\text{SG}}$. The components of the mean vector $\boldsymbol{\mu}_t$ are modeled similarly to Equation (4). The spatial dependencies are determined by the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{SG}}) = \left\{\Sigma_{i,j}(\boldsymbol{\theta}^{\text{SG}})\right\}_{i,j=1,\ldots,d}$ which is based on the *Gaussian variogram model* (see for example Gelfand et al., 2010, Chapter 3) $\gamma(h; \eta, \varsigma, \rho) \coloneqq \varsigma\left\{1 - \exp\left(-\frac{h^2}{\rho^2}\right)\right\} + \eta \mathbb{1}_{(0,\infty)}(h)$. Then the variance is given as $\sigma^2 = \lim_{h\to\infty} \gamma(h; \eta, \varsigma, \rho) = \eta + \varsigma$ and we model $\Sigma_{i,j}(\boldsymbol{\theta}^{\text{SG}}) \coloneqq \sigma^2 - \gamma(D_{i,j}; \eta, \varsigma, \rho)$, where $\boldsymbol{\theta}^{\text{SG}} = (\eta, \varsigma, \rho)'$. Here $D_{i,j}$ are the distances between the station pairs (i, j), $i, j = 1, \ldots, d$. We implicitly make a stationarity assumption. Parameter estimation and prediction for the spatial Gaussian model are outlined in Web Appendix B.

Comparing the spatial R-vine and the spatial Gaussian model we use the same mean function, however the distribution of the residuals is modeled differently. In the case of the spatial R-vine model we utilize skew-t marginals and an R-vine copula compared to Gaussian marginals and a Gauss copula for the spatial Gaussian model.

6. Model Validation and Comparison

For model comparison we determine (negatively oriented) continuous ranked probability scores (CRPS) (see Gneiting and Raftery, 2007, Section 4.2). Negatively oriented means that smaller scores indicate a better fit. The scores will allow for an adequate comparative model validation. In the following we consider averaged continuous ranked probability scores (Table 2, $\overline{\text{CRPS}}$), percentaged model outperformance (Table 2, %) and a new concept called log-score difference plots (Figure 4). We additionally study a spatial R-vine model allowing only for Gaussian pair-copulas. It will be called *Gaussian spatial R-vine model* (GSV).

Averaged scores. To see which model provides better predictions we compare the averaged continuous ranked probability scores ($\overline{\text{CRPS}}$) in Table 2, where we average over time. Consideration of the averaged scores yields a preference for the spatial R-vine models.

[Table 2 about here.]

Percentaged outperformance. Furthermore, Table 2 compares the models based on percentaged outperformance. For all stations in the validation data set we count for how many points in time one model yields a lower score than the other. For more than two thirds of all stations of the validation data set and for a share of more than 60% of all temperature predictions under consideration we observe an outperformance of the spatial R-vine models over the spatial Gaussian model (see SV $\overset{\%}{\succ}$ SG and GSV $\overset{\%}{\succ}$ SG). We see, that the separate modeling of marginal distributions and dependency structure contributes to a distinct improvement in the model fit and prediction capabilities. Moreover, the modeling of non-Gaussian dependencies (see SV $\overset{\%}{\succ}$ GSV) yields further improvement.

Log-score difference plots. It is possible that the model outperformance depends on the time, i.e. there may be time intervals in which one model yields better results than the other. To be able to detect such time dependencies we consider Figure 4. We call these plots log-score difference plots, since they plot the difference of the (negatively oriented) log-scores of two models against the corresponding time points. The figure shows log-score difference plots of the continuous ranked probability scores averaged over all 19 stations of the validation data set, comparing the spatial R-vine model (SV) to the spatial Gaussian model (SG) and to the Gaussian spatial R-vine model (GSV), respectively. For both plots we observe similar temporal patterns and time intervals towards the end of each year, where the models assuming a Gaussian dependency structure consistently yield lower scores than the spatial R-vine model. Overall consideration of the remainder of the year however shows preference of the spatial R-vine model over the spatial Gaussian model.

[Figure 4 about here.]

7. Discussion

An extensive analysis of an ordinary (truncated) R-vine copula fitted to the training data led to a new model for spatial dependencies, the spatial R-vine model. The investigation of relationships between Kendall's τ's occurring in the R-vine copula and the associated distances and elevation differences proposed different tree-wise model specifications for the first pair-copula parameters. We found that the explanatory power of the elevation differences is comparatively small, whereas the station distances are able to explain the respective dependencies to a large extent. Therefore, we selected a model accounting for all distances between the observation stations, which are associated with the corresponding bivariate copulas of the R-vine copula specification. Moreover, a model specification for the second copula parameters needed for the large share of Student-t copulas was applied to reduce the necessary number of parameters further. This resulted in the modeling of strong tail dependencies in the lower trees, which distinguishes our spatial R-vine model from classical Gaussian approaches.

All in all the selected model specifications led to a distinct reduction in the number of parameters. In the case of our example data set, the 733 parameters needed in the original truncated R-vine copula model could be replaced by 41 parameters in the spatial R-vine model. This reduction is also mirrored in the computation time for the full maximum likelihood estimation in both models. Whereas the estimation for the truncated R-vine took about 3.7 days, this time could be reduced to 14.3 hours for our spatial R-vine model and only 28 minutes in the case of the sequential estimation. The computations were performed on a 2.6 GHz AMD Opteron processor. To get an idea about the scale of the computation time for longer time periods (N) and for more spatial locations (d), we performed sequential estimation on subsets of the training data (copula data). Web Figure 7 plots the computation

time (in minutes) for $14, 18, \ldots, 50$ and 54 spatial locations and (a) 365, (b) 730 and (c) 1093 points in time, respectively. The black line indicating the average over the different temporal scales shows an approximately linear trend of the computation time. Therefore we conclude that sequential estimation is also applicable in much higher dimensions.

For comparison we introduced a spatial Gaussian model, which requires only three parameters. Our aim was it to show that our new approach yields better predictions, which will justify a longer computation time. A validation of the prediction results from both models in terms of continuous ranked probability scores (CRPS) yielded reasonable accuracy of our predictions, as long as the location from which we aimed to predict lay within the range of the training data. Overall consideration of the scores showed an outperformance in 63% of all considered points in time. Transformation of the maximum log-likelihood of the truncated and the spatial R-vine model to the residual level on which the spatial Gaussian model is built, allows model comparison. In particular we obtain maximum log-likelihoods (residual level) of -42515.23, -46282.46 and -49099.16 for the truncated R-vine model, the spatial R-vine model and the spatial Gaussian model, respectively. These values show a clear preference of the spatial R-vine model over the spatial Gaussian model.

With regard to future work on vine copula based models for spatial dependencies an application of our modeling approach to other types of data sets is desirable. Especially data sets where asymmetries of bivariate dependencies are observed should be in the focus of further research. Moreover, further improvement might be achieved by the inclusion of further covariates. Covariates of interest may be microclimatic variates like *urban/rural area*, *closeness to body of water* or *wind force*.

8. Supplementary Materials

Web Appendices A and B, referenced in Section 4.4 and Section 5, Web Table 1 referenced in Section 3, the Web Figures referenced in Sections 2, 4.3, 4.4 and 7, along with a soft-

ware package implementing the presented methodology are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors gratefully acknowledge the helpful comments of the referees, who further improved the manuscript. The first author likes to thank the TUM Graduate School's Graduate Center International Graduate School of Science and Engineering (IGSSE) for support. The numerical computations were performed on a Linux cluster supported by DFG grant INST 95/919-1 FUGG.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* **44,** 182–198.

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

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence* **32,** 245–268.

Bedford, T. and Cooke, R. M. (2002). Vines - a new graphical model for dependent random variables. *The Annals of Statistics* **30,** 1031–1068.

Brechmann, E. C., Czado, C., and Aas, K. (2012). Truncated regular vines in high dimensions with applications to financial data. *Canadian Journal of Statistics* **40,** 68–85.

Brechmann, E. C. and Schepsmeier, U. (2013). Modeling dependence with C- and D-Vine Copulas: The R package CDVine. *Journal of Statistical Software* **52,** 1–27.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, Lecture Notes in Statistics, pages 93–109. Berlin: Springer.

Czado, C., Brechmann, E. C., and Gruber, L. (2013). Selection of vine copulas. In Jaworski, P., Durante, F., and Härdle, W. K., editors, *Copulae in Mathematical and Quantitative Finance*. Springer.

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis* **59,** 52–69.

Erhardt, T. M. (2013). Predicting temperature time series using spatial vine copulae. Master's thesis, Technische Universität München. http://mediatum.ub.tum.de/node?id=1173363.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population. *Biometrika* **10,** 507–521.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press: Boca Raton.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12,** 347–368.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82,** 543–552.

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

Gräler, B. and Pebesma, E. (2011). The pair-copula construction for spatial data: a new approach to model spatial dependency. *Procedia Environmental Sciences* **7,** 206–211.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In Rüschendorf, L., Schweizer, B., and Taylor, M. D., editors, *Distributions with fixed marginals and related topics*, volume 28 of *Lecture Notes - Monograph Series*, pages 120–141. Institute of Mathematical Statistics.

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

Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis* **51,** 2836–2850.

Kurowicka, D. and Cooke, R. (2006). *Uncertainty analysis with high dimensional dependence modelling*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd.

Kurowicka, D. and Joe, H. (2011). *Dependence Modeling: Vine Copula Handbook*. Singapore: World Scientific.

Schepsmeier, U., Stöber, J., Brechmann, E. C., and Gräler, B. (2014). *VineCopula: Statistical inference of vine copulas*. R package version 1.3.

Simmons, L. (1990). Time-series decomposition using the sinusoidal model. *International Journal of Forecasting* **6,** 485–495.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leures marges. In *Publications de l'Institut de Statistique de L'Université de Paris, 8*, pages 229–231. Institut Henri Poincaré.

Received March 2014. *Revised October* 2014. *Accepted November* 2014.

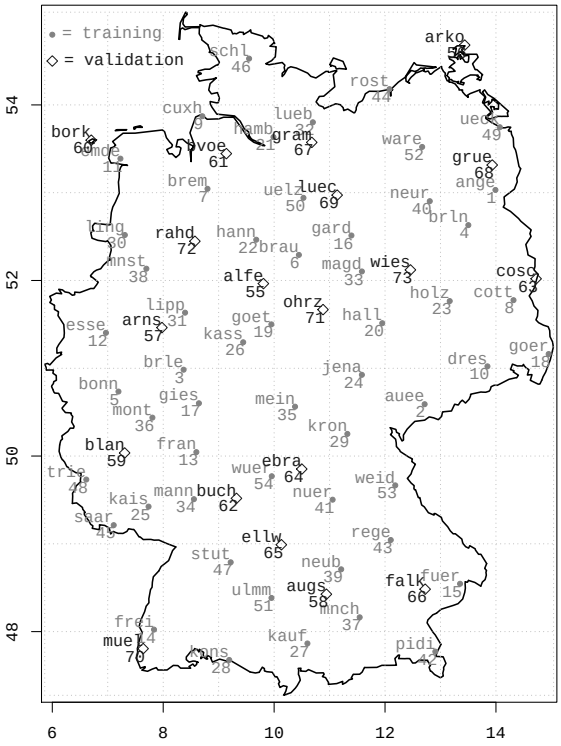

Figure 1. The 73 observation stations across Germany with ID and respective short name: Training data ($s = 1, \ldots, 54$) and validation data ($s = 55, \ldots, 73$).

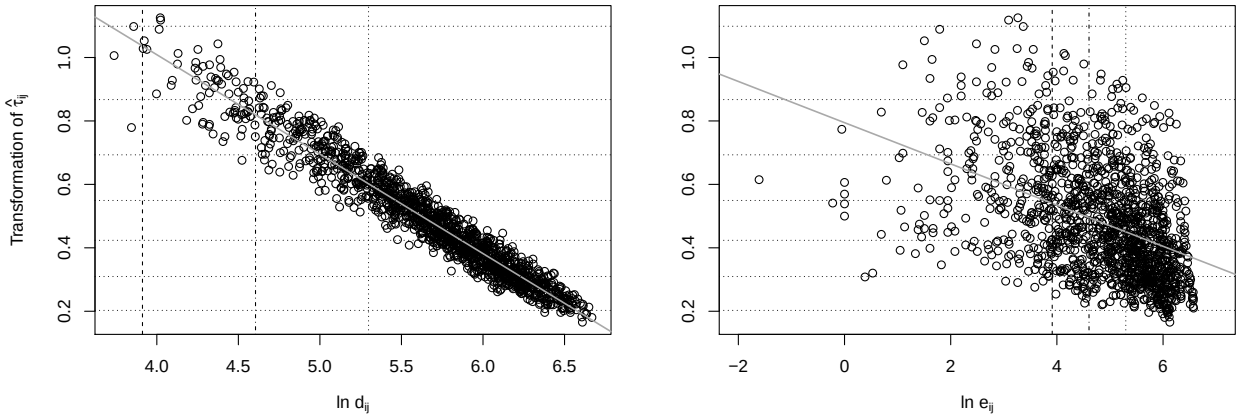

Figure 2. Relationship of Fisher z-transformed estimated Kendall's τ's $g_z(\widehat{\tau}_{i,j})$ with log-distance $\ln(D_{i,j})$ and log-elevation $\ln(E_{i,j})$, respectively. The straight gray lines depict the regression lines corresponding to the particular linear relationship. The horizontal lines help to identify the level of Kendall's τ, whereas the vertical lines indicate the three distances of 50, 100 and 200 kilometers and the three elevation differences of 50, 100 and 200 meters, respectively.

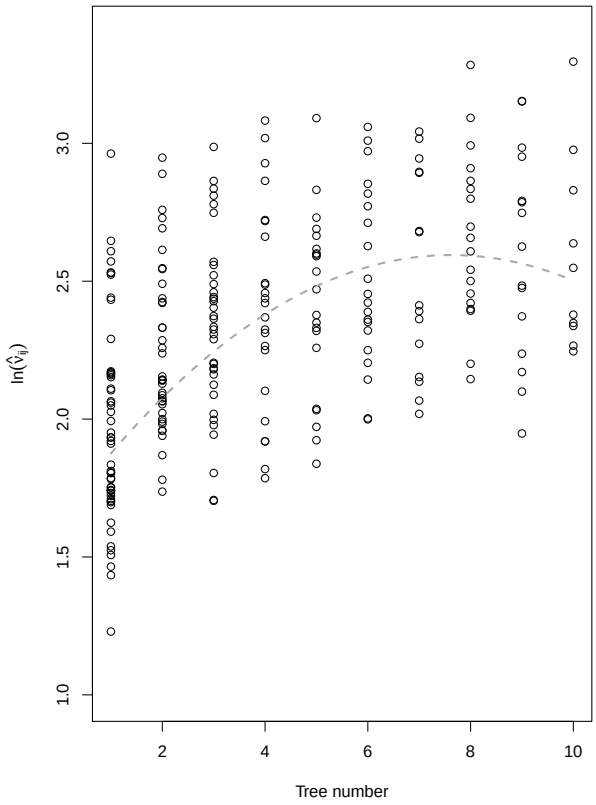

Figure 3. Plot of logarithmized estimated degree of freedom parameters $\ln\left\{\widehat{\nu}^{l}_{i(e),j(e);\mathcal{D}_e}\right\}$, for edges $e \in \mathcal{E}_l$ with Student-t copulas, against the respective tree number $l = 1, \ldots, 10$. The curve given by the model specification (10) using the parameters $\widehat{\boldsymbol{\beta}}^{\mathrm{SV}}_{\nu}$ estimated in Section 4.3 is indicated as a dashed line.

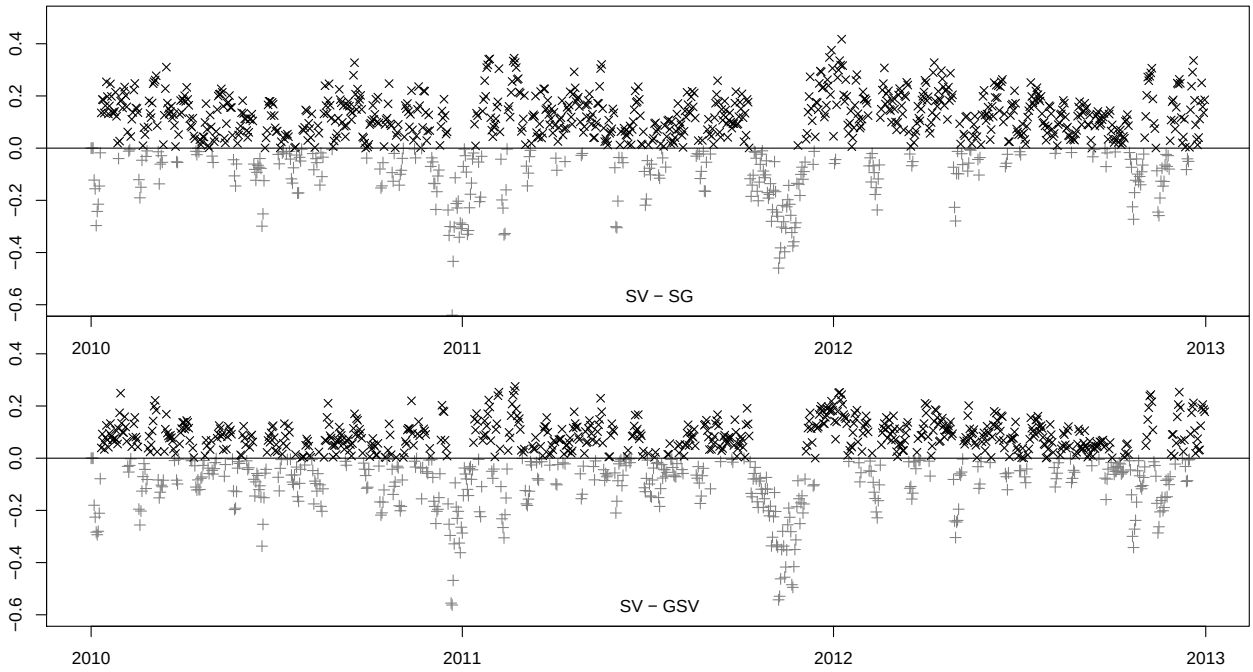

Figure 4. Log-score difference plots of the averaged continuous ranked probability scores comparing the spatial R-vine model (SV) and the Gaussian spatial R-vine model (GSV) to the corresponding averaged spatial Gaussian model (SG) scores (average over all 19 observation stations of the validation data set). Points in time where the first model has the lower average scores are marked by a black x. Points in time where the other model has the lower average scores are marked by a gray plus sign.

Table 1

Summary of the estimated structure of the truncated R-vine under consideration. Besides the numbers of the different copula families (Φ=Gaussian, t=Student-t, C=Clayton, G=Gumbel, F=Frank pair-copula) selected for each tree, the minimum and the maximum estimated Kendall's τ's and the averages over the occurring estimated second copula parameters ($\overline{\widehat{\nu}^l} := \frac{1}{\#t}\sum_{e\in\mathcal{E}_l^t}\widehat{\nu}^l_{i(e),j(e);\mathcal{D}_e}$, $\mathcal{E}_l^t := \{e \in \mathcal{E}_l : b_{i(e),j(e);\mathcal{D}_e}$ is a Student-t copula$\}$) are provided.

tree (l)	# Φ	# t	# C	# G	# F	$\min_{e\in\mathcal{E}_l}\{\widehat{\tau}^l_{i(e),j(e);\mathcal{D}_e}\}$	$\max_{e\in\mathcal{E}_l}\{\widehat{\tau}^l_{i(e),j(e);\mathcal{D}_e}\}$	$\overline{\widehat{\nu}^l}$
1	0	53	0	0	0	0.59	0.81	7.66
2	1	38	1	7	5	-0.15	0.32	9.91
3	1	35	4	6	5	-0.22	0.36	10.84
4	2	23	5	13	7	-0.18	0.30	11.87
5	1	21	9	9	9	-0.15	0.28	11.84
6	5	20	5	12	6	-0.11	0.29	12.97
7	6	15	10	10	6	-0.19	0.24	13.38
8	5	18	4	7	12	-0.10	0.19	14.89
9	7	15	6	8	9	-0.07	0.28	14.37
10	3	10	8	16	7	-0.13	0.25	14.10
Sum	**31**	**248**	**52**	**88**	**66**			

Table 2

Comparison of the averaged CRPS ($\overline{CRPS}$) of the spatial R-vine model (SV), the Gaussian spatial R-vine model (GSV) and the spatial Gaussian model (SG) and percentaged outperformance (%) in terms of CRPS over the period 01/01/2010 − 12/31/2012 *for the observation stations of the validation data set. Here we define $A \overset{\%}{\succ} B$ as the share of the points in time for which Approach A is preferred over Approach B in terms of CRPS.*

	short	$\overline{\text{CRPS}}$			%		
s	name	SV	GSV	SG	SV $\overset{\%}{\succ}$ GSV	SV $\overset{\%}{\succ}$ SG	GSV $\overset{\%}{\succ}$ SG
55	alfe	3.18	3.04	2.59	0.51	0.22	0.11
56	arko	3.17	3.22	3.42	0.61	0.68	0.69
57	arns	2.19	2.16	2.61	0.56	0.81	0.91
58	augs	2.65	2.59	2.56	0.54	0.51	0.52
59	blan	2.94	2.86	2.64	0.51	0.36	0.29
60	bork	2.26	2.37	3.25	0.66	0.94	0.96
61	bvoe	2.38	2.47	2.57	0.67	0.70	0.63
62	buch	2.50	2.51	2.61	0.58	0.65	0.63
63	cosc	2.60	2.68	2.83	0.65	0.69	0.65
64	ebra	2.33	2.31	2.59	0.59	0.74	0.79
65	ellw	3.15	2.94	2.59	0.45	0.24	0.21
66	falk	2.64	2.57	2.61	0.53	0.55	0.57
67	gram	1.80	1.93	2.55	0.72	0.93	0.93
68	grue	1.92	2.01	2.64	0.66	0.92	0.95
69	luec	2.25	2.31	2.57	0.64	0.77	0.78
70	muel	2.14	2.00	3.04	0.45	0.94	0.99
71	ohrz	3.68	3.53	2.59	0.50	0.06	0.02
72	rahd	2.38	2.49	2.65	0.69	0.74	0.66
73	wies	2.72	2.82	2.57	0.66	0.45	0.26
mean		**2.57**	2.57	2.71	0.59	**0.63**	0.61