Standardized drought indices:
A novel uni- and multivariate approach

by

Tobias M. Erhardt* and Claudia Czado
Zentrum Mathematik
Technische Universität München
Boltzmannstr. 3, 85748 Garching, Germany

August 27, 2015

Abstract

As drought is among the natural hazards which affects people and economies worldwide and often results in huge monetary losses sophisticated methods for drought monitoring and decision making are needed. Several different approaches to quantify drought have been developed during past decades. However, most of these drought indices suffer from different shortcomings and do not account for the multiple driving factors which promote drought conditions and their inter-dependencies. We provide a novel methodology for the calculation of (multivariate) drought indices, which combines the advantages of existing approaches and omits their disadvantages. Moreover, our approach benefits from the flexibility of vine copulas in modeling multivariate non-Gaussian inter-variable dependence structures. A three-variate data example is used in order to investigate drought conditions in Europe and to illustrate and reason the different modeling steps. The data analysis shows the appropriateness of the described methodology. Comparison to well-established drought indices shows the benefits of our multivariate approach. The validity of the new methodology is verified by comparing the spatial extent of historic drought events based on different drought indices. Further, we show that the assumption of non-Gaussian dependence structures is well-grounded in this real-world application.
Keywords: standardized drought indices, dependence modeling, drought modeling, vine copulas

---

*TUM International Graduate School of Science and Engineering (IGSSE)

1   Introduction

The challenging field of drought research has a long history. Scientists of different disciplines described and defined different drought concepts and tried to measure, quantify and predict drought events and their impacts. There exist several review papers trying to depict/portray the state of the art and different developments in drought modeling. One of the most recent and comprehensive ones is the review of drought concepts by Mishra and Singh (2010). They state that "drought is best characterized by multiple climatological and hydrological parameters". Different drought types like *meteorological drought* (lack of precipitation), *hydrological drought* (declining water resources), *agricultural drought* (lack of soil moisture), *socio-economic drought* (excess demand for economic good(s) due to shortfall in water supply) or *ground water drought* (decrease in groundwater recharge, levels and discharge) are driven by different variables/phenomena. Recently, there have been several attempts to develop *multivariate drought indicators* (see e.g. Kao and Govindaraju, 2010; Hao and AghaKouchak, 2013, 2014; Farahmand and AghaKouchak, 2015), combining at least two different variables. Subsequently, we motivate and present a statistically sound approach for the calculation of standardized uni- and multivariate drought indices for arbitrary (sets of) drought relevant variables. The multivariate indices use so called vine copulas to flexibly model the variable dependencies.

Copulas are explained best by Sklar's Theorem (Sklar, 1959). Let F be a multivariate (d-dimensional) distribution function and $F_1, \ldots, F_d$ the corresponding marginals. Then there exists a copula C, such that $F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d))$, where $\boldsymbol{x} = (x_1, \ldots, x_d)'$ is the realization of a (continuous) random vector $\boldsymbol{X} \in \mathbb{R}^d$. A copula itself is a d-dimensional distribution function on the unit hypercube with uniformly distributed margins. It captures all dependency information between the marginals of the corresponding multivariate distribution function. Vine copulas are d-dimensional copula constructions built on bivariate copulas only (see Aas et al., 2009; Dißmann et al., 2013). They allow very flexible modeling of non-Gaussian, asymmetric dependency structures due to their modularity.

The most popular drought indices are the Palmer Drought Severity Index (**PDSI**) (Palmer, 1965) respectively its self-calibrating version (**SC-PDSI**) (Wells et al., 2004) and the Standardized Precipitation Index (**SPI**) (McKee et al., 1993; Edwards and McKee, 1997). Drought indices in general should quantify deviations from normal conditions, i.e. they should take *seasonality* into account. Often *negative/small values reflect dry conditions and positive/high values wet conditions*. They usually *require long data records* to yield meaningful results.

The *PDSI* is calculated based on precipitation and temperature and assumes a simplifying water balance model (for details see Palmer, 1965). The major criticisms on the PDSI are its lack of applicability and comparability for different climatic regions. Some of its major shortcomings vanished with the *SC-PDSI*, whose parameters are determined based on local climatic conditions rather than on some fixed locations in the US, i.e. it allows for *spatial comparison*. One further criticism of the PDSI is its *autoregressive structure*. Present conditions depend on past conditions, however the time interval which influences the present varies across space but cannot be accessed from the model.

In contrast to the PDSI, other drought indices like the *SPI* (McKee et al., 1993; Ed-

wards and McKee, 1997) are of *probabilistic nature*. This allows risk analysis, classification and frequency analysis of drought events. Two advantages of the purely precipitation based SPI over the PDSI are its *standardization* (standard normal distribution of SPI values) and the concept of *time scales*, which allows to set the time interval which has an influence on the present (drought) conditions. The SPI methodology *can be applied to other variables* as well (see e.g. the Standardized Runoff Index of Shukla and Wood, 2008) and the standardization allows for *comparison of such standardized indices and across space and time*. A criticism is that the SPI assumes a *parametric distribution* to model the data. However, a good fit to the data (especially in the distribution tails) is never guaranteed and in fact is not possible for many locations (e.g. in the Sahara). Moreover, *temporal dependencies* in the data or those introduced through the time scale cause the fitting to be biased.

As an enhancement of the SPI the Standardized Precipitation Evapotranspiration Index (**SPEI**) (Vicente-Serrano et al., 2010) quantifies drought based on *multivariate input*. Instead of precipitation a climatic water balance (precipitation minus potential evapotranspiration) is considered to quantify dry/wet conditions. The SPEI allows for *trends* in the time series data such that these are passed on to the index (to include effects of climate change).

Kao and Govindaraju (2010) present a (to our knowledge) first multivariate copula-based drought index, the Joint Deficit Index (**JDI**). They apply it to precipitation and streamflow time-series, but application to other variables is possible. Marginals are modeled using the SPI approach. Empirical copulas are used to (non-parametrically) estimate the dependence structure of the marginals representing the different time scales of one to twelve months. Finally, the joint deficit index combines the drought information captured by different time scales using the Kendall distribution function to assess the joint probability. The results are transformed to a standard normal distribution. Note, that for meaningful estimation of empirical copulas long data records are required.

Farahmand and AghaKouchak (2015) introduce the Standardized Drought Analysis Toolbox (**SDAT**), with the aim to provide a generalized approach to derive non-parametric standardized drought indices. Based on precipitation and soil moisture time series, they present a multivariate approach to drought modeling. Enhancing the SPI idea to *bivariate data* (based on non-parametric estimation), a bivariate empirical distribution is fitted to the input data and the joint cumulative probability is transformed with the inverse CDF of a standard normal distribution. Note however, that *this approach doesn't yield a real standardization*. Usually negative values of the proposed index are more probable, since the joint cumulative probability is not uniformly distributed on $[0, 1]$.

Summarizing the lessons learned from the sophisticated drought indices revised above, we state that (univariate) drought indices should ...

PROBAB   be probabilistic (allow risk/frequency analysis and classification of drought events), i.e. no assumptions about the characteristics of the underlying system have to be made.

ARBVAR   be applicable to *arbitrary drought relevant variables*.

DRYWET   be negative/positive to indicate *dry/wet conditions*.

Table 1: Comparison of different drought indices (SC-PDSI, SPI/SPEI, JDI, SDAT) and their properties: + has this property, − doesn't have this property, ? no definite answer possible or not applicable (e.g. because the corresponding model is not probabilistic).

	SC-PDSI	SPI/SPEI	JDI	SDAT
PROBAB	−	+	+	+
ARBVAR	−	?	+	+
DRYWET	+	+	+	+
SMALLS	−	−	−	−
TRENDS	+	+	+	+
SEASON	+	+	+	−
TIMDEP	?	−	+	−
NPDIST	?	−	−	+
STCOMP	?	+	+	−
TSCALE	−	+	−	+
MULTEX	−	?	+	+

SMALLS yield meaningful results for (monthly) data records for 10 years and more (i.e. *minimum sample size* = 120).

TRENDS reflect *trends* in the input data.

SEASON model and eliminate *seasonality*.

TIMDEP model and eliminate *temporal dependencies* before a probability distribution is fitted.

NPDIST use *non-parametric distribution estimates* for the (transformed) underlying variable (better fit, computationally efficient).

STCOMP be *standardized* to enable comparison over space/time and with other indices.

TSCALE allow for computation/aggregation at different *time scales l*.

MULTEX be *extendable to multivariate input* (different types of drought).

Table 1 summarizes different drought indices and lists which characteristics they fulfill. Subsequently, we introduce a novel approach to drought modeling which addresses the above criteria step by step.

2   Data

For the purpose of application and illustration we utilize the publicly available Climatic Research Unit (CRU) time series (TS) data (version 3.21, see Jones and Harris, 2013), which is monthly climatic data on a 0.5°×0.5° (longitude × latitude) grid. We restrict this (model-calculated) data set to the area $(11°W, 32°E) \times (35°N, 71°N)$ covering most of Europe (see gray shaded area in Figure 1). This results in data for $S = 3380$ grid cells. For

the calculation of drought indices we use the variables *potential evapotranspiration* (PET), *precipitation* (PRE) and *vapor pressure deficit* (VPD) for the years 1961 to 2010 ($T = 600$ months). VPD is calculated based on mean temperature (TMP) and vapor pressure (VAP) as $\texttt{VPD} = \texttt{SVP} - \texttt{VAP}$, where $\texttt{SVP} = 6.1078 \cdot 10^{[(7.5 \cdot \texttt{TMP})/(\texttt{TMP}+237.3)]}$ is the saturated vapor pressure (see Murray, 1967). The five pixels C, N, E, SE and SW highlighted in Figure 1 are used for subsequent illustrations. Their coordinates are provided in the figure. Time series plots corresponding to these five locations of the variables PET, PRE, VPD as well as SPI and SPEI are provided in the supporting information (see Section 7.1).
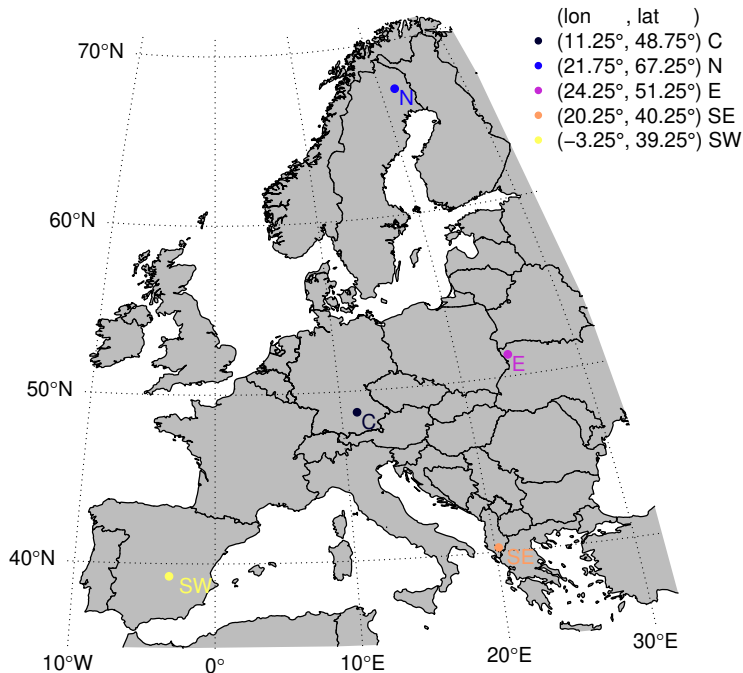

Figure 1: Study area and location of pixels used for illustration.

3 Univariate standardized indices

In a first step we seek to develop and illustrate a statistically sound (PROBAB) generalized modeling framework for (univariate) standardized indices. These indices should have the properties which were discussed in the introduction (Section 1).

3.1 Variable transformation

Let us now consider a time series x_{t_k}, $k = 1, \ldots, T$, for an arbitrary drought relevant variable (ARBVAR). Small values should always indicate dry and big values wet conditions (DRYWET). To ensure that, we change the sign of the time series beforehand if it was the other way round. Consider for instance potential evapotranspiration (PET). A high value corresponds to potentially high evaporation and transpiration, i.e. to dry conditions. For

low values the opposite is observed. Therefore we need to multiply the PET (and also the VPD) time series by -1.

Subsequent steps include a month-wise standardization of the time series. Hence, it is preferable that the distribution of the time series for each month is not skewed. To achieve that, we consider monotone and continuous transformations. Figure 2 shows the spatial variation of skewness for the month-wise time series of vapor pressure deficit (VPD). We observe negative and positive skewness and a variation over the year, which supports a month-wise modeling approach.
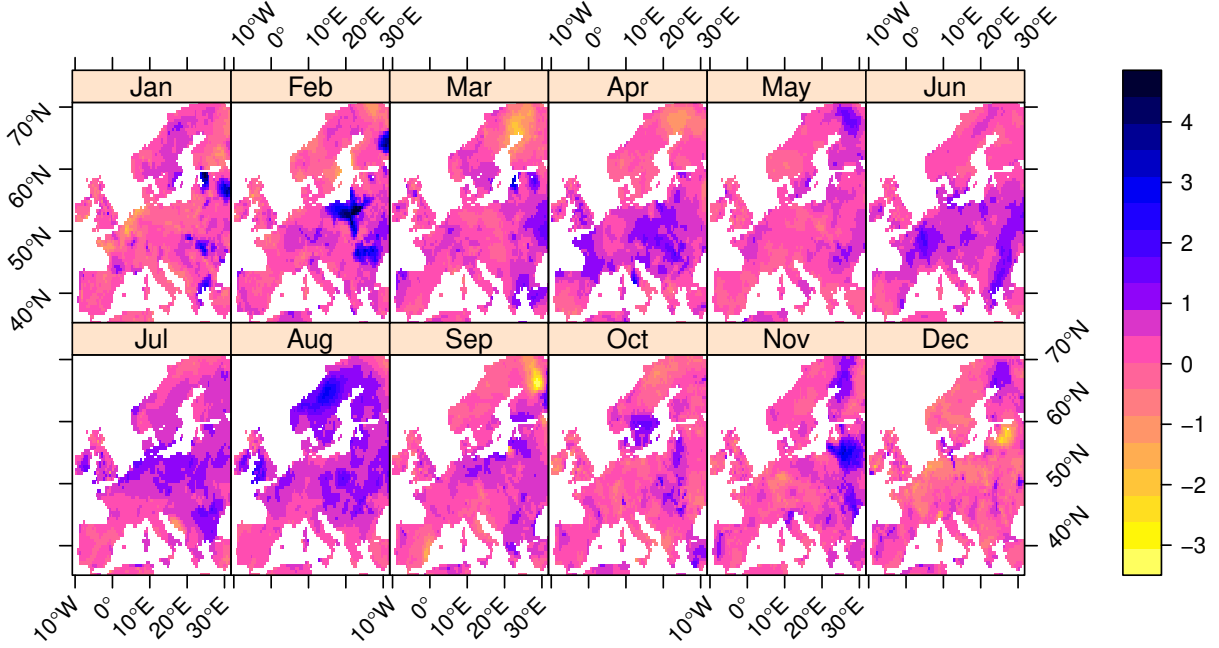

Figure 2: Empirical skewness estimates for the month-wise vapor pressure deficit (VPD) time series.

Let each time point t_k, $k = 1, \ldots, T$, be a 2-tupel (m_k, y_k), where $m_k \in \{1, \ldots, 12\}$ (1 = January, $\ldots$, 12 = December) represents the month and the integer $y_k \in \mathbb{Z}$ the year corresponding to t_k. Then we consider the month-wise time series $\boldsymbol{x}_m := (x_{t_k})_{k \in \mathcal{K}(m)} = \{x_{(m, y_k)}, k \in \mathcal{K}(m)\}$, $m = 1, \ldots, 12$, where the index set for month m is defined as $\mathcal{K}(m) := \{k : m_k = m\}$.

To eliminate/reduce skewness in the (12) month-wise time series $\boldsymbol{x}_m$, $m = 1, \ldots, 12$, we apply power transformations. An appropriate family of transformations, similar to the famous Box-Cox transformations, which is defined not only for positive values is the Yeo and Johnson (2000) transformation $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined as

$$\psi\left(\lambda, x\right) = \begin{cases} \left((x+1)^\lambda - 1\right)/\lambda & \text{if } x \geq 0, \lambda \neq 0 \\ \ln(x+1) & \text{if } x \geq 0, \lambda = 0 \\ -\left((-x+1)^{2-\lambda} - 1\right)/(2-\lambda) & \text{if } x < 0, \lambda \neq 2 \\ -\ln(-x+1) & \text{if } x < 0, \lambda = 2. \end{cases}$$
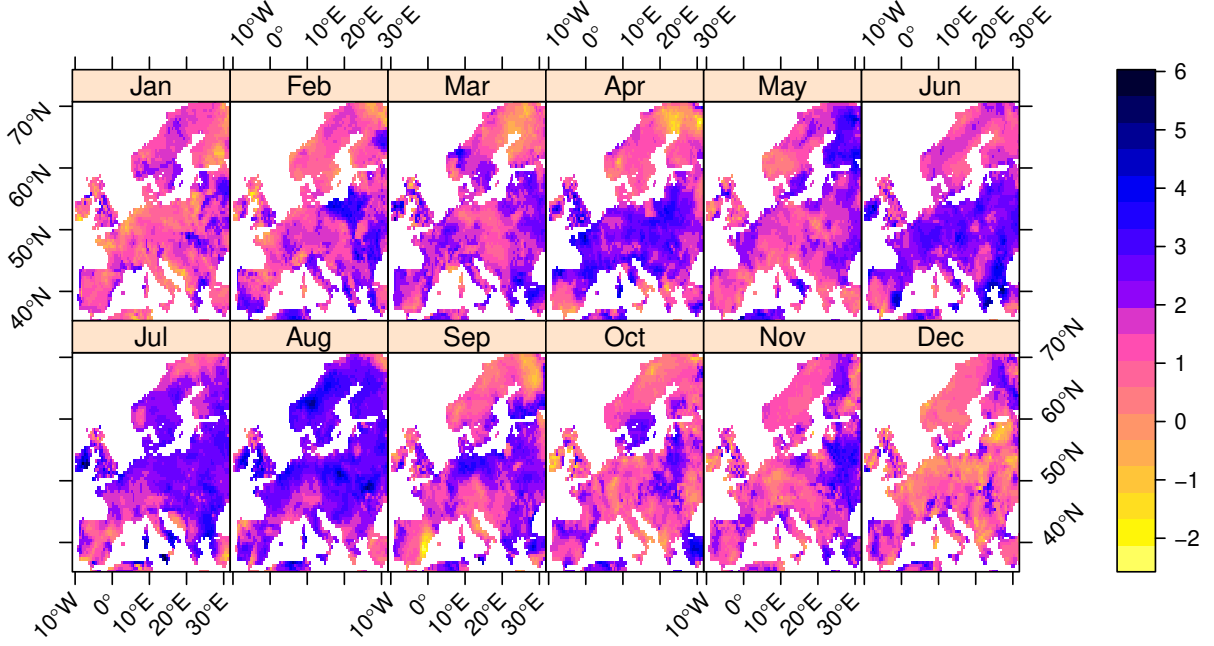
Figure 3: Yeo and Johnson transformation parameter λ for the month-wise `VPD` time series.

Figure 3 maps the Yeo and Johnson transformation parameter λ for the month-wise `VPD` time series. The observed spatial paterns resemble those observed for skewness in Figure 2.

3.2   Elimination of seasonality

Often (climatic) variables are subject to seasonal fluctuations (see e.g. `PET`, Figure 4). Moreover, they can be subject to trends (e.g. due to climate change). `TRENDS` are not removed since a drought index should be able to detect changes in drought frequency and intensity due to climate change. Since drought is considered as a (negative) deviation from 'normal' conditions (anomaly), we remove seasonality (`SEASON`). This is accounted for by month-wise modeling of the time series x_{t_k}, $k = 1, \ldots, T$. However, to ensure that the sample size (`SMALLS`) for fitting a distribution is not too small, our deseasonalization procedure allows to recompose the resulting anomalies to a single time series.

To eliminate seasonality, we model the month-wise mean μ_m separately for each of the 12 time series $\boldsymbol{x}_m$, $m = 1, \ldots, 12$. We estimate it as

$$\widehat{\mu}_m := \frac{1}{|\mathcal{K}(m)|} \sum_{k \in \mathcal{K}(m)} x_{(m, y_k)}, \quad m = 1, \ldots, 12. \tag{1}$$

Figure 4 illustrates the month-wise modeling (1) of potential evapotranspiration (`PET`). Least-squares estimation ensures that $\sum_{k \in \mathcal{K}(m)} \left(x_{(m, y_k)} - \widehat{\mu}_m \right) = 0$ for all $m = 1, \ldots, 12$. Thus also the anomalies $a_{t_k} := x_{t_k} - \widehat{\mu}_{m_k}$, $k = 1, \ldots, T$, are centered around 0 (i.e. $\sum_{k=1}^{T} a_{t_k} = 0$). Hence, seasonal deviations from the annual mean could be eliminated.
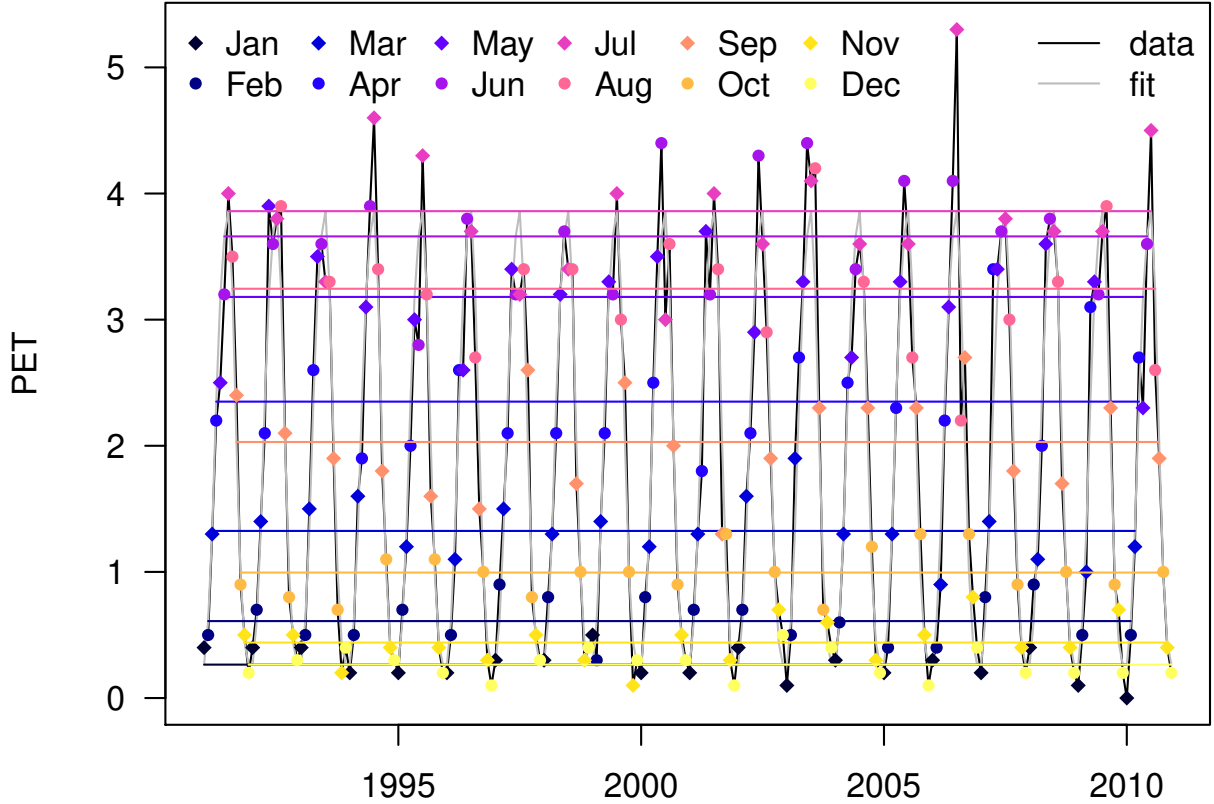
Figure 4: Month-wise modeling of `PET`: The original time series (black, 1991-2010, for pixel C as given in Figure 1) is superimposed by the corresponding month-wise mean fit (grey, cp. Equation (1)). The month-wise time series are illustrated by points colored differently for each month. The modeled month-wise means are visualized by lines in the corresponding color.

Also the variance of the time series may be subject to seasonality, i.e. in some months the time series may deviate more from its mean compared to other months. The color-coding in Figure 4 reveals inhomogeneity of the variance. To quantify this seasonal heterogeneity of the time series a_{t_k}, $k = 1, \dots, T$, we estimate month-wise standard deviations as

$$\widehat{\sigma}_m := \sqrt{\frac{1}{|\mathcal{K}(m)| - 1} \sum_{k \in \mathcal{K}(m)} a^2_{(m, y_k)}}, \quad m = 1, \dots, 12,$$

where $|\cdot|$ is the cardinality. To obtain a homogenized time series we compute the standardized anomalies (residuals) $r_{t_k} := a_{t_k} / \widehat{\sigma}_{m_k}$, $k = 1, \dots, T$.

3.3   Elimination of temporal dependencies

Apart from seasonality, time series often feature temporal dependence (`TIMDEP`). Such serial dependencies can be captured by *autoregressive moving-average models* (see e.g. Box et al. (2008)). For a (deseasonalized, homogeneous, zero-mean) time series r_{t_k}, $k = 1, \dots, T$, the autoregressive moving-average model ARMA(p, q) with AR-order $p \in \mathbb{N}_0$

and MA-order $q \in \mathbb{N}_0$ is defined as

$$r_{t_k} = \sum_{j=1}^{p} \phi_j r_{t_{k-j}} + \sum_{j=1}^{q} \theta_j \varepsilon_{t_{k-j}} + \varepsilon_{t_k},$$

where the error terms ε_{t_k} are i.i.d. $N(0, \sigma^2)$ distributed. Note, that for p or q equal to 0 the corresponding summands are neglected. For adequate choice of the orders p and q and estimates $\widehat{\phi}_j$, $j = 1, \ldots, p$, and $\widehat{\theta}_j$, $j = 1, \ldots, q$, of the corresponding parameters the model residuals $\epsilon_{t_k} \coloneqq r_{t_k} - \sum_{j=1}^{p} \widehat{\phi}_j r_{t_{k-j}} - \sum_{j=1}^{q} \widehat{\theta}_j \epsilon_{t_{k-j}}$, $k = 1, \ldots, T$, are approximately temporally independent. For the variables at hand (`PET`, `PRE`, `VPD`) $p = 1$ and $q = 0$ are an adequate choice.

3.4   Transformation to standard normal distribution

As the assumption of established standardized drought indices like SPI and SPEI of a parametric distribution model for the data performs bad, it seems appropriate to use the (non-parametric) empirical distribution (`NPDIST`) function $\widehat{F}_T(x) \coloneqq \frac{1}{T} \sum_{k=1}^{T} \mathbb{1}\{\epsilon_{t_k} \leq x\}$ of the data respectively the residuals ϵ_{t_k}, $k = 1, \ldots, T$, resulting from the previous modeling step. Here $\mathbb{1}\{\mathcal{A}\}$ is the indicator function, which equals 1 if the event $\mathcal{A}$ is true and 0 otherwise. Note that for fitting a distribution (no matter if parametric or not) to a sample ϵ_{t_k}, $k = 1, \ldots, T$, it is a critical assumption that the sample originates from the same distribution and is i.i.d. We ensured the i.i.d. assumption in the previous step by eliminating the temporal dependencies.

We use the estimated distribution $\widehat{F}_T$ to transform our residuals ϵ_{t_k}, $k = 1, \ldots, T$, to the u-scale, i.e. to be (approximately) uniformly distributed on the interval $[0, 1]$. This transformation is called probability integral transform (PIT). We calculate $u_{t_k} \coloneqq T/(T + 1)\widehat{F}_T(\epsilon_{t_k}) = \text{rank}(\epsilon_{t_k})/(T + 1)$, $k = 1, \ldots, T$. We multiply by $T/(T + 1)$ to avoid any $u_{t_k} = 1$. Further, we transform to the z-scale, calculating $z_{t_k} \coloneqq \Phi^{-1}(u_{t_k})$, $k = 1, \ldots, T$, using the inverse PIT based on the CDF Φ of a standard normal distribution. It holds that z_{t_k}, $k = 1, \ldots, T$, is (approximately) independent and identically standard normal distributed (`STCOMP`).

3.5   Standardized indices on different time scales

McKee et al. (1993) introduced the concept of time scales (`TSCALE`) to make their drought index (the SPI) applicable to different types of drought. We adopt this concept, however we perform the temporal aggregation in the end of the above described modeling process, in order not to violate the independency assumption for fitting a probability distribution to the residuals. This has also the advantage of being computationally more efficient. We need to perform the different modeling steps of Sections 3.1-3.4 only once, after that we are able to calculate the index on arbitrary time scales.

The (approximately) temporally independent standard normal distributed time series z_{t_k}, $k = 1, \ldots, T$, from above is already a standardized index with time scale $l = 1$. The normal distribution has the advantage that a sum of independent normal distributed

Table 2: Dryness and wetness categories.

	category	cumulative probability	quantile
$W4$	exceptionally wet	0.98-1.00	$+2.05 < \text{SI} < +\infty$
$W3$	extremely wet	0.95-0.98	$+1.64 < \text{SI} \le +2.05$
$W2$	severely wet	0.90-0.95	$+1.28 < \text{SI} \le +1.64$
$W1$	moderately wet	0.80-0.90	$+0.84 < \text{SI} \le +1.28$
$W0$	abnormally wet	0.70-0.80	$+0.52 < \text{SI} \le +0.84$
$D0$	abnormally dry	0.20-0.30	$-0.84 < \text{SI} \le -0.52$
$D1$	moderately dry	0.10-0.20	$-1.28 < \text{SI} \le -0.84$
$D2$	severely dry	0.05-0.10	$-1.64 < \text{SI} \le -1.28$
$D3$	extremely dry	0.02-0.05	$-2.05 < \text{SI} \le -1.64$
$D4$	exceptionally dry	0.00-0.02	$-\infty < \text{SI} \le -2.05$

random variables is again normally distributed. We use this property to calculate standardized indices for time scales $l \ge 1$. The sum $\sum_{j=1}^{l} z_{t_{k+1-j}}$ of standard normal variables is normally distributed with mean 0 and variance l. Hence, we obtain a *standardized index* with time scale l as $\text{SI}_l(t_k) := \frac{1}{\sqrt{l}} \sum_{j=1}^{l} z_{t_{k+1-j}}$, $k = 1, \ldots, T$.

To classify the values of standardized indices we use the dryness/wetness categories as defined in Table 2 based on quantiles (cp. Svoboda et al., 2002). A comparison of precipitation (`PRE`) based drought indices for different time scales is provided by Figure 5. For the selected location we identify persistent dry periods during the years 1976, $1989 - 1991$, $1992 - 1993$ and $2003 - 2004$. Whereas the index with time scale 1 identifies single (agricultural) drought months, higher time scales (e.g. 6, 12) allow to identify persistent periods of dryness (hydrological drought).

4   Multivariate standardized indices

Subsequently, we provide an extension of the methodology introduced in Section 3 to multivariate standardized drought indices (`MULTEX`). This extension is based on vine copulas (see Aas et al., 2009) used for dependency modeling of the involved variables. The dependence parameters will be estimated using a semi-parametric estimation procedure (see Genest et al., 1995). Other copula based drought indices were introduced by Farahmand and AghaKouchak (2015) and Kao and Govindaraju (2010).

4.1   Marginal models

As copulas allow separate modeling of margins and dependence structure, we first model the margins according to Sections 3.1-3.4 as in the univariate case. We transform the input data (see Section 3.1), then we eliminate seasonality (see Section 3.2) and temporal dependencies (see Section 3.3) and estimate the distribution of the remaining residuals non-parametrically (see Section 3.4). This enables transformation to the u-scale (copula data) and after that copula based dependency modeling.
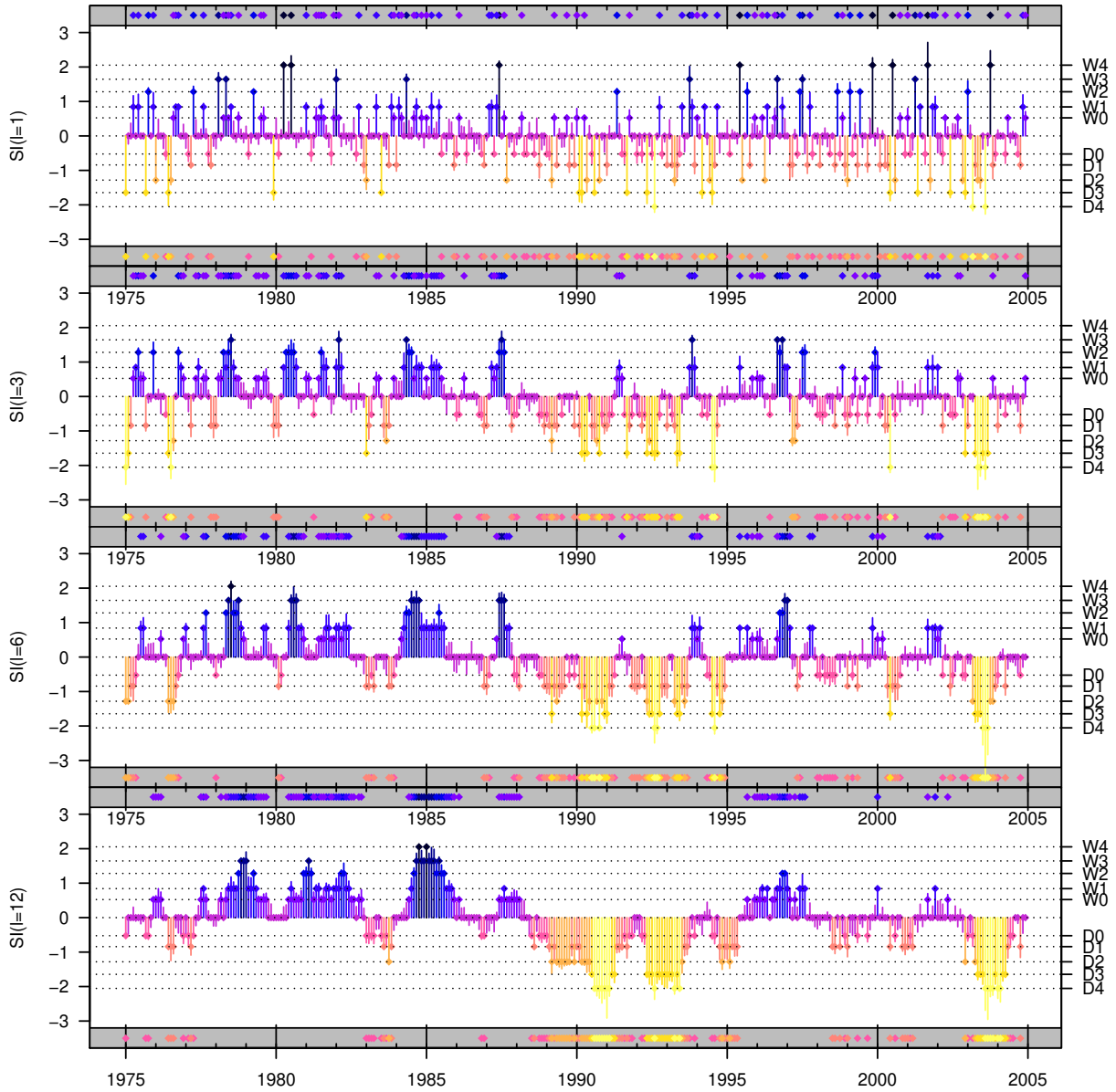
Figure 5: Time series (1975 - 2004, for pixel C Figure 1) of standardized drought index (SI) based on `PRE`, for time scales $l = 1, 3, 6$ and 12, repectively. The color-coding reflects the severity of wetness/dryness according to the different categories specified in Table 2. For better identification of dry/wet periods points at the top/bottom of the panels (colored accordingly) indicate points in time of wet/dry conditions.

4.2   Vine copula based dependency modeling

Let now $\boldsymbol{u} := (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)$ be the copula data obtained from the marginal models corresponding to d different drought relevant variables, where $\boldsymbol{u}_j = (u_{j,t_k})_{k=1,\ldots,T}$, $j = 1, \ldots, d$, and u_{j,t_k} is the copula data corresponding to variable j at time t_k. In a second (parametric) step we select and estimate a *vine copula C* for this data. We illustrate this procedure based on a $d = 3$ dimensional example. For a more general explanation of vine copulas

see Aas et al. (2009) and Dißmann et al. (2013).

T_1 $2{=}\texttt{PET}$ — $\overset{1,2}{\underset{\texttt{VPD,PET}}{}}$ — $1{=}\texttt{VPD}$ — $\overset{1,3}{\underset{\texttt{VPD,PRE}}{}}$ — $3{=}\texttt{PRE}$ $1{=}\texttt{VPD},2{=}\texttt{PET}$ — $\overset{2,3;1}{\underset{\texttt{PET,PRE;VPD}}{}}$ — $1{=}\texttt{VPD},3{=}\texttt{PRE}$ T_2

Figure 6: Selected vine tree structure.

Let now $d = 3$ and $1 = \texttt{VPD}$, $2 = \texttt{PET}$ and $3 = \texttt{PRE}$. Generally, the structure of vine copulas is organized using a nested set of trees (graphs) fulfilling certain conditions. The edges of these trees correspond to bivariate copulas which are the building blocks of the vine copula. Selecting the tree structure as given in Figure 6, we explicitly model the bivariate dependence structures (copulas) $C_{1,2}$, $C_{1,3}$ (tree T_1) and $C_{2,3;1}$ (tree T_2) for the variable pairs (VPD,PET), (VPD,PRE) and (PET,PRE) given VPD, respectively. Here $C_{2,3;1}$ denotes the pair-copula associated with the conditional distribution of the variable pair $(2, 3)$ given variable 1. Further, we select pair-copula families for the pairs above and denote their parameters as $\boldsymbol{\theta} \coloneqq (\theta_{1,2}, \theta_{1,3}, \theta_{2,3;1})$. Then the vine copula density c is given as

$$c(u_1, u_2, u_3; \boldsymbol{\theta}) = c_{1,2}(u_1, u_2; \theta_{1,2}) \cdot c_{1,3}(u_1, u_3; \theta_{1,3})$$
$$\cdot c_{2,3;1}(h_{2|1}(u_2, u_1; \theta_{1,2}), h_{3|1}(u_3, u_1; \theta_{1,2}); \theta_{2,3;1}),$$

where $c_{1,2}$, $c_{1,3}$ and $c_{2,3;1}$ are the pair-copula densities corresponding to the copulas $C_{1,2}$, $C_{1,3}$ and $C_{2,3;1}$. The involved h-functions are defined as $h_{b|a}(u_b, u_a; \theta) \coloneqq C_{b|a}(u_b|u_a; \theta)$, where $C_{b|a}$ denotes the conditional distribution function of U_b given U_a. The tree structure can be saved in a triangular, so called R-vine matrix. For the given three dimensional example a valid R-vine matrix is given as

$$\begin{pmatrix} 3 & 0 & 0 \\ 2 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{respectively} \quad \begin{pmatrix} \texttt{PRE} & 0 & 0 \\ \texttt{PET} & \texttt{PET} & 0 \\ \texttt{VPD} & \texttt{VPD} & \texttt{VPD} \end{pmatrix}.$$

Whereas the second column encodes the pair (VPD,PET), the first column contains the pairs (VPD,PRE) and (PET,PRE;VPD). Other orders of these variables are possible. For a comparison of different orders see the supporting information (Section 7.1).

For the pair-copula family selection we can choose among a variety of bivariate copula families, amongst others among the Gaussian (N), Student-t (t), Clayton (C), Gumbel (G) Frank (F) and Joe (J) family, which all feature different dependence structures and properties. Also rotated versions of the Clayton, Gumbel and Joe copula are considered to capture negative asymmetric dependencies. The pair-copulas are selected separately (according to the BIC) starting in tree T_1. Their parameters are estimated at the same time using maximum likelihood estimation. Before that a bivariate independence test (Genest and Favre, 2007) can be performed, to see if an independence copula should be selected. For more details on different (rotated) copula families and their selection we refer to Brechmann and Schepsmeier (2013).

Figure 7 visualizes for all spatial pixels under consideration which dependence structures were selected for the vine tree structure specified above (Figure 6). For the pair
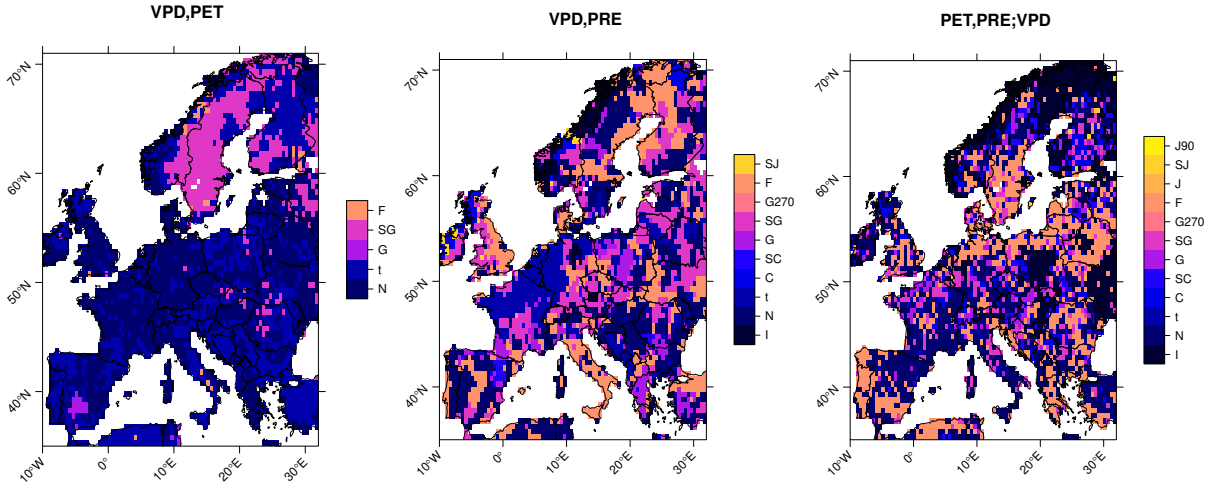
Figure 7: Spatial variation in the pair-copula families selected for the pairs specified in Figure 6.

(VPD,PET) the elliptical and symmetric Gaussian (N) and Student-t (t) copula were selected most over Europe. Where the Student-t copula was selected extreme high or low VPD and PET anomalies occur jointly, since the Student-t copula allows for dependence in the upper and lower distribution tails, so called tail-dependence. For a large area on the Iberian Peninsula (G) extreme wet conditions for both variable pairs seem to occur simultaneously, since the Gumbel copula allows for upper tail dependence. For most of Scandinavia (SG) we observe the opposite, high correlation of extreme dry conditions, since the survival/180° rotated Gumbel copula allows for lower tail dependence. For the other two (conditioned) pairs similar interpretations can be made. We observe that for most pixels non-Gaussian dependence structures were selected.

4.3   Computation of multivariate indices

Based on the previously selected vine copula C for the data $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)$, we transform $\boldsymbol{u}$ to i.i.d. uniform data on $[0, 1]$, using the so called *Rosenblatt (1952) transformation*, a multivariate probability integral transform. The Rosenblatt transform $\boldsymbol{v} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$ of $\boldsymbol{u}$ is defined as

$$
\begin{aligned}
v_{1,t_k} &:= u_{1,t_k}, \\
v_{2,t_k} &:= C_{2|1}(u_{2,t_k}|u_{1,t_k}), \\
&\cdots \\
v_{d,t_k} &:= C_{d|1,\ldots,d-1}(u_{d,t_k}|u_{1,t_k}, \ldots, u_{d-1,t_k}), \quad k = 1, \ldots, T
\end{aligned}
$$

where $C_{j|1,\ldots,j-1}$, is the conditional cumulative distribution function for variable j given the variables $1, \ldots, j-1$, for all $j = 2, \ldots, d$. For vine copulas the order of the variables is determined by the vine tree structure respectively the R-vine matrix. For details on the computation of the Rosenblatt transform for vine copulas see Schepsmeier (2015).

Generally speaking, application of the Rosenblatt transform to our d dependent variables yields independent information about dry/wet conditions captured in these vari-

ables. $\boldsymbol{v}_1$ incorporates the same information as an univariate drought index calculated according to Section 3 based on variable 1. $\boldsymbol{v}_j$, $j = 2, \ldots, d$, provide information on dry/wet conditions identified by variable j, conditioned on the dryness/wetness information provided by the previously considered variables $1, \ldots, j-1$.

For our three dimensional example from above we compute $v_{\texttt{VPD},t} = u_{\texttt{VPD},t}$, which represents the dry-/wetness information captured in the variable VPD for time point t. $v_{\texttt{PET},t} = C_{\texttt{PET}|\texttt{VPD}}(u_{\texttt{PET},t}|u_{\texttt{VPD},t})$ provides additional information based on PET knowing about VPD in that particular time point t. Its calculation involves the pair-copula $C_{\texttt{VPD},\texttt{PET}}$. The calculation of $v_{\texttt{PRE},t} = C_{\texttt{PRE}|\texttt{VPD},\texttt{PET}}(u_{\texttt{PRE},t}|u_{\texttt{VPD},t}, u_{\texttt{PET},t})$ is a bit more involved. We calculate $v_{\texttt{PRE},t} = C_{\texttt{PRE}|\texttt{PET};\texttt{VPD}}(C_{\texttt{PRE}|\texttt{VPD}}(u_{\texttt{PRE},t}|u_{\texttt{VPD},t})|C_{\texttt{PET}|\texttt{VPD}}(u_{\texttt{PET},t}|u_{\texttt{VPD},t}))$, based on the pair-copulas $C_{\texttt{PRE},\texttt{PET};\texttt{VPD}}$, $C_{\texttt{VPD},\texttt{PRE}}$ and $C_{\texttt{VPD},\texttt{PET}}$.

Subsequently we consider two different approaches to join this multivariate drought information into one index. For comparison, we provide a third approach which assumes multivariate normality.

Method $\mathcal{A}$ (aggregation)   This approach allows for a weighting with weights $\boldsymbol{w} = (w_1, \ldots, w_d)$, $w_j > 0$, for the different variables $j = 1, \ldots, d$. We calculate the *standardized multivariate index (SMI)* with time scale l as

$$\mathrm{SMI}_l^{\mathcal{A}}(\boldsymbol{w}; 1, \ldots, d)(t_k) := \frac{1}{\sqrt{l\boldsymbol{w}'\boldsymbol{w}}} \sum_{i=1}^{l} \sum_{j=1}^{d} w_j \Phi^{-1}\left(v_{j,t_{k+1-i}}\right).$$

Method $\mathcal{M}$ (multiplication)   For the second approach we exploit that the multivariate dependence structure of $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$ is represented by the independence copula $C_{\Pi}(v_1, \ldots, v_d) = \prod_{j=1}^{d} v_j$. Hence, we calculate $\widetilde{v}_{t_k} := \prod_{j=1}^{d} v_{j,t_k}$, $k = 1, \ldots, T$. To obtain a standardized (multivariate) index we proceed as in the univariate case (see Sections 3.4 and 3.5). We calculate the rank transformation $\widetilde{u}_{t_k} := \mathrm{rank}\left(\widetilde{v}_{t_k}\right)/(T+1)$, $k = 1, \ldots, T$, transform to the z-scale and calculate the SMI with time scale l as

$$\mathrm{SMI}_l^{\mathcal{M}}(\boldsymbol{1}; 1, \ldots, d)(t_k) := \frac{1}{\sqrt{l}} \sum_{i=1}^{l} \Phi^{-1}\left(\widetilde{u}_{t_{k+1-i}}\right),$$

where no weighting is allowed, i.e. $\boldsymbol{w} = \boldsymbol{1} := (1, \ldots, 1)$.

Method $\mathcal{N}$ (normal)   Let $\boldsymbol{z}$ be the marginal transformation of $\boldsymbol{u}$ to the z-scale and consider a vector of weights $\boldsymbol{w} = (w_1, \ldots, w_d)$, $w_j > 0$. Assuming $\boldsymbol{z}$ to be a sample from a zero mean multivariate normal distribution, we can conclude that the linear transformation $\boldsymbol{w}'\boldsymbol{z}$ is a sample from a zero mean univariate normal distribution. We estimate the sample variance of $\boldsymbol{w}'\boldsymbol{z}$ by $S := \frac{1}{T-1} \sum_{k=1}^{T} \left(\sum_{j=1}^{d} w_j \Phi^{-1}(u_{j,t_k})\right)^2$ and calculate a (weighted) SMI with time scale l as

$$\mathrm{SMI}_l^{\mathcal{N}}(\boldsymbol{w}; 1, \ldots, d)(t_k) := \frac{1}{\sqrt{l \cdot S}} \sum_{i=1}^{l} \sum_{j=1}^{d} w_j \Phi^{-1}\left(u_{j,t_{k+1-i}}\right).$$

Table 3: Maximum spatial extent of drought events classified as extreme ($D3$) or exceptional ($D4$) according to SPI_6, SPEI_6, $\text{SMI}_6^{\mathcal{A}}(\mathbf{1}; \texttt{VPD}, \texttt{PET}, \texttt{PRE})$ and $\text{SMI}_6^{\mathcal{M}}(\mathbf{1}; \texttt{VPD}, \texttt{PET}, \texttt{PRE})$.

| | univariate | | | | multivariate | | | |
| | SPI_6 | | SPEI_6 | | $\text{SMI}^{\mathcal{A}}$ | | $\text{SMI}^{\mathcal{M}}$ | |
event	max.	% area	max.	% area	max.	% area	max.	% area
1976	07.1976	31.0%	08.1976	28.4%	08.1976	28.7%	08.1976	24.1%
1990	03.1990	18.3%	03.1990	21.3%	05.1990	25.7%	05.1990	36.1%
2003	08.2003	21.9%	08.2003	37.2%	08.2003	50.7%	08.2003	46.9%

5 Application

To measure pair-wise dependence we use the rank-based association measure Kendall's τ (see e.g. Kendall, 1970). In Figure 8 we provide maps of Kendall's τ between the univariate drought indices $\text{SI}_6(\texttt{VPD})$, $\text{SI}_6(\texttt{PET})$ and $\text{SI}_6(\texttt{PRE})$ and the (multivariate) drought indices $\text{SMI}_6^{\mathcal{N}}(\mathbf{1}; \texttt{VPD}, \texttt{PET}, \texttt{PRE})$, $\text{SMI}_6^{\mathcal{A}}(\mathbf{1}; \texttt{VPD}, \texttt{PET}, \texttt{PRE})$, $\text{SMI}_6^{\mathcal{M}}(\mathbf{1}; \texttt{VPD}, \texttt{PET}, \texttt{PRE})$, SPI_6 and SPEI_6 on time scale 6, to see how the different variables contribute to the different drought indices and how this contribution varies over space. Whereas the $\text{SMI}^{\mathcal{N}}$ is dominated by $\texttt{PET}$ and $\texttt{VPD}$ (high Kendall's τ values all over Europe for the pairs (SMIN,SIVPD) and (SMIN,SIPET)), the other indices are stronger associated with $\texttt{PRE}$ (comparatively high Kendall's τ values for the pairs (SMIA,SIPRE), (SMIM,SIPRE), (SPI,SIPRE) and (SPEI,SIPRE)). For $\text{SMI}^{\mathcal{A}}$ and $\text{SMI}^{\mathcal{M}}$ the overall association with $\texttt{PET}$ and $\texttt{VPD}$ is stronger compared to SPI and SPEI (compare the corresponding pairs). Especially for SPI and SPEI we observe spatial differences in Kendall's τ (see all pairs involving SPI and SPEI).

To validate and compare the different drought indices we consider the three major drought events of the 30 years period $1975 - 2004$ which were observed in large parts of Europe. These droughts occured in the years 1976, 1989/90 and 2003. We summarize these events in Table 3. It gives the dates when the drought events (in terms of an extreme ($D3$) or exceptional ($D4$) drought) reached their maximum spatial extent (i.e. the month in which the area affected by a $D3$ or $D4$ drought reached it's maximum) and the corresponding percentage of area under consideration which was affected by an extreme ($D3$) or exceptional ($D4$) drought.

Figure 9 compares time series of the percentage of area affected by drought according to the different univariate and multivariate drought indices calculated following the methodology described above, as well as SPI_6 and SPEI_6. Comparing the univariate indices we see that those based on $\texttt{PET}$ and $\texttt{VPD}$ yield similar however not identical results. During the three major drought events in 1976, 1989/90 and 2003 all three univariate indices indicate extreme dry conditions for large parts of Europe. Comparison to the middle panel shows that the multivariate indices successfully combine the drought information captured in the single variables used for their calculation. For the years 1990 and 2003 abnormally high $\texttt{PET}$ and $\texttt{VPD}$ aggravate the dry conditions due to a lack of precipitation. During the years 1994, 1995, 1999 and 2000 one can see that the vine copula based indices are more conservative compared to $\text{SMI}^{\mathcal{N}}$, since they are not as much influenced by $\texttt{PET}$ and $\texttt{VPD}$. In terms of spatial extent the multivariate indices classify the drought events of
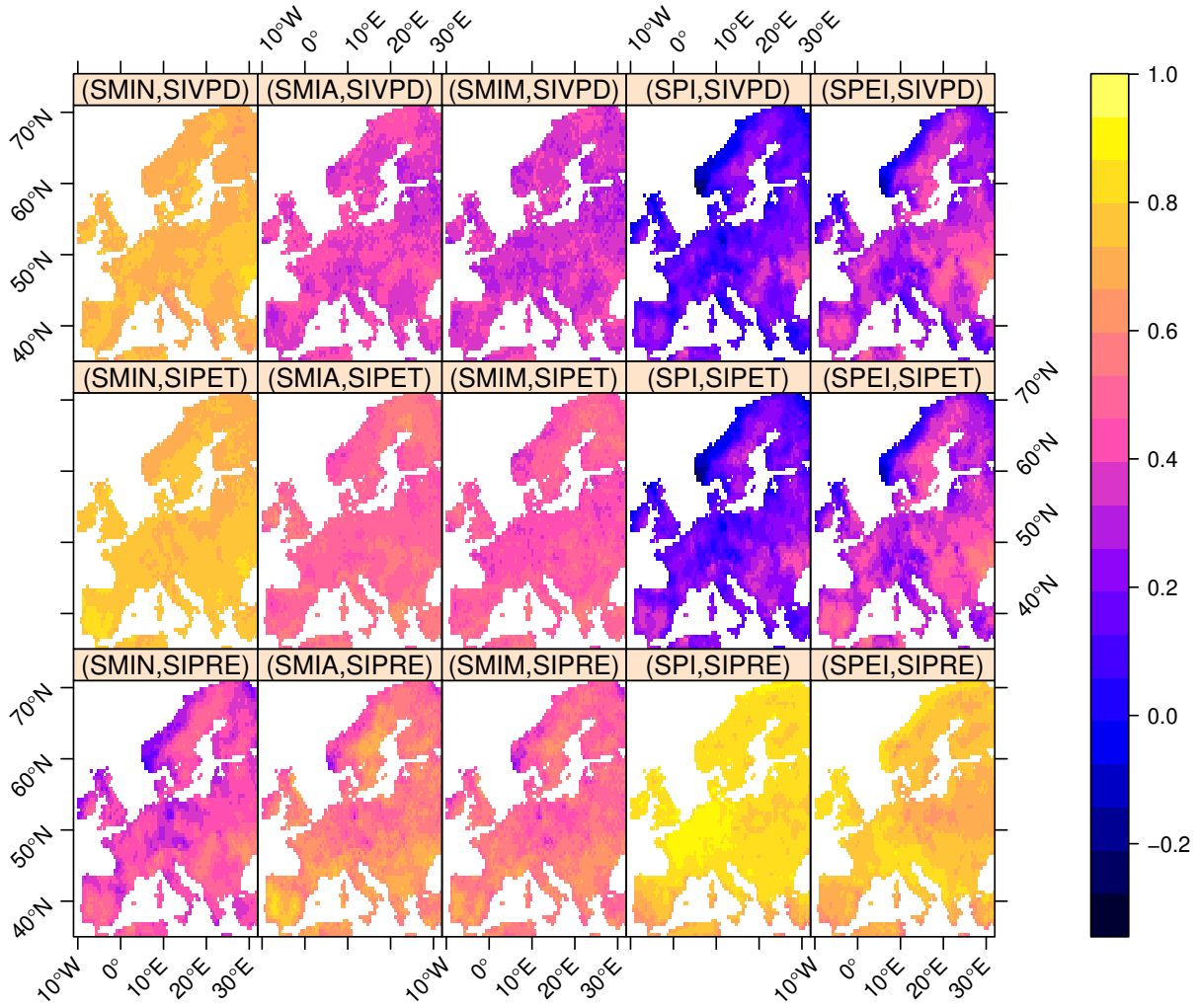
Figure 8: Maps of Kendall's τ for all combinations of the univariate drought indices $\mathrm{SI}_6(\mathtt{VPD})$ (SIVPD), $\mathrm{SI}_6(\mathtt{PET})$ (SIPET) and $\mathrm{SI}_6(\mathtt{PRE})$ (SIPRE) with the indices $\mathrm{SMI}_6^{\mathcal{N}}(\mathbf{1}; \mathtt{VPD}, \mathtt{PET}, \mathtt{PRE})$ (SMIN), $\mathrm{SMI}_6^{\mathcal{A}}(\mathbf{1}; \mathtt{VPD}, \mathtt{PET}, \mathtt{PRE})$ (SMIA), $\mathrm{SMI}_6^{\mathcal{M}}(\mathbf{1}; \mathtt{VPD}, \mathtt{PET}, \mathtt{PRE})$ (SMIM), SPI_6 (SPI) and SPEI_6 (SPEI).

1990 and 2003 as more severe compared to SPI and SPEI.

6   Conclusions and outlook

Comparison of the advantages and disadvantages of existing drought indices and the flexibility of vine copulas in modeling multivariate dependence structures led to a novel and flexibly applicable approach to calculate drought indices based on arbitrary sets of drought relevant variables. This approach involves several well reasoned modeling steps which we summarize in Figure 10.

Taking several drought drivers and their dependencies into account at the same time our novel approach enables flexible modeling of different drought types and allows tailoring
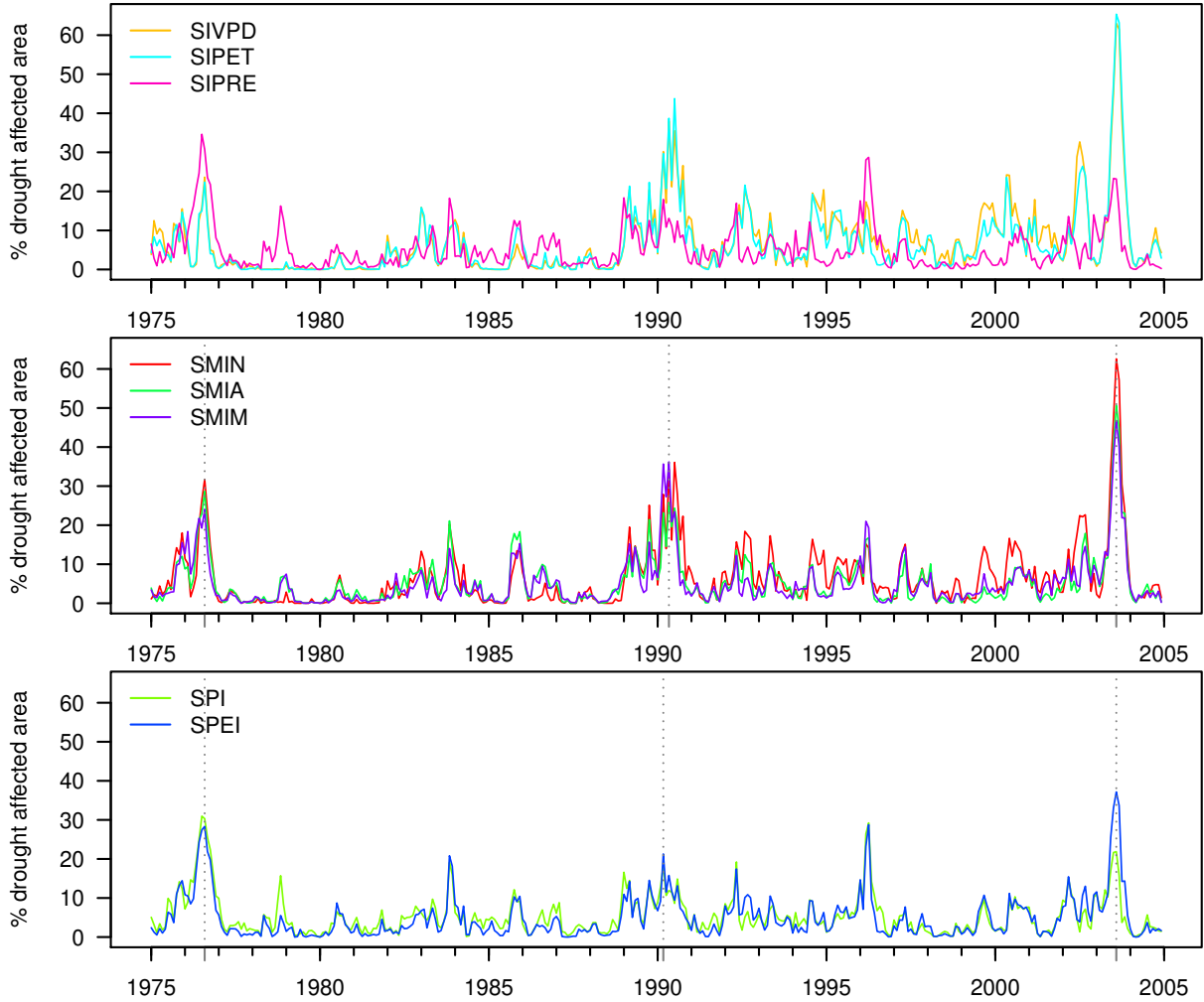
Figure 9: Percentage of area affected by a $D3$ or $D4$ drought according to $\mathrm{SI}_6(\mathtt{VPD})$, $\mathrm{SI}_6(\mathtt{PET})$ and $\mathrm{SI}_6(\mathtt{PRE})$ (upper panel), $\mathrm{SMI}_6^{\mathcal{N}}(\mathbf{1}; \mathtt{VPD}, \mathtt{PET}, \mathtt{PRE})$, $\mathrm{SMI}_6^{\mathcal{A}}(\mathbf{1}; \mathtt{VPD}, \mathtt{PET}, \mathtt{PRE})$ and $\mathrm{SMI}_6^{\mathcal{M}}(\mathbf{1}; \mathtt{VPD}, \mathtt{PET}, \mathtt{PRE})$ (middle panel), and SPI_6 and SPEI_6 (lower panel).

of drought indices to specific applications. An example would be the application of the novel methodology in the field of ecology. Multivariate drought indices based on selected variables could be calibrated to tree ring data to find good models for the response of tree growth to climatic conditions. Moreover, the presented approach for the calculation of severity indices is not restricted to drought. Applications to model for example the degree of contamination of a water body due to different contaminants are feasible.

7 Supporting information

7.1 Accompanying figures and analyses

We provide figures and further analyses of the data at hand, visualizing/complementing the presented methodology for drought index calculation. We address the following issues:
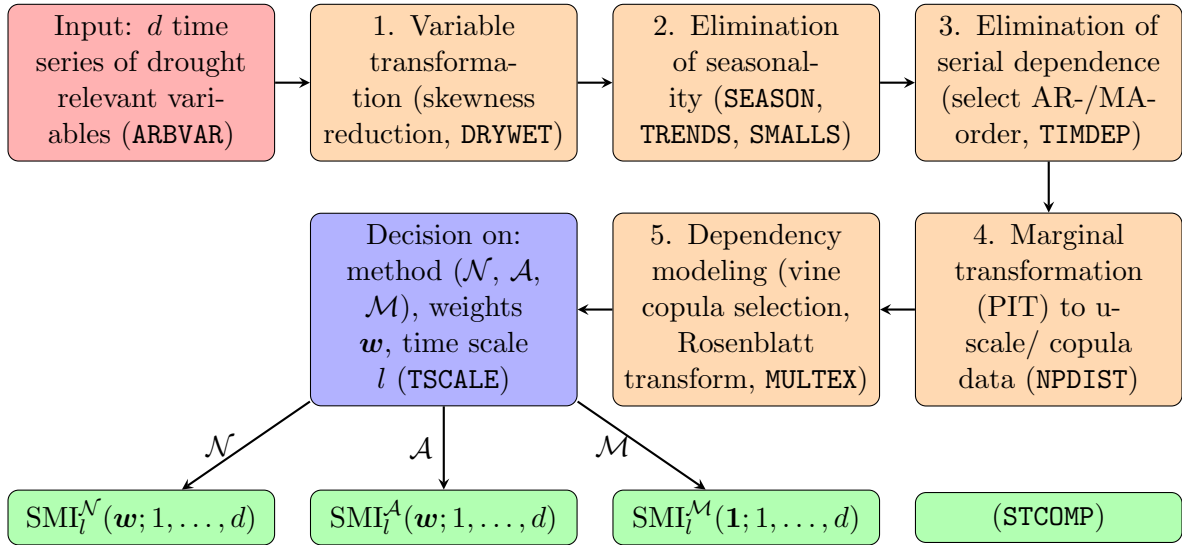
Figure 10: Modeling steps for multivariate drought index calculation.

1. Visualization of the data and its features

2. Testing of multivariate normality

3. Visualization of the area affected by drought according to the different indices

4. Visualization of the inter-index association

5. Visualization of drought index time series for selected locations

6. Comparison of different variable orders for the calculation of multivariate drought indices

7. Effect of trends on multivariate drought indices

7.2   Software and data

Moreover, we provide an `R` software package (`SIndices`, version 1.0) which is an implementation of the presented methodology. It comes along with a detailed manual. Further, we provide the `R`-code which was used to produce all results presented in the article and the supporting information. The Climatic Research Unit (CRU) time series (TS) data (version 3.21, see Jones and Harris, 2013) on which all examples and computations are based can be obtained from `http://dx.doi.org/10.5285/D0E1585D-3417-485F-87AE-4FCECF10A992`.

Acknowledgments

load the CRU data set we used the `raster` package (Hijmans, 2015). To handle spatial and spatio-temporal data we used the packages `sp` (Pebesma and Bivand, 2005) and `spacetime` (Pebesma, 2012), respectively. To work with time series we used the package `xts` (Ryan and Ulrich, 2014). To calculate SPI and SPEI we used the `SPEI` package (Beguería and Vicente-Serrano, 2013). For dependency modeling we used the `VineCopula` package (Schepsmeier et al., 2015). Empirical skewness estimates were calculated using the package `moments` (Komsta and Novomestky, 2015). For the Yeo and Johnson transformation we used the package `car` (Fox and Weisberg, 2011).

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics 44*(2), 182–198.

Beguería, S. and S. M. Vicente-Serrano (2013). *SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index*. R package version 1.6.

Box, G., G. Jenkins, and G. Reinsel (2008). *Time Series Analysis: Forecasting and Control* (4th ed.). Wiley Series in Probability and Statistics. Wiley.

Brechmann, E. C. and U. Schepsmeier (2013). Modeling dependence with C- and D-Vine Copulas: The R package CDVine. *Journal of Statistical Software 52*(3), 1–27.

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

Edwards, D. C. and T. B. McKee (1997). Characteristics of 20th century drought in the United States at multiple time scales. Atmospheric Science Paper No. 634, Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523-1371.

Farahmand, A. and A. AghaKouchak (2015). A generalized framework for deriving non-parametric standardized drought indicators. *Advances in Water Resources 76*, 140–145.

Fox, J. and S. Weisberg (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage. R package version 2.0-25.

Genest, C. and A.-C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering 12*(4), 347–368.

Genest, C., K. Ghoudi, and L.-P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika 82*(3), 543–552.

Hao, Z. and A. AghaKouchak (2013). Multivariate standardized drought index: A parametric multi-index model. *Advances in Water Resources 57*, 12–18.

Hao, Z. and A. AghaKouchak (2014). A nonparametric multivariate multi-index drought monitoring framework. *Journal of Hydrometeorology 15*, 89–101.

Hijmans, R. J. (2015). *raster: Geographic Data Analysis and Modeling*. R package version 2.4-15.

Jones, P. and I. Harris (2013). *Climatic Research Unit (CRU) Time-Series (TS) Version 3.21 of High Resolution Gridded Data of Month-by-month Variation in Climate (Jan. 1901-Dec. 2012)*. University of East Anglia Climatic Research Unit. NCAS British Atmospheric Data Centre, 24th September 2013. `http://dx.doi.org/10.5285/D0E1585D-3417-485F-87AE-4FCECF10A992`.

Kao, S.-C. and R. S. Govindaraju (2010). A copula-based joint deficit index for droughts. *Journal of Hydrology 380*(1–2), 121–134.

Kendall, M. G. (1970). *Rank Correlation Methods* (4th ed.). London: Griffin.

Komsta, L. and F. Novomestky (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14.

McKee, T. B., N. J. Doesken, and J. Kleist (1993, January 17-22). The relationship of drought frequency and duration to time scales. In *Eighth Conference on Applied Climatology*, Anaheim California, pp. 179–184. American Meteorological Society.

Mishra, A. K. and V. P. Singh (2010). A review of drought concepts. *Journal of Hydrology 391*(1–2), 202–216.

Murray, F. W. (1967). On the computation of saturation vapor pressure. *Journal of Applied Meteorology 6*(1), 203–204.

Palmer, W. C. (1965, February). Meteorological drought. Reserach Paper No. 45, US Department of Commerce, U.S. Weather Bureau, Washington, D.C.

Pebesma, E. (2012). spacetime: Spatio-temporal data in r. *Journal of Statistical Software 51*(7). R package version 1.1-4.

Pebesma, E. and R. Bivand (2005). Classes and methods for spatial data in R. *R News 5*(2), 9–13. R package version 1.1-1.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*(3), 470–472.

Ryan, J. A. and J. M. Ulrich (2014). *xts: eXtensible Time Series*. R package version 0.9-7.

Schepsmeier, U. (2015). Efficient information based goodness-of-fit tests for vine copula models with fixed margins: A comprehensive review. *Journal of Multivariate Analysis*.

Schepsmeier, U., J. Stoeber, E. Brechmann, B. Graeler, T. Nagler, and T. Erhardt (2015). *VineCopula: Statistical Inference of Vine Copulas.* R package version 1.6.

Shukla, S. and A. W. Wood (2008). Use of a standardized runoff index for characterizing hydrologic drought. *Geophysical Research Letters 35*(2).

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*, pp. 229–231. Institut Henri Poincaré.

Svoboda, M., D. LeComte, M. Hayes, R. Heim, K. Gleason, J. Angel, B. Rippey, R. Tinker, M. Palecki, D. Stooksbury, D. Miskus, and S. Stephens (2002). The drought monitor. *Bulletin of the American Meteorological Society 83*(April), 1181–1190.

Vicente-Serrano, S. M., S. Beguería, and J. I. López-Moreno (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of Climate 23*(7), 1696–1718.

Wells, N., S. Goddard, and M. J. Hayes (2004). A self-calibrating palmer drought severity index. *Journal of Climate 17*, 2335–2351.

Yeo, I.-K. and R. A. Johnson (2000). A new family of power transformations to improve normality or symmetry. *Biometrika 87*(4), 954–959.