Technische Universität München
Fakultät für Mathematik
Lehrstuhl für Mathematische Statistik

Development of Vine Copula based Drought Indices and Model Evaluation under the Presence of Non-Stationarity

Tobias Michael Erhardt

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Matthias Scherer
Prüfende/-r der Dissertation:	1.	Prof. Claudia Czado, Ph.D.
	2.	Prof. Peter X.K. Song, Ph.D. University of Michigan

Die Dissertation wurde am 14.11.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 16.02.2017 angenommen.

Zusammenfassung

In den vergangenen Jahrzehnten wurden zahlreiche verschiedene Ansätze zur Quantifizierung des Schweregrades von Dürren entwickelt. Jedoch haben die meisten dieser Dürreindizes unterschiedliche Mängel, berücksichtigen nur ein oder zwei Faktoren die Dürre begünstigen und vernachlässigen deren wechselseitige Abhängigkeiten. Im ersten Teil der Dissertation präsentieren wir eine neuartige Methodik zur Berechnung von (multivariaten) Dürreindizes, welche die Vorteile bestehender Ansätze kombiniert und deren Nachteile umgeht. Sie kann flexibel in verschiedensten Anwendungen eingesetzt werden, um verschiedene Arten von Dürre basierend auf benutzerdefinierten, Dürre-relevanten Variablen zu modellieren. Sie profitiert von der Flexibilität von Vine Copulas bei der Modellierung multivariater nicht-Gaußscher Abhängigkeitsstrukturen zwischen verschiedenen Variablen. Basierend auf einem dreidimensionalen Datensatz entwickeln wir einen beispielhaften agrometeorologischen Dürreindex. Eine Analyse der Daten veranschaulicht und rechtfertigt die beschriebene Methodik. Eine Validierung des exemplarischen multivariaten agrometeorologischen Dürreindexes mit Hilfe von beobachtetem Sojabohnenertrag bestätigt die Validität und das Potenzial der Methodik. Der Vergleich mit etablierten Dürreindizes zeigt die Überlegenheit unseres multivariaten Ansatzes.

Verschiedene Disziplinen verfolgen das Ziel, Modelle zu entwickeln, die bestimmte Phänomene so genau wie möglich charakterisieren. Die Klimawissenschaft als Paradebeispiel ist daran interessiert die zeitliche Entwicklung des Klimas zu modellieren. Um verschiedene Modelle zu vergleichen und zu verbessern, ist Methodik für eine faire Modellevaluierung unerlässlich. Da Modelle und Vorhersagen eines Phänomens für gewöhnlich mit Unsicherheit behaftet sind, sind korrekte Bewertungsregeln (proper scoring rules) und korrekte Divergenzfunktionen (proper divergence functions), welche Werkzeuge sind die diese Unsicherheit berücksichtigen, eine geeignete Wahl für die Modellevaluierung. In der Gegenwart von Nicht-Stationarität wird eine derartige Modellevaluierung jedoch schwierig, da sich die Eigenschaften des Phänomens von Interesse verändern. Der zweite Teil der Dissertation liefert Methodik zur Modellevaluation im Kontext von nichtstationären Zeitreihen. Die neue Methodik nimmt an, dass die Zeitreihen in kleinen, zeitlich versetzten (gleitenden) Zeitfenstern stationär sind. Diese gleitenden Fenster, welche basierend auf einer Changepoint-Analyse ausgewählt werden, werden verwendet, um die Unsicherheit des Phänomens/Modells für die entsprechenden Zeitpunkte zu beschreiben. Dies resultiert im Konzept der gleitenden Scores und Divergenzen, welches eine zeitliche Beurteilung der Modell-Performance ermöglicht. Die Vorzüge der vorgeschlagenen Methodik werden anhand einer Simulations- und einer Fallstudie illustriert.

Abstract

During past decades, many different approaches to quantify drought severity have been developed. However, most of these drought indices suffer from different shortcomings, account only for one or two driving factors which promote drought conditions and neglect their interdependencies. In the first part of the thesis, we provide a novel methodology for the calculation of (multivariate) drought indices, which combines the advantages of existing approaches and omits their disadvantages. It can be used flexibly in different applications to model different drought types based on user-selected, drought relevant variables. It benefits from the flexibility of vine copulas in modeling multivariate non-Gaussian inter-variable dependence structures. Based on a three-variate data set, an exemplary agro-meteorological drought index is developed. The data analysis illustrates and justifies the described methodology. A validation of the exemplary multivariate agro-meteorological drought index against observed soybean yield affirms the validity and capabilities of the novel approach. Comparison to established drought indices shows the superiority of our multivariate approach.

Different disciplines pursue the aim to develop models which characterize certain phenomena as accurately as possible. Climatology is a prime example, where the temporal evolution of the climate is modeled. In order to compare and improve different models, methodology for a fair model evaluation is indispensable. As models and forecasts of a phenomenon usually are associated with uncertainty, proper scoring rules and proper divergence functions, which are tools that account for this kind of uncertainty, are an adequate choice for model evaluation. However, under the presence of non-stationarity, such a model evaluation becomes challenging, as the characteristics of the phenomenon of interest change. The second part of the thesis provides methodology for model evaluation in the context of non-stationary time series. The novel methodology assumes stationarity of the time series in small moving time windows. These moving windows, which are selected based on a changepoint analysis, are used to characterize the uncertainty of the phenomenon/model for the corresponding time instances. This leads to the concept of moving scores and divergences, which allows a temporal assessment of the model performance. The merits of the proposed methodology are illustrated based on a simulation and a case study.

Acknowledgment

First and foremost I would like to express my sincere gratitude to my Ph.D. advisor Prof. Claudia Czado, Ph.D., who has been guiding and mentoring me since the time I worked on my Master's thesis under her supervision. I thank Prof. Claudia Czado for providing me with an enthralling topic for my Ph.D. research, for her unremitting support and her excellent supervision. The advancement of my skills benefited from her competent guidance throughout the last few years. Our fruitful discussions and her constructive feedback helped me in the process of conducting my own research, resulting in this thesis.

As a joint effort, Prof. Czado, Prof. Dr. Annette Menzel and Dr. rer. nat. Christian Zang had proposed the project "Drought modeling and monitoring by novel statistical and analytical methods" for the Focus Area Water of the TUM International Graduate School of Science and Engineering (IGSSE). I was happy to become one of the doctoral researchers working on the scientific problems pursued by this interdisciplinary project team. I am thankful to the whole team for the collaboration and the helpful feedback I got during the numerous meetings and joint events. Moreover, I acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), without which this work would not have been possible.

I would also like to thank Thordis L. Thorarinsdottir, Ph.D., for hosting me at the Norwegian Computing Center for a research stay of three months. I especially thank her for the discussions we had and for supplying me with new ideas and inspiration for my work.

Further appreciation is directed at the editors and the anonymous referees of my publications and also to the referees of this thesis for all their efforts and their valuable feedback.

Last but not least, I thank my family and my friends. I am particularly grateful for my parents' and my fiancée's indispensable support and encouragement during my time as a Ph.D. candidate.

Contents

1 Introduction 1

2 Preliminaries 5
- 2.1 Standardized drought indices 5
- 2.2 Multivariate drought indices 9
- 2.3 Dependence modeling with copulas 10
  - 2.3.1 Copulas 10
  - 2.3.2 Vine copulas 14
  - 2.3.3 Statistical inference for vine copulas 19
  - 2.3.4 The multivariate probability integral transform for vine copulas 20
- 2.4 Validation metrics 21
- 2.5 Proper scoring rules and divergence functions 22
  - 2.5.1 Proper scoring rules 23
  - 2.5.2 Proper divergence functions 26
- 2.6 Traditional evaluation measures 30
- 2.7 Detection of multiple changepoints 31

3 Data 37
- 3.1 ECMWF Atmospheric Reanalysis of the 20th Century 37
- 3.2 Arkansas soybean yield 38
- 3.3 E-OBS gridded data set 40
- 3.4 ENSEMBLES Regional Climate Model simulations 42

4 Novel Drought Indices 43
- 4.1 Lessons learned from established drought indices 43
- 4.2 Outline of the modeling approach 45
- 4.3 Data example: ERA-20C/Arkansas 46
- 4.4 Univariate standardized indices 48
  - 4.4.1 Identification of dry and wet conditions 48
  - 4.4.2 Elimination of seasonality 48
  - 4.4.3 Elimination of temporal dependencies 51
  - 4.4.4 Transformation to standard normal distribution 54
  - 4.4.5 Standardized indices on different time scales 54
- 4.5 Multivariate standardized indices 58
  - 4.5.1 Marginal models 58
  - 4.5.2 Vine copula based dependence modeling 58
  - 4.5.3 Computation of multivariate indices 60

4.6 Simulation study on the relevance of variable order 64
4.7 Application 67
4.8 Conclusions 74

5 Proper evaluation of non-stationary time series models 77
5.1 Introduction 77
5.2 Review: Evaluation of Regional Climate Models 79
5.3 Methodology 82
5.3.1 Discussion of naive evaluation approaches 82
5.3.2 Moving scores and divergences 83
5.3.3 Selection of moving windows 86
5.4 Organization of simulation studies 94
5.5 Simulation study: Changepoint scenario 96
5.6 Simulation study: Trend scenario 115
5.7 Simulation study: Periodicity scenario 134
5.8 Summary of simulation study results 153
5.9 Case study: Evaluation of Regional Climate Models 154
5.9.1 Overall model assessment 154
5.9.2 Temporal evaluation 155
5.9.3 Spatial evaluation 157
5.9.4 Comparison of results to an evaluation based on linear trends 159
5.10 Conclusions and outlook 162

Appendix 165
A List of recurring acronyms 165
B Evaluation of Regional Climate Models: Further results 167

Bibliography 171

1 Introduction

Years after the establishment of the Intergovernmental Panel on Climate Change (IPCC) in 1988, *climate change* has slowly become one of the hot topics concerning our modern day society. Ever since 1990, the IPCC has published several reports on the current state of knowledge on climate change. Its latest Assessment Report (AR5, see IPCC, 2014, for the synthesis report) consolidates our current understanding of climate change, its causes and impacts, and discusses possible adaptation and mitigation strategies. As a contribution to all the endeavors to better understand (different aspects of) climate change, this thesis in the area of applied statistics is concerned with two different topics of relevance in the wide field of climate change research. These two topics are *drought* (respectively drought modeling) and the *evaluation of climate models* (respectively model evaluation under the presence of non-stationarity; see also Flato et al., 2013, for Chapter 9 of the contribution of Working Group I to the Fifth Assessment Report of the IPCC on the evaluation of climate models). The first part of the following introduction on drought modeling and the development of novel drought indices is based on Erhardt and Czado (2016).

Drought modeling and the development of novel drought indices 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 (see e.g. Mishra and Singh, 2010; AghaKouchak et al., 2015). *Drought indices* aim to quantify how dryness conditions evolve over time (based on time series of drought-relevant variables) and enable a classification of the severity of drought events. 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). Despite the fact, that there already exist plenty of different drought indices (see also Section 2.1 and Section 4.1 for a detailed discussion of established drought indices), the development of novel drought indices is still a vibrant field of research. This is especially due to the fact, that "drought is best characterized by multiple climatological and hydrological parameters" (Mishra and Singh, 2010) and that 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)

- *ground water drought* (decrease in groundwater recharge, levels and discharge)

are driven by different variables. Recently, there have been several attempts to develop *multivariate drought indices* (see e.g. Kao and Govindaraju, 2010; Hao and AghaKouchak, 2013, 2014; Farahmand and AghaKouchak, 2015), i.e. measures that summarize the dryness information captured in at least two different variables in one indicator of drought severity (see also Section 2.2). In this thesis (Chapter 4), we motivate and present a flexible and statistically sound approach for the calculation of standardized uni- and multivariate drought indices. The novel (multivariate) method allows the end-user to decide which type(s) of drought to investigate and which variables are relevant for her or his specific application. In contrast to earlier methods, it allows to incorporate more than two variables. Inter-variable dependencies are modeled based on vine copulas (see e.g. Aas et al., 2009, as well as Section 2.3, where we provide the necessary background on (vine) copulas). Flexible modeling of the full multivariate distribution of interest is crucial in order to account for the joint occurrence of extremes of different drought drivers.

Model evaluation under the presence of non-stationarity Climate models (which are a mathematical description of certain processes in the Earth's climate system) are the main tool to learn about possible future changes of the climate. Also the discussion on climate change in the Fifth Assessment Report of the IPCC (see IPCC, 2014, for the synthesis report) is based on projections of future states of the climate which assume different scenarios of greenhouse gas and air pollutant emissions and land use. A multitude of climate models is used to simulate future states of the climate (that is, variables like surface temperature and precipitation, amongst others). In order to evaluate the accuracy of different climate models, simulations from these models are compared to historical observations. One way to perform such a comparison is to compare simulations to observations separately for each time unit (and each spatial unit). Such a point-wise comparison (e.g. based on one of the metrics discussed in Hyndman and Koehler, 2006) is obviously the most straightforward approach to model evaluation. However, this approach might be rather less adequate for the evaluation of climate models based on a daily or even finer temporal resolution. As climate models aim at modeling the long-term evolution of the climate, they can not provide accurate simulations/forecasts on a daily basis, they rather intend to model/simulate the characteristics of the climate for longer time periods (several consecutive days/weeks). In other words, a climate model respectively a simulation/forecast from a climate model for a specific time instance is associated with uncertainty. A point-wise evaluation neglects this uncertainty. Proper scoring rules/scores (see e.g. Gneiting and Raftery, 2007) are a possible remedy. Assuming a distribution for the model/forecast, a scoring rule evaluates how well the observation of the modeled/forecast quantity fits to this distribution. One can further argue, that also the observation is associated with uncertainty. Then, proper divergence functions/divergences (see Thorarinsdottir et al., 2013) can be used to compare a distribution for the observed quantity to the model/forecast distributions. However, considering these options from a practical point of view, three questions arise:

1. How can we assess the model/forecast distribution, if all we have is one time series of realizations from the model/deterministic forecasts?

2. How can we assess the distribution of the observed quantity, if all we have is one time series of realizations of that quantity?

3. Can we assume one and the same distribution for different time instances, that is, are the distributions stationary?

To answer these questions we cling to our application of climate model evaluation. Under the presence of climate change and seasonality, it is obvious, that we have to negate the third question. This makes it more difficult to answer the first two questions. Our answer assumes, that the characteristics of the variables of interest (modeled by the climate model) change only gradually, and can be considered as (approximately) stationary for short time windows. Then, for each of these time windows, we can construct empirical distributions based on the corresponding realizations. This idea is the basis for the novel evaluation approaches under the presence of non-stationarity introduced in this thesis (Chapter 5). It is not restricted to the evaluation of climate models. Hence, we introduce it in a general setting. The (moving) time windows, for which we assume stationarity are selected based on a changepoint detection algorithm (see Killick et al., 2012). We propose and compare three different window selection strategies. Based on the samples/empirical distributions corresponding to these (moving) windows we compute time series of (moving) scores and divergences. This allows to assess the model performance over time.

Outline of the thesis In Chapter 2 we provide some background on standardized and multivariate drought indices. Being the heart of the novel drought indices which we propose, we further introduce (vine) copulas. Furthermore, we introduce some metrics which will be used to validate the novel drought indices. As the backbone of our novel evaluation technique, we introduce proper scoring rules and divergence functions. For comparison we give an overview of traditional evaluation measures used in a time series context. Moreover, we provide an introduction to the problem of changepoint detection, with a focus on the PELT (Pruned Exact Linear Time) method (see Killick et al., 2012).

After providing the necessary theoretical background on which this thesis is built (Chapter 2), we introduce the data sets used to illustrate the new methodology presented in this thesis (Chapter 3). Finally, Chapter 4 elaborates on the novel standardized uni- and multivariate drought indices. Their development is discussed in detail and illustrated with an example. Chapter 5 covers the novel technique for model evaluation under the presence of non-stationarity. In a simulation and a case study, we show the merits of the proposed methodology.

2 Preliminaries

2.1 Standardized drought indices

Among the multitude of methods applied to quantify drought (see e.g. Mishra and Singh, 2010, and reference therein), there has been a trend towards *standardized drought indices* (see e.g. Bachmair et al., 2016). Following the original suggestion of the Standardized Precipitation Index (SPI) by McKee et al. (1993) (see also Edwards and McKee, 1997) the concept of standardized drought indices has been used for/extended to various drought-relevant variables (see e.g. Bachmair et al., 2016, and reference therein). We list the corresponding indices in chronological order:

SPI	Standardized Precipitation Index (McKee et al., 1993)
SRI	Standardized Runoff Index (Shukla and Wood, 2008)
SPEI	Standardized Precipitation Evapotranspiration Index (Vicente-Serrano et al., 2010)
SSI	Standardized Streamflow Index (Vicente-Serrano et al., 2012)
SGI	Standardised Groundwater level Index (Bloomfield and Marchant, 2013)
SSI	Standardized Soil moisture Index (AghaKouchak, 2014)
SMRI	Standardized Snow Melt and Rain Index (Staudinger et al., 2014)
SPDI	Standardized Palmer Drought Index (Ma et al., 2014)

Subsequently, we outline in a general setting how standardized drought indices are calculated using the SPI-method introduced by McKee et al. (1993). A (detailed) description of the mathematical procedure behind the SPI can be found in Edwards and McKee (1997). To illustrate the single steps of the standardized drought index computation in Figures 2.1–2.3 we use a time series of monthly precipitation aggregates obtained from the Deutscher Wetterdienst (2015). We consider the precipitation (PRE) time series for Regensburg (Germany) for the 30 year period 1951–1980.

Computation of standardized drought indices according to McKee et al. (1993) Let now x_{t_k}, $k = 1, \ldots, N$ be a monthly time series of a drought-relevant variable (e.g. precipitation). Then we can consider the time index t_k as a 2-tupel (m_k, y_k), where $m_k \in \{1, \ldots, 12\}$ (1 =January, ..., 12 =December) represents the month and $y_k \in \mathbb{Z}$ the year corresponding to t_k. Using this notation, the computation of a standardized drought index (following the SPI-method, McKee et al., 1993) comprises the following steps in the given order:

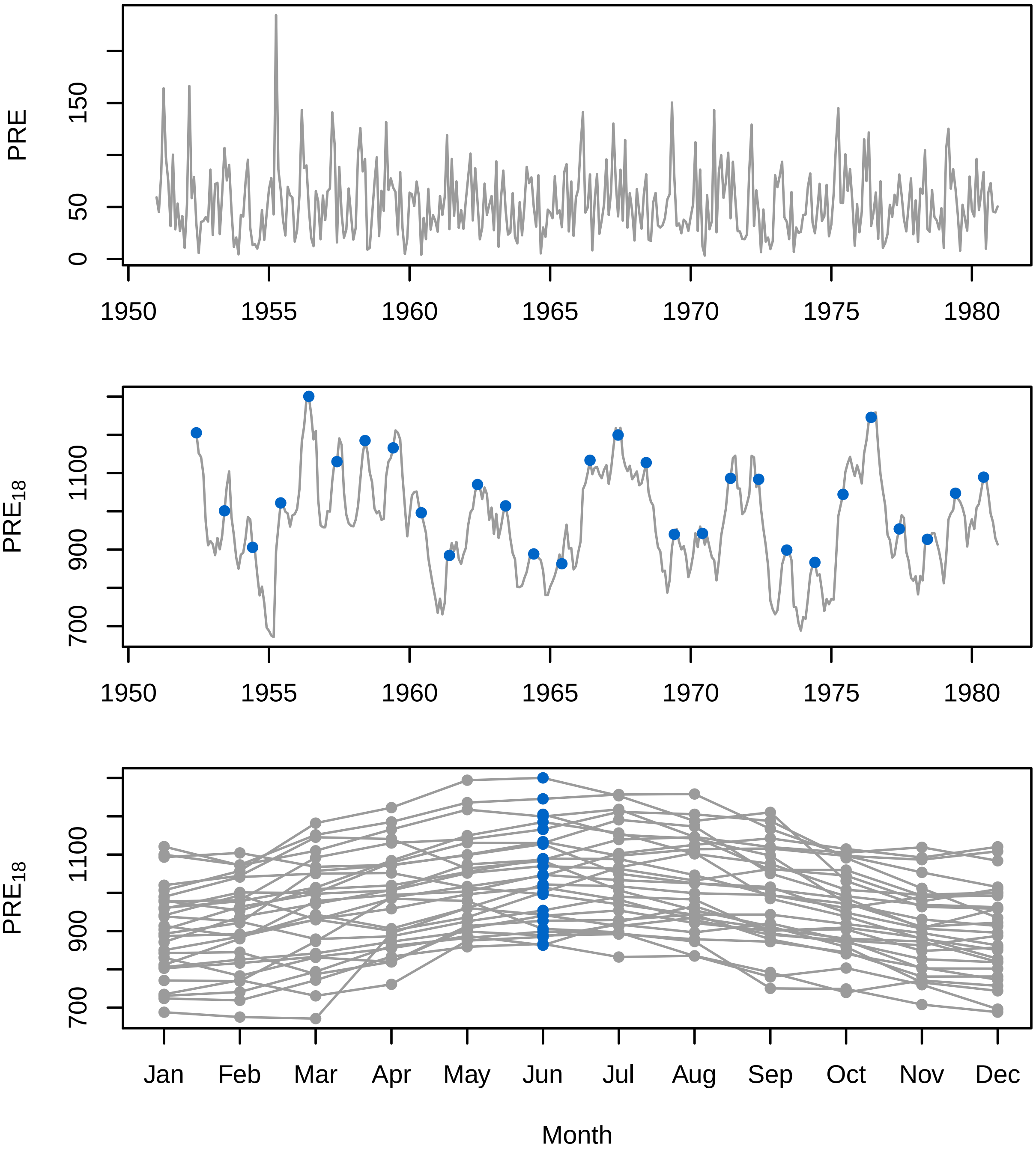


Figure 2.1: Illustration of Steps 1 and 2 of the computation of standardized drought indices in Section 2.1. The original precipitation (PRE) time series (upper panel) is aggregated at time scale $\ell = 18$ (middle panel) and the 12 month-wise sub-series are extracted (lower panel). The blue dots correspond to the aggregated PRE_{18} observations for June.

1. For a selected *time scale/aggregation period* $\ell \in \mathbb{N}$ calculate the *aggregated time series* x_{ℓ,t_k} as $x_{\ell,t_k} = \sum_{j=0}^{\ell-1} x_{t_{k-j}}$, $k = \ell, \ldots, N$.

2. From the aggregated time series x_{ℓ,t_k}, $k = \ell, \ldots, N$, extract the 12 *month-wise sub-series* $\boldsymbol{x}_{\ell,m} \coloneqq (x_{\ell,t_k})_{k \in \mathcal{K}(m)} = \left\{x_{\ell,(m,y_k)}, k \in \mathcal{K}(m)\right\}$, $m = 1, \ldots, 12$, where $\mathcal{K}(m) \coloneqq$

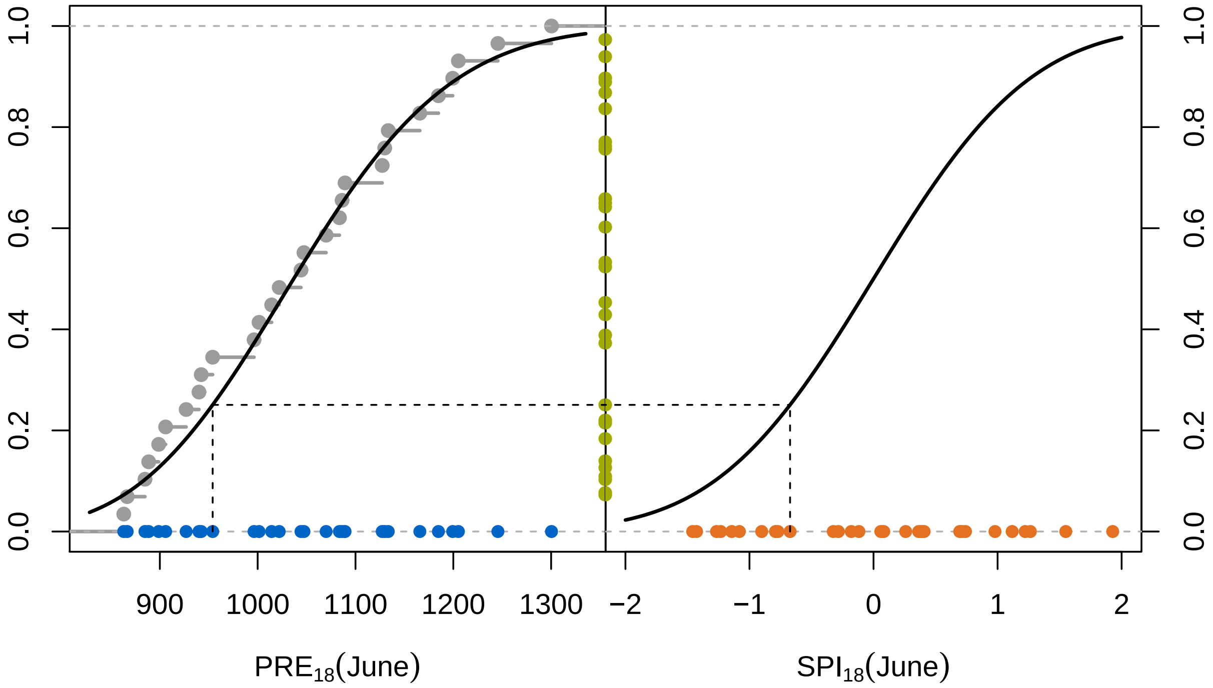


Figure 2.2: Illustration of Steps 3 and 4 of the computation of standardized drought indices in Section 2.1. The CDF of a gamma distribution (black line, left panel) is fitted to the month-wise sub-series of the (at time scale $\ell = 18$) aggregated precipitation (PRE$_{18}$) time series corresponding to the June observations (blue dots). Based on the fitted CDF each observation (blue dots) is first transformed to a standard uniform distribution (green dots) and then transformed further (orange dots) using a standard normal distribution CDF (black line, right panel). Then the orange dots represent the Standardized Precipitation Index (SPI$_{18}$) values corresponding to the month of June.

$\{k : m_k = m\}$ defines the set of all time indices corresponding to a particular month m.

3. To each of the 12 month-wise sub-series $\boldsymbol{x}_{\ell,m}$, $m = 1, \ldots, 12$, separately fit a parametric probability distribution with CDF $F(x; \boldsymbol{\theta}_m)$ parametrized by a (set of) parameters $\boldsymbol{\theta}_m$. (This step incorporates the selection of an adequate distribution family and the computation of the corresponding parameter estimates $\widehat{\boldsymbol{\theta}}_m$, $m = 1, \ldots, 12$.)

4. Using the probability distributions estimated in the previous step, transform each month-wise sub-series $\boldsymbol{x}_{\ell,m}$, $m = 1, \ldots, 12$, separately to a normal distribution. Hence, the standardized drought index time series is obtained as $\mathrm{SDI}_\ell(t_k) := \Phi^{-1}\Big(F(x_{\ell,t_k}; \widehat{\boldsymbol{\theta}}_{m_k})\Big)$, $k = \ell, \ldots, N$, where Φ is the cumulative distribution function (CDF) corresponding to the standard normal distribution.

The choice of the distribution family in Step 3 depends on the variable under consideration. Candidate distributions which were considered for (at least one of) the standardized indices listed in the beginning of the section are the gamma, log-logistic, Pearson type III, generalized extreme value, beta, log-normal, normal, Weibull and the generalized Pareto distribution. Also non-parametric approaches were considered for the transformation to a standard normal distribution. For instance AghaKouchak (2014) used the empirical Gringorten plotting position (Gringorten,

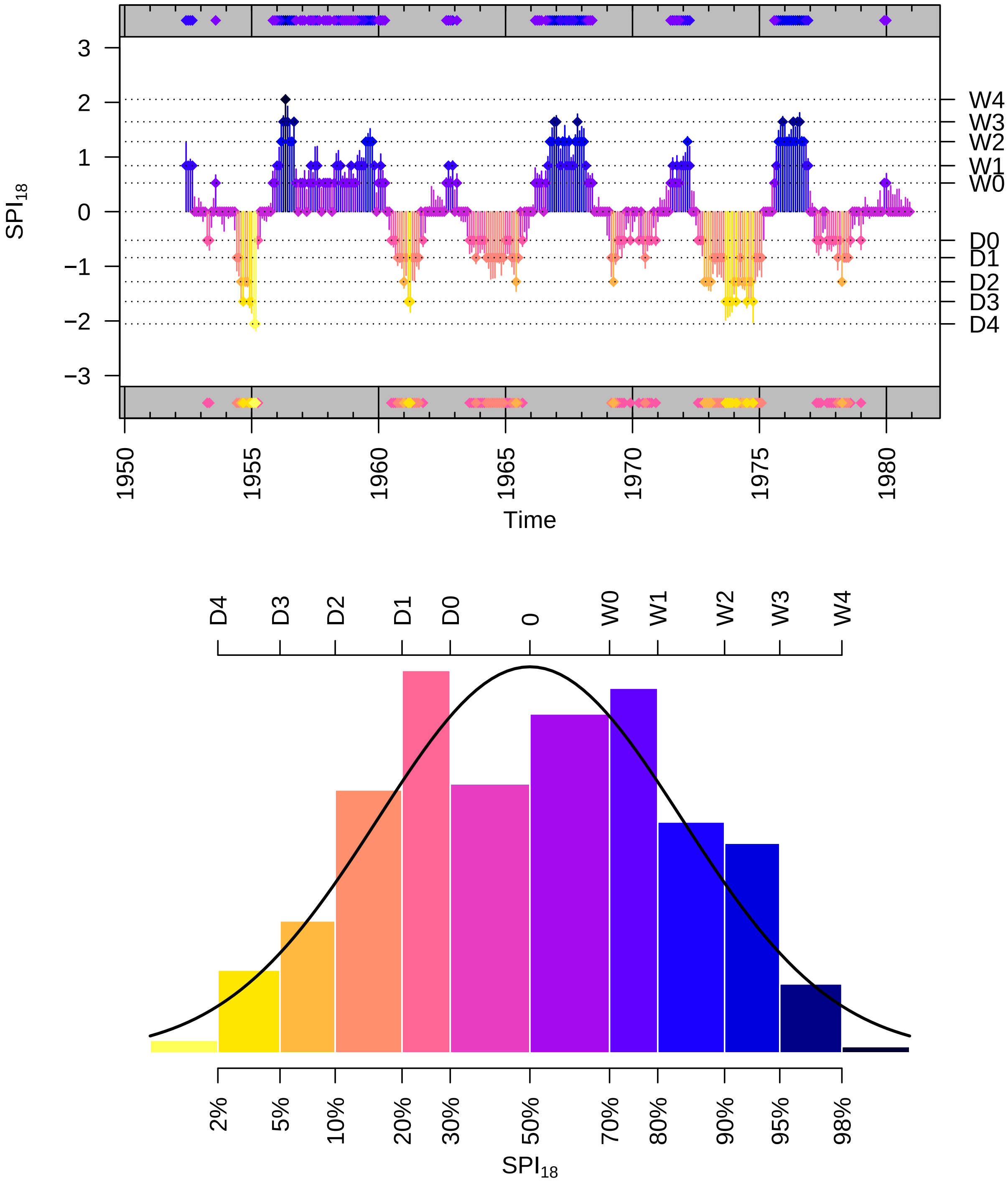


Figure 2.3: Illustration of the computation of standardized drought indices in Section 2.1. The upper panel shows the SPI_{18} (Standardized Precipitation Index, time scale 18) time series for Regensburg. The color-coding reflects the severity of dry-/wetness according to the different (D/W) categories specified in Table 2.1 (which are defined based on certain quantiles of the standard normal distribution). For better identification of dry/wet periods (upper panel) points at the bottom/top indicate points in time of dry/wet conditions. A histogram of the SPI_{18} realizations (lower panel) shows the grouping into the different categories. The PDF of a standard normal distribution is indicated for comparison.

1963), and Bloomfield and Marchant (2013) considered a rank transformation instead of fitting a parametric distribution.

Figures 2.1–2.3 illustrate the computation of standardized drought indices outlined above in the case of a precipitation time series and for a time scale of 18. Figure 2.1 illustrates the aggregation of the precipitation time series for $\ell = 18$ and the extraction of the 12 monthly sub-series. It shows that a differentiation between different months is meaningful, as the characteristics of the time series varies with the season. Figure 2.2 illustrates the month-wise transformation of the aggregated time series to a standard normal distribution, for the month of June. Here a gamma distribution was employed. From the plot it becomes clear that the result of the transformation depends very much on the chosen distribution family and on the length of the time series (sample size). The transformation results only in an approximately standard normal distributed sample. Figure 2.3 visualizes the resulting Standardized Precipitation Index (SPI_{18}) and the grouping of its realizations into certain dry-/wetness categories (see Table 2.1). The figure identifies the most severe drought conditions during the years 1954/55, 1961 and 1973/74. The histogram which summarizes the SPI_{18} realizations into dry-/wetness categories shows a general lack in the approximation of a standard normal distribution.

Classification of standardized drought indices To classify the values of standardized drought indices we use the dry-/wetness categories as defined in Table 2.1, which are based on quantiles of the standard normal distribution (cp. Svoboda et al., 2002).

Table 2.1: Dryness (D) and wetness (W) categories for standardized drought indices (SI), defined based on certain quantiles of a standard normal distribution.

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

2.2 Multivariate drought indices

Besides the trend towards the development and favored application of standardized drought indices (see Section 2.1), in recent years there has also been a trend towards drought indices which aim to quantify drought based on multiple input variables. The review paper of Hao and Singh (2015) discusses different approaches undertaken to join drought information captured in different variables. They differentiate between indices that are obtained based on

- blending different drought indicators,
- a water balance model,

- latent variables computed from observed variables,
- linear combinations of drought indicators,
- joint distributions (copulas), and
- principal component analysis (PCA).

In the literature there is no unified naming convention for these different types of (multivariate) drought indices. Such indices have been termed aggregate, combined, composite, comprehensive, hybrid, integrated, joint, multi-scalar and multivariate (see Hao and Singh, 2015, and reference therein). In this thesis we are particularly interested in drought indices which take variable interdependencies into consideration by modeling a joint, multivariate distribution. Throughout the thesis we will use the naming convention *multivariate drought indices* to address this class of drought indices.

2.3 Dependence modeling with copulas

In this section, we provide the necessary background on vine copulas, which is required to fully understand the modeling ideas behind the novel multivariate drought indices introduced in this thesis. Also known as pair-copula constructions, vine copulas allow to model highly flexible and asymmetric dependence structures in dimension $d > 2$, by constructing a d-dimensional copula based on bivariate building blocks, so called pair-copulas. Subsequently, we introduce copulas in general (Section 2.3.1), explain vine copulas (Section 2.3.2) with a focus on so called canonical vine copulas as well as their inference (Section 2.3.3), and show how a multivariate probability integral transformation can be calculated for this special class of copulas (Section 2.3.4).

2.3.1 Copulas

What is a copula? Considering the general setting of $d \geq 2$ dimensions, a *d-dimensional copula C* is a d-dimensional (cumulative) distribution function (CDF) on the d-dimensional unit hypercube $[0,1]^d$ with the univariate margins following a standard uniform distribution. We denote the corresponding copula density by c.

How can copulas be used to model dependencies? To explain the use of copulas in dependence modeling, we consider a setting where we are interested in modeling the dependence of $d \geq 2$ random variables $X_1, \ldots, X_d$ with marginal distributions $F_1, \ldots, F_d$ and joint multivariate distribution F. The famous theorem of Sklar (1959) states that in this setting there exists a copula C, such that

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{2.1}$$

In the case of an absolutely continuous CDF F the copula C is unique. From Equation (2.1) we see, that the copula C captures the dependence between the margins of the multivariate distribution F. Hence, Sklar's Theorem allows a separation of the modeling of univariate margins and multivariate dependence structure.

How does this work in practice? Let us now consider an i.i.d. sample $(x_{1,k}, \ldots, x_{d,k})$, $k = 1, \ldots, N$, from an unknown d-variate distribution function F. In order to model F, we now can make use of the above addressed separation (Sklar's Theorem/Equation (2.1)).

1. We transform the margins to so called *copula data* using the *probability integral transformations* (PIT)
$$u_{j,k} = F_j(x_{j,k}), \qquad j = 1, \dots, d, \quad k = 1, \dots, N, \tag{2.2}$$
where F_j, $j = 1, \dots, d$, are the univariate marginal CDFs corresponding to F.

2. We model the dependence of the copula data $(u_{1,k}, \dots, u_{d,k})$, $k = 1, \dots, N$, by means of a d-dimensional copula C.

However, the marginal distributions F_j, $j = 1, \dots, d$, in the first step are usually unknown. Hence, they have to be estimated, either parametrically or non-parametrically. If the marginal distributions in the above *two-step procedure* are estimated using parametric marginal models, we speak of the *inference functions for margins (IFM)* method (see Joe and Xu, 1996). If they are estimated based on empirical distributions we call the above procedure *semi-parametric* (see Genest et al., 1995).

How do we model copulas? It is common practice in dependence modeling to model the copula C in Equation (2.1) using a parametric copula $C(\cdot; \boldsymbol{\theta})$, parametrized by one or more parameters, here summarized in a vector $\boldsymbol{\theta}$. The literature provides a wide range of parametric copula families (see e.g. Joe, 2014, Chapter 4).

Which are the most popular parametric copula families? One of the most popular classes of copulas are *elliptical copulas*. Considering an arbitrary elliptical multivariate distribution F with known marginal distributions F_j, $j = 1, \dots, d$, elliptical copulas (see Fang et al., 2002; Frahm et al., 2003) are derived easily from Equation (2.1) provided by Sklar's Theorem. Substituting x_j in Equation (2.1) by $F_j^{-1}(u_j)$ for all $j = 1, \dots, d$, results in the formula

$$C(u_1, \dots, u_d) = F\left(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)\right),$$

which yields the copula corresponding to our arbitrary elliptical multivariate distribution F. Popular elliptical copulas are the *Gaussian* and the *Student t copula* derived from the multivariate Gaussian and Student t distribution, respectively. Hence, their bivariate versions are both parametrized by a correlation parameter $\rho \in (-1, 1)$ and the Student t copula has an additional degrees of freedom parameter $\nu > 0$.

Another popular copula class are called *Archimedean copulas*. We briefly introduce them in a two-dimensional setting. They are discussed in more detail for example by Joe (2001) and Nelsen (2006). A bivariate Archimedean copula is defined by

$$C(u_1, u_2) := \varphi^{[-1]}\left(\varphi(u_1) + \varphi(u_2)\right),$$

where the function $\varphi : [0, 1] \to [0, \infty]$ called *generator* has to be continuous, convex, strictly decreasing, with $\varphi(1) = 0$ and the pseudo-inverse $\varphi^{[-1]}$ is defined as

$$\varphi^{[-1]}(u) := \begin{cases} \varphi^{-1}(u) & \text{for } 0 \leq u \leq \varphi(0) \\ 0 & \text{for } \varphi(0) < u \leq \infty. \end{cases}$$

Popular examples of Archimedean copulas (each parametrized by one parameter θ) are the

Clayton $\left[\varphi(u) = \frac{1}{\theta}(u^{-\theta} - 1), \theta \in (0, \infty)\right]$,

Gumbel $\left[\varphi(u) = (-\ln u)^{\theta}, \theta \in [1, \infty)\right]$, *Frank* $\left[\varphi(u) = -\ln \frac{\exp(-\theta u)-1}{\exp(-\theta)-1}, \theta \in (-\infty, \infty)\backslash\{0\}\right]$, and

Joe $\left[\varphi(u) = -\ln\left(1-(1-u)^\theta\right), \theta \in [1, \infty)\right]$

copula (with the corresponding generators stated in brackets).

Another important copula, which is a limiting case of all copulas stated above is the *independence copula*. The d-dimensional independence copula is defined as

$$C(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j, \tag{2.3}$$

with density $c(u_1, \ldots, u_d) = 1$, for $(u_1, \ldots, u_d) \in [0,1]^d$. It can be used to model independence. Due to its CDF being a product it is also known as *product copula* and often denoted as Π.

What is the interpretation of the copula parameters? The copula parameters are often also called dependence parameters, as they characterize the magnitude of association among the variables under consideration. The parameters of (bivariate) elliptical and Archimedean copulas can for instance be related to the association measure Kendall's τ (see e.g. Embrechts et al., 2003, for mathematical expressions of these relationships). The association measure Kendall's τ (see e.g. Kruskal, 1958) for a variable pair is defined as

$$\begin{aligned}\tau &:= \mathrm{P}((X_1 - Y_1)(X_2 - Y_2) > 0) - \mathrm{P}((X_1 - Y_1)(X_2 - Y_2) < 0)\\ &= 2\,\mathrm{P}((X_1 - Y_1)(X_2 - Y_2) > 0) - 1,\end{aligned}$$

where the random variable pairs (X_1, X_2) and (Y_1, Y_2) are i.i.d. One important property of Kendall's τ is, that it depends only on the copula C corresponding to the variable pair under consideration. It can be written as

$$\tau = 4 \int_{[0,1]^2} C(u_1, u_2) \mathrm{d}C(u_1, u_2) - 1$$

(see e.g. Nelsen, 2006, Chapter 5).

How do copulas specify dependence in bivariate distribution tails? As we have seen, there are plenty of different parametric copula families, whose parameters influence the magnitude of dependence. Moreover, their functional shapes differ. This results in different behavior of these families in terms of the joint occurrence of extreme values. This behavior can be characterized using tail dependence coefficients (see e.g. Joe, 1993; Nelsen, 2006, Section 5.4). Let $(X_1, X_2) \sim F$ with corresponding copula C. Then the *lower tail dependence coefficient* is defined as the limit

$$\lambda_{\text{lower}} := \lim_{u \searrow 0} \mathrm{P}(X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)) = \lim_{u \searrow 0} \frac{C(u,u)}{u}.$$

Accordingly, the *upper tail dependence coefficient* is given by

$$\lambda_{\text{upper}} := \lim_{u \nearrow 1} \mathrm{P}(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)) = \lim_{u \nearrow 1} \frac{1 - 2u + C(u,u)}{1-u}.$$

If the limits λ_{lower} and λ_{upper} exist and lie in $\in (0,1]$, then the copula C has *lower* and/or *upper tail dependence*, respectively. Table 2.2 summaries which of the above addressed copulas have upper and/or lower tail dependence.

Table 2.2: Summary of the presence of upper/lower tail dependence for Gaussian (N), Student-*t* (t), Clayton (C), Gumbel (G), Frank (F) and Joe (J) copulas.

	N	t	C	G	F	J
upper	no	yes	no	yes	no	yes
lower	no	yes	yes	no	no	no

How do different copula families look like? What are their distributional characteristics? Here, we provide a visualization of how the dependence structures modeled by bivariate Gaussian (N), Student-*t* (t), Clayton (C), Gumbel (G), Frank (F) and Joe (J) copulas differ. Figure 2.4 provides contour plots which visualize the densities of bivariate distributions obtained from Equation (2.1), where the corresponding copulas from above were combined with standard normal margins. The provided plots show that both elliptical copulas (N and t) as well as the Frank copula (F) are symmetric with respect to both diagonals. Also the different tail behavior summarized in Table 2.2 can be observed. For more details on different copula families we refer to Joe (2014, Chapter 4).

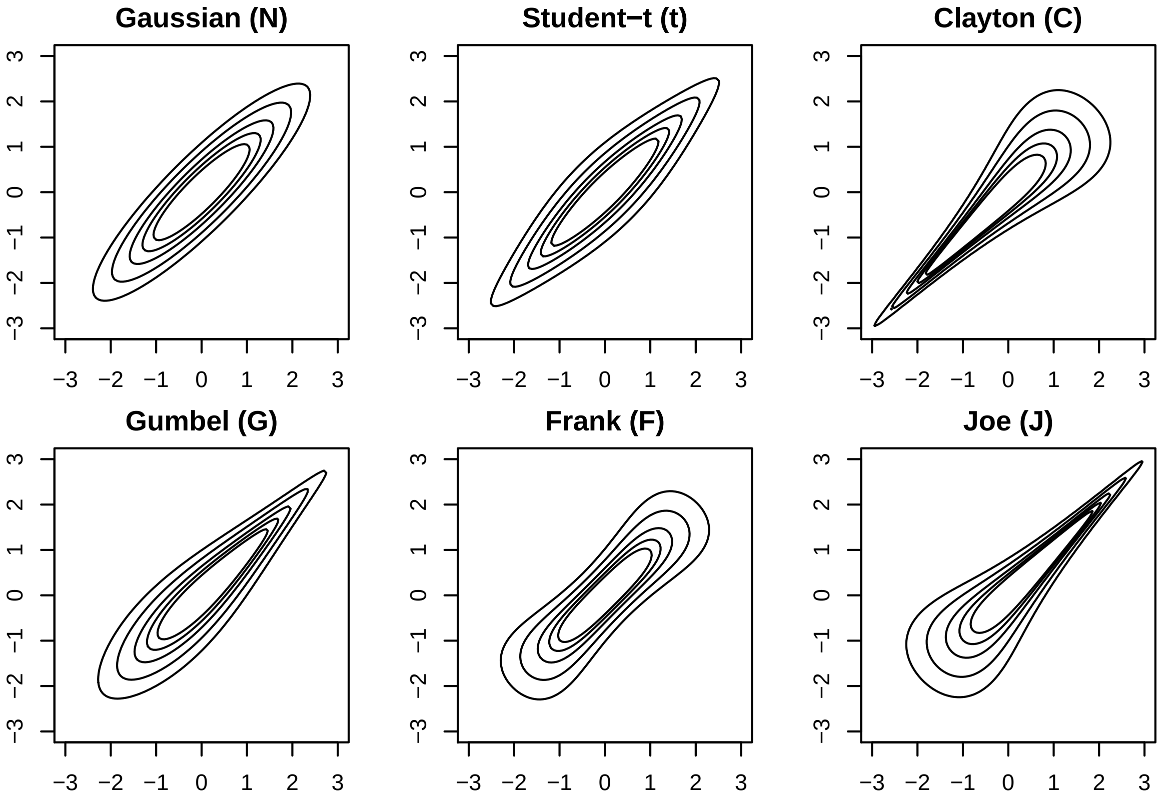


Figure 2.4: Visualization of bivariate Gaussian (N), Student-*t* (t), Clayton (C), Gumbel (G), Frank (F) and Joe (J) copulas: Density contour plots of bivariate distributions obtained from Equation (2.1) (Sklar's Theorem), where the corresponding copulas were combined with standard normal margins. For all copulas the (first) parameter was chosen such that Kendall's $\tau = 0.7$. The degrees of freedom parameter of the Student-*t* copula was fixed to $\nu = 5$.

How can (asymmetric) copulas be used to obtain copulas which model the opposite tail behavior or negative dependence? Many copulas are asymmetric (e.g. in terms of tail dependence) and often their parametrization is restricted to positive dependence. By (counterclockwise) rotation of a bivariate copula C we can obtain a copula which models negative dependence or the opposite tail behavior (see e.g. Joe, 1993). The corresponding copulas rotated by 90, 180 and 270 degrees are defined for $(u_1, u_2) \in [0,1]^2$ as

- $C^{90}(u_1, u_2) := u_2 - C(1-u_1, u_2)$ $[c^{90}(u_1, u_2) := c(1-u_1, u_2)]$,
- $C^{180}(u_1, u_2) := u_1 + u_2 - 1 + C(1-u_1, 1-u_2)$ $[c^{180}(u_1, u_2) := c(1-u_1, 1-u_2)]$,
- $C^{270}(u_1, u_2) := u_1 - C(u_1, 1-u_2)$ $[c^{270}(u_1, u_2) := c(u_1, 1-u_2)]$,

respectively, where the corresponding copula densities are stated in brackets. The copula rotated by 180 degrees is also called *survival copula*.

How do we select a copula and estimate its parameters? In order to select an appropriate parametric copula $C(\cdot;\boldsymbol{\theta})$ modeling the dependence of a given i.i.d. sample $(x_{1,k}, \ldots, x_{d,k})$, $k = 1, \ldots, N$, from an unknown d-variate distribution function F, we can make use of the above introduced two-step modeling procedure. Using either parametric marginal models or empirical estimates of the univariate marginal distributions we can transform our observations to pseudo copula data $\boldsymbol{u} := \{(u_{1,k}, \ldots, u_{d,k}), k = 1, \ldots, N\}$, as shown in Equation (2.2). Then, maximization of the pseudo log-likelihood

$$\ell_{\text{pseudo}}(\boldsymbol{\theta}; \boldsymbol{u}) = \sum_{k=1}^{N} \log c(u_{1,k}, \ldots, u_{d,k}; \boldsymbol{\theta}) \tag{2.4}$$

yields an estimate $\widehat{\boldsymbol{\theta}}$ for the parameter(s) of a specific parametric copula with density $c(\cdot;\boldsymbol{\theta})$. This allows to compare the fit of different candidate copula families to the copula data. For the comparison/selection, criteria like the Akaike information criterion

$$\text{AIC} = -2\ell_{\text{pseudo}}(\widehat{\boldsymbol{\theta}}; \boldsymbol{u}) + 2(\#\text{parameters}) \tag{2.5}$$

or the Bayesian information criterion

$$\text{BIC} = -2\ell_{\text{pseudo}}(\widehat{\boldsymbol{\theta}}; \boldsymbol{u}) + (\log N)(\#\text{parameters}) \tag{2.6}$$

can be used, where small values of these criteria are preferred. For further reading about copula selection and parameter estimation we refer to Joe (2014, Chapters 1, 5).

2.3.2 Vine copulas

Vine copulas are d-dimensional copula constructions composed of *pair-copulas* (bivariate copulas) only. Hence, they are also often referred to as *pair-copula constructions (PCCs)*. Due to their modularity, they allow very flexible modeling of non-Gaussian, asymmetric dependence structures, as the choice for the different bivariate building blocks can be made among a wide variety of different copula families. Pair-copula constructions were initially considered by Joe (1996), where the construction was based on distribution functions. Later Bedford and Cooke (2001, 2002), rediscovered such constructions. They systematized them based on densities. Inspired by the early work on PCCs, Aas et al. (2009) complemented the existing theory on the construction of such multivariate distributions with practical results, including methodology for statistical inference.

Pair-copula constructions We introduce pair-copula constructions (PCCs) in a general d-dimensional setting ($d \geq 2$). Let us consider a set of random variables $X_1, \ldots, X_d$, with joint probability density function f, univariate marginal distributions $F_1, \ldots, F_d$ and corresponding densities $f_1, \ldots, f_d$. Moreover, we denote conditional densities corresponding to conditional (multivariate) margins as $f_{\mathcal{I}|\mathcal{J}}$, where $\mathcal{I}$ and $\mathcal{J}$ are disjoint subsets of the index set $\{1, \ldots, d\}$. Then, apart from re-labeling, f can be decomposed uniquely into the following product of one marginal density and $d-1$ conditional densities $f_{j|1,\ldots,j-1}$, $j = 2, \ldots, d$:

$$f(x_1, \ldots, x_d) = f_1(x_1) \cdot f_{2|1}(x_2|x_1) \cdot f_{3|1,2}(x_3|x_1, x_2) \cdots f_{d|1,\ldots,d-1}(x_d|x_1, \ldots, x_{d-1}) \tag{2.7}$$

Moreover, Equation (2.1) from Sklar's Theorem allows to express conditional densities $f_{i|\{j\}\cup\mathcal{D}}$, $i, j \in \{1, \ldots, d\}$, $i \neq j$, $\mathcal{D} \subseteq \{1, \ldots, d\} \setminus \{i, j\}$, as they occur in Equation (2.7) as

$$\begin{aligned} f_{i|\{j\}\cup\mathcal{D}}(x_i|x_k, k \in \{j\} \cup \mathcal{D}) &= \frac{f_{i,j|\mathcal{D}}(x_i, x_j|x_k, k \in \mathcal{D})}{f_{j|\mathcal{D}}(x_j|x_k, k \in \mathcal{D})} \\ &= \frac{\partial^2}{\partial x_i \partial x_j} F_{i,j|\mathcal{D}}(x_i, x_j|x_k, k \in \mathcal{D}) \frac{1}{f_{j|\mathcal{D}}(x_j|x_k, k \in \mathcal{D})} \\ &= c_{i,j;\mathcal{D}}\left(F_{i|\mathcal{D}}(x_i|x_k, k \in \mathcal{D}), F_{j|\mathcal{D}}(x_j|x_k, k \in \mathcal{D})|x_k, k \in \mathcal{D}\right) \qquad (2.8) \\ &\quad \cdot f_{i|\mathcal{D}}(x_i|x_k, k \in \mathcal{D}), \end{aligned}$$

where $c_{i,j;\mathcal{D}}$ is the density of the copula associated with the conditional distribution $F_{i,j|\mathcal{D}}$. For $\mathcal{D} = \emptyset$ Equation (2.8) simplifies to

$$f_{i|j}(x_i|x_j) = c_{i,j}\left(F_i(x_i), F_j(x_j)\right) f_i(x_i). \tag{2.9}$$

Hence, replacement of the conditional densities in Equation (2.7) by recursive application of Equation (2.8) allows to write/decompose our multivariate density f as/into a product of bivariate copula densities and marginal densities. Such a decomposition is called *pair-copula construction (PCC)*. Note however, that the recursive application of Equation (2.8) can be done in different ways, i.e. there is not a unique decomposition.

Now, we illustrate the above procedure in three dimensions. For $d = 3$ Equation (2.7) reduces to

$$f(x_1, x_2, x_3) = f_1(x_1) f_{2|1}(x_2|x_1) f_{3|1,2}(x_3|x_1, x_2).$$

Exploiting Equation (2.9), the second term in the product equals

$$f_{2|1}(x_2|x_1) = c_{1,2}\left(F_1(x_1), F_2(x_2)\right) f_2(x_j). \tag{2.10}$$

The third term $f_{3|1,2}(x_3|x_1, x_2)$ can be decomposed in two different ways by recursive application of Equation (2.8). We obtain

$$\begin{aligned} f_{3|1,2}(x_3|x_1, x_2) &= c_{2,3;1}\left(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)|x_1\right) c_{1,3}\left(F_1(x_1), F_3(x_3)\right) f_3(x_3) \\ &= c_{1,3;2}\left(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)|x_2\right) c_{2,3}\left(F_2(x_2), F_3(x_3)\right) f_3(x_3). \end{aligned}$$

Hence, omitting the arguments, we obtain the two PCCs

$$\begin{aligned} f(x_1, x_2, x_3) &= c_{2,3;1} c_{1,3} c_{1,2} f_1 f_2 f_3 \\ &= c_{1,3;2} c_{2,3} c_{1,2} f_1 f_2 f_3. \end{aligned}$$

Computation of conditional distribution functions from pair-copulas Until now we have not addressed how we obtain the arguments of the pair-copulas composing a PCC (see e.g. Equation (2.8)). We provide a general expression how such conditional distribution functions of the form $F_{i|\{j\}\cup\mathcal{D}}(x_i|x_k, k \in \{j\} \cup \mathcal{D})$ are computed. As before (see Equation (2.8)) we consider indices $i, j \in \{1, \dots, d\}$, $i \neq j$, and an arbitrary index set $\mathcal{D} \subseteq \{1, \dots, d\} \setminus \{i, j\}$. Following Joe (1996) we obtain

$$F_{i|\{j\}\cup\mathcal{D}}(x_i|x_k, k \in \{j\} \cup \mathcal{D}) = \frac{\partial\, C_{i,j;\mathcal{D}}\left(F_{i|\mathcal{D}}(x_i|x_k, k \in \mathcal{D}), F_{j|\mathcal{D}}(x_j|x_k, k \in \mathcal{D})|x_k, k \in \mathcal{D}\right)}{\partial F_{j|\mathcal{D}}(x_j|x_k, k \in \mathcal{D})}, \tag{2.11}$$

where $C_{i,j;\mathcal{D}}$ is the bivariate copula associated with the conditional distribution $F_{i,j|\mathcal{D}}$. For $\mathcal{D} = \emptyset$ Equation (2.11) simplifies to

$$F_{i|j}(x_i|x_j) = \frac{\partial\, C_{i,j}\left(F_i(x_i), F_j(x_j)\right)}{\partial F_j(x_j)}. \tag{2.12}$$

Similar to Equation (2.8), Equation (2.11) hast to be applied recursively in order to calculate $F_{i|\{j\}\cup\mathcal{D}}(x_i|x_k, k \in \{j\} \cup \mathcal{D})$ for $\mathcal{D} \neq \emptyset$.

Simplifying assumption As we have seen, PCC densities are composed of several pair-copula densities of the form $c_{i,j;\mathcal{D}}$ (see Equation (2.8)) and the computation of their arguments requires repeated evaluation of derivatives of $C_{i,j;\mathcal{D}}$ (see Equation (2.11)). In order to facilitate the work with PCCs, we consider the commonly used *simplifying assumption*

$$\begin{aligned} C_{i,j;\mathcal{D}}(\cdot,\cdot|x_k, k \in \mathcal{D}) &= C_{i,j;\mathcal{D}}(\cdot,\cdot), \\ c_{i,j;\mathcal{D}}(\cdot,\cdot|x_k, k \in \mathcal{D}) &= c_{i,j;\mathcal{D}}(\cdot,\cdot). \end{aligned}$$

It is assumed that pair-copulas associated with conditional distributions $F_{i,j;\mathcal{D}}$ do not depend on the conditioning values x_k, $k \in \mathcal{D}$. In the remainder of the thesis, we make this simplifying assumption. Hence, we drop the corresponding conditioning terms in Equations (2.8) and (2.11).

h-functions As the computation of simplified PCCs (i.e. PCCs with simplifying assumption) involves repeated evaluation of partial derivatives of copulas $C_{i,j;\mathcal{D}}$, Aas et al. (2009) defined so called *h-functions*

$$h_{i|j;\mathcal{D}}\left(v_i|v_j\right) \coloneqq \frac{\partial}{\partial v_j} C_{i,j;\mathcal{D}}\left(v_i, v_j\right). \tag{2.13}$$

Application of Equation (2.11) for the conditional distributions $C_{i|\{j\}\cup\mathcal{D}}$ and utilization of the h-function notation (2.13) yields

$$C_{i|\{j\}\cup\mathcal{D}}\left(u_i|u_k, k \in \{j\} \cup \mathcal{D}\right) = h_{i|j;\mathcal{D}}\left(C_{i|\mathcal{D}}(u_i|u_k, k \in \mathcal{D})|C_{j|\mathcal{D}}(u_j|u_k, k \in \mathcal{D})\right). \tag{2.14}$$

Hence, conditional distributions $C_{i|\{j\}\cup\mathcal{D}}$ can be evaluated by nesting h-functions.

Regular vines Now we introduce vines which are a tool that helps to obtain/organize valid PCCs. Generally speaking, the structure of (d-dimensional) PCCs is organized using a nested set of trees $\mathcal{T}_1, \dots, \mathcal{T}_{d-1}$ (a tree is a connected graph without cycles) fulfilling certain conditions. The trees $\mathcal{T}_k$, $k = 1, \dots, d-1$, are nested in the sense that edges of a tree become nodes in the subsequent tree. We formalize this in the following.

A d-dimensional *regular vine (R-vine) tree structure* $\mathscr{V}$ (see Bedford and Cooke, 2001, 2002) is defined as follows:

(a) $\mathscr{V} = (\mathcal{T}_1, \dots, \mathcal{T}_{d-1})$

(b) $\mathcal{T}_1 = (\mathcal{N}_1, \mathcal{E}_1)$ is a tree with node set $\mathcal{N}_1 = \{1, \dots, d\}$ and edge set $\mathcal{E}_1$.

(c) $\mathcal{T}_k = (\mathcal{N}_k, \mathcal{E}_k)$ is a tree with node set $\mathcal{N}_k = \mathcal{E}_{k-1}$ and edge set $\mathcal{E}_k$, for all $k = 2, \dots, d-1$.

(d) For all $k = 2, \dots, d-1$, two nodes $a, b \in \mathcal{N}_k$ in tree $\mathcal{T}_k$ may only be joined by an edge $e \in \mathcal{E}_k$, if the corresponding edges $a, b \in \mathcal{E}_{k-1}$ share a common node in $\mathcal{N}_{k-1}$ (*proximity condition*).

Note, that for all $k = 1, \dots, d-1$ it holds that each edge set $\mathcal{E}_k$ consists of $d-k$ edges due to $\mathcal{T}_k$ being a tree. Hence, the number of edges in a d-dimensional R-vine tree structure sums up to $d(d-1)/2$. From the above definition it becomes clear, that with increasing dimension the number of valid tree structures grows drastically.

Subsequently, we label the edges following the scheme $i, j; \mathcal{D}_k$. We call $\{i, j\}$ *conditioned set* and the set $\mathcal{D}_k$ consisting of $k-1$ elements *conditioning set*. Figure 2.5 provides an exemplary 5-dimensional R-vine tree structure, where the labeling of nodes and edges follows this scheme.

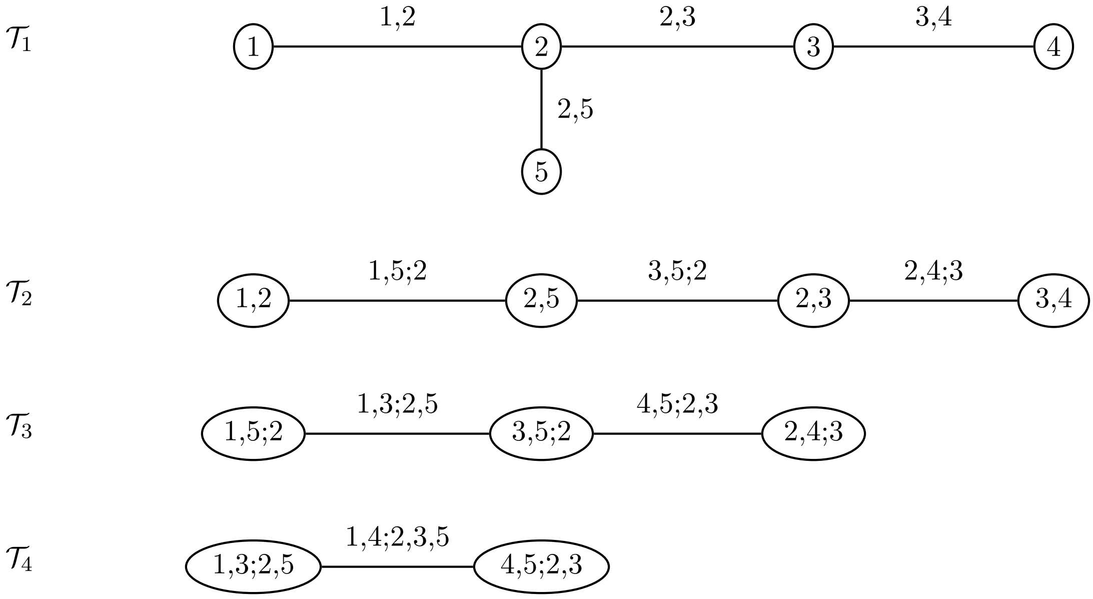


Figure 2.5: Exemplary 5-dimensional R-vine tree structure.

Canonical vines The literature (see e.g. Aas et al., 2009) features two sub-classes of regular vines, which exhibit special tree structures, namely canonical (C-) and drawable (D-) vines. For our application in Chapter 4 we focus on canonical vines, as their structure is of particular interest for this application.

For *canonical vines (C-vines)* each tree $\mathcal{T}_k$, $k = 1, \dots, d-1$, has a star like structure, i.e. one node is linked to all remaining nodes. That allows to order the variables under consideration by importance. The *variable order* determines for each tree, which variable plays the role of the root variable, i.e. the variable which occurs in all tree edges in the conditioned set. Hence, for variables named $1, 2, \dots, d$ and order $(1, 2, \dots, d)$, variable 1 is the root of tree $\mathcal{T}_1$, variable 2 is the root of tree $\mathcal{T}_2$ and so on. The last variable d remains as there are only $d-1$ trees. To further illustrate C-vines, Figure 2.6 provides a graphical representation of a d-dimensional C-vine.

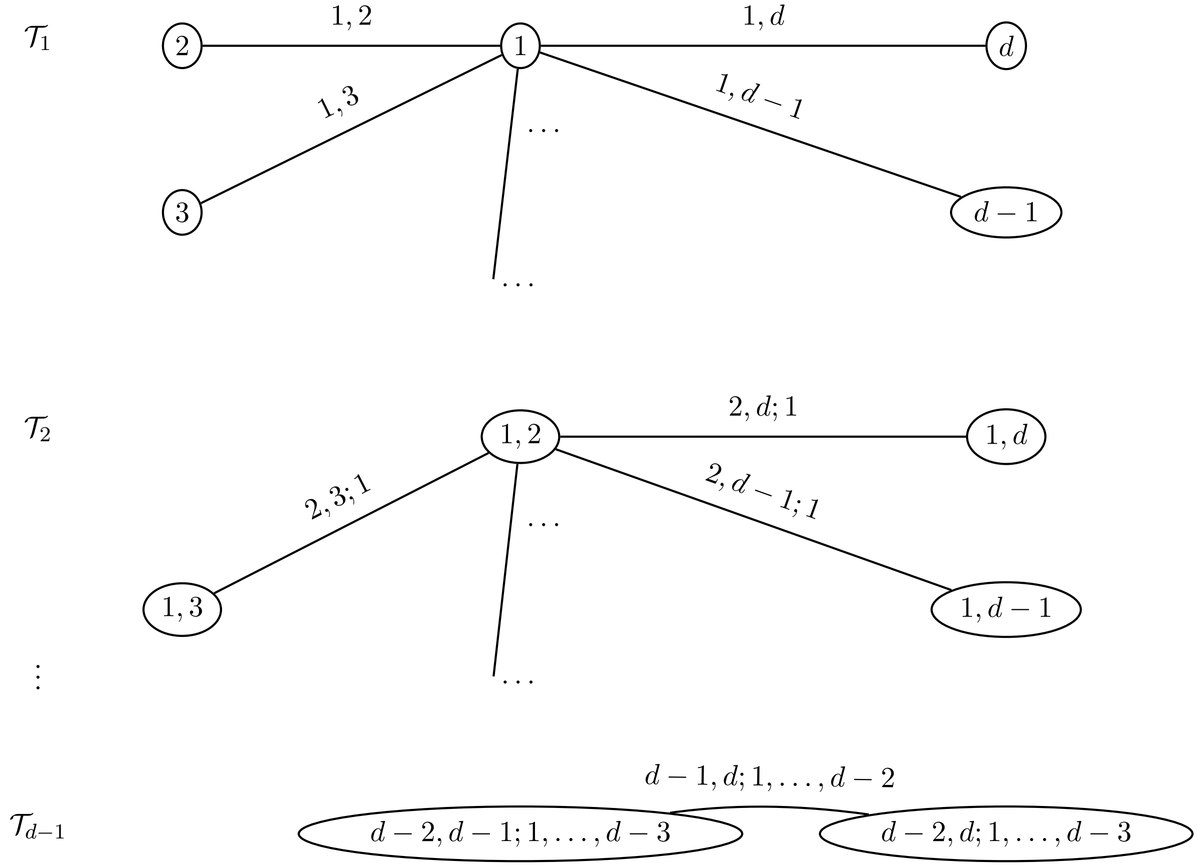


Figure 2.6: Illustration of d-dimensional C-vine tree structure.

Vine copulas As already indicated above, vines allow to construct valid d-dimensional copulas (distributions) based on $d(d-1)/2$ pair-copulas (and d univariate marginal distributions), see also Bedford and Cooke (2001, 2002). Hence, to obtain a *regular vine copula* we associate a bivariate copula (pair-copula) to each edge of an R-vine tree structure. The corresponding d-dimensional copula density is then obtained as the product of the pair-copula densities. Note, that here we will use parametric pair-copulas (see also Section 2.3.1). Thus, full specification of a parametric vine copula also requires specification of the corresponding pair-copula parameters. Subsequently, we treat this construction principle in more detail for the R-vine sub-class of C-vines. For a more general treatment we refer the reader to Czado (2010) and Kurowicka and Joe (2011).

Canonical vine copulas Now we outline the construction of a *canonical vine copula* in d dimensions. For this, we start from the canonical vine tree structure shown in Figure 2.6, corresponding to variables named $1, 2, \dots, d$ and variable order $(1, 2, \dots, d)$. Then a canonical vine copula is specified by assigning a parametric pair-copula $C_{i,j;\mathcal{D}}(\cdot, \cdot; \boldsymbol{\theta}_{i,j;\mathcal{D}})$ (with parameters $\boldsymbol{\theta}_{i,j;\mathcal{D}}$) to each edge $\{i, j; \mathcal{D}\}$ occurring in the vine trees $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{d-1}$. We list below, which edges occur for each tree $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{d-1}$:

($\mathcal{T}_1$) $\{1, 2\}$, …, $\{1, d\}$

($\mathcal{T}_2$) $\{2, 3; 1\}$, …, $\{2, d; 1\}$

$\vdots$

($\mathcal{T}_{d-1}$) $\{d-1, d; 1, \ldots, d-2\}$

The corresponding d-dimensional *canonical vine copula density* c is given as the product

$$c(u_1, \ldots, u_d; \boldsymbol{\theta}) = \prod_{i=1}^{d-1} \prod_{j=1}^{d-i} c_{i,i+j;\mathcal{D}_i} \left(C_{i|\mathcal{D}_i}(u_i | u_k, k \in \mathcal{D}_i; \boldsymbol{\theta}), C_{i+j|\mathcal{D}_i}(u_{i+j} | u_k, k \in \mathcal{D}_i; \boldsymbol{\theta}); \boldsymbol{\theta}_{i,i+j;\mathcal{D}_i} \right), \tag{2.15}$$

of the pair-copula densities $c_{i,i+j;\mathcal{D}_i}$ associated to the edges of all $d-1$ trees, where $\mathcal{D}_1 := \emptyset$, $\mathcal{D}_i := \{1 \ldots, i-1\}$, $i = 2, \ldots, d-1$, are the conditioning sets corresponding to the trees $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{d-1}$, respectively, the conditional distributions $C_{i|\mathcal{D}_i}$ and $C_{i+j|\mathcal{D}_i}$ can be obtained recursively using h-functions (see Equation (2.14)), and the pair-copula parameters are summarized in a vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_{i,j;\mathcal{D}})_{i=1,\ldots,d-1;j=1,\ldots,d-i}$. The outer product in Equation (2.15) runs through the trees $\mathcal{T}_i$, $i = 1, \ldots, d-1$. The inner product runs through all $d-i$ tree edges for each tree $\mathcal{T}_i$, $i = 1, \ldots, d-1$. Note, that Equation (2.15) can be used for arbitrary variable orders, after relabeling the variables according to their order.

To further illustrate canonical vine copulas we consider the three-dimensional case (i.e. $d = 3$). In this case we need to specify pair-copulas $C_{1,2}$, $C_{1,3}$ (tree $\mathcal{T}_1$) and $C_{2,3;1}$ (tree $\mathcal{T}_2$) for the variable pairs $(1,2)$, $(1,3)$ and $(2,3)$ given 1, respectively. Moreover, the corresponding parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{1,3}, \boldsymbol{\theta}_{2,3;1})$ have to be determined. Then the vine copula density c is given as

$$\begin{aligned} c(u_1, u_2, u_3; \boldsymbol{\theta}) = & \, c_{1,2}(u_1, u_2; \boldsymbol{\theta}_{1,2}) \cdot c_{1,3}(u_1, u_3; \boldsymbol{\theta}_{1,3}) \\ & \cdot c_{2,3;1}(h_{2|1}(u_2|u_1; \boldsymbol{\theta}_{1,2}), h_{3|1}(u_3|u_1; \boldsymbol{\theta}_{1,3}); \boldsymbol{\theta}_{2,3;1}), \end{aligned}$$

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 $C_{2,3;1}$.

Simulation from vine copulas A simulation algorithm for the general class of R-vine copulas is provided by Dißmann et al. (2013). Algorithms for the two sub-classes of D- and C-vine copulas are treated in Aas et al. (2009). Refer also to Joe (2014, Section 6.14) for a detailed discussion of all three algorithms. All these sampling algorithms are based on one common idea. To obtain a sample $u_1, \ldots, u_d$ from a d-dimensional vine copula C, we start with a sample $v_1, \ldots, v_d$, with each v_j, $j = 1, \ldots, d$, sampled i.i.d. from a standard uniform distribution. Then we fix $u_1 = v_1$ and successively apply inverse probability integral transforms $u_j = C^{-1}_{j|1,\ldots,j-1}(v_j|u_1, \ldots, u_{j-1})$, $j = 2, \ldots, d$, based on conditional distributions $C_{j|1,\ldots,j-1}$ specified by the vine copula C. Note, that the numbering of the variables $1, \ldots, d$ depends on the vine tree structure at hand.

2.3.3 Statistical inference for vine copulas

For the inference of vine copulas we assume a d-dimensional i.i.d. sample $(u_{1,k}, \ldots, u_{d,k})$, $k = 1, \ldots, N$, from a vine copula C. We summarize the sample as $\boldsymbol{u} := \{(u_{1,k}, \ldots, u_{d,k}), k = 1, \ldots, N\}$. As we have seen in Section 2.3.2 a model for the vine copula C consists of three components,

- a vine tree structure,
- pair-copulas (corresponding to the vine tree edges), and
- corresponding (pair-copula) parameters.

Thus, different settings for the inference of a vine copula model are possible:

(S1) The tree structure and pair-copulas are fixed/pre-selected. The parameters are unknown.

(S2) The tree structure is fixed/pre-selected. The pair-copulas and the corresponding parameters have to be specified.

(S3) All three components are unknown.

Sequential and maximum likelihood based parameter estimation in Setting (S1) is treated in detail by Aas et al. (2009) in the context of C- and D-vine copulas and by Dißmann et al. (2013) for R-vine copulas. For C-vine copulas maximum likelihood estimation of the parameters is easily achieved by maximizing the pseudo log-likelihood (2.4) based on the C-vine copula density (2.15). Note, that the pseudo log-likelihood of vine copulas can be written as a sum of logarithmized pair-copula densities. Two different sequential model selection approaches which deal with Setting (S3) (where also the vine tree structure and the pair-copulas have to be determined) are proposed in Kurowicka and Joe (2011) and by Dißmann et al. (2013).

The setting of particular interest in this thesis is Setting (S2). For the canonical vine copula based drought indices in Chapter 4 we pre-select a canonical vine tree structure based on a specific variable order. Appropriate pair-copulas and their parameters need to be inferred from the data. The pair-copulas in such a setting are usually selected sequentially, starting in the first tree ($\mathcal{T}_1$). In subsequent trees ($\mathcal{T}_2, \dots$), so called *pseudo observations* can be used to determine suitable pair-copulas corresponding to conditioned variable pairs. Pseudo observations corresponding to the pair-copula arguments are computed using Equation (2.14). To decide among a range of candidate pair-copulas, bivariate maximum-likelihood estimation (see Equation (2.4)) is performed for all pair-copula families under consideration (for each relevant pair of (pseudo) observations). The resulting pair-copula fits are then compared based on criteria like the Akaike information criterion (2.5) or the Bayesian information criterion (2.6). Suitable pair-copula parameters are given by the sequential parameter estimates corresponding to the selected families. Note that before a pair-copula is selected for a (conditioned) variable pair a bivariate independence test (see Genest and Favre, 2007) can be performed to see if an independence copula (2.3) should be selected instead.

2.3.4 The multivariate probability integral transform for vine copulas

The computation of univariate probability integral transforms (see Equation (2.2)) is a recurring task in copula based dependence modeling. Rosenblatt (1952) introduced a multivariate analog of such a transformation, the so called *Rosenblatt transform*. Here, we introduce the Rosenblatt transform for vine copulas. Let us consider a sample $u_1, \dots, u_d$ from a d-dimensional vine copula C. Then, the Rosenblatt transform $v_1, \dots, v_d$ of $u_1, \dots, u_d$ is defined as

$$\begin{aligned} v_1 &:= u_1, \\ v_2 &:= C_{2|1}(u_2|u_1), \\ &\vdots \\ v_d &:= C_{d|1,\dots,d-1}(u_d|u_1, \dots, u_{d-1}), \end{aligned} \tag{2.16}$$

where $C_{j|1,\dots,j-1}$ is the conditional cumulative distribution function for variable j given the variables $1, \dots, j-1$, for all $j = 2, \dots, d$. Then, the Rosenblatt transform $v_1, \dots, v_d$ is i.i.d. standard uniform distributed.

The computation of the Rosenblatt transform for C- and D-vine copulas is treated in Aas et al. (2009). For details on the computation of the Rosenblatt transform for R-vine copulas see Schepsmeier (2015). We elaborate on the Rosenblatt transform for C-vine copulas. For them, the order of the variables in the transformation (2.16) is determined by the selected order of root variables. For C-vine copulas the Rosenblatt transform (2.16) can be computed based on h-functions (2.13) using the following algorithm:

```
1. Set $v_1 = u_1$.
2. For $i \leftarrow 2, \dots, d$
    (a) Set $v_i = u_i$.
    (b) For $j \leftarrow 1, \dots, i-1$ set $v_i = h_{i|j;1,\dots,j-1}(v_i|v_j)$.
```

2.4 Validation metrics

To validate the drought indices developed in Chapter 4 we consider different methods and metrics. In a (binary) setting, where we model a phenomenon that differentiates if an *event of interest* occurred or not, the *probability of detection (POD)*, the *false alarm ratio (FAR)* and the *critical success index (CSI)* (see e.g. Wilks, 2011) are metrics of interest. For instance, these metrics were used in Hao and AghaKouchak (2014) for the validation of drought indices.

Validation metrics for phenomena with binary outcome Generally speaking, the metrics addressed above allow to assess the capability of a model to correctly *detect* a certain event *occurrence*. For this assessment, two time series $x_t^{\mathrm{o}}, x_t^{\mathrm{d}} \in \{0, 1\}$, $t = 1, \dots, N$, representing if the event of interest *occurred* (o)/was *detected* (d) or not are compared. For any time instance t, $x_t^{\mathrm{o}} = 1$ means event occurrence, $x_t^{\mathrm{d}} = 1$ means event detection, $x_t^{\mathrm{o}} = 0$ means that the event did not occur and $x_t^{\mathrm{d}} = 0$ means that the event was not detected. For any t, four different scenarios are possible:

1. The event occurred and was detected.
2. The event occurred but was not detected.
3. The event did not occur but was detected.
4. The event neither occurred nor was it detected.

We define number counts of the first three scenarios as follows:

$$\begin{aligned} H &= \#\{t \in \{1, \dots, N\} : x_t^{\mathrm{o}} = 1, x_t^{\mathrm{d}} = 1\} \\ M &= \#\{t \in \{1, \dots, N\} : x_t^{\mathrm{o}} = 1, x_t^{\mathrm{d}} = 0\} \\ F &= \#\{t \in \{1, \dots, N\} : x_t^{\mathrm{o}} = 0, x_t^{\mathrm{d}} = 1\} \end{aligned}$$

Hence, the probability of detection is defined as

$$\mathrm{POD} \coloneqq H/(H + M). \tag{2.17}$$

It provides the share of correctly identified event occurrence. That is, a high value indicates good performance. The false alarm ratio is defined as

$$\mathrm{FAR} \coloneqq F/(H + F). \tag{2.18}$$

It gives the portion of incorrectly detected events. Thus, a low value indicates good performance. Moreover, the critical success index is defined as

$$\mathrm{CSI} \coloneqq H/(H+M+F). \tag{2.19}$$

It is a measure which indicates good performance (high value) if event occurrence is mostly detected correctly and bad performance (low value) if event occurrence and detection do not match in most cases.

Application of POD, FAR and CSI for continuous variables We are further interested in the setting, where the phenomenon of interest is quantified in terms of continuous variables. Hence, we consider a time series $z_t^{\mathrm{o}} \in \mathbb{R}$, $t = 1, \dots, N$, of observations from the phenomenon (occurrence) and a time series $z_t^{\mathrm{d}} \in \mathbb{R}$, $t = 1, \dots, N$, representing the model of the phenomenon (detection). In order to differentiate, if the event of interest occurred/was detected or not a *threshold* $\vartheta \in \mathbb{R}$ which determines occurrence/detection has to be specified. By convention we decide, that non-exceedance of the threshold ϑ corresponds to occurrence/detection. Then we can translate the time series $z_t^{\mathrm{o}}, z_t^{\mathrm{d}} \in \mathbb{R}$, $t = 1, \dots, N$, to binary time series $x_t^{\mathrm{o}}, x_t^{\mathrm{d}} \in \{0, 1\}$, $t = 1, \dots, N$, containing information if the event of interest occurred (o)/was detected (d) or not. For all $t = 1, \dots, N$ we obtain

$$x_t^{\mathrm{o}} = \begin{cases} 1 & \text{if } z_t^{\mathrm{o}} \leq \vartheta \text{ (event occurrence)}, \\ 0 & \text{else}, \end{cases}$$

and

$$x_t^{\mathrm{d}} = \begin{cases} 1 & \text{if } z_t^{\mathrm{d}} \leq \vartheta \text{ (event detection)}, \\ 0 & \text{else}. \end{cases}$$

This threshold-based translation of the originally continuous time series to binary time series allows computations of the probability of detection (POD), the false alarm ratio (FAR) and the critical success index (CSI) according to Equations (2.17)–(2.19) from above.

2.5 Proper scoring rules and divergence functions

In Chapter 5 we investigate novel evaluation techniques for the assessment of model performance in a time series context. In this section, we provide the necessary background on proper scoring rules (see Gneiting and Raftery, 2007) and proper divergence functions (Thorarinsdottir et al., 2013). To outline the required theory behind these two concepts, we consider the following setup: Let $\mathcal{F}$ be a convex class of probability measures on a sample space Ω. Moreover, consider an (observed) *phenomenon* with the random outcome

- Y with (unknown) distribution $G \in \mathcal{F}$ and realization $y \in \Omega$.

Further, we consider a *model/forecast* for Y given through a random variable

- X with (modeled) distribution $F \in \mathcal{F}$ and realization $x \in \Omega$.

In practice we often face the setting that we have access to samples y_j, $j = 1, \dots, m$, $m \in \mathbb{N}$, and x_j, $j = 1, \dots, n$, $n \in \mathbb{N}$, from the distributions $G, F \in \mathcal{F}$, respectively, rather than knowing their actual parametric representation. Hence, we are interested in the corresponding empirical distribution functions defined as

$$\widehat{G}^m_{\boldsymbol{y}}(z) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{y_j \le z\} \qquad \text{and} \qquad \widehat{F}^n_{\boldsymbol{x}}(z) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{x_j \le z\}, \tag{2.20}$$

respectively, where $\boldsymbol{y} \coloneqq (y_1, \ldots, y_m)$, $\boldsymbol{x} \coloneqq (x_1, \ldots, x_n)$ and the indicator function $\mathbb{1}\{\mathcal{A}\}$ equals 1 if the event $\mathcal{A}$ is true and 0 otherwise.

2.5.1 Proper scoring rules

A *scoring rule* is a function

$$s : \mathcal{F} \times \Omega \to \mathbb{R}, \tag{2.21}$$

that assigns a score to a pair (F, y) of a distribution F and a realization $y \in \Omega$ of a phenomenon with (random) outcome Y. It can be interpreted as a distance measure between the realization y and a model/forecast distribution F for Y.

For $Y \sim G$ we denote the expectation of the score $s(F, Y)$ as

$$S(F, G) \coloneqq \mathbb{E}_G\left[s(F, Y)\right]. \tag{2.22}$$

Then, we call the scoring rule (2.21) *(negatively oriented) proper scoring rule* if

$$S(G, G) \le S(F, G) \quad \text{for all } F, G \in \mathcal{F}.$$

Hence, a scoring rule (2.21) is proper if we expect the mean of the random score $s(F, Y)$ to be minimized if our forecast/model F equals the true distribution G of Y.

Two examples for proper scoring rules For continuous variables (sample space $\Omega = \mathbb{R}$), popular examples of proper scoring rules are the *Squared Error (SE)* score

$$s_{\mathrm{SE}}(F, y) \coloneqq \left(\mathbb{E}_F\left[X\right] - y\right)^2, \quad \text{where } X \sim F, \tag{2.23}$$

and the *Continuous Ranked Probability Score (CRPS)*

$$s_{\mathrm{CRPS}}(F, y) \coloneqq \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\left\{z \ge y\right\}\right)^2 \mathrm{d}z, \tag{2.24}$$

where $\mathbb{1}\left\{z \ge y\right\}$ equals 1 if $z \ge y$ and 0 otherwise. In case of the normal distribution there exists an analytic solution of the integral in (2.24). If F is the normal distribution with mean μ and standard deviation σ, then it holds

$$s_{\mathrm{CRPS}}(\mathcal{N}(\mu, \sigma^2), y) = -\sigma\left\{\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right]\right\}, \tag{2.25}$$

where φ and Φ are the probability density function (PDF) and the cumulative distribution function (CDF) of a standard normal distribution (see Gneiting and Raftery, 2007). As in general there are no analytic solutions for the CRPS formula (2.24), Gneiting and Raftery (2007) suggest an alternative expression of the CRPS in terms of expectations. For $\widetilde{X} \sim F$ an independent copy of $X \sim F$, the CRPS can be obtained using the equation

$$s_{\mathrm{CRPS}}(F, y) = \mathbb{E}_F\left[|X - y|\right] - \frac{1}{2}\,\mathbb{E}_F\left[\left|X - \widetilde{X}\right|\right]. \tag{2.26}$$

Whereas the CRPS is of interest if the model/forecast is given in terms of a CDF F, there exist also (proper) scoring rules which are calculated based on the corresponding PDF f (see e.g. Gneiting and Raftery, 2007, Section 4.1). Besides scoring rules for continuous variables, there are also scoring rules for categorial variables. For more details on different scoring rules we refer the reader to Gneiting and Raftery (2007).

Scores for deterministic models/forecasts Before we discuss the above introduced scoring rules from a practical point of view and illustrate them with an example, we are interested in the special case when the model/forecast is deterministic. In that case the model/forecast is not subject to uncertainty, it corresponds to a constant random variable $X = x$, which follows a degenerate distribution with CDF $F_x(z) := \mathbb{1}\{x \leq z\}$. The SE score (2.23) simplifies to the *squared error*

$$s_{\mathrm{SE}}(x, y) := (x - y)^2 = s_{\mathrm{SE}}(F_x, y). \tag{2.27}$$

From Equation (2.26) it follows, that the degenerated CRPS equals the *absolute error*

$$s_{\mathrm{AE}}(x, y) := |x - y| = s_{\mathrm{CRPS}}(F_x, y). \tag{2.28}$$

Practical implications Let us now discuss the practical implications of the Squared Error score (2.23) and the CRPS (2.24). Evaluation based on the SE score assesses the deviation of a realization y of Y from the mean $\mathbb{E}_F[X]$ of the forecast/model distiribution F. The SE score does not consider higher order moments of the distribution F for the evaluation. The CRPS however, considers not only the distance of the realization y from the "center" of the distribution F. On the contrary, it measures the distance of y to the whole distribution F.

Illustration To illustrate the computation of the different scores and differences between them, we consider the following example: Let the phenomenon of interest (with outcome Y) have the realization $y = 0$. Further, let us consider the distributions

$$\begin{aligned} F_1 &= \mathcal{N}(0, 4/9), \\ F_2 &= \mathcal{N}(0, 1/9), \\ F_3 &= \mathcal{N}(1/2, 1/9), \end{aligned} \tag{2.29}$$

as candidate models/forecasts for the outcome of the phenomenon. Figure 2.7 illustrates the example. The upper left panel shows the densities f_1, f_2 and f_3 corresponding to F_1, F_2 and F_3, respectively. Since the densities are symmetric, their means coincide with their modes. The means are visualized in the middle left panel and compared to the realization $y = 0$. We see that the means for F_1 and F_2 equal the realization $y = 0$. The deviation of the mean for F_3 is indicated by a dash-dotted horizontal line. This illustrates the computation of the SE scores, which are equal to the squared difference between the realization y and the mean of the model/forecast distribution. Hence, the SE score is not able to distinguish between F_1 and F_2, which have the same mean ($s_{\mathrm{SE}}(F_1, y) = s_{\mathrm{SE}}(F_2, y) = 0$). However, these two distributions differ in terms of uncertainty. The CRPS takes this into consideration (see upper right and middle right panel) and is able to distinguish between F_1 and F_2 ($s_{\mathrm{CRPS}}(F_1, y) = 0.16$, $s_{\mathrm{CRPS}}(F_2, y) = 0.08$). F_2 has the better (lower) score, as it is more certain (sharp) about the location of $y = 0$. The differences (in terms of CRPS) between the three models can be identified best by looking at the curves $(F_i(z) - \mathbb{1}\{z \geq y\})^2$, $i = 1, 2, 3$ (bottom left panel). The CRPS corresponds to the area under these curves. As F_3 is not centered around $y = 0$, it has a higher SE score ($s_{\mathrm{SE}}(F_3, y) = 0.25$). Also the CRPS of F_3 is higher compared to F_1 and F_2 ($s_{\mathrm{CRPS}}(F_3, y) = 0.33$).

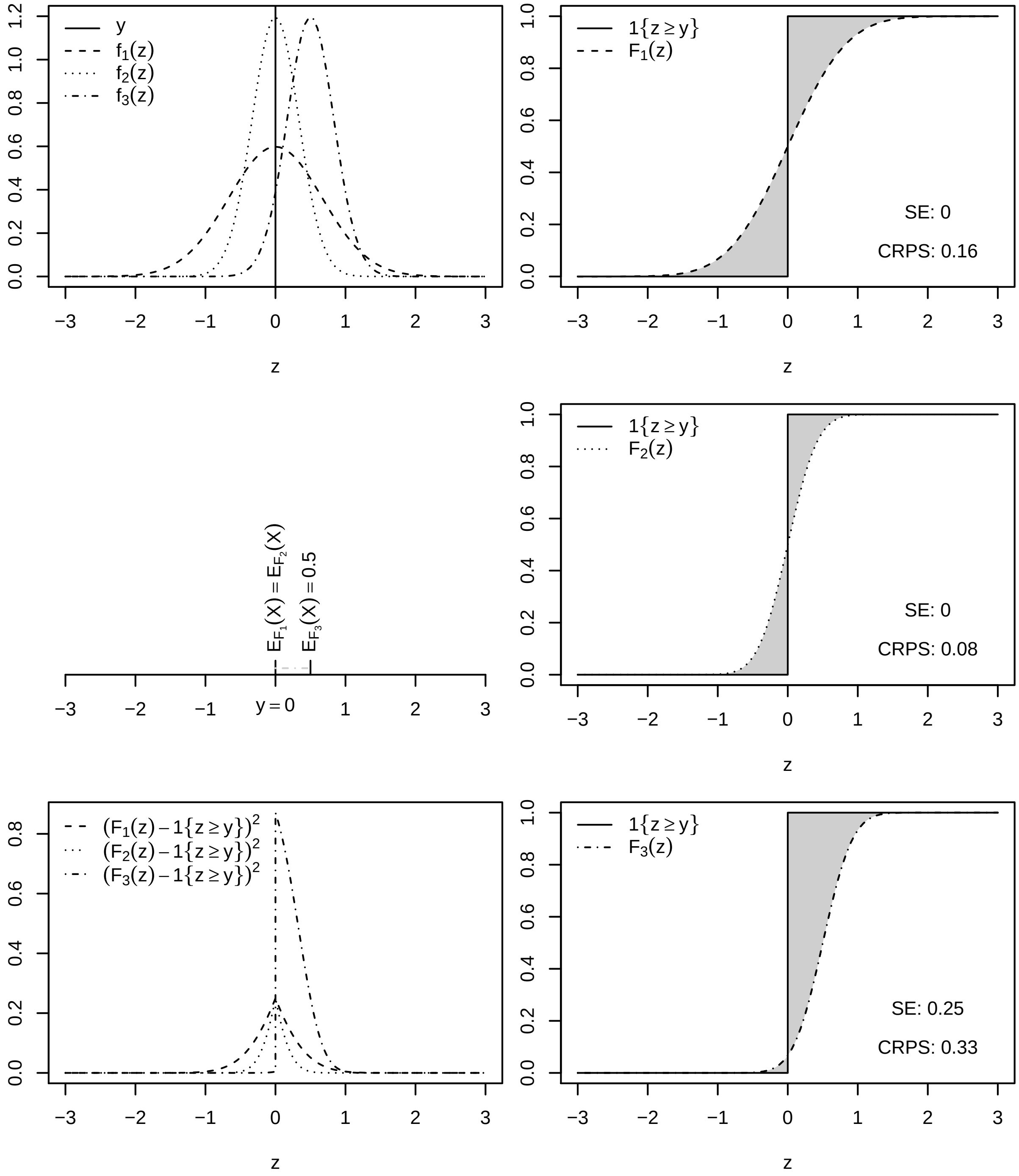


Figure 2.7: Illustration of score computation (realization $y = 0$ and models $F_1 = \mathcal{N}(0, 4/9)$, $F_2 = \mathcal{N}(0, 1/9)$ and $F_3 = \mathcal{N}(1/2, 1/9)$): Comparison of realization y with densities f_1, f_2, f_3 corresponding to the models (CDFs) F_1, F_2 and F_3 (upper left). Comparison of model means and realization (middle left). Computation of CRPS (bottom left). Comparison of $\mathbb{1}\{z \geq y\}$ with CDFs F_1, F_2 and F_3, respectively, and corresponding SE scores and CRPS (right panel).

Scoring based on samples from the model/forecast distribution In practice we do not always have a (parametric) model/forecast distribution F, but rather a sample $\boldsymbol{x} := (x_1, \dots, x_n)$ of size $n \in \mathbb{N}$ from F. In that setting, we consider *sample versions of scoring functions*. Using the empirical distribution $\widehat{F}_{\boldsymbol{x}}^n$ corresponding to $\boldsymbol{x}$ (see Equation (2.20)) they are defined as

$$s(\boldsymbol{x}, y) := s(\widehat{F}_{\boldsymbol{x}}^n, y). \tag{2.30}$$

Sample versions of the SE score and the CRPS can easily be derived from Equations (2.23) and (2.26), respectively, by replacing the occurring expectations with their corresponding sample version. The *sample SE score* is given by

$$s_{\mathrm{SE}}(\boldsymbol{x}, y) = \left(\frac{1}{n} \sum_{j=1}^{n} x_j - y \right)^2, \tag{2.31}$$

and the *sample CRPS* by

$$s_{\mathrm{CRPS}}(\boldsymbol{x}, y) = \frac{1}{n} \sum_{j=1}^{n} |x_j - y| - \frac{1}{2n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} |x_j - x_k|. \tag{2.32}$$

If the sample consists only of one element ($n = 1$), the sample SE score degenerates to the squared error (2.27) and the sample CRPS to the absolute error (2.28).

2.5.2 Proper divergence functions

Divergence functions are functions

$$d : \mathcal{F} \times \mathcal{F} \to [0, \infty] \tag{2.33}$$

which fulfill the property $d(F, F) = 0$ for all $F \in \mathcal{F}$. Considering a distribution $G \in \mathcal{F}$ that describes the random outcome of a phenomenon on the one hand and a distribution $F \in \mathcal{F}$ that is supposed to model/forecast the very same outcome of the phenomenon on the other hand, $d(F, G)$ can be used to judge how well the outcome of the phenomenon is modeled/forecast by F. Hence, divergences are distance measures between distributions $F, G \in \mathcal{F}$.

To introduce the concept of propriety for divergence functions (see Thorarinsdottir et al., 2013), we consider the following quantities. For $k \in \mathbb{N}$ let $Y_1, \dots, Y_k \sim G \in \mathcal{F}$ (independent) and $\boldsymbol{Y} := (Y_1, \dots, Y_k)$. The corresponding (random) empirical distribution function is given by $\widehat{G}_{\boldsymbol{Y}}^k(z) := 1/k \sum_{j=1}^{k} \mathbb{1}\{Y_j \leq z\}$. Further, we denote the expectation of the divergence $d(F, \widehat{G}_{\boldsymbol{Y}}^k)$ as

$$D^k(F, G) := \mathbb{E}_G \left[d(F, \widehat{G}_{\boldsymbol{Y}}^k) \right].$$

Then a divergence function d in (2.33) is called *k-proper*, if for all $F \in \mathcal{F}$ it holds that

$$D^k(G, G) \leq D^k(F, G).$$

Hence, having observed a sample of size $k \in \mathbb{N}$ from $G \in \mathcal{F}$ (the phenomenon of interest), a k-proper divergence $d(F, \widehat{G}_{\boldsymbol{Y}}^k)$ is supposed to be minimized in expectation if the forecast/model F equals the true distribution G.

Construction of divergence functions from proper scoring rules Divergence functions can for example be constructed using score expectations (2.22) of proper scoring rules s, if it holds that $|S(G,G)| < \infty$ for all $G \in \mathcal{F}$. *Score divergences* are then defined as the difference

$$d(F,G) \coloneqq S(F,G) - S(G,G). \tag{2.34}$$

Thorarinsdottir et al. (2013) show that score divergences constructed as in (2.34) are k-proper for all $k \in \mathbb{N}$.

Two examples for score divergences Now, we consider two divergence functions for continuous variables (sample space $\Omega = \mathbb{R}$), the *Mean Value (MV) divergence*

$$d_{\mathrm{MV}}(F,G) \coloneqq \left(\mathbb{E}_F\left[X\right] - \mathbb{E}_G\left[Y\right]\right)^2, \quad \text{where } X \sim F \text{ and } Y \sim G, \tag{2.35}$$

and the *Integrated Quadratic (IQ) distance*

$$d_{\mathrm{IQ}}(F,G) \coloneqq \int_{-\infty}^{\infty} \left(F(z) - G(z)\right)^2 \mathrm{d}z. \tag{2.36}$$

Whereas the Mean Value divergence (2.35) is the score divergence corresponding to the Squared Error score (2.23), the Integrated Quadratic distance (2.36) is the score divergence derived from the Continuous Ranked Probability Score (2.24). Hence, both divergence functions (2.35) and (2.36) are k-proper for all $k \in \mathbb{N}$. In analogy to Equation (2.26) the Integrated Quadratic distance (2.36) can be rewritten in terms of expectations (see Thorarinsdottir et al., 2013) as

$$d_{\mathrm{IQ}}(F,G) = \mathbb{E}_{F,G}\left[|X - Y|\right] - \frac{1}{2}\left(\mathbb{E}_F\left[\left|X - \widetilde{X}\right|\right] + \mathbb{E}_G\left[\left|Y - \widetilde{Y}\right|\right]\right), \tag{2.37}$$

where $\widetilde{X} \sim F$ and $\widetilde{Y} \sim G$ are independent copies of $X \sim F$ and $Y \sim G$, respectively. Again, there are also divergence functions for categorial variables. For more details see Thorarinsdottir et al. (2013).

Divergences for deterministic models/forecasts Before we further discuss and illustrate the Mean Value divergence (2.35) and the Integrated Quadratic distance (2.36), we are again interested in their degenerated versions. To obtain them we assume this time, that both distributions (outcome of the phenomenon and model/forecast) are deterministic. That is, we consider two constant random variables $Y = y$ and $X = x$ with degenerate CDFs $G_y(z) \coloneqq \mathbb{1}\left\{y \leq z\right\}$ and $F_x(z) \coloneqq \mathbb{1}\left\{x \leq z\right\}$, respectively. Then, the Mean Value divergence (2.35) reduces to the *squared error*

$$d_{\mathrm{MV}}(F_x, G_y) = (x - y)^2 = s_{\mathrm{SE}}(x,y). \tag{2.38}$$

From Equation (2.37) it follows, that the degenerated IQ distance equals the *absolute error*

$$d_{\mathrm{IQ}}(F_x, G_y) = |x - y| = s_{\mathrm{AE}}(x,y). \tag{2.39}$$

Practical implications From a practical perspective, the Mean Value divergence (2.35) considers only the distance between the means of two distributions F and G and completely ignores higher order moments of F and G. The Integrated Quadratic distance (2.36) however measures the distance between $F(z)$ and $G(z)$ for all $z \in \mathbb{R}$ and hence also considers higher order structures of the distributions F and G.

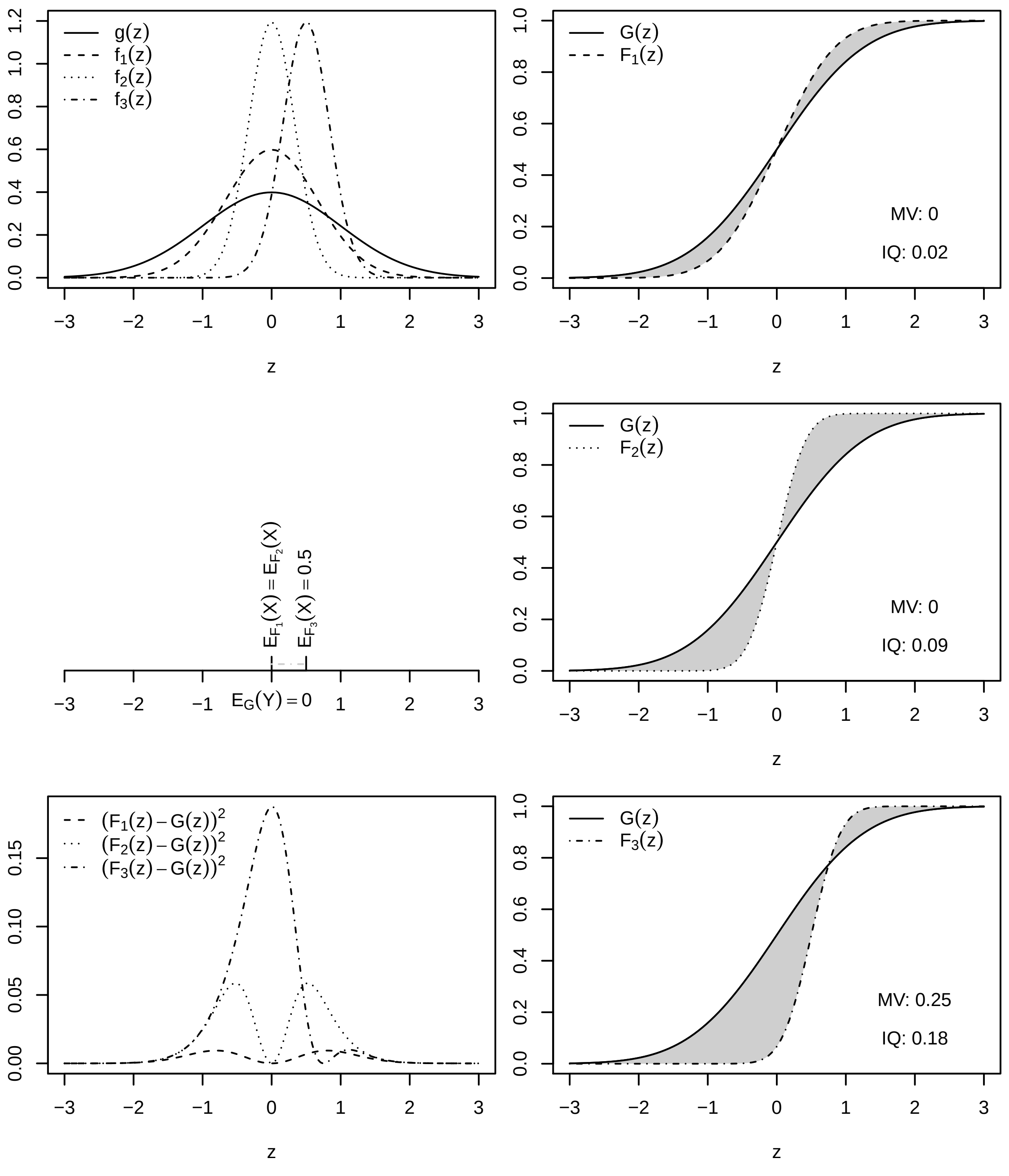


Figure 2.8: Illustration of divergence function computation (outcome $Y \sim G$, $G = \mathcal{N}(0,1)$, and models/forecasts $F_1 = \mathcal{N}(0,4/9)$, $F_2 = \mathcal{N}(0,1/9)$ and $F_3 = \mathcal{N}(1/2,1/9)$): Comparison of densities g, f_1, f_2, f_3 corresponding to G, F_1, F_2 and F_3 (upper left). Comparison of model means and the mean of the outcome (middle left). Computation of IQ distances (bottom left). Comparison of outcome CDF G with model/forecast CDFs F_1, F_2 and F_3, respectively, and corresponding MV divergences and IQ distances (right panel).

Illustration To illustrate the computation of and the differences between the two divergence functions (2.35) and (2.36), we consider the following example: Let the outcome of the phenomenon of interest be $Y \sim G$ with $G = \mathcal{N}(0,1)$. Further, again consider the model/forecast distributions F_1, F_2 and F_3 (2.29) from the example in Section 2.5.1. Figure 2.8 illustrates the example. The upper left panel compares the densities f_1, f_2 and f_3 corresponding to F_1, F_2 and F_3, respectively, to the density g corresponding to the distribution G of the outcome. Since the densities are symmetric, their means coincide with their modes. The means corresponding to F_1, F_2 and F_3 are visualized in the middle left panel and compared to the mean of the outcome. We see that the means for F_1 and F_2 equal that of the outcome $Y \sim G$. The deviation of the mean for F_3 is indicated by a dash-dotted horizontal line. This illustrates the computation of the MV divergences, which are equal to the squared difference between the means of the outcome and the model/forecast distribution. Hence, the MV divergence does not allow to differentiate between F_1 and F_2, which have the same mean ($d_{\mathrm{MV}}(F_1, G) = d_{\mathrm{MV}}(F_2, G) = 0$). However, F_1 and F_2 differ in terms of uncertainty. The IQ distance takes this into consideration (compare upper right and middle right panel) and is able to distinguish between F_1 and F_2 ($d_{\mathrm{IQ}}(F_1, G) = 0.02$, $d_{\mathrm{IQ}}(F_2, G) = 0.09$). F_1 has the better (lower) divergence, as its variance is closer to the variance of G. The differences (in terms of IQ distances) between the three models can be identified best by looking at the curves $(F_i(z) - G(z))^2$, $i = 1, 2, 3$ (bottom left panel). The IQ distance corresponds to the area under these curves. While the curves corresponding to F_1 and F_2 are symmetric around 0, which is the center of the distributions G, F_1 and F_2, the curve corresponding to F_3 is asymmetric. Both divergences penalize F_3 for not being centered around 0 ($\mathbb{E}_{F_3}[X] \neq \mathbb{E}_G[Y]$), the IQ distance accounts also for the wrong specified variance ($\mathrm{Var}_{F_3}[X] \neq \mathrm{Var}_G[Y]$). As a result, the divergences for F_3 are higher (worse) compared to F_1 and F_2 ($d_{\mathrm{MV}}(F_3, G) = 0.25$, $d_{\mathrm{IQ}}(F_3, G) = 0.18$).

Computation of divergences based on samples In practice we usually/often do not have explicit outcome distributions G and model/forecast distributions F, but rather samples $\boldsymbol{y} := (y_1, \ldots, y_m)$, $m \in \mathbb{N}$, from G and $\boldsymbol{x} := (x_1, \ldots, x_n)$, $n \in \mathbb{N}$, from F, respectively. In that setting, we consider *sample versions of divergence functions.* Using the empirical distributions $\widehat{G}^m_{\boldsymbol{y}}$ and $\widehat{F}^n_{\boldsymbol{x}}$ (see Equation (2.20)) they are defined as

$$d(\boldsymbol{x}, \boldsymbol{y}) := d(\widehat{F}^n_{\boldsymbol{x}}, \widehat{G}^m_{\boldsymbol{y}}). \tag{2.40}$$

Sample versions of the MV divergence and the IQ distance can easily be derived from Equations (2.35) and (2.37), respectively, by replacing the occurring expectations with their corresponding sample version. The *sample MV divergence* is given by

$$d_{\mathrm{MV}}(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{1}{n} \sum_{j=1}^{n} x_j - \frac{1}{m} \sum_{j=1}^{m} y_j \right)^2, \tag{2.41}$$

and the *sample IQ distance* by

$$d_{\mathrm{IQ}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{mn} \sum_{k=1}^{m} \sum_{j=1}^{n} |x_j - y_k| - \frac{1}{2} \left(\frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} |x_j - x_k| + \frac{1}{m^2} \sum_{j=1}^{m} \sum_{k=1}^{m} |y_j - y_k| \right). \tag{2.42}$$

If the samples consists only of one element ($m = n = 1$), the sample MV divergence degenerates to the squared error (2.27) and the sample IQ distance to the absolute error (2.28). If $m = 1$ ($n \in \mathbb{N}$), it holds $d_{\mathrm{MV}}(\boldsymbol{x}, y) = s_{\mathrm{SE}}(\boldsymbol{x}, y)$ and $d_{\mathrm{IQ}}(\boldsymbol{x}, y) = s_{\mathrm{CRPS}}(\boldsymbol{x}, y)$.

2.6 Traditional evaluation measures

Hyndman and Koehler (2006) provide a comprehensive review of different *traditional evaluation measures* which can be used for the evaluation of forecasts/models. To list the corresponding definitions of these measures, we consider time series of observations y_t, $t = 1, \ldots, N$ from the phenomenon of interest and corresponding model output/forecasts x_t, $t = 1, \ldots, N$. Then we can define a *time series of model/forecast errors* as

$$e_t \coloneqq y_t - x_t, \quad t = 1, \ldots, N.$$

Scale-dependent measures The most popular scale dependent measures are the sample *Mean Square Error*

$$\widehat{\mathrm{MSE}} \coloneqq \frac{1}{N} \sum_{t=1}^{N} e_t^2, \tag{2.43}$$

the sample *Root Mean Square Error*

$$\widehat{\mathrm{RMSE}} \coloneqq \sqrt{\frac{1}{N} \sum_{t=1}^{N} e_t^2},$$

the sample *Mean Absolute Error*

$$\widehat{\mathrm{MAE}} \coloneqq \frac{1}{N} \sum_{t=1}^{N} |e_t|, \tag{2.44}$$

and the sample *Median Absolute Error*

$$\widehat{\mathrm{MdAE}} \coloneqq \underset{t=1,\ldots,N}{\mathrm{median}} \left(|e_t| \right),$$

where $\mathrm{median}_{t=1,\ldots,N} \left(\cdot \right)$ denotes the sample median. As their scale depends on the data, the measures above are not suitable to compare methods based on different types of data sets.

Measures based on percentage errors If we scale the errors e_t by $100/y_t$, we obtain scale independent *percentage errors*

$$p_t = \frac{100 e_t}{y_t}, \quad t = 1, \ldots, N.$$

Commonly used measures based on percentage errors are the sample *Mean Absolute Percentage Error*

$$\widehat{\mathrm{MAPE}} \coloneqq \frac{1}{N} \sum_{t=1}^{N} |p_t|,$$

the sample *Median Absolute Percentage Error*

$$\widehat{\mathrm{MdAPE}} \coloneqq \underset{t=1,\ldots,N}{\mathrm{median}} \left(|p_t| \right),$$

the sample *Root Mean Square Percentage Error*

$$\widehat{\mathrm{RMSPE}} \coloneqq \sqrt{\frac{1}{N} \sum_{t=1}^{N} p_t^2},$$

and the sample *Root Median Square Percentage Error*

$$\widehat{\mathrm{RMdSPE}} \coloneqq \sqrt{\underset{t=1,\ldots,N}{\mathrm{median}} \left(p_t^2 \right)}.$$

Measures based on relative errors One other way of scaling the errors e_t, $t = 1, \dots, N$, is to scale by errors e_t^*, $t = 1, \dots, N$, corresponding to a benchmark model/method. The subsequent measure definitions are based on the relative errors

$$r_t = \frac{e_t}{e_t^*}, \quad t = 1, \dots, N:$$

Some measures based on relative errors are the sample *Mean Relative Absolute Error*

$$\widehat{\mathrm{MRAE}} := \frac{1}{N} \sum_{t=1}^{N} |r_t|,$$

the sample *Median Relative Absolute Error*

$$\widehat{\mathrm{MdRAE}} := \underset{t=1,\dots,N}{\mathrm{median}} \left(|r_t|\right),$$

and the sample *Geometric Mean Relative Absolute Error*

$$\widehat{\mathrm{GMRAE}} := \left(\prod_{t=1}^{N} |r_t| \right)^{\frac{1}{N}}.$$

Relative measures Besides measures based on relative errors it is also possible to consider relative measures. Let $\widehat{\mathrm{EM}}$ denote one of the evaluation measures introduced above and $\widehat{\mathrm{EM}}^*$ the same measure applied for a benchmark model. Then, the corresponding *relative measure* (sample version) is defined as

$$\widehat{\mathrm{RelEM}} := \frac{\widehat{\mathrm{EM}}}{\widehat{\mathrm{EM}}^*}.$$

The sample Relative Mean Absolute Error $\widehat{\mathrm{RelMAE}} := \widehat{\mathrm{MAE}}/\widehat{\mathrm{MAE}}^*$ is an example of a relative measure.

Measures based on scaled errors Another method to obtain evaluation measures is to use scaled errors of the form

$$q_t := \frac{e_t}{\frac{1}{N-1} \sum_{j=2}^{N} |y_j - y_{j-1}|}, \quad t = 1, \dots, N.$$

Hence, measure like the sample *Mean Absolute Scaled Error*

$$\widehat{\mathrm{MASE}} := \frac{1}{N} \sum_{t=1}^{N} |q_t|,$$

can be defined in analogy to the measures defined above.

2.7 Detection of multiple changepoints

The detection of multiple changepoints and hence, the segmentation of time series into stationary segments is an active field of research. Many different algorithms have been suggested (see e.g. the changepoint repository (`http://changepoint.info`) provided by Killick et al., 2012). In Chapter 5 we are interested in segmenting non-stationary time series into segments where stationarity can be assumed (approximately), in order to be able to access the ability of models/forecasts to reproduce the distributional characteristics of the (observed) outcome of a phenomenon. In this section, we provide the necessary background on multiple changepoint detection. After giving a general introduction to the problem of detecting multiple changepoints by minimization of a target function, we provide more detail for the so called PELT (Pruned Exact Linear Time) method (Killick et al., 2012). We focus on this method, as it is exact (in terms of minimization of the cost function). Moreover, under certain conditions, the computational cost of the algorithm depends linear on the number of observations of the time series. The most important of these conditions is the assumption that the number of changepoints increases linearly with increasing length of the time series, which is met by many applications. The subsequent introduction is based on Killick et al. (2012).

What is a changepoint analysis? A *changepoint analysis* concerns the detection of one or multiple so called changepoints in a time series $y_1, \dots, y_N$. Generally speaking, a time instance $\tau \in \{1, \dots, N-1\}$ is considered a *changepoint*, if the statistical properties of the sub-series $y_1, \dots, y_\tau$ and $y_{\tau+1}, \dots, y_N$ differ.

Notation To provide a more general explanation of how multiple changepoints can be detected, we have to introduce some more notation. We can interpret a time series $y_1, \dots, y_N$ as an *ordered sequence* of data points

$$\boldsymbol{y}_{1:N} := (y_1, \dots, y_N).$$

For $1 \leq s \leq t \leq N$, we denote *sub-sequences* of $\boldsymbol{y}_{1:N}$ by $\boldsymbol{y}_{s:t} := (y_s, \dots, y_t)$. Further, we denote by

$$\boldsymbol{\tau}_{1:m} := (\tau_1, \dots, \tau_m)$$

the ordered sequence of $m \in \{0, \dots, N-1\}$ *changepoints* of $\boldsymbol{y}_{1:N}$, where $\tau_j \in \mathbb{N}$, $1 \leq \tau_j \leq N-1$, $j = 1, \dots, m$. As $\boldsymbol{\tau}_{1:m}$ is ordered, it holds that $\tau_i < \tau_j$ for $1 \leq i < j \leq m$. Defining $\tau_0 := 0$ and $\tau_{m+1} := N$, the changepoints $\boldsymbol{\tau}_{1:m}$ split the sequence $\boldsymbol{y}_{1:N}$ into the $m+1$ *segments*

$$\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}, \quad j = 0, \dots, m,$$

with *segment lengths* $(\tau_{j+1} - \tau_j)$, $j = 0, \dots, m$. We call

$$\boldsymbol{\tau}_{0:(m+1)} := (\tau_0, \boldsymbol{\tau}_{1:m}, \tau_{m+1})$$

an $(m+1)$-*segmentation* of $\boldsymbol{y}_{1:N}$. Furthermore, we denote the *set of all possible* $(m+1)$-*segmentations* of an ordered sequence $\boldsymbol{y}_{1:s}$ by

$$\mathcal{T}_s^m := \{\boldsymbol{\tau} : \boldsymbol{\tau} \text{ is a valid } (m+1)\text{-segmentation of } \boldsymbol{y}_{1:s}\},$$

for all $s = 1, \dots, N$.

How can we identify changepoints? One option to detect multiple changepoints is based on the minimization of a *target function*

$$\sum_{j=0}^{m} \mathcal{C}(\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}) + \kappa g(m). \tag{2.45}$$

The target function is a sum of *cost functions* $\mathcal{C} : \mathbb{R}^n \to \mathbb{R}$, $n \in \mathbb{N}$, which for all $j = 0, \dots, m$ assign a cost to the segment $\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}$, and a *penalty term* $\kappa g(m)$, composed out of a constant κ and a function g in m.

A changepoint detection algorithm which is minimizing (2.45), that considers all possible $(m+1)$-segmentations $\mathcal{T}_N^m$, for all possible numbers of changepoints $m = 0, \dots, N-1$ is called *exact*. The *Binary Segmentation Algorithm* (see Scott and Knott, 1974) is an example for a popular algorithm, which is not exact. A common choice for the cost function $\mathcal{C}$ is (twice) the negative log-likelihood. The penalty term is supposed to prevent overfitting by penalizing the number of changepoints m. Most often, the common choice of $g(m) = m$ is used in combination with a κ that corresponds to one of the popular information criteria like the Akaike information criterion (AIC; $\kappa = 2p$) or the Bayesian information criterion (BIC; $\kappa = p\ln(N)$). These penalties depend on p, the number of additional parameters needed per additional segment. Note, that the additional changepoint, which is required per segment, is also counted as an additional parameter.

How can the changepoint problem (2.45) be solved in an algorithmic fashion? The *Optimal Partitioning (OP)* method (see Jackson et al. (2005) and also Killick et al. (2012)) is an exact changepoint detection algorithm. It requires, that for the penalty term of (2.45) it holds that $g(m) = m$. In that case the minimum of (2.45) for an ordered sequence $\boldsymbol{y}_{1:s}$, $s \le N$, is given by

$$M(s) := \min_{m<s} \min_{\boldsymbol{\tau} \in \mathcal{T}_s^m} \left\{ \sum_{j=0}^{m} \mathcal{C}(\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}) + \kappa m \right\}. \tag{2.46}$$

The OP method is based on the fact, that $M(s)$ can be calculated recursively. Setting $M(0) := -\kappa$, the recursion is obtained easily by rewriting Equation (2.46) as

$$\begin{aligned} M(s) &= \min \left\{ \mathcal{C}(\boldsymbol{y}_{1:s}), \min_{0<t<s} \left[\min_{m<t} \min_{\boldsymbol{\tau} \in \mathcal{T}_t^m} \left(\sum_{j=0}^{m} \mathcal{C}(\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}) + \kappa m \right) + \mathcal{C}(\boldsymbol{y}_{(t+1):s}) + \kappa \right] \right\} \\ &= \min_{t<s} \left\{ M(t) + \mathcal{C}(\boldsymbol{y}_{(t+1):s}) + \kappa \right\}. \end{aligned} \tag{2.47}$$

Hence, starting from $s = 1$, the OP algorithm successively detects the changepoints of the sub-sequences $\boldsymbol{y}_{1:s}$. In each step s the algorithm computes $M(s)$. Further it detects the last changepoint before s as

$$\tau_s := \operatorname*{argmin}_{t<s} \left\{ M(t) + \mathcal{C}(\boldsymbol{y}_{(t+1):s}) + \kappa \right\}.$$

Starting with an empty set $\boldsymbol{\tau}(0) := \emptyset$, the algorithm recursively stores the changepoints for $\boldsymbol{y}_{1:s}$ in the set $\boldsymbol{\tau}(s)$, by adding the detected changepoint τ_s to the changepoints $\boldsymbol{\tau}(\tau_s)$ which were detected for the sub-sequence $\boldsymbol{y}_{1:\tau_s}$. Hence, the algorithm sets $\boldsymbol{\tau}(s) := (\boldsymbol{\tau}(\tau_s), \tau_s)$. Finally, after the last iteration ($s = N$), $\boldsymbol{\tau}(N)$ contains the changepoints detected for $\boldsymbol{y}_{1:N}$.

Can this algorithm be improved? Killick et al. (2012) found that the performance of the OP algorithm can be improved by reducing the set of time instances t considered in the minimization (2.47) in order to avoid irrelevant computations. This technique called *pruning* is based on the following result: Considering time instances $t < s < T$ and a constant K such that

$$\mathcal{C}(\boldsymbol{y}_{(t+1):s}) + \mathcal{C}(\boldsymbol{y}_{(s+1):T}) + K \le \mathcal{C}(\boldsymbol{y}_{(t+1):T}), \tag{2.48}$$

the time instance t can not be the latest changepoint before T, if

$$M(t) + \mathcal{C}(\boldsymbol{y}_{(t+1):s}) + K \geq M(s). \tag{2.49}$$

While Equation (2.48) ensures, that the introduction of an additional changepoint does not increase the total cost, the condition given by Equation (2.49) identifies time instances whose selection as a changepoint would increase the value of the target function (2.45) of our minimization. Note, that if we select the (scaled) negative log-likelihood for the cost function $\mathcal{C}$, we can select $K = 0$. This is because splitting $\boldsymbol{y}_{(t+1):T}$ in Equation (2.48) into two segments allows different parameters for both segments, which can only lead to a smaller (overall) negative log-likelihood.

How is the pruning technique implemented in the changepoint detection algorithm? For a selected cost function $\mathcal{C}$, a constant K which fulfills (2.48) and a penalty constant κ, we provide the PELT (Pruned Exact Linear Time) algorithm (see Killick et al., 2012) for the detection of changepoints of an ordered sequence $\boldsymbol{y}_{1:N}$:

- Set $M(0) := -\kappa$, $\boldsymbol{\tau}(0) := \emptyset$ and $\mathcal{S}(1) := \{0\}$.
- For $s = 1, \dots, N$ iterate
  1. $\tau_s = \operatorname{argmin}_{t \in \mathcal{S}(s)} \left\{ M(t) + \mathcal{C}(\boldsymbol{y}_{(t+1):s}) + \kappa \right\}$,
  2. $M(s) = M(\tau_s) + \mathcal{C}(\boldsymbol{y}_{(\tau_s+1):s}) + \kappa$,
  3. $\boldsymbol{\tau}(s) = (\boldsymbol{\tau}(\tau_s), \tau_s)$,
  4. $\mathcal{S}(s+1) = \left\{ s \cup \{ t \in \mathcal{S}(s) : M(t) + \mathcal{C}(\boldsymbol{y}_{(t+1):s}) + K < M(s) \} \right\}$.

Step 1 of the algorithm detects the last changepoint τ_s before s, step 2 calculates the minimum of the target function $M(s)$ for the sub-sequence $\boldsymbol{y}_{1:s}$, step 3 gathers all detected changepoints for $\boldsymbol{y}_{1:s}$ and step 4 determines the changepoint candidates for the next iteration. Again, $\boldsymbol{\tau}(N)$ contains the changepoints detected for $\boldsymbol{y}_{1:N}$.

Which cost function $\mathcal{C}$ can we use if we want to detect changes in mean and variance? Let us assume that the time series observations $y_1, \dots, y_N$ come from a normal distribution $\mathcal{N}(\mu_j, \sigma_j^2)$, where either the (unknown) mean μ_j and/or the (unknown) variance σ_j^2 change after certain (unknown) time instances (changepoints) $\boldsymbol{\tau}_{1:m}$. Then we can select for the cost function $\mathcal{C}$ twice the negative log-likelihood corresponding to a normal distribution.

Subsequently, we derive the cost $\mathcal{C}(\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}})$ of a segment $\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}$ of an arbitrary $(m+1)$-segmentation $\boldsymbol{\tau}_{0:(m+1)}$ of the ordered sequence $\boldsymbol{y}_{1:N}$. The cost

$$\begin{aligned} \mathcal{C}(\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}) &:= -2 \sum_{i=\tau_j+1}^{\tau_{j+1}} \left[-\frac{1}{2} \ln\left(2\pi\widehat{\sigma}_j^2\right) - \frac{(y_i - \widehat{\mu}_j)^2}{2\widehat{\sigma}_j^2} \right] \\ &= (\tau_{j+1} - \tau_j) \ln\left(2\pi\widehat{\sigma}_j^2\right) + \frac{1}{\widehat{\sigma}_j^2} \sum_{i=\tau_j+1}^{\tau_{j+1}} (y_i - \widehat{\mu}_j)^2 \\ &= (\tau_{j+1} - \tau_j) \left[\ln\left(2\pi\widehat{\sigma}_j^2\right) + \frac{\left(\widehat{\sigma}_j^2\right)^{-1}}{(\tau_{j+1} - \tau_j)} \sum_{i=\tau_j+1}^{\tau_{j+1}} (y_i - \widehat{\mu}_j)^2 \right] \end{aligned} \tag{2.50}$$

corresponds to twice the negative log-likelihood for a normal distribution $\mathcal{N}(\mu_j, \sigma_j^2)$, where the unknown parameters μ_j and σ_j^2 are replaced by their maximum likelihood estimators

$$\widehat{\mu}_j = \frac{1}{\tau_{j+1} - \tau_j} \sum_{i=\tau_j+1}^{\tau_{j+1}} y_i \tag{2.51}$$

and

$$\widehat{\sigma}_j^2 = \frac{1}{\tau_{j+1} - \tau_j} \sum_{i=\tau_j+1}^{\tau_{j+1}} \left(y_i - \widehat{\mu}_j\right)^2 , \tag{2.52}$$

respectively. Plugging in expressions (2.51) and (2.52) for $\widehat{\mu}_j$ and $\widehat{\sigma}_j^2$ in the cost function (2.50) yields

$$\mathcal{C}(\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}) \coloneqq (\tau_{j+1} - \tau_j) \left\{ \ln \left[\frac{2\pi}{\tau_{j+1} - \tau_j} \sum_{i=\tau_j+1}^{\tau_{j+1}} \left(y_i - \frac{\sum_{k=\tau_j+1}^{\tau_{j+1}} y_k}{\tau_{j+1} - \tau_j} \right)^2 \right] + 1 \right\} . \tag{2.53}$$

Hence, we can detect changes in mean and variance using the PELT algorithm with $K = 0$ and cost function defined by Equation (2.53). Note, that the estimation of the variance requires the PELT algorithm to enforce a minimum segment length $(\tau_{j+1} - \tau_j)$ of 2 for all segments $\boldsymbol{y}_{(\tau_j+1):\tau_{j+1}}$, $j = 0, \ldots, m$, of all considered $(m + 1)$-segmentations $\boldsymbol{\tau}_{0:(m+1)}$.

3 Data

This chapter introduces the different data sets used throughout the thesis. We provide (background) information on the data sets used to illustrate the novel methodology for drought index computation (Section 3.1) and the data used in an evaluation study of such drought indices (Section 3.2). Moreover, we outline the data sets used to illustrate the evaluation of Regional Climate Models (RCMs) using novel evaluation techniques (Sections 3.3–3.4).

3.1 ECMWF Atmospheric Reanalysis of the 20th Century

To illustrate the novel methodology for drought index computation (and its different modeling steps) presented in Chapter 4 we utilize the publicly available *ECMWF Atmospheric Reanalysis of the 20th Century (ERA-20C)* data (European Centre for Medium-Range Weather Forecasts, 2014). The data set is a reanalysis of the weather observed on the earth's surface during the years 1900–2010. It is generated using coupled atmosphere, land-surface and ocean-wave models and intended to describe the spatial and temporal development of the atmosphere, the weather on the land-surface and ocean waves. It is available on different spatial and temporal resolutions and consists of manifold variables. We consider

- time series of monthly means (temporal resolution) and
- $0.125° \times 0.125°$ (longitude/latitude) grids (spatial resolution).

The variables of interest are

- 2 metre temperature (T) [in K],
- total precipitation (P) [in m] and
- volumetric soil water (W) [in $\mathrm{m}^3\mathrm{m}^{-3}$] in the top soil layer (0–7 cm),

with variable units stated in square brackets. Using the method proposed by Thornthwaite (1948) we additionally compute the variable *potential evapotranspiration* (PET) [in mm] based on the temperature data. Further, we compute the so called climatic water balance (B) [in mm] (see e.g. Vicente-Serrano et al., 2010) as difference between P and PET.

3.2 Arkansas soybean yield

For the validation of the novel drought indices (see Section 4.7), we consider *soybean yield* (see National Agricultural Statistics Service, United States Department of Agriculture, 2015) for selected counties of the U.S. state of Arkansas. To be able to observe the effect of droughts in this data we consider only the yield [in bu/acre] from non-irrigated fields. To have complete time series (without missing values), we restrict us to the yearly yield over the period 1972–2001 (see Figure 3.2) for the 27 counties Arkansas, Ashley, Chicot, Clay, Craighead, Crittenden, Cross, Desha, Drew, Greene, Independence, Jackson, Jefferson, Lawrence, Lee, Lincoln, Lonoke, Mississippi, Monroe, Phillips, Poinsett, Prairie, Pulaski, Randolph, Saint Francis, White and Woodruff (see Figure 3.1). Arkansas county (which is used for illustration throughout the article) is highlighted in Figures 3.1 and 3.2.

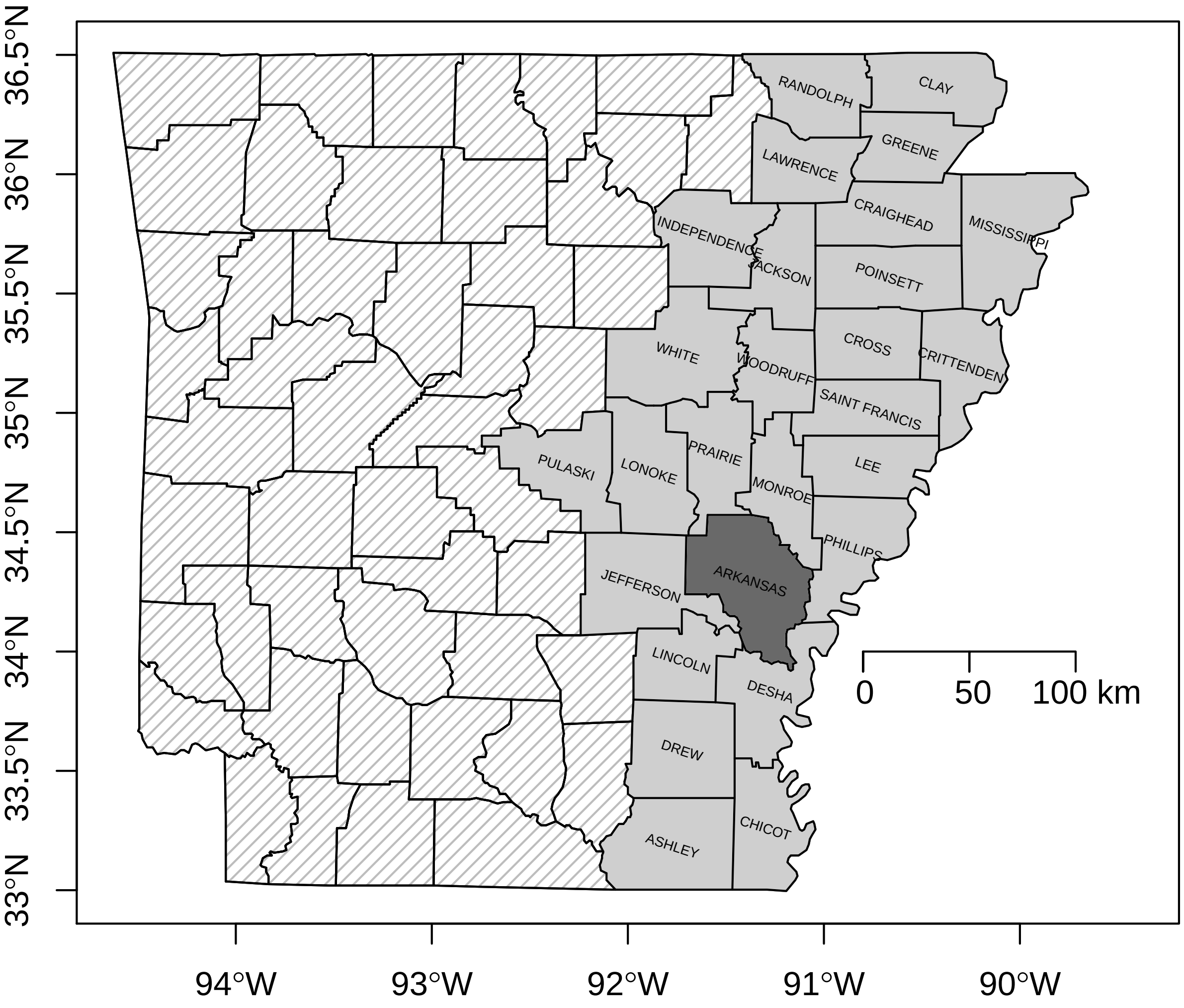


Figure 3.1: Counties of the U.S. state of Arkansas where soybean yield data (non-irrigated) is available for all years 1972–2001. Arkansas county is highlighted.

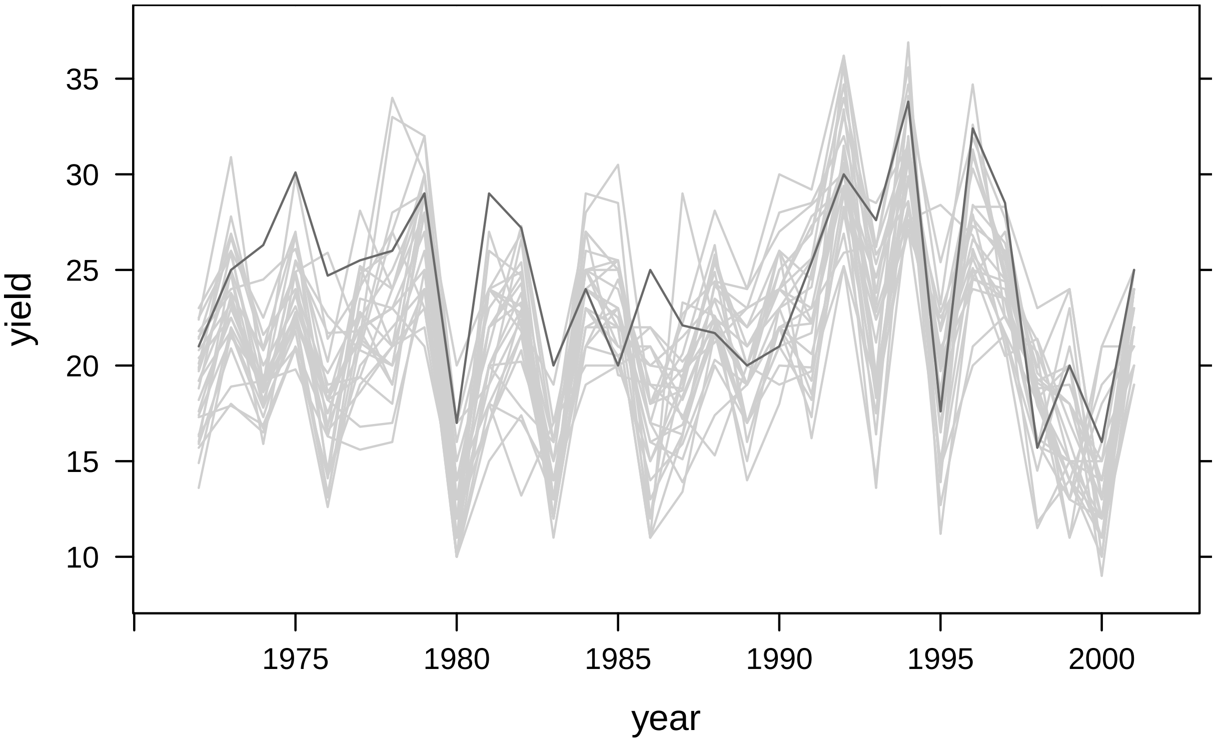


Figure 3.2: Soybean yield time series for all 27 selected counties of the U.S. state of Arkansas. The time series for Arkansas county is highlighted.

3.3 E-OBS gridded data set

For an application (Section 5.9) of the moving score/divergence methodology presented in Chapter 5 we use the high-resolution gridded *E-OBS data set* (version 13.1, Haylock et al., 2008) as a reference data set, which was developed in the course of the *ENSEMBLES project* (see van der Linden and Mitchell, 2009). The data set is a gridded data set derived/interpolated from station observations. It covers the European continent (only land, 25°N–75°N × 40°W–75°E) and is available for different grids and spatial resolutions (longitude-latitude grids: 0.25°, 0.5°; rotated pole grids with North Pole at 39.25°N, 162°W: 0.22°, 0.44°). We use version 13.1 of the data set, which provides daily values of the variables

- precipitation sum [in mm]
- mean temperature [in °C]
- minimum temperature [in °C]
- maximum temperature [in °C]

for the years 1950–2015, with variable units stated in square brackets.

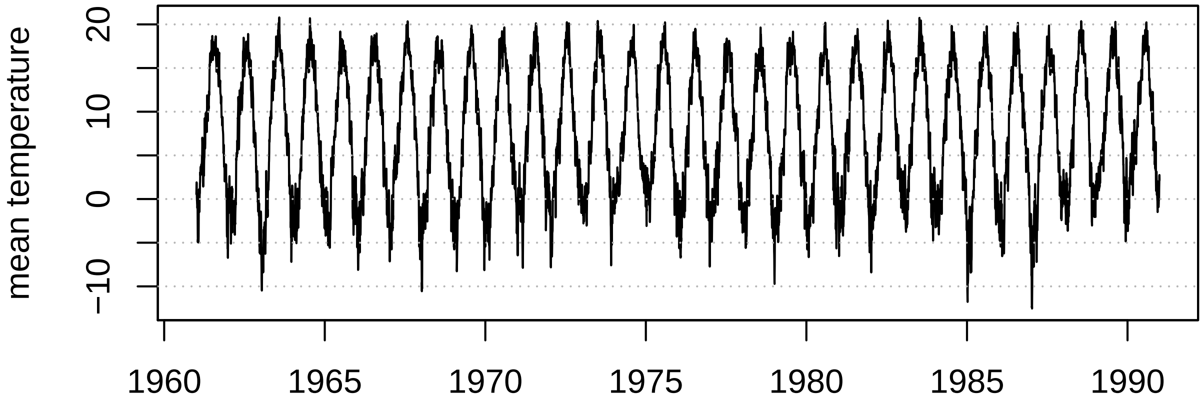


Figure 3.3: E-OBS: Time series (1961–1990) of spatial mean temperature average. The average is taken over the area indicated in Figure 3.4.

As in Section 5.9 we are particularly interested in the mean temperatures, we illustrate (Figure 3.3) the spatial average of the mean temperature time series over the rectangular region visualized in Figure 3.4 for the time period 1961–1990. On the contrary, Figure 3.4 illustrates the temporal average over the same time period.

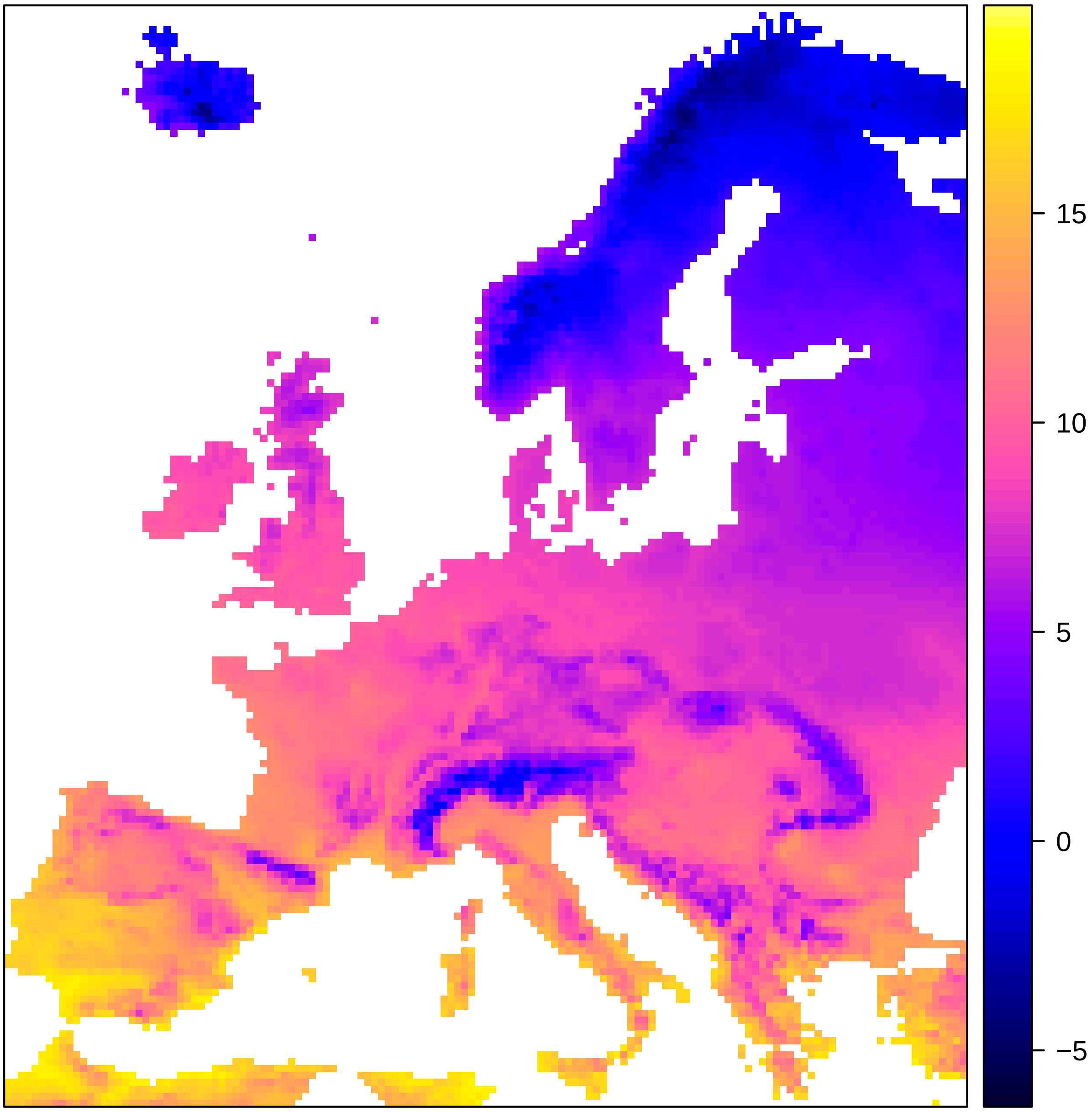


Figure 3.4: E-OBS: Map (Europe) of temporal mean temperature average. The average is taken over the time period 1961–1990.

3.4 ENSEMBLES Regional Climate Model simulations

In a case study in Section 5.9, we evaluate Regional Climate Models (RCMs) using the newly introduced evaluation method presented in Chapter 5 (Section 5.3). For that, we consider a selection of the RCM simulations which were carried out as part of the *ENSEMBLES project* (see van der Linden and Mitchell, 2009).

The purpose of *Regional Climate Models (RCMs)* is to dynamically downscale gridded climate data (which provides the *boundary conditions* for the RCM/is used to *drive* the RCM) with a coarse spatial resolution (say $\sim$ 200–300 km) to obtain information/simulations of the climate on a much finer resolution ($\sim$ 25–50 km). Due to their complexity and the related computational burden RCMs are usually applied for a certain (small) area of interest (e.g. Europe). The output of an RCM is manifold. For climate studies, near-surface (2 meter) temperature as well as precipitation are variables of particular interest.

As part of the *ENSEMBLES project* (van der Linden and Mitchell, 2009), ensembles of RCM simulations from different (European) climate research institutes were compiled. In total 16 climate research institutes provided simulations from their individual RCMs. In an initial 40-year experiment (1961–2000) covering Europe, the *RCMs were driven by the ERA-40 reanalysis data set* (from the European Centre for Medium-Range Weather Forecasts, cp. ERA-20C data set outlined in Section 3.1). This experiment was used to evaluate the RCMs (see also Section 5.2). Further, *RCM experiments driven by Global Climate Model (GCM) output* were conducted to create an ensemble of regional climate change projections for Europe. These experiments cover the time period 1951–2050 (some of them cover 1951–2100). Whereas the different involved climate research institutes considered different driving GCMs, they almost all considered the same emission scenario. Most of the RCM simulations are available on a spatial resolution of 25 kilometers, some of them are also available for 50 kilometers.

For our case study in Section 5.9 we consider only the output variable *2 meter temperature*. For the control period 1961–1990 and a spatial resolution of 25 kilometers, we compare the RCMs of

- the Danish Meteorological Institute (*DMI*),
- the Royal Netherlands Meteorological Institute (*KNMI*),
- the Max-Planck-Institute for Meteorology (*MPI*),
- the Swedish Meteorological and Hydrological Institute (*SMHI*).

The corresponding model acronyms are

- DMI-HIRHAM,
- KNMI-RACMO2,
- MPI-M-REMO,
- SMHIRCA.

In Section 5.9, we will refer to these different models using only the acronym of the corresponding institute. We evaluate these four models for both, ERA-40 and GCM boundary conditions, which yields in total eight different model outputs for our comparison. Note, that all four RCMs were driven by the same GCM called *ECHAM5*. To differentiate between the two different boundary conditions, we will hence use the acronyms *ERA-40* and *ECHAM5*.

4 Novel Drought Indices

As we have seen in the introduction (see Chapter 1), there already exist several different drought indices, often with a focus on a specific application area. There have been numerous attempts to provide multivariate drought indices that outrun established (univariate) indices. We summarize the lessons we have learned from established drought indices (Section 4.1). Based on these insights we provide novel methodology for the flexible computation of uni- and multivariate drought indices (Sections 4.4 and 4.5). The multivariate approach utilizes vine copulas (see Section 2.3.2) to model the dependence among several drought relevant variables. In a simulation study (Section 4.6) we investigate some characteristics of the proposed multivariate indices. In an application (Section 4.7), we illustrate the novel methodology and it its merits. This chapter is mainly based on Erhardt and Czado (2016).

4.1 Lessons learned from established drought indices

Among the plethora of different drought indicators and indices a few have gained widespread popularity (see e.g. Mishra and Singh, 2010, for an overview of commonly used drought indices). Clearly, 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) are the most popular ones. Due to their properties they have been used in multitudinous drought studies and have been included in drought monitoring and early warning systems (Bachmair et al., 2016). However, also some of their properties have been criticized (Wells et al., 2004; Farahmand and AghaKouchak, 2015). The vivid scientific discussion about these well established approaches and other recent approaches to multivariate drought quantification (multivariate drought indices, see Section 2.2) provides us with valuable insights for the development of novel (multivariate) indices with the aim to improve existing approaches from a theoretical and practical point of view. Subsequently, we discuss these insights and summarize the lessons we have learned.

As the climate differs for different regions across the globe, the occurrence of drought depends on the local (climate) conditions. Moreover, these local conditions vary with the seasons. Hence, drought indices should quantify deviations from (local) normal conditions, i.e. they should account for *seasonality.* By convention, for many established drought indices, *negative/small values reflect dry conditions and positive/high values wet conditions.* The computation of most established drought indices usually *requires long data records* to yield meaningful results.

The *Palmer Drought Severity Index* (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 U.S., 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, many other drought indices like the *Standardized Precipitation Index* (SPI) (McKee et al., 1993; Edwards and McKee, 1997) do not make any assumptions about the physics of the water cycle and have an *interpretation in terms of probability*. 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 Section 2.1) 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 (see e.g. Farahmand and AghaKouchak, 2015). 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 not only based on precipitation. Instead a climatic water balance (precipitation minus potential evapotranspiration) is considered to quantify dry/wet conditions. Using the SPEI, temperature *trends* (climate change) are passed on to the index.

Kao and Govindaraju (2010) present a (to our knowledge) first multivariate (see Section 2.2) copula-based (see Section 2.3.1) 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 empirical Kendall distribution function (see e.g. Barbe et al., 1996) to assess the joint probability. The results are transformed to a standard normal distribution. Note, that for meaningful estimation of the empirical Kendall distribution in high dimensions (here multiples of 12) long data records are required. As in such high dimensions the Kendall distribution becomes almost degenerate at 0 (see e.g. Brechmann, 2014), the quantification/distinction of dry conditions may become problematic.

Hao and AghaKouchak (2013, 2014) introduce a bivariate parametric and non-parametric version of the *Multivariate Standardized Drought Index* (**MSDI**), respectively, by enhancing the SPI idea to bivariate data (in their example precipitation and soil moisture time series). While the parametric MSDI models the bivariate distribution based on copulas (see Section 2.3.1), the non-parametric MSDI uses a bivariate empirical distribution. To obtain the MSDI the joint cumulative probability is transformed with the inverse CDF of a standard normal distribution. Note however, that this approach does not yield a real standardization. Usually, negative values of the proposed index are more probable, since the joint cumulative probability is not uniformly distributed on $[0, 1]$. Moreover, Farahmand and AghaKouchak (2015) provide a Standardized Drought Analysis Toolbox (**SDAT**), which allows computation of non-parametric standardized univariate and non-parametric bivariate (MSDI) drought indices.

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

Table 4.1: Comparison of different drought indices (SC-PDSI, SPI/SPEI, univariate SDAT, JDI, MSDI) 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 a statistical model).

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

INTERP be interpretable in terms of probability (frequency analysis/classification of droughts).

ARBVAR be applicable to *arbitrary drought relevant variables*.

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

SMALLS yield meaningful results for (monthly) data records of 10 years (120 months) and more (*small sample size*).

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* (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 have *a multivariate extension* (different types of drought).

Table 4.1 summarizes which of these characteristics the above addressed indices fulfill.

4.2 Outline of the modeling approach

In the subsequent sections (Sections 4.4 and 4.5), we introduce a novel approach to drought modeling. Its main idea is to combine a set of drought-relevant variables into a single drought index and to simultaneously take the inter-variable dependencies at hand into account. To model these dependencies we consider the very flexible class of vine copulas, which we introduce in Section 2.3.2.

The novel methodology for drought index computation considers the above introduced (Section 4.1) desired characteristics of drought indices step-by-step as shown in Figure 4.1. The proposed procedure starts off with d time-series of user-selected, drought-relevant variables (red box). It consists of two logical blocks. The first block (orange boxes, see Section 4.4) deals with modeling each of the selected variables separately (univariate/marginal modeling). It results in a transformation of each original time-series to an i.i.d. sample of a standard uniform distribution (uniform margins). Further transformation to the standard normal distribution yields univariate standardized indices (green box). The second block (blue boxes, see Section 4.5) deals with the (vine copula based) dependence modeling of the uniform margins resulting from block one. It provides methodology to combine several variables into a multivariate standardized index (see Section 2.2).

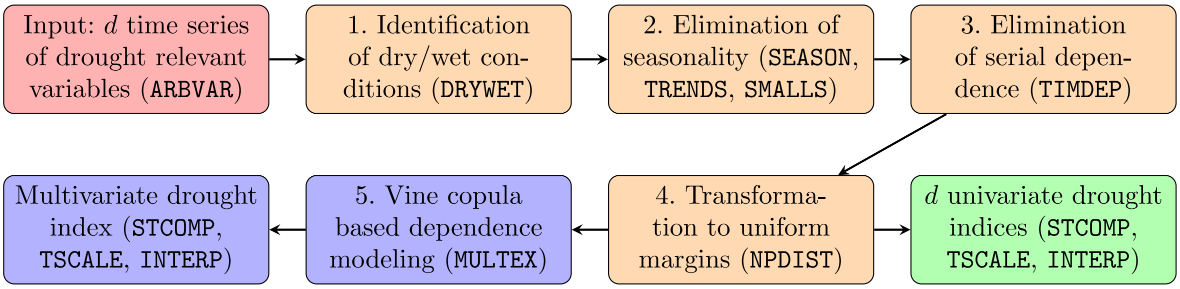

Input: d time series of drought relevant variables (ARBVAR)
1. Identification of dry/wet conditions (DRYWET)
2. Elimination of seasonality (SEASON, TRENDS, SMALLS)
3. Elimination of serial dependence (TIMDEP)
4. Transformation to uniform margins (NPDIST)
d univariate drought indices (STCOMP, TSCALE, INTERP)
5. Vine copula based dependence modeling (MULTEX)
Multivariate drought index (STCOMP, TSCALE, INTERP)

Figure 4.1: Modeling steps for uni-/multivariate drought index calculation.

4.3 Data example: ERA-20C/Arkansas

In the subsequent sections, we elaborate on the modeling/transformation steps addressed in Figure 4.1. To illustrate the different steps we consider time series of monthly means of the three variables

T 2 metre temperature,

B climatic water balance,

W volumetric soil water (top layer),

(derived) from the ERA-20C data introduced in Section 3.1. As the introduction of the novel methodology in Sections 4.4 and 4.5 will be followed by a small evaluation of the proposed indices, which will be based on the soybean crop yield data for 27 counties of the U.S. state of Arkansas (see Section 3.2/Figures 3.1–3.2), we consider spatial aggregates of the above time series to the same 27 counties over the period 1972–2001. For illustration we provide time series of T, B and W for the county Arkansas (in the state of Arkansas) in Figure 4.2. Subsequently, if a figure or table refers to this data example we indicate this in the corresponding caption. If the caption starts with

ERA-20C/Arkansas (27),

the figure/table refers to the full data set consisting of the T, B and W time series for all 27 counties under consideration. If it starts with

ERA-20C/Arkansas (county),

it refers only to the time series for Arkansas county.

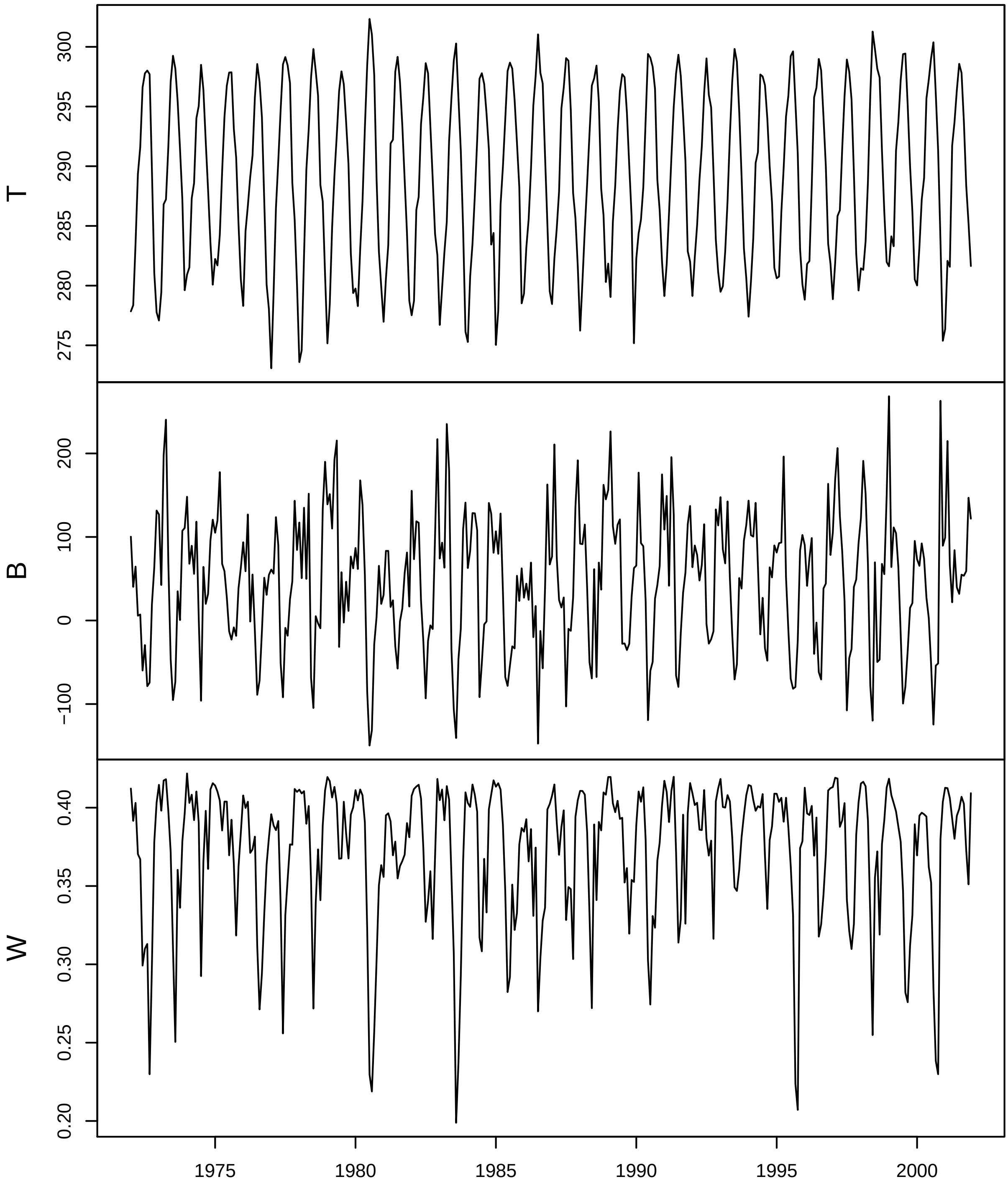


Figure 4.2: ERA-20C/Arkansas (county): Temperature (T), climatic water balance (B; precipitation minus potential evapotranspiration) and (top layer) volumetric soil water (W) time series (1972–2001).

4.4 Univariate standardized indices

In a first step we develop a statistically sound and generalized modeling framework for monthly time series of drought-relevant variables. Simultaneously, this serves as marginal model in our vine copula based modeling framework (see Section 4.5) and as algorithm for the calculation of univariate standardized indices. These indices will have the properties which were discussed in Section 4.1.

4.4.1 Identification of dry and wet conditions

Let us now consider a time series x_{t_k}, $k = 1, \ldots, N$, for an arbitrary drought-relevant variable (ARBVAR). By convention, small values should always indicate dry and big values wet conditions (DRYWET). To ensure that, we change the sign of the time series if it was the other way round. Consider for instance the three time series for temperature (T), climatic water balance (B) and soil water volume (W) provided in Figure 4.2. We observe minima of B and W at the same time when temperature peaks occur. High temperatures mean high evapotranspiration, which leads to dry conditions. For low values the opposite is observed. Therefore we need to multiply the T time series by -1. In the case of the variables B and W this is not required, as low climatic water balance and soil water volume correspond to dry conditions.

4.4.2 Elimination of seasonality

The monthly time series of interest x_{t_k}, $k = 1, \ldots, N$, typically show seasonal variations in their mean, variance and skewness. This becomes clear from our exemplary time series provided in Figure 4.2, if we decompose each of these time series into 12 month-wise sub-series, as visualized in Figure 4.3.

To eliminate this seasonal heterogeneity (SEASON), subsequent steps include a month-wise standardization of the time series. As illustrated in Figure 4.3, we decompose the time series x_{t_k}, $k = 1, \ldots, N$, month-wise into 12 separate series, one for each month of the year: Let each time point t_k, $k = 1, \ldots, N$, be a 2-tupel (m_k, y_k), where $m_k \in \{1, \ldots, 12\}$ (1 = January, …, 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*

$$\boldsymbol{x}_m := (x_{t_k})_{k \in \mathcal{K}(m)} = \left\{ x_{(m,y_k)}, k \in \mathcal{K}(m) \right\}, \quad m = 1, \ldots, 12,$$

where the index set for month m is defined as $\mathcal{K}(m) := \{k : m_k = m\}$.

To enable a standardization of the month-wise time series $\boldsymbol{x}_m$, $m = 1, \ldots, 12$, we first eliminate the skewness which may as well vary depending on the season (see e.g. the empirical skewness estimates in Figure 4.3). To achieve that, we consider continuous, monotonic increasing 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

$$\psi(\lambda, x) = \begin{cases} \left((x+1)^{\lambda} - 1 \right) / \lambda & \text{if } x \geq 0, \lambda \neq 0 \\ \ln(x+1) & \text{if } x \geq 0, \lambda = 0 \\ -\left((-x+1)^{2-\lambda} - 1 \right) / (2-\lambda) & \text{if } x < 0, \lambda \neq 2 \\ -\ln(-x+1) & \text{if } x < 0, \lambda = 2. \end{cases}$$

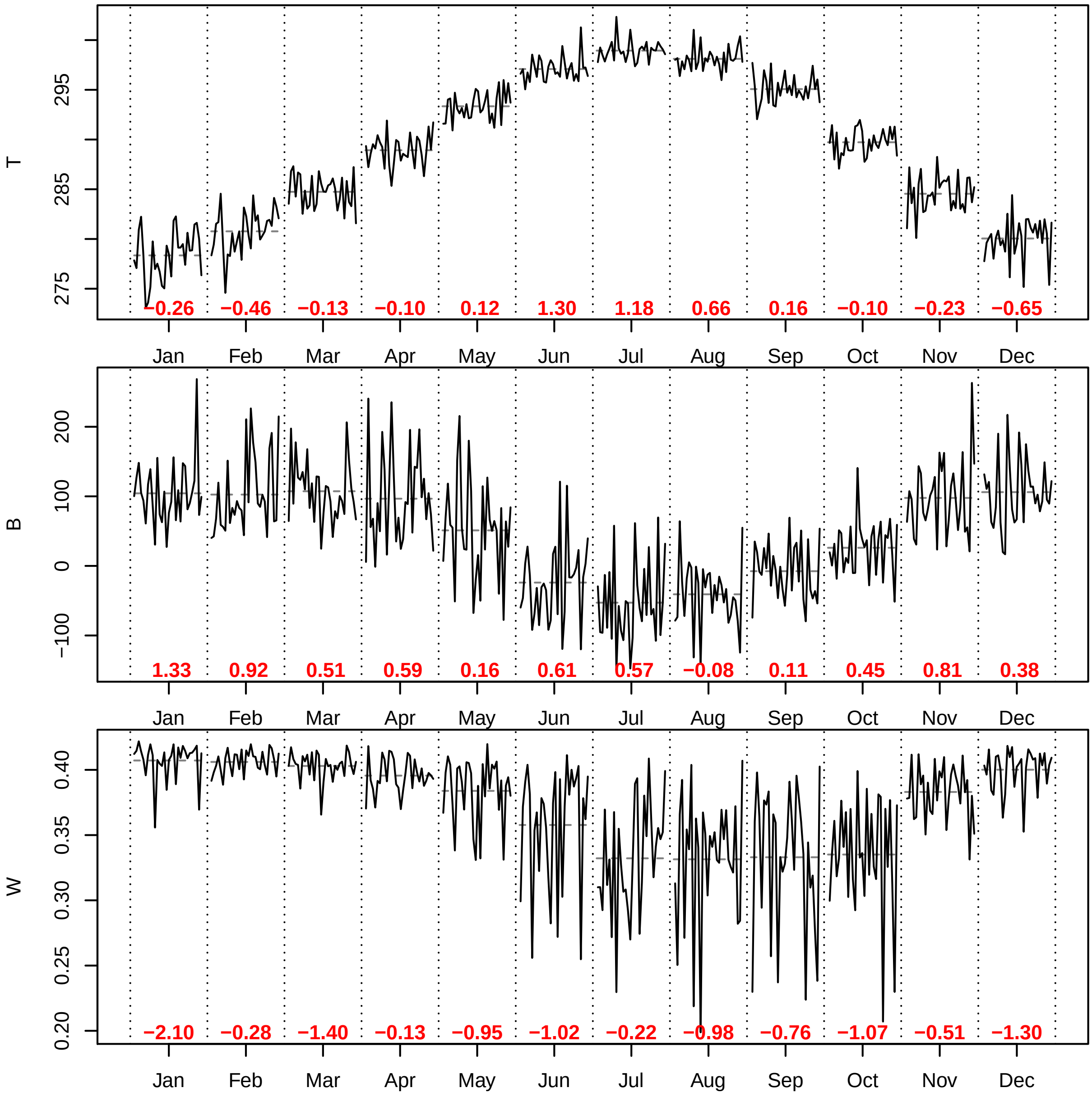


Figure 4.3: ERA-20C/Arkansas (county): Month-wise T, B and W time series (1972–2001). The gray dashed lines indicate the month-wise means. The red numbers are month-wise empirical skewness estimates.

For each month $m = 1, \dots, 12$ separately we estimate a parameter λ_m using maximum likelihood estimation (see Yeo and Johnson, 2000) ($\widehat{\lambda}_m$, $m = 1, \dots, 12$, are the corresponding estimates) and denote the transformed month-wise time series by

$$\widetilde{\boldsymbol{x}}_m := \left\{ \widetilde{x}_{t_k} = \widetilde{x}_{(m,y_k)} = \psi\left(\widehat{\lambda}_m, x_{(m,y_k)}\right), k \in \mathcal{K}(m) \right\}, \quad m = 1, \dots, 12. \tag{4.1}$$

Figure 4.4 provides maps depicting the estimates $\widehat{\lambda}_m$, $m = 1, \dots, 12$, for the three variables T, B and W for all selected 27 counties of the state of Arkansas.

Since drought is considered as a (negative) deviation from ‘normal’ conditions (anomaly), we model and subtract the month-wise mean μ_m separately for each of the 12 time series $\widetilde{\boldsymbol{x}}_m$,

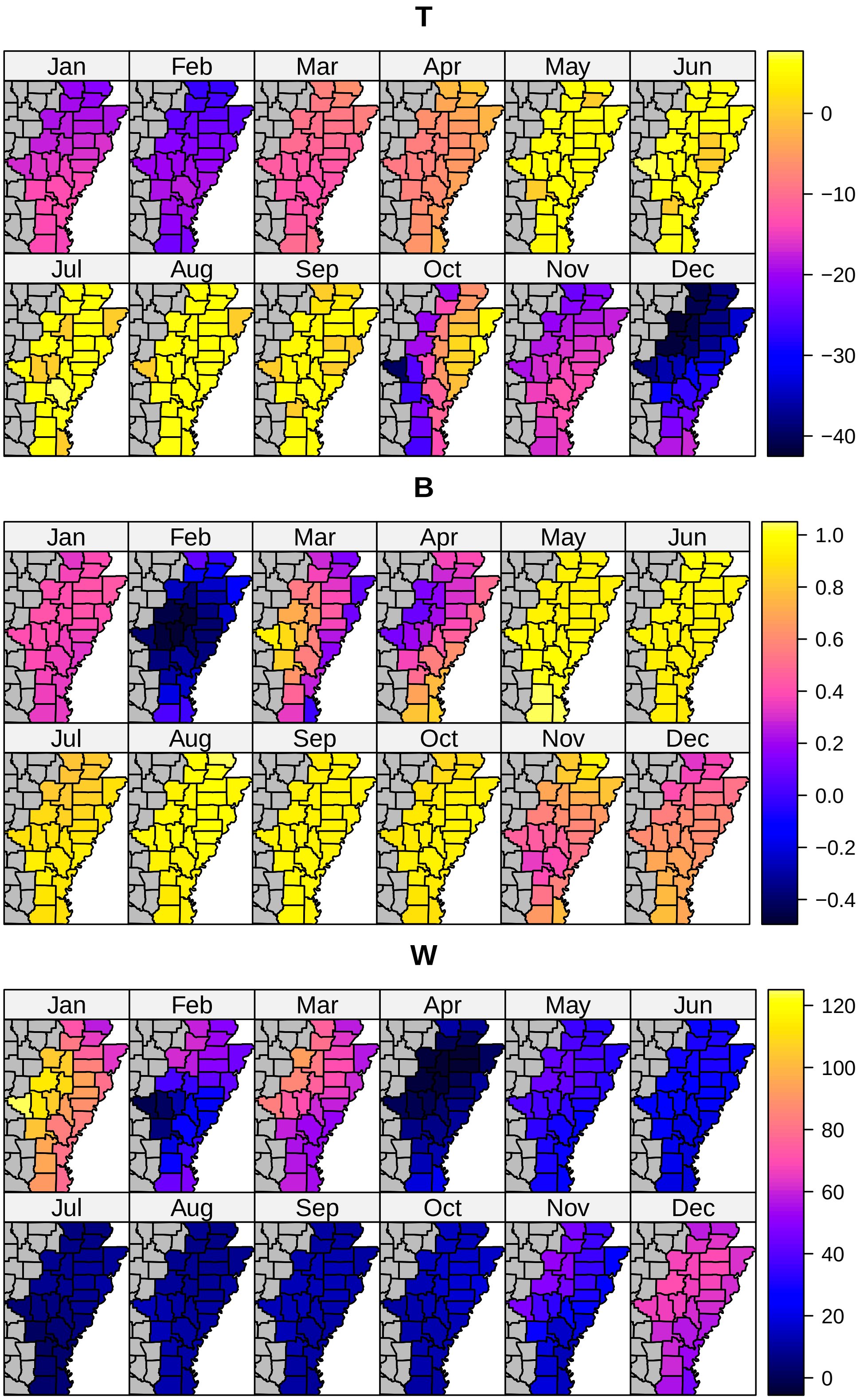


Figure 4.4: ERA-20C/Arkansas (27): Month-wise maps of Yeo and Johnson transformation parameters λ_m, $m = 1, \dots, 12$, for the month-wise T (top), B (middle) and W (bottom) time series.

$m = 1, \dots, 12$ given by Equation (4.1). We estimate it as the sample mean

$$\widehat{\mu}_m := \frac{1}{|\mathcal{K}(m)|} \sum_{k \in \mathcal{K}(m)} \widetilde{x}_{(m,y_k)}, \quad m = 1, \dots, 12.$$

This ensures that $\sum_{k \in \mathcal{K}(m)} \left(\widetilde{x}_{(m,y_k)} - \widehat{\mu}_m \right) = 0$ for each month $m = 1, \dots, 12$. Thus also the re-composed time series of *anomalies* $a_{t_k} := \widetilde{x}_{t_k} - \widehat{\mu}_{m_k}$, $k = 1, \dots, N$, is centered around 0. Hence, seasonal deviations from the annual mean could be eliminated.

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. To quantify this seasonal heterogeneity of the time series a_{t_k}, $k = 1, \dots, N$, we estimate month-wise standard deviations based on the unbiased sample variance 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. Finally we can return to the full (re-composed) monthly time series of anomalies and obtain a homogenized time series by computing the standardized residuals

$$r_{t_k} := a_{t_k} / \widehat{\sigma}_{m_k}, \quad k = 1, \dots, N.$$

We call r_{t_k}, $k = 1, \dots, N$, *deseasonalized time series.* The deseasonalized time series corresponding to the time series T, B and W provided in Figure 4.2 are shown in Figure 4.5.

The re-composition into a single time series ensures that the sample size (`SMALLS`) for fitting a distribution is twelve times bigger and that only one distribution has to be fitted compared to fitting one distribution to each of the twelve month-wise time series, how it is done for example in case of classical standardized drought indices (see Section 2.1). Note moreover, that during the three steps described above, only monotonic increasing transformations were applied to the month-wise time series, i.e. their (month-wise) ranking did not change. Hence, the resulting residuals r_{t_k}, $k = 1, \dots, N$, can be interpreted as deviations from 'normal' conditions which are comparable across different months, due to the standardization.

Besides seasonality, climatic variables 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.

4.4.3 Elimination of temporal dependencies

Apart from seasonality, time series often feature temporal dependence (`TIMDEP`). Generally 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}, $k = 1, \dots, N$, 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

$$r_{t_k} = \sum_{j=1}^{p} \phi_j r_{t_{k-j}} + \sum_{j=1}^{q} \theta_j \varepsilon_{t_{k-j}} + \varepsilon_{t_k}, \tag{4.2}$$

where the error terms ε_{t_k} are independent and identically normal distributed with mean 0 and variance σ^2, that is i.i.d. $N(0, \sigma^2)$. Note, that for p or q equal to 0 the corresponding summands are neglected and the model is denoted as MA(q) or AR(p), respectively. For adequate choice

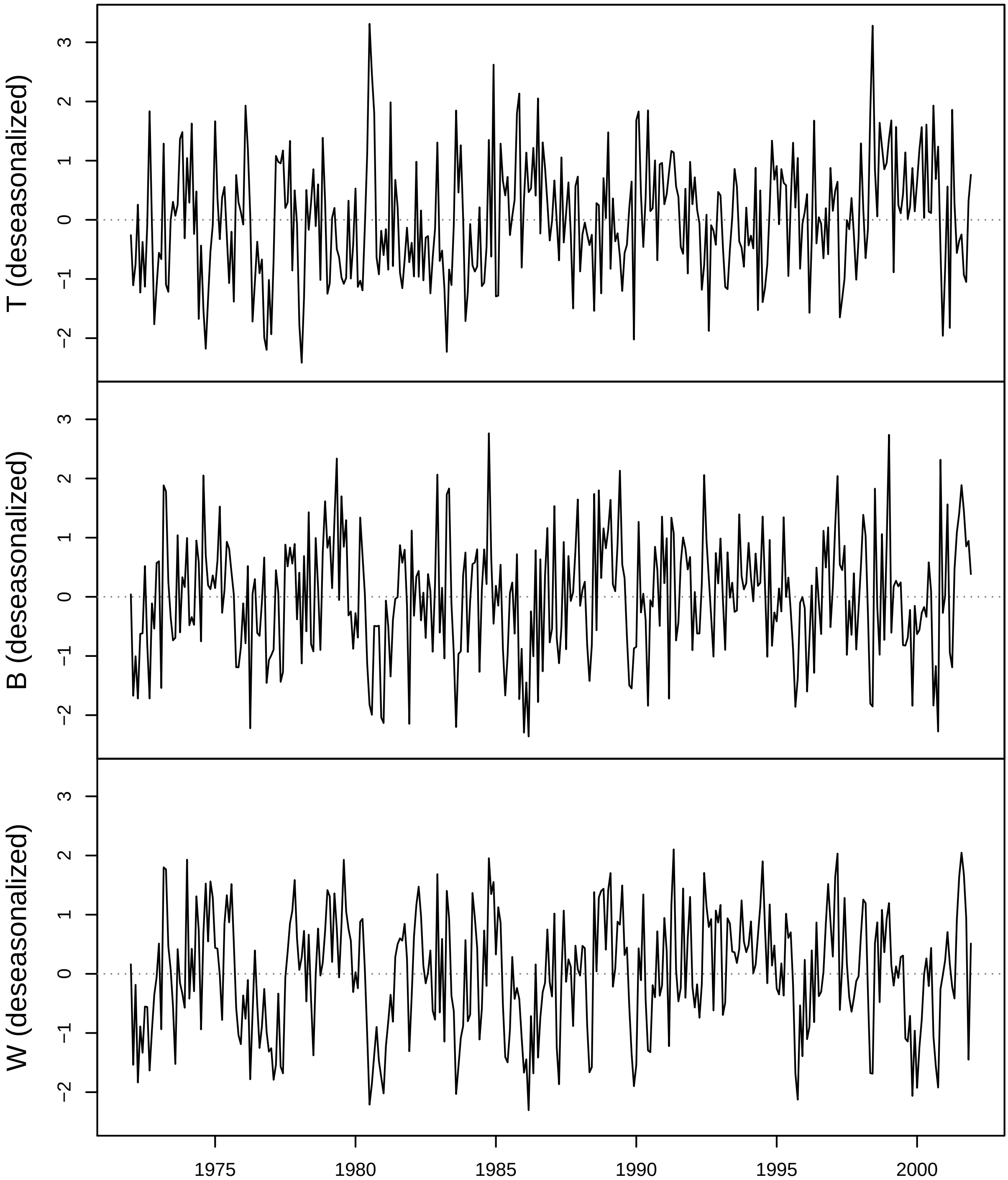


Figure 4.5: ERA-20C/Arkansas (county): Deseasonalized T, B and W time series (1972–2001).

of the orders p and q in the ARMA(p, q) model (4.2) and estimates $\widehat{\phi}_j$, $j = 1, \ldots, p$, and $\widehat{\theta}_j$, $j = 1, \ldots, q$, of the corresponding parameters the model residuals

$$\epsilon_{t_k} := r_{t_k} - \sum_{j=1}^{p} \widehat{\phi}_j r_{t_{k-j}} - \sum_{j=1}^{q} \widehat{\theta}_j \epsilon_{t_{k-j}}, \quad k = 1, \ldots, N, \tag{4.3}$$

are approximately temporally independent. Hence, serial dependencies are filtered out and the residuals ϵ_{t_k}, $k = 1, \ldots, N$, contain information on the contribution of each month to the current dry-/wetness conditions.

Note that in the context of monthly climate data most often an AR(1) model is sufficient, as shown in Figure 4.6 for the time series of our data example. For the deseasonalized variables T, B and W Ljung-Box tests (at a significance level of $\alpha = 5\%$) reject temporal independence (p-values close to 0) for almost all counties (with some exceptions for the variable B). However, the same tests do not reject temporal independence of the residuals of AR(1) models for these deseasonalized variables for any of the variables or counties (p-values above 0.2).

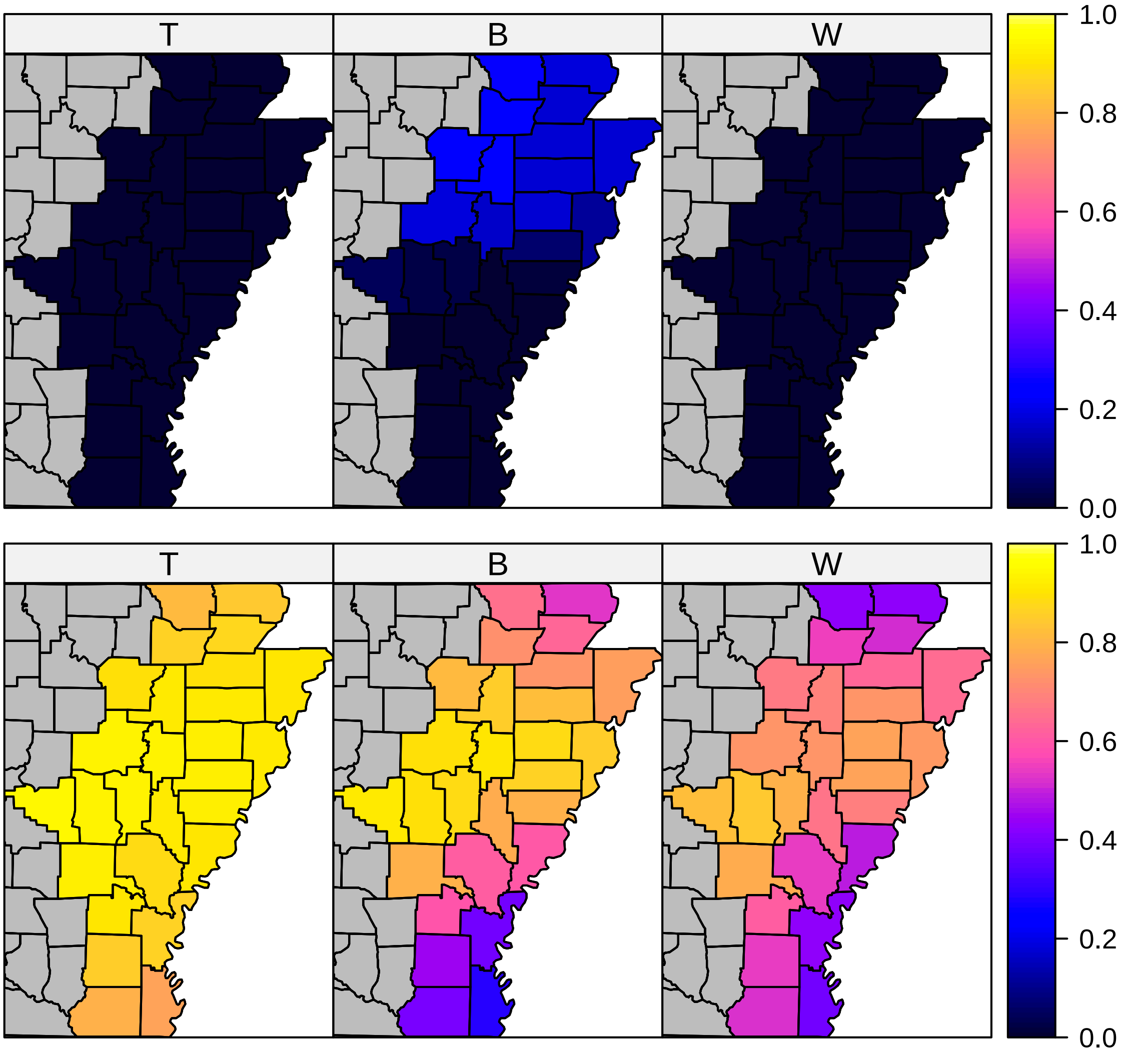


Figure 4.6: ERA-20C/Arkansas (27): Ljung-Box test for serial independence; Maps of p-values for deseasonalized T, B and W time series (top) and corresponding AR(1) residuals (bottom).

4.4.4 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

$$\widehat{F}_N(x) := \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}\{\epsilon_{t_k} \le x\} \tag{4.4}$$

of the data respectively the residuals ϵ_{t_k}, $k = 1, \ldots, N$, (see Equation (4.3)) 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}, $k = 1, \ldots, N$, it is a critical assumption that the components of the sample are independent from each other and come from one and the same distribution. We ensured this independent and identical distribution (i.i.d.) assumption in the previous step by eliminating the temporal dependencies.

We use the estimated distribution $\widehat{F}_N$ from Equation (4.4) to transform our residuals ϵ_{t_k}, $k = 1, \ldots, N$, to copula data (*u-scale*), i.e. to be uniformly distributed on the interval $[0, 1]$ (see Equation (2.2) Section 2.3.1). We calculate the (scaled) probability integral transform (PIT)

$$u_{t_k} := N/(N+1)\widehat{F}_N\left(\epsilon_{t_k}\right) = \operatorname{rank}\left(\epsilon_{t_k}\right)/(N+1), \quad k = 1, \ldots, N,$$

where we multiply by $N/(N + 1)$ to avoid any $u_{t_k} = 1$. Further, we transform to the *z-scale*, calculating

$$z_{t_k} := \Phi^{-1}\left(u_{t_k}\right), \quad k = 1, \ldots, N, \tag{4.5}$$

using the inverse PIT based on the CDF Φ of a standard normal distribution. It holds that z_{t_k}, $k = 1, \ldots, N$, is (approximately) independent and identically standard normal distributed (`STCOMP`).

4.4.5 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 (see Section 2.1). 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 independence 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 4.4.1–4.4.4 only once, after that we are able to calculate the index on arbitrary time scales. Subsequently, we denote the time scale by ℓ and indicate it as a subscript along with a standardized index.

The (approximately) temporally independent, standard normal distributed time series z_{t_k}, $k = 1, \ldots, N$, calculated according to Equation (4.5), is already a standardized index with time scale $\ell = 1$. Hence we can write the corresponding index as $\mathrm{SI}_1(t_k) := z_{t_k}$, $k = 1, \ldots, N$. The normal distribution has the advantage that a sum of (independent) normal distributed random variables is again normally distributed. We use this property to calculate standardized indices for time scales $\ell \ge 1$. The sum $\sum_{j=0}^{\ell-1} z_{t_{k-j}}$ of standard normal variables is normally distributed with mean 0 and variance ℓ. Hence, we obtain a *standardized index* with time scale ℓ as

$$\mathrm{SI}_\ell(t_k) := \frac{1}{\sqrt{\ell}}\sum_{j=0}^{\ell-1} z_{t_{k-j}}, \quad k = 1, \ldots, N. \tag{4.6}$$

Following our convention from Section 4.4.1 a small (negative) index value of the index (4.6) indicates dryness/drought and a big (positive) value indicates wetness. To classify the values of standardized indices (like the SPI, SPEI and our standardized indices proposed in Sections 4.4 and 4.5) we use the dry- and wetness categories defined in Table 2.1.

Figure 4.7 depicts the univariate standardized drought indices based on the climatic water balance B for time scales $\ell = 1, 6, 12$, for Arkansas county (1972–2001). We identify persistent dry periods during the years 1972, 1976/77, 1980/81, 1986, 1996 and 2000. The drought events during the years 1980/81, 1986 can be considered as exceptionally dry (D4), according to the classification of Table 2.1. 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).

Figure 4.8 depicts the univariate drought indices $\mathrm{SI}_6(T)$, $\mathrm{SI}_6(B)$ and $\mathrm{SI}_6(W)$ for Arkansas county (1972–2001). We observe that the different variables carry different information about dry-/wetness conditions. Whereas the drought events during the years 1980/81, 1986 and 2000 were accompanied and intensified by high temperatures, the temperature based index $\mathrm{SI}_6(T)$ alone does not indicate drought during the other events (1972, 1976/77 and 1996) identified by $\mathrm{SI}_6(B)$ and $\mathrm{SI}_6(W)$. Hence, it might be of value to combine the dry-/wetness information captured in different drought relevant variables in one multivariate index, as we propose in the subsequent section.

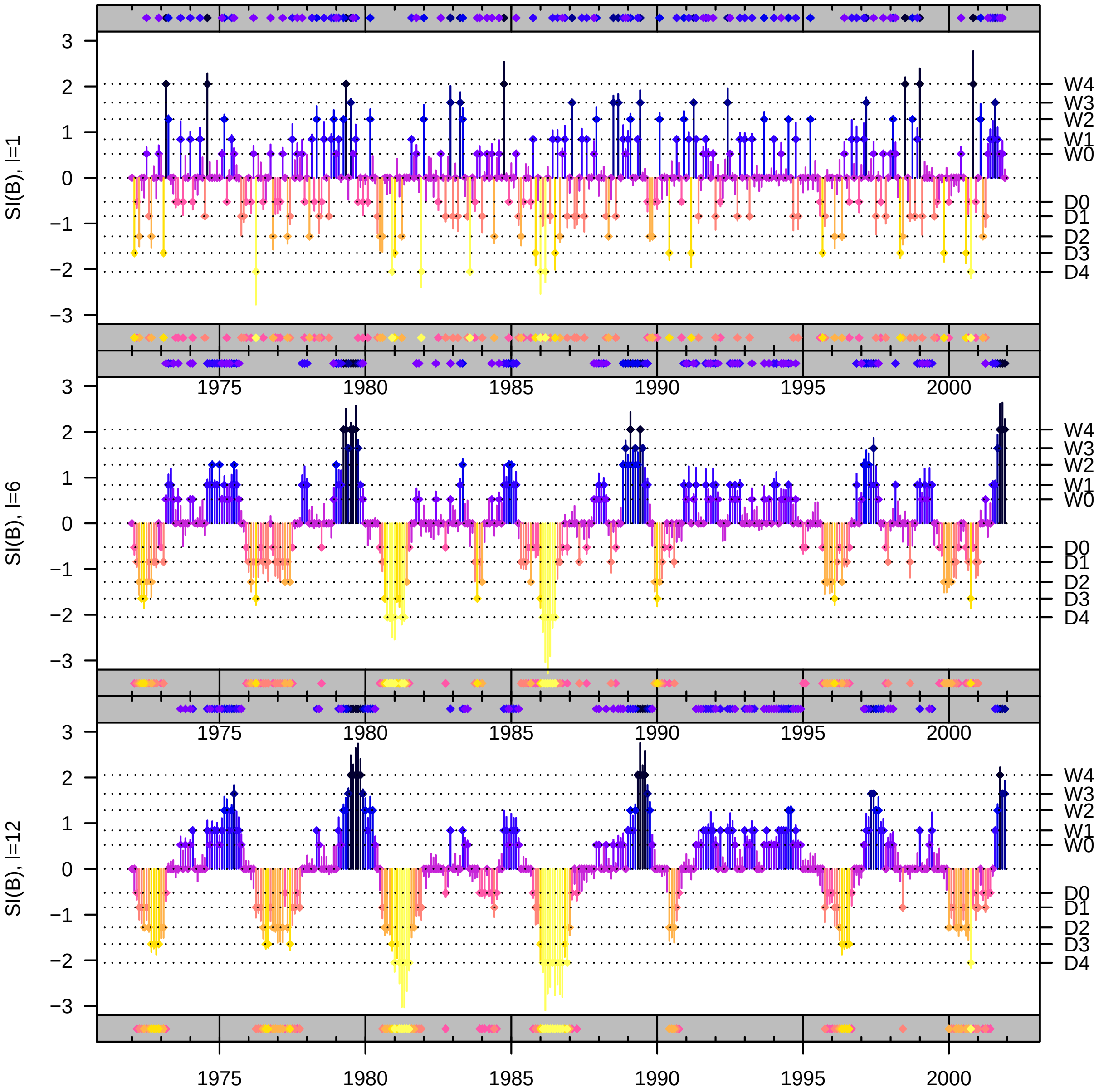


Figure 4.7: ERA-20C/Arkansas (county): Time series of the standardized indices $\mathrm{SI}_1(B)$, $\mathrm{SI}_6(B)$ and $\mathrm{SI}_{12}(B)$ (different time scales). The color-coding reflects the severity of wet-/dryness according to the different categories specified in Table 2.1. For better identification of dry/wet periods points at the top/bottom of the panels (colored accordingly) indicate points in time of wet/dry conditions.

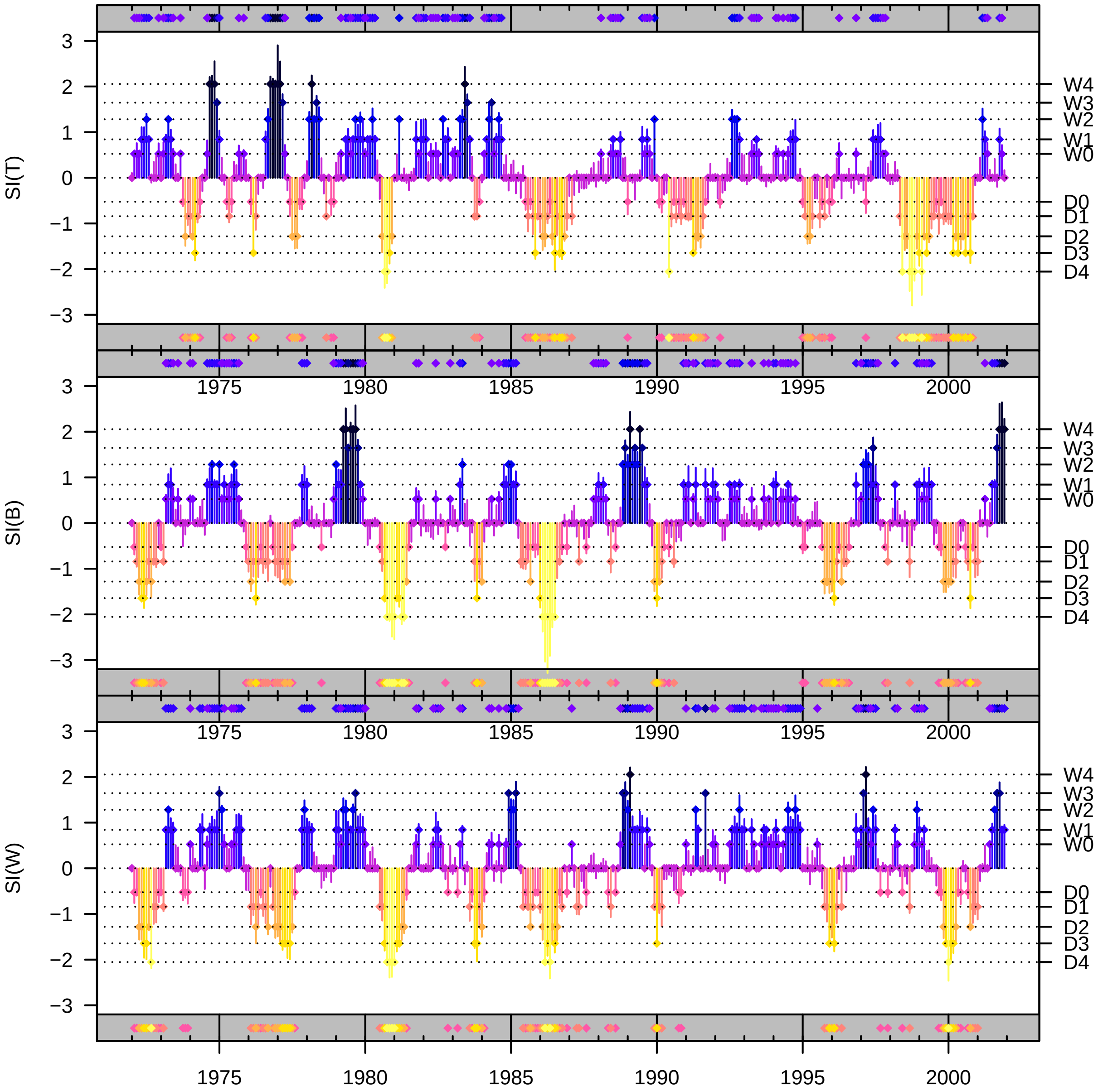


Figure 4.8: ERA-20C/Arkansas (county): Time series of the standardized indices $\mathrm{SI}_6(T)$, $\mathrm{SI}_6(B)$ and $\mathrm{SI}_6(W)$ (different variables). The color-coding reflects the severity of wet-/dryness according to the different categories specified in Table 2.1. For better identification of dry/wet periods points at the top/bottom of the panels (colored accordingly) indicate points in time of wet/dry conditions.

4.5 Multivariate standardized indices

Subsequently, we provide an extension of the methodology introduced in Section 4.4 to multivariate standardized drought indices (`MULTEX`), i.e. to drought indices that summarize the dryness information captured in multiple (possibly dependent) variables (see Section 2.2). This extension is based on vine copulas (see e.g. Aas et al., 2009) which we introduce briefly in Section 2.3.2. Here, vine copulas are utilized to flexibly model the dependence of several drought relevant variables. The dependence parameters are estimated after separate modelling of the univariate margins, using a semi-parametric estimation procedure (see Genest et al., 1995, and Section 2.3.1). Other copula based drought indices were introduced by Hao and AghaKouchak (2013) and Kao and Govindaraju (2010), as we have seen in Section 4.1.

4.5.1 Marginal models

As copulas allow separate modeling of margins and dependence structure (see Section 2.3.1), we first model the margins according to Sections 4.4.1–4.4.4 as in the univariate case. If required we change the sign of the input data (see Section 4.4.1), then we eliminate seasonality (see Section 4.4.2) and temporal dependencies (see Section 4.4.3) and estimate a non-parametric distribution of the residuals (see Section 4.4.4). This enables transformation to copula data and after that copula based dependence modeling.

4.5.2 Vine copula based dependence 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 $\boldsymbol{u}_j = (u_{j,t_k})_{k=1,\ldots,N}$, $j = 1, \ldots, d$, 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. For an introduction to vine copulas see Aas et al. (2009) and reference therein, as well as Section 2.3.2.

Vine tree structure selection The proposed multivariate indices utilize canonical vines (C-vines) to organize the pair-copula construction (vine copula, see Section 2.3.2) which is supposed to model the inter-variable dependencies of d drought relevant variables. The star like structured C-vine trees allow to order the involved variables by importance. Having selected a specific variable order, the corresponding C-vine tree structure tells us, which variable pairs have to be modeled explicitly in the next step, using (parametric) pair-copulas.

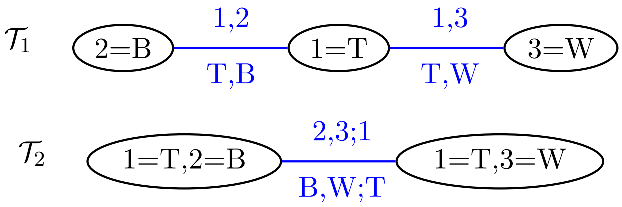


Figure 4.9: Selected vine tree structure for the three variables temperature (T), climatic water balance (B) and volumetric soil water (W).

To further illustrate the C-vine tree structure selection we return to our three-variate data example (i.e. $d = 3$). As in Section 4.4 we consider the three variables temperature (T), climatic water balance (B) and volumetric soil water (W). For these three variables we select the tree structure as given in Figure 4.9, i.e. we select the variable order (T, B, W). We decide for the given order, as temperature (1=T) has direct impact on the climatic water balance (2=B) and indirect impact on the water availability in the soil (3=W), as the soil water varies depending on (2=B) precipitation and evapotranspiration (climatic water balance). Such a derivation of the variable order based on physical arguments might not always be possible. Note however, that in low dimensions all possible variable orders can be compared in a validation study of the resulting multivariate drought indices to find the best model (see also Section 4.7). The above selected vine tree structure explicitly models the pair-copulas for the variable pairs (T, B), (T, W) and (B, W) given T, which can also be seen in Figure 4.9.

Pair-copula selection and parameter estimation In a next step we select pair-copula families for all variable pairs occuring in the above chosen C-vine tree structure. For the pair-copula family selection we can choose among a variety of bivariate copula families. In the following we restrict us to 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 (see Section 2.3.1). We select the pair-copulas sequentially (see Section 2.3.3) based on the Bayesian information criterion (2.6) starting in the first tree $\mathcal{T}_1$, which incorporates only unconditioned pairs. The parameters of each pair-copula are estimated along with the pair-copula family selection using maximum likelihood estimation. Before that we test for pair-wise independence (Genest and Favre, 2007), to see if an independence copula should be selected. For more details on the selection of pair-copula families and estimation of the corresponding parameters see Sections 2.3.1 and 2.3.3.

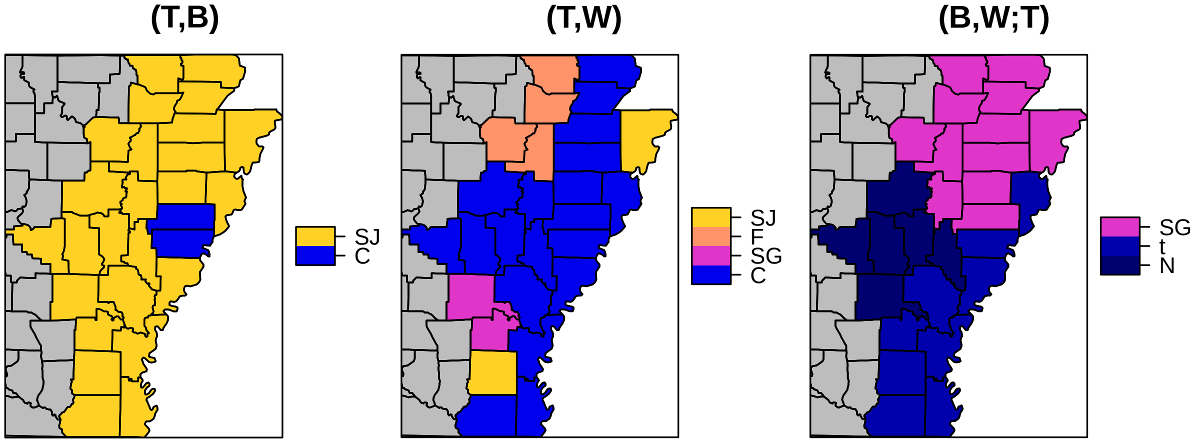


Figure 4.10: ERA-20C/Arkansas (27): Selected pair-copula families for the variable pairs (T, B), (T, W) and $(B, W; T)$ corresponding to the selected vine tree structure as specified in Figure 4.9.

Returning to our data example (see also Section 4.2) and the C-vine tree structure selected above (see Figure 4.9), we have to select pair-copula families for the pairs (T, B), (T, W) and $(B, W; T)$. Figures 4.10 and 4.11 show which pair-copulas/parameters were selected/estimated for the 27 counties of Arkansas under consideration. For the pair (T, B) for almost all counties the survival Joe copula was selected. This copula family exhibits lower tail-dependence which seems meaningful from a practical perspective, as high temperatures[1] negatively affect the climatic water balance B (higher evapotranspiration), but low temperatures do not necessarily mean that

[1] The temperatures were multiplied by -1 as low values are supposed to correspond to dry conditions.

B will be high (e.g. if it did not rain in a long time). For the pair (T, W) mostly Clayton, survival Joe and survival Gumbel copulas (all lower tail-dependent) were selected. For the conditional pair $(B, W; T)$ the map is divided into three parts. Whereas in the north survival Gumbel copulas are selected, Gaussian copulas (no tail-dependence) are favored in central Arkansas and Student-t copulas (lower and upper tail-dependence) in the south-east. From Figure 4.11 we see that the pair-copula parameters vary slightly across different counties. Whereas the unconditioned variable pairs (T, B) and (T, W) are rather weakly associated, a higher association is modeled for the conditioned pair $(B, W; T)$. For more details about the addressed copula families and their characteristics like tail-dependence we refer to Section 2.3.1 and Table 2.2.

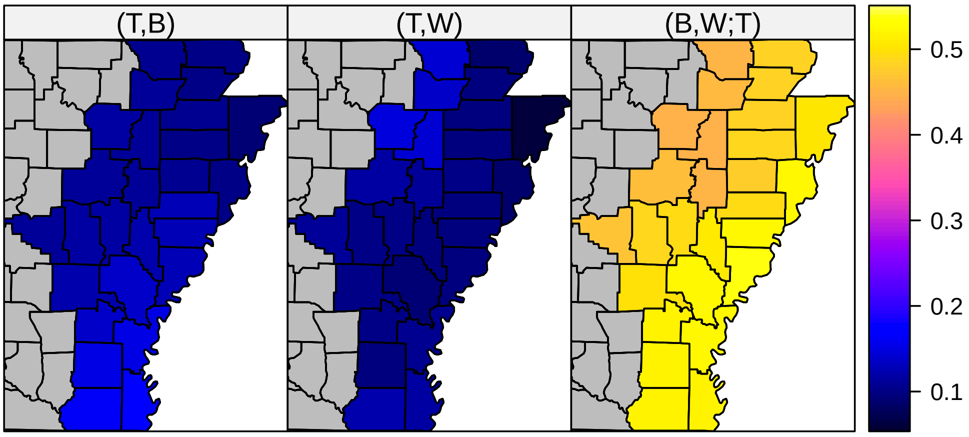


Figure 4.11: ERA-20C/Arkansas (27): Kendall's τ values corresponding to the estimated pair-copula parameters for the variable pairs (T, B), (T, W) and $(B, W; T)$ corresponding to the selected vine tree structure as specified in Figure 4.9.

4.5.3 Computation of multivariate indices

In the previous step we showed how we select/estimate a C-vine copula for the copula data $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)$. The copula data $\boldsymbol{u}$ was obtained as a result of the marginal modeling described in Section 4.4, where $\boldsymbol{u}_j = (u_{j,t_k})_{k=1,\ldots,N}$, $j = 1, \ldots, d$, and u_{j,t_k} is the copula data corresponding to variable j at time t_k. For the C-vine copula we selected/estimated

(a) a variable order and hence a C-vine tree structure,

(b) pair-copula families and

(c) corresponding pair-copula parameters.

Based on the C-vine copula, which is specified by the model components (a)–(c), we now eliminate the dependencies in the copula data $\boldsymbol{u}$, i.e. we transform to independent, uniform data on $[0, 1]^d$. This is achieved by a *Rosenblatt (1952) transformation*, a multivariate probability integral transform. For details on the computation of the Rosenblatt transform for vine copulas see Section 2.3.4 as well as Aas et al. (2009) and Schepsmeier (2015). We calculate the Rosenblatt transformation $\boldsymbol{v} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$ of $\boldsymbol{u}$ using Equation (2.16) as

$$\begin{aligned} v_{1,t_k} &:= u_{1,t_k}, \\ v_{2,t_k} &:= C_{2|1}(u_{2,t_k} | u_{1,t_k}), \end{aligned}$$

$$\vdots$$

$$v_{d,t_k} := C_{d|1,\dots,d-1}(u_{d,t_k}|u_{1,t_k},\dots,u_{d-1,t_k}), \quad k = 1,\dots,N,$$

where for all $j = 2,\dots,d$, $C_{j|1,\dots,j-1}$ is the conditional CDF for variable j given the variables $1,\dots,j-1$. Note that for canonical vine copulas the order of the variables $1,\dots,d$ is given by the selected order of root variables.

In the context of our application, the Rosenblatt transformation of d dependent drought relevant variables (which initially were transformed to copula data) yields independent information about dry/wet conditions captured in these variables. $\boldsymbol{v}_1$ incorporates the same information as an univariate drought index calculated according to Section 4.4 based on the first variable. $\boldsymbol{v}_j$, $j = 2,\dots,d$, provide information on dry/wet conditions identified by variable j, conditioned on a certain state of the previously considered variables $1,\dots,j-1$.

For our three dimensional example, where the variable order was selected as (T, B, W), we discuss the Rosenblatt transformation in the following. First, we compute $v_{T,t} = u_{T,t}$, which represents the dry-/wetness information captured in the variable temperature (T) for time point t. $v_{B,t} = C_{B|T}(u_{B,t}|u_{T,t})$ provides additional information on how extreme the observed climatic water balance (B) in time point t is, given the corresponding realization of T. The calculation of $v_{B,t}$ involves the pair-copula $C_{T,B}$. $v_{W,t} = C_{W|T,B}(u_{W,t}|u_{T,t}, u_{B,t})$ captures further drought information based on the variable volumetric soil water (W) given the realizations of T and B, which was not already captured in $v_{T,t}$ and $v_{B,t}$. Its calculation is a bit more involved. We calculate $v_{W,t} = C_{W|B;T}(C_{W|T}(u_{W,t}|u_{T,t})|C_{B|T}(u_{B,t}|u_{T,t}))$, based on the pair-copulas $C_{B,W;T}$, $C_{T,W}$ and $C_{T,B}$.

Subsequently we consider two different approaches to join multivariate drought information into one index. For comparison, we provide a third approach which ignores the inter-variable dependence and assumes multivariate normality.

Method $\mathcal{A}$ (aggregation) We define the *aggregated* standardized index with time scale ℓ as

$$\mathrm{SI}_\ell^{\mathcal{A}}(1,\dots,d)(t_k) := \frac{1}{\sqrt{\ell \cdot d}} \sum_{i=0}^{\ell-1} \sum_{j=1}^{d} \Phi^{-1}\left(v_{j,t_{k-i}}\right). \tag{4.7}$$

Hence, we first transform the elements of the multivariate probability integral transform to a standard normal distribution, using the inverse of the CDF Φ of a standard normal distribution. Then we aggregate the dry-/wetness information captured in these d different variables by summing them up. After aggregation for a certain time scale ℓ (compare Equation (4.6) in Section 4.4.5) we divide the sum by $\sqrt{\ell \cdot d}$, which yields the desired standardization.

Method $\mathcal{M}$ (multiplication) For the second approach we exploit that the multivariate dependence structure of $\boldsymbol{v} = (\boldsymbol{v}_1,\dots,\boldsymbol{v}_d)$ is represented by the independence copula $C(v_1,\dots,v_d) = \prod_{j=1}^d v_j$. Hence, we calculate the product $\widetilde{v}_{t_k} := \prod_{j=1}^d v_{j,t_k}$ for all $k = 1,\dots,N$. To obtain a standardized index we proceed as in the univariate case (see Sections 4.4.4 and 4.4.5). We calculate the rank transformation $\widetilde{u}_{t_k} := \mathrm{rank}\left(\widetilde{v}_{t_k}\right)/(N+1)$, $k = 1,\dots,N$, transform to the z-scale and aggregate for a selected time scale ℓ. Following this procedure, the *multiplicative* standardized index with time scale ℓ is defined as

$$\mathrm{SI}_\ell^{\mathcal{M}}(1,\dots,d)(t_k) := \frac{1}{\sqrt{\ell}} \sum_{i=0}^{\ell-1} \Phi^{-1}\left(\widetilde{u}_{t_{k-i}}\right). \tag{4.8}$$

Method $\mathcal{N}$ (normal) For the third approach which does not eliminate the inter-variable dependence, we consider $\boldsymbol{z}$ to be the marginal transformation of $\boldsymbol{u}$ to the z-scale. Assuming $\boldsymbol{z}$ to be a sample from a zero mean multivariate normal distribution, we can conclude that the linear transformation $\mathbf{1}'\boldsymbol{z}$, where $\mathbf{1} = (1, \ldots, 1)$, is a sample from a zero mean univariate normal distribution. We estimate the sample variance of $\mathbf{1}'\boldsymbol{z}$ by

$$S := \frac{1}{N-1} \sum_{k=1}^{N} \left(\sum_{j=1}^{d} \Phi^{-1}(u_{j,t_k}) \right)^2$$

and calculate a standardized index with time scale ℓ as

$$\mathrm{SI}_{\ell}^{\mathcal{N}}(1, \ldots, d)(t_k) := \frac{1}{\sqrt{\ell \cdot S}} \sum_{i=0}^{\ell-1} \sum_{j=1}^{d} \Phi^{-1}\left(u_{j,t_{k-i}}\right). \tag{4.9}$$

We return to our three-variate data example to illustrate the three different methods. Figure 4.12 depicts the multivariate drought indices $\mathrm{SI}_6^{\mathcal{N}}(T, B, W)$, $\mathrm{SI}_6^{\mathcal{A}}(T, B, W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T, B, W)$ for Arkansas county (1972–2001). Slight differences between the three approaches can be observed. Whereas the timing of dry/wet events matches, there are differences in the detected intensities.

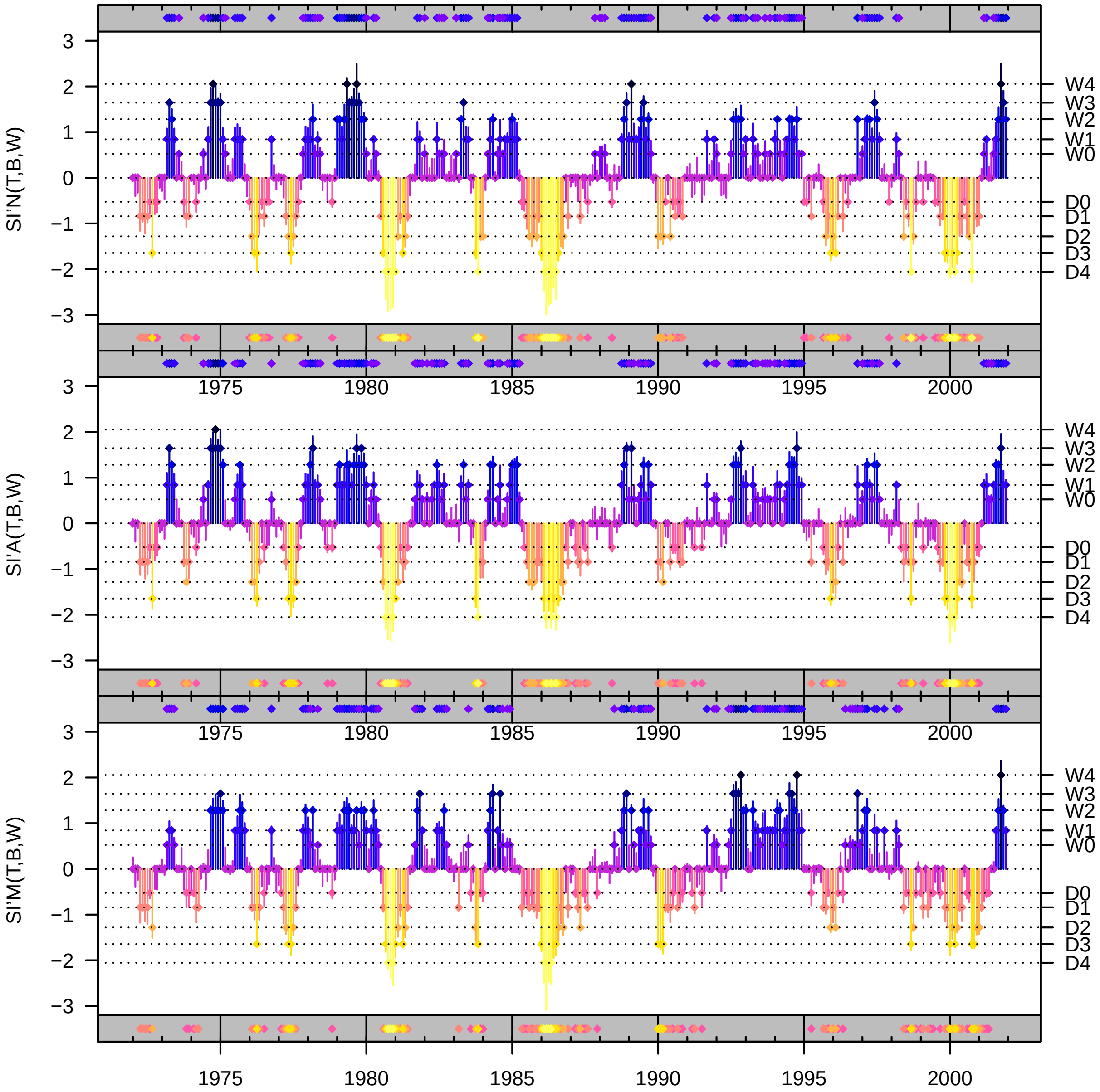


Figure 4.12: ERA-20C/Arkansas (county): Time series of the standardized indices $\mathrm{SI}_6^{\mathcal{N}}(T, B, W)$, $\mathrm{SI}_6^{\mathcal{A}}(T, B, W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T, B, W)$ (comparison of different methods). The color-coding reflects the severity of wet-/dryness according to the different categories specified in Table 2.1. 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.

4.6 Simulation study on the relevance of variable order

The above presented vine copula based multivariate drought indices (methods $\mathcal{A}$ (aggregation, see Equation (4.7)) and $\mathcal{M}$ (multiplication, see Equation (4.8))) require specification of a variable order. However, the practical implications of this variable order are not clear from a theoretical point of view. Hence, we investigate the effect of the order on the drought modeling capabilities of such indices in a simulation study. Our simulation study for three-variate indices considers different (artificial) scenarios. For the simulation of (artificial) "drought observations" we make the following assumptions:

(A1) 3 drought driving variables carry all information about dry-/wetness conditions.

(A2) These 3 variables are realizations of a specific three-variate C-vine copula (true vine copula) with a specific variable order.

(A3) A three-variate drought index based on the true vine copula from (A2) (and method $\mathcal{A}$ or $\mathcal{M}$, respectively) fully specifies the dry-/wetness conditions, i.e. we consider it as observed standardized drought index.

Hence, considering variables A, B, C, we simulate ($N = 12{,}000$ observations, i.e. thousand years of monthly observations) from a specific C-vine copula with variable order (A, B, C) (O0), and calculate the observed drought index based on the same (true) vine copula (A2). We do this separately for methods $\mathcal{A}$ and $\mathcal{M}$, respectively, as there can be only one true/observed drought index. As pair-copulas for the true vine copula we select a survival Joe copula for the pair $(1, 2)$, a Clayton copula for the pair $(1, 3)$ and a Gumbel copula for the pair $(2, 3; 1)$. The corresponding parameters are specified in terms of Kendall's τ (see Section 2.3.1). We consider the eight different scenarios

(T1) (0.1,0.1,0.1), (T2) (0.1,0.1,0.5), (T3) (0.1,0.5,0.1), (T4) (0.5,0.1,0.1),

(T5) (0.1,0.5,0.5), (T6) (0.5,0.1,0.5), (T7) (0.5,0.5,0.1), (T8) (0.5,0.5,0.5),

for the triplet ($\tau_{1,2}$, $\tau_{1,3}$, $\tau_{2,3;1}$). The Kendall's τ values 0.1 and 0.5 are motivated by the example in Section 4.5.2 (see Figure 4.11). To investigate the effect of variable order, we select/estimate vine copulas as discussed in Section 4.5.2, now based on the simulated variables and for all 6 possible variable orders

(O0) (A,B,C), (O1) (A,C,B), (O2) (B,A,C),

(O3) (B,C,A), (O4) (C,A,B), (O5) (C,B,A),

and calculate the corresponding multivariate indices for comparison with the "observed" index.

To validate the drought indices originating from orders (O0)–(O5) and seeing the effect of different magnitudes of pair-wise association (T1)–(T8) we consider four different validation metrics. The first metric is the *Pearson correlation (COR)* to assess the overall performance of a drought index in quantifying the dry-/wetness conditions given by the observed index. Further considered metrics are the *probability of detection (POD)*, the *false alarm ratio (FAR)* and the *critical success index (CSI)* (see Section 2.4 for a mathematical description of these metrics). These metrics allow to assess the drought detection capabilities of modeled (*detected*) drought indices in comparison to an *observed* (true) drought index.

POD provides the share of correctly identified droughts (i.e. a high value indicates good performance),

FAR gives the portion of incorrectly detected droughts (i.e. a low value indicates good performance) and

CSI is a measure which indicates good performance (high value) if observed drought is detected correctly and bad performance (low value) if observation and detection do not match in most cases.

In order to compute POD, FAR and CSI (according to Section 2.4) we have to specify a threshold ϑ for the drought indices, which decides if a drought occurred/was detected or not. Considering the quantile-based (standard normal distribution) classification of drought events given in Table 2.1, selection of one of the quantile-based thresholds given in Table 4.2 seems adequate. For the further analysis in this section, we select the threshold $\vartheta = -0.52$ which corresponds to the 0.3-quantile of the standard normal distribution.

Table 4.2: Candidate thresholds ϑ for drought occurrence/detection, based on certain quantiles of a standard normal distribution. The corresponding dryness (D) categories for standardized drought indices (SI) are stated along with the thresholds. The function Φ^{-1} denotes the quantile function of the standard normal distribution.

ϑ	quantile	categories	(in words)
$\Phi^{-1}(0.30) = -0.52$	0.30	D0–D4	abnormally dry or worse
$\Phi^{-1}(0.20) = -0.84$	0.20	D1–D4	moderately dry or worse
$\Phi^{-1}(0.10) = -1.28$	0.10	D2–D4	severely dry or worse
$\Phi^{-1}(0.05) = -1.64$	0.05	D3–D4	extremely dry or worse
$\Phi^{-1}(0.02) = -2.05$	0.02	D4	exceptionally dry

Table 4.3 shows the results of the simulation study (using the 0.3-quantile of the standard normal distribution as a threshold). Subsequently we summarize the results. Generally, we observe that the variable order matters. For some ($< 10\%$) order-Kendall's τ combinations (O-T combinations) we observe COR < 0.5, POD < 0.5 and FAR > 0.5. However, in the majority ($> 60\%$) of all considered O-T combinations we observe COR ≥ 0.8, POD ≥ 0.7 and FAR ≤ 0.3. We observe that for high association (T8) drought indices for both methods perform considerably worse for misspecified variable order compared to low association (T1), according to all measures under consideration. Method $\mathcal{M}$ appears to yield more stable results. Moreover, we find that (with a few exceptions) the orders (O1) and (O2) which differ from (O0) comparatively less provide better indices compared to (O3)–(O5). Comparing the different specifications (T1)–(T8) in more detail yields, that order misspecifications for pairs with small association (0.1) have less impact on index performance compared to those with high association (0.5). We conclude that it is not meaningful to combine only highly associated variables into a multivariate drought index, as in that case additional variables do not carry any additional information but rather add noise to the index.

Table 4.3: Simulation study results (validation against "observed" drought index): Pearson correlations (COR), probability of detection (POD), false alarm ratio (FAR) and critical success index (CSI) for drought indices calculated based on different methods ($\mathcal{A}$ and $\mathcal{M}$), variable orders (O0–O5) and different pair-wise magnitudes of variable association (T1–T8). POD, FAR and CSI are computed using a threshold $\vartheta = -0.52$, which distinguishes if abnormally dry or even worse conditions (D1–D4) occured/were detected or not.

	$\mathcal{A}$ (aggregation)						$\mathcal{M}$ (multiplication)					
	(O0)	(O1)	(O2)	(O3)	(O4)	(O5)	(O0)	(O1)	(O2)	(O3)	(O4)	(O5)
COR												
(T1)	1.00	0.99	0.99	0.97	0.97	0.96	1.00	0.99	0.99	0.97	0.97	0.96
(T2)	1.00	0.80	0.99	0.98	0.76	0.78	1.00	0.80	0.99	0.98	0.76	0.78
(T3)	1.00	0.99	0.98	0.77	0.74	0.74	1.00	0.99	0.98	0.78	0.75	0.75
(T4)	1.00	0.99	0.80	0.80	0.97	0.80	1.00	0.99	0.80	0.80	0.96	0.80
(T5)	1.00	0.80	0.98	0.73	0.48	0.53	1.00	0.80	0.98	0.75	0.52	0.51
(T6)	1.00	0.80	0.80	0.92	0.77	0.86	1.00	0.80	0.80	0.89	0.74	0.83
(T7)	1.00	0.99	0.78	0.60	0.73	0.56	1.00	0.99	0.79	0.61	0.74	0.58
(T8)	1.00	0.80	0.78	0.68	0.45	0.44	1.00	0.80	0.80	0.71	0.50	0.50
POD												
(T1)	1.00	0.95	0.92	0.88	0.89	0.86	0.99	0.96	0.92	0.88	0.89	0.87
(T2)	0.99	0.72	0.92	0.90	0.69	0.70	0.99	0.78	0.92	0.90	0.74	0.76
(T3)	0.99	0.95	0.92	0.65	0.62	0.62	0.99	0.96	0.90	0.66	0.65	0.65
(T4)	0.99	0.95	0.66	0.66	0.89	0.67	0.99	0.96	0.68	0.68	0.88	0.68
(T5)	0.99	0.73	0.90	0.64	0.46	0.53	0.99	0.78	0.90	0.65	0.54	0.58
(T6)	0.99	0.73	0.68	0.81	0.70	0.76	0.99	0.78	0.69	0.80	0.73	0.74
(T7)	0.99	0.95	0.65	0.52	0.61	0.48	0.99	0.96	0.66	0.55	0.62	0.53
(T8)	0.99	0.73	0.66	0.58	0.45	0.43	0.99	0.78	0.68	0.63	0.51	0.50
FAR												
(T1)	0.01	0.05	0.08	0.12	0.11	0.13	0.01	0.04	0.08	0.12	0.11	0.13
(T2)	0.00	0.27	0.07	0.09	0.30	0.29	0.01	0.22	0.08	0.10	0.26	0.24
(T3)	0.01	0.05	0.10	0.35	0.37	0.37	0.01	0.04	0.10	0.34	0.35	0.35
(T4)	0.01	0.05	0.33	0.33	0.10	0.33	0.01	0.04	0.32	0.32	0.12	0.32
(T5)	0.01	0.27	0.09	0.37	0.53	0.46	0.01	0.22	0.10	0.35	0.46	0.42
(T6)	0.00	0.27	0.33	0.19	0.31	0.26	0.01	0.22	0.31	0.20	0.27	0.26
(T7)	0.01	0.05	0.35	0.49	0.40	0.51	0.01	0.04	0.34	0.45	0.38	0.47
(T8)	0.01	0.27	0.34	0.42	0.55	0.57	0.01	0.22	0.32	0.37	0.49	0.50
CSI												
(T1)	0.99	0.90	0.85	0.78	0.80	0.76	0.99	0.92	0.85	0.79	0.81	0.77
(T2)	0.99	0.57	0.86	0.83	0.53	0.55	0.99	0.64	0.85	0.83	0.59	0.61
(T3)	0.98	0.91	0.84	0.48	0.45	0.45	0.98	0.92	0.82	0.50	0.48	0.48
(T4)	0.99	0.90	0.50	0.50	0.81	0.51	0.99	0.92	0.52	0.52	0.79	0.52
(T5)	0.99	0.57	0.83	0.46	0.30	0.37	0.98	0.64	0.81	0.48	0.37	0.40
(T6)	0.99	0.57	0.51	0.68	0.53	0.60	0.99	0.64	0.53	0.66	0.58	0.58
(T7)	0.98	0.90	0.48	0.35	0.43	0.32	0.99	0.92	0.49	0.38	0.45	0.36
(T8)	0.98	0.57	0.49	0.41	0.29	0.27	0.98	0.64	0.52	0.46	0.34	0.33

4.7 Application, validation and comparison to established standardized drought indices

In this section we resume the data example considered throughout Sections 4.3–4.5. We provide an application and illustration of the novel methodology in the context of agro-meteorological drought detection. We compute three-variate standardized (agro-meteorological) drought indices and further uni- and bivariate standardized drought indices based on other established approaches (see references given below) for comparison. Motivated by the application at hand (validation data), we consider the variables temperature (T), climatic water balance (B) and volumetric soil water (W) for the computation of agro-meteorological drought indices, as they are expected to have an influence on crop yield and mirror dry/wet conditions. We validate the drought indices using the Arkansas soybean yield data introduced in Section 3.2.

Selection of drought indices for comparison and validation based on soybean yield For our application we compare the eight drought indices

$\mathrm{SPI}_6(P)$	Standardized Precipitation Index, McKee et al. (1993); Edwards and McKee (1997),
$\mathrm{SPEI}_6(B)$	Standardized Precipitation Evapotranspiration Index, Vicente-Serrano et al. (2010),
$\mathrm{SI}_6(B)$	Standardized Index, univariate, see Equation (4.6) in Section 4.4,
$\mathrm{SDAT}_6(B)$	Standardized Drought Analysis Toolbox (SDAT), univariate, Farahmand and AghaKouchak (2015),
$\mathrm{SDAT}_6(B,W)$	SDAT/Multivariate Standardized Drought Index (MSDI), bivariate, Hao and AghaKouchak (2014),
$\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$	Standardized Index (method $\mathcal{N}$/normal), trivariate, see Equation (4.9) in Section 4.5,
$\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$	Standardized Index (method $\mathcal{A}$/aggregation), trivariate, see Equation (4.7) in Section 4.5,
$\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$	Standardized Index (method $\mathcal{M}$/multiplication), trivariate, see Equation (4.8) in Section 4.5,

all for time scale 6, where the variables (and their order) used for the computation are stated in parentheses. We reason our decision for these eight indices as follows. Among the univariate indices we consider the well-established SPI and SPEI, to see if replacing precipitation (P) by the climatic water balance (B) yields improvements for drought quantification. We compare to two further B based univariate indices, the $\mathrm{SI}_6(B)$ based on our novel methodology and $\mathrm{SDAT}_6(B)$ (see Farahmand and AghaKouchak, 2015) to see if these more recent approaches yield any improvements for drought event identification. As B and W quantify water availability, thus appear to be more important than T and SDAT/MSDI is restricted to the bivariate case, we further consider $\mathrm{SDAT}_6(B,W)$. Finally, we are interested in validating the novel multivariate indices and identifying differences between $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$, $\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$. The variable order (T,B,W) was selected based on a comparative validation of all 6 possible variable orders. For the results of this comparison see the discussion and Table 4.4 below.

The selection of time scale 6 is again motivated by the nature of the validation data. As we have only yearly yield observations and drought can affect the yield only during the growing season we make the following assumption (based on information about the growing season from the Arkansas Soybean Promotion Board, 2011). We assume that the growing season for soybeans lasts for six months and approximately ends at the end of October. Hence, for the validation it seems meaningful to standardize the soybean yield (i.e. transform it to a standard normal distribution) and compare it to the October values of standardized drought indices based on time scale 6, as these values are meant to reflect the drought conditions during the respective growing seasons.

Illustration of drought indices and comparison with standardized soybean yield Figure 4.13 compares the standardized drought indices under consideration with the standardized soybean yield for Arkansas county (1972–2001). Comparison of the univariate (left panel) and multivariate indices (right panel) shows a similar temporal evolution of the different indices and in many cases it is difficult to visually detect differences. However, especially in the extremes differences become clear. $\mathrm{SDAT}_6(B)$ appears to be more conservative compared to the other univariate indices. It never detects an exceptional drought (i.e. values below the 0.02-quantile of the standard normal distribution). The smallest possible value of $\mathrm{SDAT}_6(B)$ corresponds to the $1/(30+1)$-quantile of the standard normal distribution. That is due to the length of the data record (30 years) and the month-wise transformation to the z-scale. $\mathrm{SDAT}_6(B, W)$ "prefers" to indicate dry over wet conditions (see upper plot in right panel). It is not standardized, as already discussed in Section 4.1 for the multivariate SDAT indices in general. Comparison with the standardized soybean yield (grey points) shows that the drought indices are mostly able to reflect the dry-/wetness information captured in the validation data. For some time instances the multivariate indices capture the dry-/wetness conditions better (see e.g. 1994, 1998). To clearly differentiate between the different indices, further numerical (non-visual) validation (including all counties of Arkansas where validation data is available) is required.

Numerical validation of selected drought indices For a numerical comparison and validation of the different indices, we now consider the four metrics COR, POD (2.17), FAR (2.18) and CSI (2.19) already applied in our simulation study in Section 4.6 (see also Section 2.4). We calculate these metrics using only the October values of the indices (as stated above) separately for each of the 27 counties under consideration. As a threshold ϑ, to distinguish if drought occurred/was detected or not, we choose the 0.1-quantile of the standard normal distribution, i.e. $\vartheta = \Phi^{-1}(0.1) = -1.28$ (cp. Table 4.2). This threshold corresponds to severe or even worse drought conditions (D2–D4). Hence, we validate the capability of the indices under consideration to reliably identify the 10% worst drought conditions.

In order to choose the best variable order for our novel multivariate drought indices (methods $\mathcal{M}$ and $\mathcal{A}$), Table 4.4 provides a comparison of the corresponding indices for all 6 possible variable orders and shows that the selected variable order (T, B, W) yields the best results in terms of Pearson correlations (COR) and critical success index (CSI).

Table 4.5 summarizes the validation results for the eight selected indices. The table provides averages of the four considered metrics over all 27 counties under consideration. Maps showing the results county-wise are provided in Figures 4.14–4.17. The results for COR show that inclusion of more than one variable yields improvements. Comparison of the univariate indices yields clear benefits of using the climatic water balance (B) instead of precipitation (P). While $\mathrm{SPEI}_6(B)$ and $\mathrm{SDAT}_6(B)$ show similar performance, $\mathrm{SI}_6(B)$ yields better results in terms of POD, FAR and CSI. The novel three-variate indices seem to detect D2–D4 droughts better compared to the other indices (POD), where $\mathrm{SI}_6^{\mathcal{N}}(T, B, W)$ performs best. Only $\mathrm{SI}_6^{\mathcal{M}}(T, B, W)$ yields a lower average FAR than the univariate $\mathrm{SI}_6(B)$. CSI ranks all four indices based on our novel methodology better than the remaining four. Comparing only our three multivariate approaches (based on POD, FAR and CSI) $\mathrm{SI}^{\mathcal{M}}$ is preferred over $\mathrm{SI}^{\mathcal{A}}$ and $\mathrm{SI}^{\mathcal{A}}$ over $\mathrm{SI}^{\mathcal{N}}$.

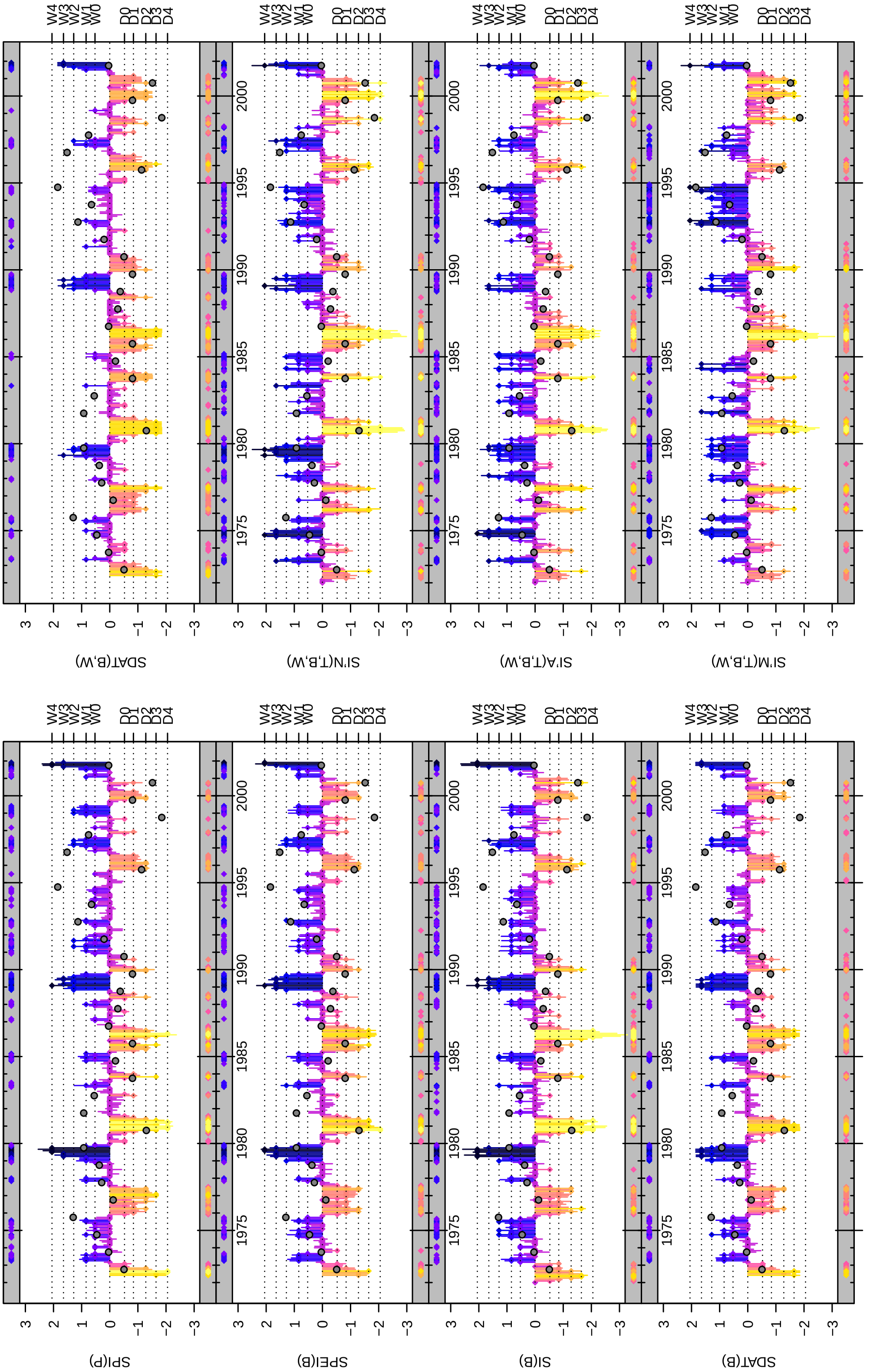


Figure 4.13: ERA-20C/Arkansas (county): Comparison of $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$ and $\mathrm{SDAT}_6(B)$ time series (left panel), as well as $\mathrm{SDAT}_6(B, W)$, $\mathrm{SI}_6^{\mathcal{N}}(T, B, W)$, $\mathrm{SI}_6^{\mathcal{A}}(T, B, W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T, B, W)$ time series (right panel), with standardized soybean yield (points).

Table 4.4: ERA-20C/Arkansas (27): Standardized soybean yield based validation of the drought indices $\mathrm{SI}_6^{\mathcal{A}}$ and $\mathrm{SI}_6^{\mathcal{M}}$ for different variable orders. Only the October values of the indices (time scale 6) are used for the validation as they mirror the dry-/wetness conditions in the corresponding growing seasons. The provided values are averages of the Pearson correlations (COR), probability of detection (POD), false alarm ratio (FAR) and critical success index (CSI) calculated separately for each of the 27 counties of Arkansas under consideration. For the calculation of POD, FAR and CSI we use $\vartheta = \Phi^{-1}(0.1) = -1.28$ as a threshold to distinguish if a severe or even worse drought (D2–D4) occurred/was detected or not.

		Variable order					
Index	Metric	(T, B, W)	(T, W, B)	(B, T, W)	(B, W, T)	(W, T, B)	(W, B, T)
$\mathrm{SI}_6^{\mathcal{A}}$	COR	0.48	0.43	0.46	0.46	0.44	0.44
	POD	0.70	0.75	0.65	0.65	0.81	0.80
	FAR	0.47	0.52	0.53	0.53	0.56	0.56
	CSI	0.44	0.42	0.37	0.37	0.40	0.40
$\mathrm{SI}_6^{\mathcal{M}}$	COR	0.50	0.43	0.48	0.48	0.43	0.43
	POD	0.69	0.56	0.68	0.68	0.56	0.56
	FAR	0.38	0.51	0.42	0.42	0.58	0.59
	CSI	0.49	0.38	0.45	0.45	0.33	0.33

Table 4.5: ERA-20C/Arkansas (27): Standardized soybean yield based validation of the drought indices $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$, $\mathrm{SDAT}_6(B)$, $\mathrm{SDAT}_6(B, W)$, $\mathrm{SI}_6^{\mathcal{N}}(T, B, W)$, $\mathrm{SI}_6^{\mathcal{A}}(T, B, W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T, B, W)$. Only the October values of the indices (time scale 6) are used for the validation as they mirror the dry-/wetness conditions in the corresponding growing seasons. The provided values are averages of the Pearson correlations (COR), probability of detection (POD), false alarm ratio (FAR) and critical success index (CSI) calculated separately for each of the 27 counties of Arkansas under consideration. For the calculation of POD, FAR and CSI we use $\vartheta = \Phi^{-1}(0.1) = -1.28$ as a threshold to distinguish if a severe or even worse drought (D2–D4) occurred/was detected or not.

		COR	POD	FAR	CSI
univariate	$\mathrm{SPI}_6(P)$	0.38	0.30	0.65	0.20
	$\mathrm{SPEI}_6(B)$	0.46	0.43	0.48	0.31
	$\mathrm{SI}_6(B)$	0.46	0.60	0.45	0.41
	$\mathrm{SDAT}_6(B)$	0.46	0.44	0.58	0.30
bivariate	$\mathrm{SDAT}_6(B, W)$	0.49	0.65	0.58	0.35
trivariate	$\mathrm{SI}_6^{\mathcal{N}}(T, B, W)$	0.49	0.81	0.55	0.41
	$\mathrm{SI}_6^{\mathcal{A}}(T, B, W)$	0.48	0.70	0.47	0.44
	$\mathrm{SI}_6^{\mathcal{M}}(T, B, W)$	0.50	0.69	0.38	0.49

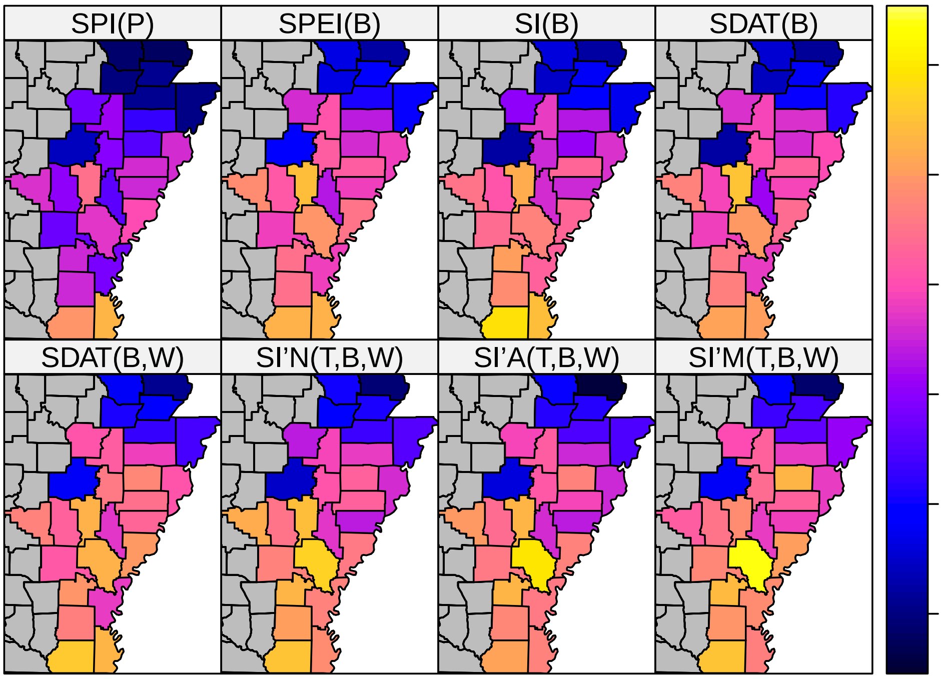


Figure 4.14: ERA-20C/Arkansas (27): Pearson correlations of the October values of $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$, $\mathrm{SDAT}_6(B)$, $\mathrm{SDAT}_6(B,W)$, $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$, $\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$ with the standardized soybean yield.

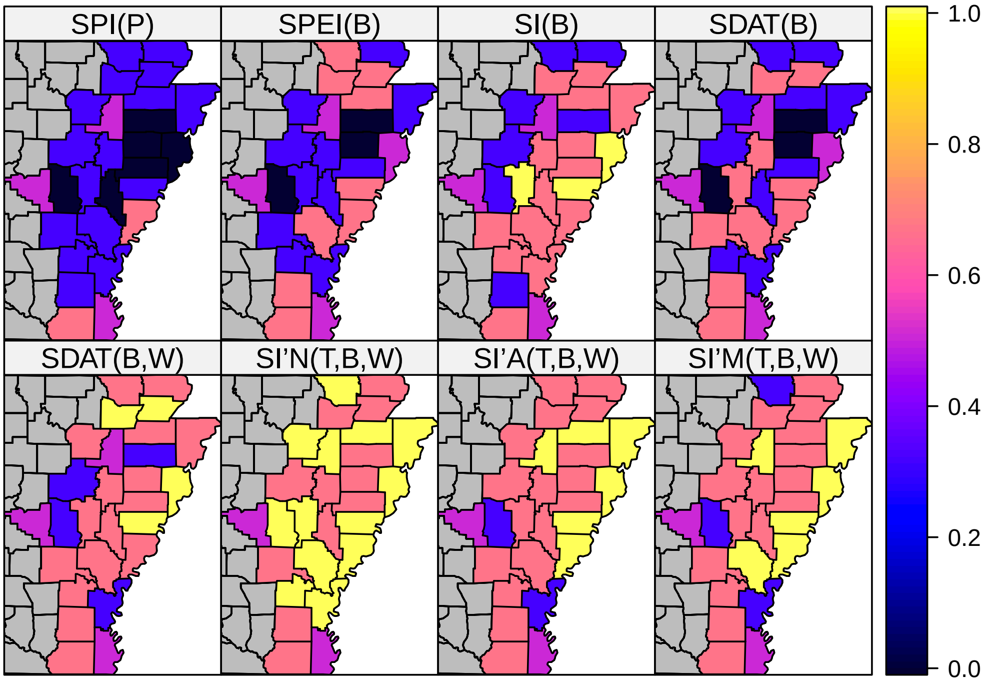


Figure 4.15: ERA-20C/Arkansas (27): Probability of detection (POD) of a severe or even worse drought (D2–D4) for $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$, $\mathrm{SDAT}_6(B)$, $\mathrm{SDAT}_6(B,W)$, $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$, $\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$.

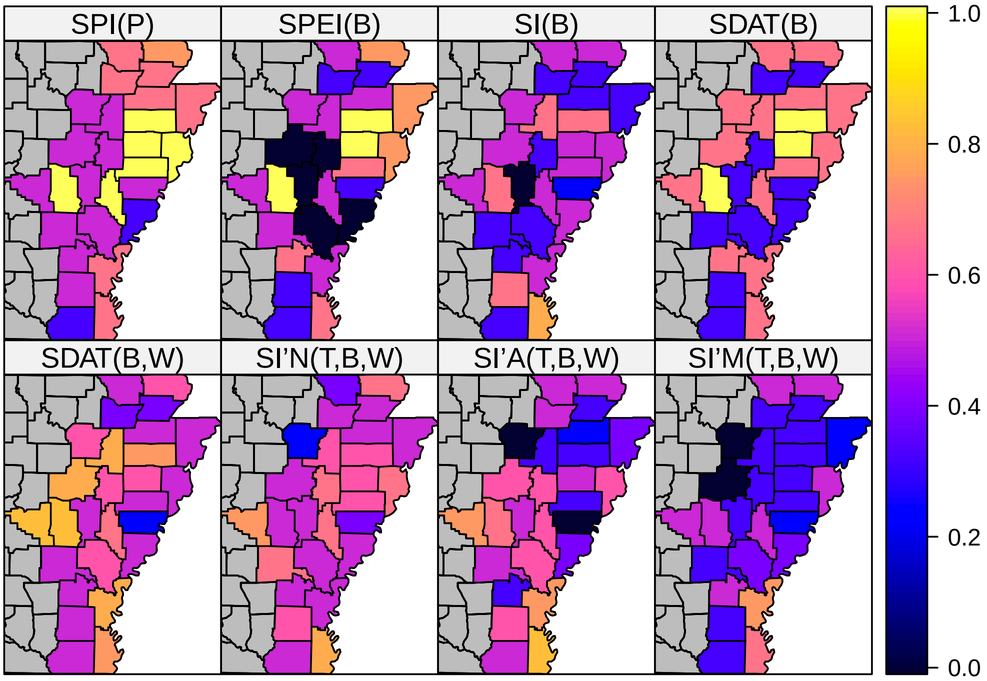


Figure 4.16: ERA-20C/Arkansas (27): False alarm ratio (FAR) for a severe or even worse drought (D2–D4) for $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$, $\mathrm{SDAT}_6(B)$, $\mathrm{SDAT}_6(B,W)$, $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$, $\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$.

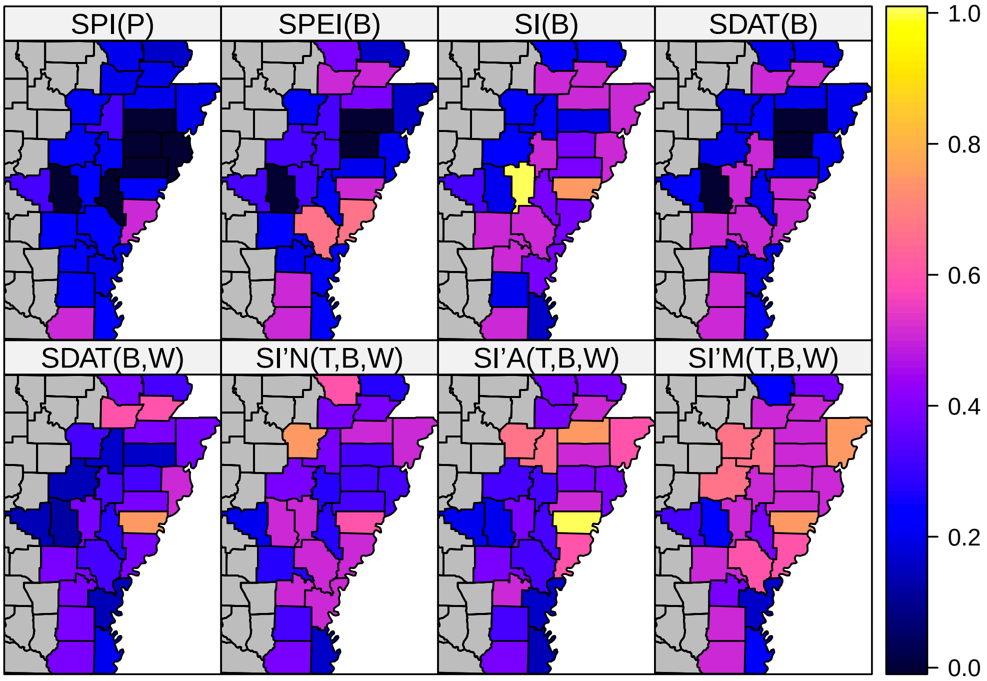


Figure 4.17: ERA-20C/Arkansas (27): Critical success index (CSI) of a severe or even worse drought (D2–D4) for $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$, $\mathrm{SDAT}_6(B)$, $\mathrm{SDAT}_6(B,W)$, $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$, $\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$.

Illustration of drought periods Considering once more severe or worse (D2–D4) drought conditions (cp. Tables 2.1 and 4.2), we are interested to see if the different drought indices identify different onset and termination of droughts or even different drought periods. For that purpose Figure 4.18 provides time series for each index (different colors) showing if a drought was detected (square) or not (or if the index could not be computed, cross). Most of the time the different indices detected the same drought events. However, we observe differences in drought onset (e.g. 1977) and termination (e.g. 2000). During 1998 the multivariate indices identify drought events which were not identified by the univariate indices. The numbers at the bottom right indicate the percentage of months where a D2–D4 drought was identified by the corresponding indices. Per definition this number should be 0.1, as we consider D2–D4 droughts (cp. Table 4.2). We see that $\mathrm{SDAT}_6(B,W)$ and $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$ have a tendency to decide too often in favor of drought.

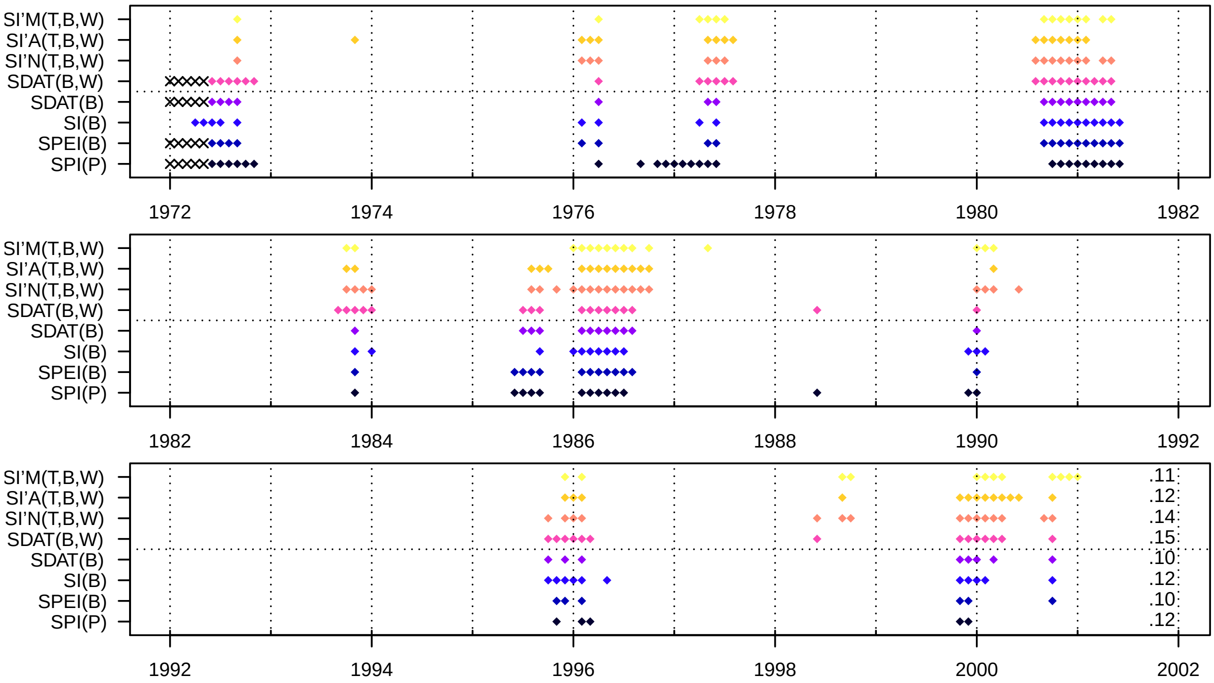


Figure 4.18: ERA-20C/Arkansas (county): Time series (1972–2001) indicating when severe or worse drought conditions (D2–D4) were detected based on $\mathrm{SPI}_6(P)$, $\mathrm{SPEI}_6(B)$, $\mathrm{SI}_6(B)$, $\mathrm{SDAT}_6(B)$, $\mathrm{SDAT}_6(B,W)$, $\mathrm{SI}_6^{\mathcal{N}}(T,B,W)$, $\mathrm{SI}_6^{\mathcal{A}}(T,B,W)$ and $\mathrm{SI}_6^{\mathcal{M}}(T,B,W)$.

4.8 Conclusions and outlook

We propose novel, vine copula based methodology for the computation of multivariate drought indices. This approach involves several well reasoned modeling steps (separated into univariate marginal and multivariate dependence modeling) which are summarized in Figure 4.1. Comparison to existing/established drought indices based on theoretical arguments and application of the presented methodology for the development and validation of an agro-meteorological drought index based on a three-variate data example show the benefits of the novel class of (multivariate) indices. We summarize the results in the following.

Advances in drought modeling (theoretical point of view)

1. More than two different variables can be combined in a multivariate drought index which accounts for inter-variable dependence (`MULTEX`).
2. Inter-variable dependence is accounted for/modeled in a very flexible fashion (based on vine copulas).
3. Multivariate drought indices (see also JDI and MSDI) can account for different drought types (e.g. meteorological and hydrological drought) at the same time (`MULTEX`).
4. Flexible modeling approaches (see also SDAT) allow application specific development of new drought indices based on user-selected variables (`ARBVAR`, `SEASON`, `TIMDEP`, `NPDIST`).
5. Compared to fitting 12 month-wise probability distributions as in SPI, SPEI, MSDI and SDAT, consideration (re-composition) of the full monthly time series provides a 12 times bigger sample size. Hence, stable results (drought indices) can be obtained for much shorter observation periods (like for instance 10 years) and the severity of droughts occurring in different months (cp. SDAT) can be distinguished unambiguously (`SMALLS`).
6. Proper standardization of multivariate indices (cp. MSDI) allows comparison over space, time and with other indices (`STCOMP`).
7. Aggregation at different time scales at the end of the modeling procedure (instead of the beginning, cp. SPI, etc.) avoids introduction of additional serial dependence and allows more efficient computation of a particular drought index on different time scales (`TSCALE`).

Application based comparison and validation The soybean yield based validation of the novel methodology and comparison to other drought indices in Section 4.7 yielded the following results:

1. The novel methodology is able to identify dry/wet conditions.
2. Our novel univariate approach outperforms established univariate approaches.
3. Compared to univariate indices, multivariate indices improve drought quantification.
4. For our application (agro-meteorological index), the validation indicates improvements of the novel multivariate approaches over other approaches in terms of probability of detection (POD) of drought events. The multiplication based method ($\mathrm{SI}^{\mathcal{M}}$) positively stands out in terms of false alarms. Taking drought detection capabilities and the chance of a false alarm into consideration simultaneously (CSI) the validation suggests to rank the multivariate indices under consideration as 1. $\mathrm{SI}^{\mathcal{M}}$, 2. $\mathrm{SI}^{\mathcal{A}}$, 3. $\mathrm{SI}^{\mathcal{N}}$, 4. SDAT, where $\mathrm{SI}^{\mathcal{M}}$ is the best.

Recommendations for the development of drought indices based on the presented methodology We do not recommend to use method $\mathcal{N}$ (which we introduced only for the purpose of comparison) as it might lead to an increased number of false alarms due to double accounting for the "same" drought information (no consideration of dependencies between variables). Based on only one specific application, we can not give any clear recommendation for either method $\mathcal{M}$ or $\mathcal{A}$. For the agro-meteorological drought index presented in Section 4.7 however we favor method $\mathcal{M}$. Further, we emphasize that the choice of variable order is crucial as it was detected in a simulation study in Section 4.6. For the development of novel drought indices based on the presented methodology we recommend to choose the order of variables with care. It should be validated against appropriate observations/data. If no secondary data is available for such a validation, however the data records used to calculate the drought indices of interest are long, the validation can be conducted based on a share of the data (i.e. the data can be split into training and validation data). For three-variate (four-variate) indices such validation allows to compare all 6 (24) possible variable orders and hence selection of the most appropriate one. Moreover, the simulation study advises to consider drought relevant variables which are not highly dependent, i.e. provide different drought information (cp. AghaKouchak et al., 2015).

Further applications of the presented methodology Our novel approach enables tailoring of drought indices to specific applications. The review paper of AghaKouchak et al. (2015) outlines several avenues for such indices, where combination of drought information from different sources is promising. Such sources can be different satellite data sets (e.g. soil moisture, land surface temperature, relative humidity, etc.) or remote sensing products which quantify vegetation conditions. Also, existing climate-based drought indices can be combined with other variables in a multivariate index. For specific applications see AghaKouchak et al. (2015) and reference therein.

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.

5 Proper evaluation of non-stationary time series models

5.1 Introduction

Many different disciplines like for instance meteorology, climatology, finance and economics provide forecasts of future states of a specific variable of interest/phenomenon as well as corresponding methodology. It has been argued extensively (see e.g. Gneiting and Raftery, 2007) that such forecasts should take the form of probability distributions, since each forecast is subject to uncertainty. In meteorology and climatology forecasts are often based on simulation models. The simulations are referred to as *Numerical Weather Prediction (NWP)*, which are often based on *Global Climate Models (GCMs)* or *Regional Climate Models (RCMs)*. Usually several simulation runs from these models with different initial conditions/perturbations are performed, which result in an *ensemble* of forecasts. These ensembles can be interpreted as realizations of a predictive probability distribution. However, the ensemble size is restricted due to limitations with respect to available resources to perform these computationally demanding numerical simulations. Hence, these (small) ensembles not necessarily are a good representation of the full predictive probability distribution and in practice a (weighted) average of such ensembles is considered as a deterministic forecast. Thus, a decision maker often faces the situation where the performance of several different models can only be compared based on time series output of these models. In the context of climate change and a natural, deterministic, periodic oscillation of weather/climate, these time series are usually non-stationary, which should be taken into consideration for decision making respectively model evaluation.

Hyndman and Koehler (2006) give a comprehensive review of different *traditional measures* which can be used for the evaluation of such models. For exact definitions of these measures see also Section 2.6. Besides the frequently used scale-dependent measures Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MdAE), Hyndman and Koehler (2006) discuss the flaws of measures based on percentage/relative errors as well as relative measures, and reason why measures based on scaled errors should be used preferably. However, all of these approaches do not account for higher order structures and properties of the forecast phenomenon, as they only allow for a point-wise comparison.

Thorarinsdottir et al. (2013) argue that

> Any (. . .) evaluation procedure ought to provide a quantitative assessment of the compatibility of the simulation model, and the real world phenomena it is meant to represent, in a manner that encourages careful assessment and integrity.

Such methods (see also Section 2.5) are given by *proper scoring rules* (see Gneiting and Raftery, 2007) and *proper divergence functions* (see Thorarinsdottir et al., 2013), where the concept of *propriety* formalizes what is meant by "careful assessment and integrity". Whereas proper scoring rules are used to compare a probability distribution to a (deterministic/point) observation, proper divergence functions are used to compare two distributions, one for the model and one for the observed phenomenon.

Coming back to the setting where we want to compare the performance of several models based on one realization/simulation of a time series for each model and a time series of realizations from the modeled phenomenon, we do not have a probability distribution for each time instance but only a single value. If the time series were stationary (that is, free of trends and periodic behavior, homogeneous with respect to variability, etc.), then we could consider the whole time series as a sample of a distribution and use proper scoring rules to assess the model accuracy in each time instance for which we have observed the phenomenon of interest, or proper divergence functions to get an overall picture of the model accuracy. However, in practice one is more concerned with the non-stationary setting, where forecasting is more challenging. Think of predicting the climate in 100 years from now under the presence of climate change. In this chapter (Section 5.3) we provide methodology for the proper evaluation of models/forecasts in a non-stationary context. The novel methodology assumes stationarity on small (moving) windows of the time series and applies proper scoring rules and divergence functions based on these windows, resulting in time series of "moving" scores and divergences. These moving scores/divergences enable model evaluation in the presence of non-stationarity and allow to assess the model capabilities/credibility over time. To learn how different approaches to calculate these moving scores and divergences perform in different settings we conduct simulation studies (see Sections 5.4–5.8) covering a wide range of cases which are relevant in practice. As a case study (Section 5.9) we evaluate the daily mean surface temperature output of four Regional Climate Models (see the *ENSEMBLES project*: van der Linden and Mitchell, 2009) on a fine resolution grid covering Europe. Moreover, in the subsequent section (Section 5.2), we review some evaluation studies of Regional Climate Models, with a focus on the ENSEMBLES project.

5.2 Review: Evaluation of Regional Climate Models

The *ENSEMBLES project* (van der Linden and Mitchell, 2009) is an example where climate research institutes from all over Europe compiled a big data set of simulations from their individual Regional Climate Models (RCMs). One goal of this project was it to evaluate and compare these different models (see e.g. Kjellström et al., 2010; Lorenz and Jacob, 2010). Further evaluation studies of these models for Scandinavia were conducted by Landgren et al. (2014). Another extensive evaluation of regional climate models for a region in Canada was performed by Eum et al. (2012). In this section we are interested to see, which evaluation techniques were applied in the studies addressed above. Further, we discuss the disadvantages of the applied approaches. Table 5.1 summarizes for the four evaluation studies addressed above, which variables were considered, how many models were compared, which boundary conditions were used to simulate from the RCMs, which reference period was considered for the evaluation, which locations/regions were considered, which reference data set (observations) was considered for comparison and which evaluation methods were applied.

The study of **Lorenz and Jacob (2010)** evaluates and compares 15 RCMs for Europe based on comparing linear trends in near-surface temperature to linear trends observed in a reference data set. As reference data set they consider the E-OBS data set (see Section 3.3). For their study, they divide Europe into eight sub-regions. As reference period they consider 1961–2000. Besides comparing trends obtained after annual aggregation of the time series, they differentiate between different seasons, as it is common practice in climate science. They differentiate between the seasons

`DJF` *winter* (December, January, February)

`MAM` *spring* (March, April, May)

`JJA` *summer* (June, July, August)

`SON` *autumn* (September, October, November)

each consisting of three months. Obviously, the study focuses only on one aspect/characteristic of the models under consideration. However, many other aspects are important to be able to judge the adequacy of such models. Hence, the evaluation does not yield a full assessment of the model performance. The way in which the study accounts for seasonality is a oversimplification of the problem. Splitting the year into four fixed seasons is somewhat arbitrary and requires a subjective decision on how important each season is for the comparison.

Kjellström et al. (2010) made an attempt to evaluate RCMs using estimates of full probability distributions. For the output of 16 RCMs (using boundary conditions from the ERA-40 reanalysis) as well as for the reference data, the gridded E-OBS data set (see Section 3.3), they constructed empirical estimates of probability density functions by binning the data into a certain number of bins. For this, they use the same eight aggregation areas and distinguish between the same seasons as Lorenz and Jacob (2010). As reference period they consider the years 1961–1990. Apart from minimum and maximum temperatures they perform their evaluation also for the precipitation output of the RCMs. In comparison to Lorenz and Jacob (2010), the evaluation approach of Kjellström et al. (2010) is more holistic, as it considers full probability distributions and not only a single model characteristic. However, Kjellström et al. (2010) are aware, that their approach is “associated with a large degree of subjectivity”, as it requires several subjective choices (bin width, metric for comparison, seasons, aggregation areas, etc.). Our novel

Table 5.1: Overview of the four evaluation studies by Lorenz and Jacob (2010), Kjellström et al. (2010), Landgren et al. (2014) and Eum et al. (2012): considered variables, number of compared models, boundary conditions used to simulate from the RCMs, reference period used for evaluation, considered locations/regions, reference data set (observations) used for comparison, methods applied for evaluation.

	Lorenz and Jacob (2010)	Kjellström et al. (2010)	Landgren et al. (2014)	Eum et al. (2012)
Variables	temperature	minimum and maximum temperature, precipitation	temperature, precipitation	minimum and maximum temperature, precipitation
# models	15	16	25	3
Boundary conditions	ERA-40 reanalysis	ERA-40 reanalysis	different driving GCMs	ERA-40 and NCEP reanalysis
Reference period	1961–2000	1961–1990	1981–2000, 2001–2012	1979–2001
Reference region	Europe (divided into 8 regions, the so called PRUDENCE regions)	Europe (PRUDENCE regions)	Scandinavia (5 selected locations)	Southern area in the Canadian provinces Quebec and Ontario
Reference data	E-OBS (see Section 3.3)	E-OBS	2 reanalysis (ERA-40, ERA-Interim) and a hindcast (NORA10) data set, observations (E-OBS, station data)	Gridded observational data
Methods	Comparison of annual/seasonal linear trends	Comparison of empirical probability density function estimates	Root Mean Square Deviation (RMSD) and a measure of Inter-Annual Variability (IAV)	5 different metrics assessing different characteristics of the models/forecasts

approach (introduced in Section 5.3) is less subjective as it applies proper scores/divergences, and does not require to choose a bin width or splitting of the year into seasons.

For their evaluation, **Landgren et al. (2014)** consider temperature and precipitation outputs from in total 25 RCMs (using boundary conditions from Global Climate Model (GCM) simulations). They restrict their evaluation to five selected locations in Scandinavia and consider two reference periods 1981–2000 and 2001–2012. Besides the E-OBS data set (see Section 3.3), they consider also two reanalysis datasets (ERA-40, ERA-Interim), a hindcast data set (NORA10) as well as station data as reference data. After monthly aggregation of the time series under consideration, they rank the different RCMs based on the Root Mean Square Deviation (RMSD) between model output and reference data on the one hand and a measure for difference in Inter-Annual Variability (IAV) on the other hand. Again only two selected, specific model characteristics are considered and there is no objective criterion which helps to decide on the relative importance of the two measures. Compared to the other approaches above, this approach has the advantage, that it does not require to split the data set into different seasons.

Eum et al. (2012) evaluate the Canadian RCM for a region in southern Canada, based on the years 1979–2001. They compare different versions of the Canadian RCM driven by different reanalysis products. Their comparison is based on the variables minimum/maximum temperature and precipitation. As reference data they use a gridded observational data set. Model performance is judged based on the five attributes

(i) Relative Absolute Mean Error (RAME) on a daily time scale,

(ii) discrepancy in the annual variability of monthly means,

(iii) discrepancy in the spatial pattern of mean values in a certain region,

(iv) discrepancy between 0.1 and 0.9 quantiles ("extremes") of daily observations and model output,

(v) differences in long-term linear trends.

This is an attempt to evaluate different features/moments of the probabilistic forecast, instead of considering the full distribution, which again comes along with the difficulty how to judge the relative importance of each attribute. Again, these considerations have to be made distinguishing between different seasons. All of this can be avoided with our novel (moving window) approach for model evaluation, as it considers the full distribution in time windows which are chosen such that the seasonal differences are evaluated in an appropriate manner.

5.3 Methodology

To make the concept of proper evaluation (see Section 2.5) available for the evaluation of time series models in a non-stationary context, we subsequently propose novel methodology. In comparison to a purely point-wise evaluation (see also Section 2.6) the novel methodology also considers higher order structures of time series. It enables to jointly evaluate different model characteristics and avoids subjective choices on the importance of these characteristics. Moreover, it is designed to deal with seasonality and avoids a separate consideration of different selected seasons.

In order to introduce the novel methodology for proper evaluation in a time series context ($t = 1, \dots, N$), we have the following setup (compare Section 2.5): Let $\mathcal{F}$ be a convex class of probability measures on a sample space Ω. Then we consider an (observed) *phenomenon* with the random outcomes

- Y_t with (unknown) distributions $G_t \in \mathcal{F}$ and realizations $y_t \in \Omega$ ($t = 1, \dots, N$).

Further, we consider a *model/forecast* for Y_t, $t = 1, \dots, N$, given through random variables

- X_t with (modeled) distributions $F_t \in \mathcal{F}$ and realizations $x_t \in \Omega$ ($t = 1, \dots, N$).

5.3.1 Discussion of naive evaluation approaches

For now, let us further assume, that G_t and F_t, $t = 1, \dots, N$, are both parametric distributions ($G_{\boldsymbol{\vartheta}_t}$ and $F_{\boldsymbol{\theta}_t}$), parametrized by (time varying) parameters $\boldsymbol{\vartheta}_t \in \mathbb{R}^q$ and $\boldsymbol{\theta}_t \in \mathbb{R}^p$, respectively. Hence, for $t = 1, \dots, N$ we have

$$Y_t \sim G_{\boldsymbol{\vartheta}_t} \qquad \text{and} \qquad X_t \sim F_{\boldsymbol{\theta}_t}.$$

Evaluation under a stationarity (ST) assumption First, we consider the stationary case, in which the parameters $\boldsymbol{\vartheta}_t$ and $\boldsymbol{\theta}_t$ are constant. That is, $\boldsymbol{\vartheta}_t = \boldsymbol{\vartheta}$ and $\boldsymbol{\theta}_t = \boldsymbol{\theta}$, and $Y_t \sim G_{\boldsymbol{\vartheta}}$ and $X_t \sim F_{\boldsymbol{\theta}}$ for all $t = 1, \dots, N$. Hence, in order to evaluate the model X_t for Y_t, we can use *ST-scores*

$$s^{\mathrm{ST}}(F_{\boldsymbol{\theta}}, y_t), \quad t = 1, \dots, N, \tag{5.1}$$

and the *ST-divergence*

$$d^{\mathrm{ST}}(F_{\boldsymbol{\theta}}, G_{\boldsymbol{\vartheta}}). \tag{5.2}$$

The prefix ST is used, as these scores and divergences are computed under a stationarity assumption. While ST-scores provide a time series of scores and allow an assessment for every time instance $t = 1, \dots, N$, the ST-divergence allows only for an overall model evaluation. As in practice, we might not know the parametric distributions $F_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\vartheta}}$, we can use sample versions of the ST-scores (5.1) and ST-divergence (5.2) instead. Using all realizations $\boldsymbol{x} = (x_1, \dots, x_N)$ of X_t, $t = 1, \dots, N$, and $\boldsymbol{y} = (y_1, \dots, y_N)$ of Y_t, $t = 1, \dots, N$, the *sample ST-scores* are given by

$$s^{\mathrm{ST}}(\boldsymbol{x}, y_t), \quad t = 1, \dots, N, \tag{5.3}$$

applying sample scores (2.30) and the *sample ST-divergence* is given by

$$d^{\mathrm{ST}}(\boldsymbol{x}, \boldsymbol{y}) \tag{5.4}$$

applying the sample divergence (2.40).

However, in practice the stationary case is not of so much interest, as one usually has to deal with the non-stationary case, where the parameters $\boldsymbol{\vartheta}_t$ and $\boldsymbol{\theta}_t$ are not constant. Then the assumption that Y_t and X_t are identically distributed for all $t = 0, \dots, N$ according to distributions $G_{\boldsymbol{\vartheta}}, F_{\boldsymbol{\theta}} \in \mathcal{F}$, respectively, is not realistic. Hence, ST-score (5.1) and ST-divergence (5.2) based evaluation is inappropriate.

Point-wise (PW) evaluation based on degenerated scores/divergences Let us now consider the non-stationary case. In Section 2.5 we have discussed degenerated scores and divergences, which arise in the case of deterministic models/forecasts, where we have only one realization x_t from our model X_t for each $t = 1, \dots, N$. As we have seen in Section 2.5, the degenerated scores and divergences correspond to degenerated sample based scores (2.30) and divergences (2.40) where the samples $\boldsymbol{x}$ and $\boldsymbol{y}$ consist of only one element each. Essentially, degenerated scores and divergences result in a point-wise (PW) model/forecast evaluation, which is conducted based on *point-wise scores (PW-scores)*

$$s^{\mathrm{PW}}(x_t, y_t), \quad t = 1, \dots, N, \tag{5.5}$$

and *point-wise divergences (PW-divergences)*

$$d^{\mathrm{PW}}(x_t, y_t), \quad t = 1, \dots, N. \tag{5.6}$$

Our introduction to proper scoring rules and divergence functions in Section 2.5 showed that the squared error $s_{\mathrm{SE}}(x_t, y_t) = (x_t - y_t)^2$ (compare Equation (2.27)) occurs as the degenerated SE score (2.27) and the degenerated MV divergence (2.38), and the absolute error $s_{\mathrm{AE}}(x_t, y_t) = |x_t - y_t|$ (compare Equation (2.28)) occurs as the degenerated CRPS (2.28) and the degenerated IQ distance (2.39). Hence it holds, that

$$s^{\mathrm{PW}}_{\mathrm{SE}}(x_t, y_t) = d^{\mathrm{PW}}_{\mathrm{MV}}(x_t, y_t) = s_{\mathrm{SE}}(x_t, y_t) = (x_t - y_t)^2$$

and

$$s^{\mathrm{PW}}_{\mathrm{CRPS}}(x_t, y_t) = d^{\mathrm{PW}}_{\mathrm{IQ}}(x_t, y_t) = s_{\mathrm{AE}}(x_t, y_t) = |x_t - y_t|,$$

for all $t = 1, \dots, N$. Note, that the means of these point-wise scores/divergences correspond to the sample Mean Square Error (2.43) and the sample Mean Absolute Error (2.44), respectively.

For a comprehensive model evaluation, point-wise scores (5.5) and divergences (5.6) are also not satisfying, since they do not account for higher order structures and properties of the observed/modeled phenomenon. They evaluate only how similar phenomenon and model/forecast behave in terms of their (temporarily varying) mean. Such point-wise model evaluation completely neglects differences in higher order moments (e.g. variance). A point-wise score/divergence treats a discrepancy in x_t and y_t that occurs due to a falsely specified model/forecast mean in the same way as such a discrepancy that occurs due to a high uncertainty of the phenomenon.

5.3.2 Moving scores and divergences

To deal with non-stationarity and to consider higher order structures of the observed phenomenon in the evaluation we suggest the following. We assume, that the phenomenon Y_t and its model/forecast X_t are (approximately) stationary for small time intervals. Then, for each time instance $t = 1, \dots, N$ (*window location*), we can select integers $\delta_t^-, \delta_t^+ \in \{0, \dots, N-1\}$ which determine the width of a *moving (time) window*

$$\mathcal{W}(t) = \{t - \delta_t^-, \dots, t, \dots, t + \delta_t^+\} \subset \{1, \dots, N\},$$

such that it completely lies in the observation period and that Y_s and X_s are (approximately) stationary for all time points $s \in \mathcal{W}(t)$. Thus, we assume that for $s \in \mathcal{W}(t)$, Y_s and X_s are distributed according to distributions G_t and F_t (depending on the window location t), respectively. Hence, we can consider the *theoretical moving scores*

$$s(F_t, y_t) \tag{5.7}$$

and the *theoretical moving divergences*

$$d(F_t, G_t), \tag{5.8}$$

in order to evaluate the model/forecast X_t for the phenomenon Y_t, for a specific time instance $t = 1, \dots, N$. As in practice we usually do not know F_t and G_t, we instead consider sample versions (see Equations (2.30) and (2.40)) of the above moving scores/divergences, based on the sub-samples $\boldsymbol{x}_{\mathcal{W}(t)} := \{x_s : s \in \mathcal{W}(t)\}$ from the realizations x_s, $s = 1, \dots, N$, of X_s, $s = 1, \dots, N$, and the sub-samples $\boldsymbol{y}_{\mathcal{W}(t)} := \{y_s : s \in \mathcal{W}(t)\}$ from the realizations y_s, $s = 1, \dots, N$, of Y_s, $s = 1, \dots, N$. Hence, we consider the (sample) *moving scores*

$$s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t) \tag{5.9}$$

and the (sample) *moving divergences*

$$d(\boldsymbol{x}_{\mathcal{W}(t)}, \boldsymbol{y}_{\mathcal{W}(t)}), \tag{5.10}$$

to evaluate the model/forecast X_t for the phenomenon Y_t, for a specific time instance $t = 1, \dots, N$. In contrast to the theoretical moving scores (5.7) and divergences (5.8), the moving scores (5.9) and divergences (5.10) are *empirical scores and divergences*, since they are sample-based. Same holds for the PW- and ST-scores and divergences introduced in Section 5.3.1.

Appropriateness of moving score/divergence averages for model evaluation In order to rank different models based on time series $s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t)$, $t = 1, \dots, N$, of empirical scores and time series $d(\boldsymbol{x}_{\mathcal{W}(t)}, \boldsymbol{y}_{\mathcal{W}(t)})$, $t = 1, \dots, N$, of empirical divergences, we consider averages

$$\frac{1}{N} \sum_{t=1}^{N} s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t) \tag{5.11}$$

and

$$\frac{1}{N} \sum_{t=1}^{N} d(\boldsymbol{x}_{\mathcal{W}(t)}, \boldsymbol{y}_{\mathcal{W}(t)}), \tag{5.12}$$

of these time series over all available time instances $t = 1, \dots, N$. Since in (5.11), the scores corresponding to the observations $y_1, \dots, y_N$ are all weighted equally, the propriety of the evaluation metric (5.11) is warranted if s is a proper scoring rule (see Theorem 1 in Gneiting and Ranjan, 2011). On the contrary, the propriety of the evaluation metric (5.12) is not necessarily warranted. The proof of Theorem 2 in Thorarinsdottir et al. (2013) demonstrates that consideration of a score divergence $d(\boldsymbol{x}_{\mathcal{W}(t)}, \boldsymbol{y}_{\mathcal{W}(t)})$ to evaluate the model/forecast X_t for the phenomenon Y_t for a specific time instance $t = 1, \dots, N$, is equivalent to considering the average

$$\frac{1}{|\mathcal{W}(t)|} \sum_{j \in \mathcal{W}(t)} s(\boldsymbol{x}_{\mathcal{W}(t)}, y_j).$$

of the corresponding scores. Thus, in the case of score divergences, model evaluation based on the metric (5.12) is equivalent to using the metric

$$\frac{1}{N}\sum_{t=1}^{N}\frac{1}{|\mathcal{W}(t)|}\sum_{j\in\mathcal{W}(t)} s(\boldsymbol{x}_{\mathcal{W}(t)}, y_j).$$

This equation can be rewritten as

$$\frac{1}{N}\sum_{t=1}^{N}\sum_{k:t\in\mathcal{W}(k)}\frac{1}{|\mathcal{W}(k)|}s(\boldsymbol{x}_{\mathcal{W}(k)}, y_t). \tag{5.13}$$

From the equivalent representation (5.13) of the evaluation metric (5.12) it follows that if the windows $\mathcal{W}(t)$, $t = 1, \ldots, N$, are overlapping (e.g. $\mathcal{W}(s) \cap \mathcal{W}(t) \neq \emptyset$ for $s \neq t$ such that $\mathcal{W}(s) \neq \mathcal{W}(t)$) or of varying size $|\mathcal{W}(t)|$, $t = 1, \ldots, N$, the observations $y_1, \ldots, y_N$ in (5.13) are not necessarily weighted equally. Hence, the metric (5.12) does not necessarily adopt the propriety property from the (proper) score divergences $d(\boldsymbol{x}_{\mathcal{W}(t)}, \boldsymbol{y}_{\mathcal{W}(t)})$, $t = 1, \ldots, N$. Along with the introduction of different selection approaches for the windows $\mathcal{W}(t)$, $t = 1, \ldots, N$, in Section 5.3.3, we will discuss if and how gravely the propriety property is violated by the metric (5.12).

Discussion of the moving window size Bearing in mind the discussion of Section 5.3.1, it becomes clear, that the moving score/divergence methodology is a compromise between a point-wise evaluation (which does not account for higher order structures) and an evaluation which wants to make use of all available realizations of X_t and Y_t, $t = 1, \ldots, N$, at the same time, ignoring non-stationarity. Let us discuss the trade-off between having small and big moving windows $\mathcal{W}(t)$: In order not to violate the *stationarity assumption*, we have to keep the moving windows small enough. Looking at the formulas (2.32) and (2.42) for the computation of sample CRPS and sample IQ distances, respectively, we see that their usage in a moving window based evaluation results in a *computation time* which grows quadratically with increasing window width $|\mathcal{W}(t)|$. Hence, if we have to deal with many long time series for several models and possibly many different spatial locations, a moving window based evaluation may become infeasible if the moving windows are too big. However, if the moving windows are too small (small sample size), the samples $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ coming from these small windows $\mathcal{W}(t)$ can only achieve an *inaccurate approximation* of the true distributions F_t and G_t, respectively. We illustrate this problem with a small simulation study. We simulate samples $\boldsymbol{y}_n = (y_1, \ldots, y_n)$ and $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ with different sample sizes $n = 1, \ldots, 50$ from $Y \sim \mathcal{N}(0, 1)$ and $X \sim \mathcal{N}(0, 1)$, repectively. Then we compute the (sample) IQ distances $d_{\mathrm{IQ}}(\boldsymbol{x}_n, \boldsymbol{y}_n)$ defined in Equation (2.42). We repeat this 1000 times and calculate the averages of the obtained IQ distances. The results are visualized in Figure 5.1. Since both Y and X have the same distribution, we know that the theoretical value of the IQ distance should be $d_{\mathrm{IQ}}(\mathcal{N}(0, 1), \mathcal{N}(0, 1)) = 0$. We see that the approximation of the true IQ distance by the sample divergences gets better with increasing sample size n. While big improvements are observed for small sample sizes ($n \leq 10$), the convergence slows down for bigger sample sizes.

We see that there are reasons supporting both, small and big window sizes of the moving windows $\mathcal{W}(t)$. In the subsequent section we introduce three different window selection strategies, which all try to make a compromise between small and big windows, where two of these strategies take the time varying stationarity behavior of the time series of interest into consideration.

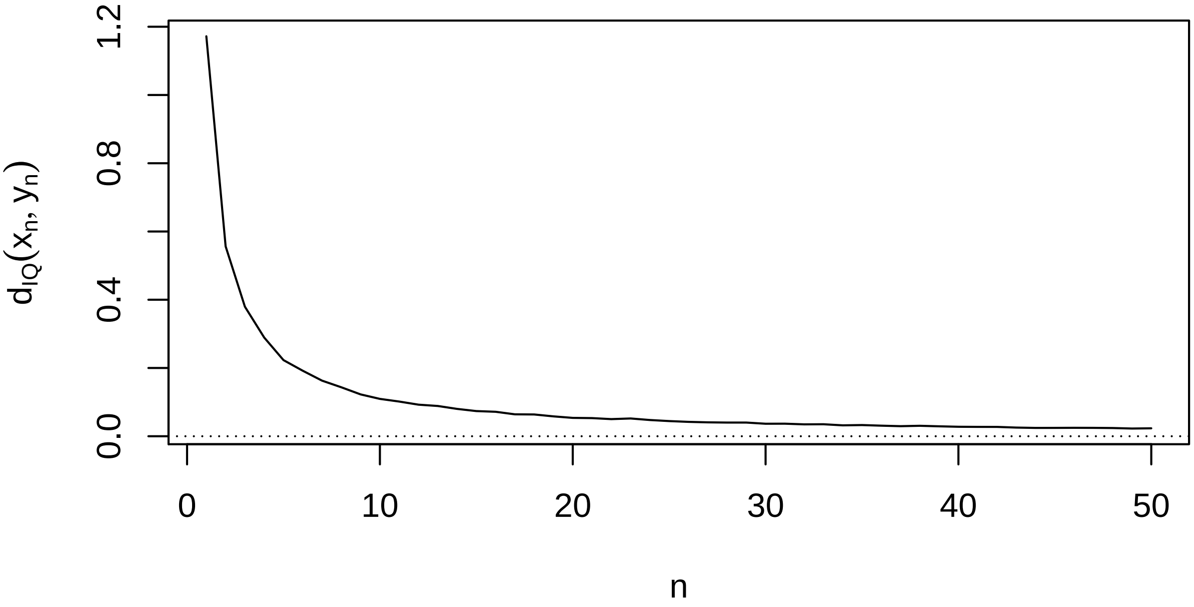


Figure 5.1: Approximation of theoretical IQ distance (dotted line) by sample IQ distances $d_{\text{IQ}}(\boldsymbol{x}_n, \boldsymbol{y}_n)$ (solid line) for different sample sizes n.

5.3.3 Selection of moving windows

So far, we have not discussed, how we choose the moving windows $\mathcal{W}(t) \subset \{1, \ldots, N\}$, $t = 1, \ldots, N$. We emphasize, that the appropriateness of the different, subsequently introduced approaches depends very much on the characteristics of the phenomenon of interest. Some approaches might be more adequate, if there are clear changepoints in the statistical behavior of the phenomenon of interest. Others are a better choice, if the characteristics of the phenomenon of interest change gradually. As moving scores and divergences are introduced in order to have a tool for the comparison of several models for a certain phenomenon and as the comparison is always based on one and the same time series of realizations $y_1, \ldots, y_N$, it is meaningful to determine the moving windows $\mathcal{W}(t) \subset \{1, \ldots, N\}$, $t = 1, \ldots, N$, based on the very same time series $y_1, \ldots, y_N$.

Subsequently, we introduce three different approaches on how to specify the moving windows, which are all based on a changepoint analysis (see Section 2.7). In a first step, $m < N - 1$ changepoints $\boldsymbol{\tau}_{1:m} = (\tau_1, \ldots, \tau_m)$ of $y_1, \ldots, y_N$ are detected using the PELT method introduced in Section 2.7. We use the cost function given by Equation (2.53), which is based on the assumption that the y_t come from a normal distribution. As it assumes varying means and variances for different segments of the time series, it allows to detect changes in both mean and variance. Since the cost function (2.53) is twice a negative log-likelihood, the constant K in the PELT algorithm is set to $K = 0$. For the penalty constant κ we choose $\kappa = p \ln(N)$ which corresponds to the penalty of the Bayesian information criterion (BIC). The number of parameters p for each additional segment is selected as $p = 3$, as we count one parameter for the mean, one for the variance and one for the changepoint. Moreover, we demand that the minimum segment length is 11 (see our discussion in Section 5.3.2 and Figure 5.1), that is it must hold $(\tau_{j+1} - \tau_j) > 10$, $j = 0, \ldots, m$. In a second step, (different types of) moving windows are specified, based on the selected $(m+1)$-segmentation $\boldsymbol{\tau}_{0:(m+1)} = (0, \boldsymbol{\tau}_{1:m}, N)$:

Overlapping windows with fixed width (OF) For the first approach, we consider moving time windows $\mathcal{W}^{\text{OF}}(t)$ of a fixed width $\omega^{\text{OF}} := \left|\mathcal{W}^{\text{OF}}(t)\right|$, as long as these windows are located "far enough" from the start and the end of the time interval $\{1, \ldots, N\}$. Below, we specify more precisely, what is meant by not "far enough". We will refer to the corresponding time instances (window locations) t which are not "far enough" as *border case*. As we center the (symmetric) windows $\mathcal{W}^{\text{OF}}(t)$ around their window location t, they are overlapping. To calculate the *fixed window width* ω^{OF}, we first calculate the *median segment length*

$$\lambda := \underset{j=0,\ldots,m}{\text{median}} \left(\tau_{j+1} - \tau_j\right),$$

where $\text{median}_{\mathcal{I}}(\cdot)$ denotes the sample median of quantities indexed by the index set $\mathcal{I}$. Then, we compute the *(fixed) window width parameter* defined as

$$\delta^{\text{OF}} := \left\lfloor \frac{\lambda - 1}{2} \right\rfloor,$$

where $\lfloor\cdot\rfloor$ rounds a number to its next smaller integer. Defining the window width parameter $\delta_t^{\text{OF}} := \delta^{\text{OF}}$, for all $t = 1 + \delta^{\text{OF}}, \ldots, N - \delta^{\text{OF}}$, we ensure that the moving windows defined as

$$\mathcal{W}^{\text{OF}}(t) := \left\{t - \delta_t^{\text{OF}}, \ldots, t, \ldots, t + \delta_t^{\text{OF}}\right\}, \tag{5.14}$$

are symmetric with *fixed window width*

$$\omega^{\text{OF}} = 2\delta^{\text{OF}} + 1,$$

for all $t = 1 + \delta^{\text{OF}}, \ldots, N - \delta^{\text{OF}}$. The windows $\mathcal{W}^{\text{OF}}(t)$ for $t = 1, \ldots, \delta^{\text{OF}}$ and $t = N - \delta^{\text{OF}} + 1, \ldots, N$ are those which are not "far enough" from the start and the end of the time interval $\{1, \ldots, N\}$ (border case). Hence, we define the *window width parameter* as $\delta_t^{\text{OF}} := t - 1$, for $t = 1, \ldots, \delta^{\text{OF}}$, and as $\delta_t^{\text{OF}} := N - t$, for $t = N - \delta^{\text{OF}} + 1, \ldots, N$. Like that it is guaranteed that $\mathcal{W}^{\text{OF}}(t) \subset \{1, \ldots, N\}$. All in all, the *overlapping windows with fixed width (OF-windows)* are defined by Equation (5.14), where

$$\delta_t^{\text{OF}} := \begin{cases} t - 1, & \text{for } t = 1, \ldots, \delta^{\text{OF}}, \\ \delta^{\text{OF}}, & \text{for } t = 1 + \delta^{\text{OF}}, \ldots, N - \delta^{\text{OF}}, \\ N - t, & \text{for } t = N - \delta^{\text{OF}} + 1, \ldots, N. \end{cases}$$

The corresponding window widths are obtained as $\omega_t^{\text{OF}} = 2\delta_t^{\text{OF}} + 1$. Moving scores and divergences obtained based on OF-windows will also be referred to as *OF-scores* and *OF-divergences*, respectively.

The top panel of Figure 5.2 illustrates two time series (y_t and x_t, $t = 1, \ldots, N$) of realizations from a phenomenon Y_t and a corresponding model/forecast X_t. Moreover, it shows which changepoints $\boldsymbol{\tau}_{1:m}$ were detected by the PELT algorithm. The middle panel shows, which window width was selected by the OF-method for each time instance $t = 1, \ldots, N$. The corresponding moving windows $\mathcal{W}^{\text{OF}}(t_1)$ and $\mathcal{W}^{\text{OF}}(t_2)$ for two selected time instances t_1 and t_2 are illustrated as gray boxes in the upper panel. We clearly see, that both windows are symmetric/centered around their locations t_1 and t_2. Both windows have the same width. Due to their fixed width it can happen that these windows contain more than one changepoint. The lower panel shows for both OF-windows (located at t_1 and t_2), how the empirical CDFs for both the phenomenon ($\widehat{G}_t$) and the model/forecast ($\widehat{F}_t$) differ from each other and how much they deviate from the observation (y_t) in the window location (t). Comparing the empirical CDFs $\widehat{F}_{t_1}$ and $\widehat{F}_{t_2}$, it is

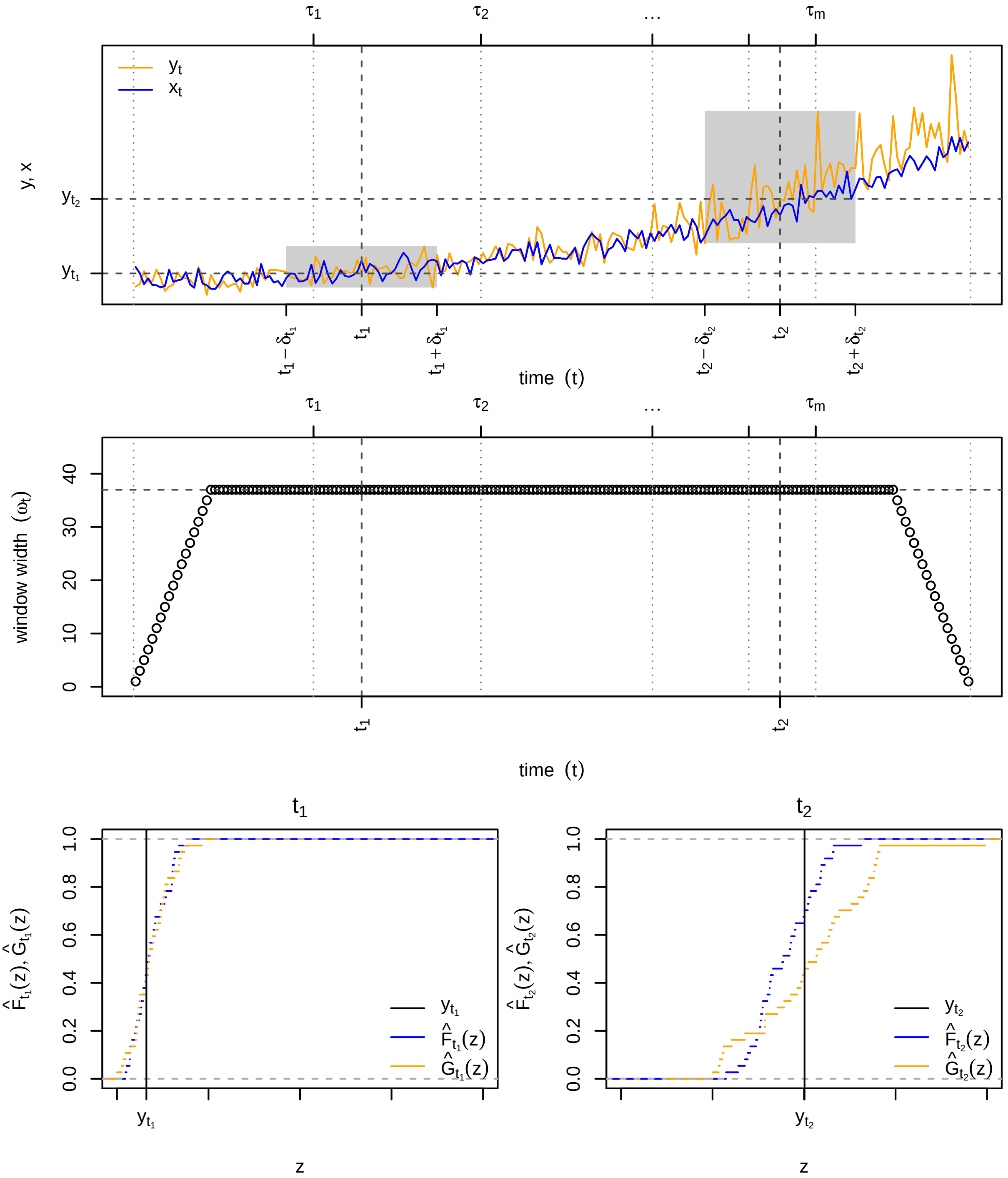


Figure 5.2: Illustration of moving window methodology for overlapping windows with fixed width (OF): Realizations y_t (orange) and x_t (blue), $t = 1, \ldots, N$ (of phenomenon Y_t and corresponding model/forecast X_t), detected changepoints $\boldsymbol{\tau}_{1:m}$, and moving windows $\mathcal{W}^{\mathrm{OF}}(t_1)$ and $\mathcal{W}^{\mathrm{OF}}(t_2)$ (gray) for two selected time instances t_1 and t_2 (upper panel). Selected window width (middle panel). Empirical CDFs for phenomenon ($\widehat{G}_t$) and model/forecast ($\widehat{F}_t$) based on window $\mathcal{W}^{\mathrm{OF}}(t)$, and observation ($y_t$) in the window location (t), for the time instances t_1 (lower left panel) and t_2 (lower right panel).

obvious that for t_1 we can expect a smaller moving CRPS, since $\widehat{F}_{t_1}$ is much steeper and centered better around the corresponding observation. Moreover, we observe that also the moving IQ distance for t_1 will be much smaller than that for t_2, as the empirical CDFs $\widehat{F}_t$ and $\widehat{G}_t$ are much closer to each other for t_1.

To close our introduction of the OF-window selection approach, we discuss if the utilization of moving divergence averages (5.12) based on OF-windows and score divergences violates the propriety property or not. Consideration of Equation (5.13), which is equivalent to the metric (5.12), reveals immediately that the metric (5.12) is non-proper, as the OF-windows $\mathcal{W}^{\mathrm{OF}}(t)$, $t = 1, \ldots, N$, are overlapping (and for the border case their width varies). To see if propriety holds at least approximately, we reconsider (5.13) under the assumption that for all time instances $t = 1, \ldots, N$ all scores $s(\boldsymbol{x}_{\mathcal{W}(k)}, y_t)$, $k \in \{j : t \in \mathcal{W}(j)\}$, corresponding to the observation y_t approximately equal the score $s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t)$. Then we can rewrite (5.13) as

$$\frac{1}{N}\sum_{t=1}^{N} \Omega(t) s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t),$$

with weights defined as

$$\Omega(t) := \sum_{k : t \in \mathcal{W}(k)} \frac{1}{|\mathcal{W}(k)|}. \tag{5.15}$$

Since the weights (5.15) are obtained based on an approximation we also refer to them as *approximate weights*. For the time instances $t = 1 + 2\delta^{\mathrm{OF}}, \ldots, N - 2\delta^{\mathrm{OF}}$ the weights (5.15) are constant and equal 1, as each of these time instances falls in exactly ω^{OF} many windows and each window has the same fixed window width ω^{OF}. From this we conclude, that if the considered time series is long (N large) in comparison to the fixed window width $\omega^{\mathrm{OF}} = 2\delta^{\mathrm{OF}} + 1$, the metric (5.12) is at least approximately proper.

Overlapping windows with varying width (OV) For the second approach, we consider moving time windows $\mathcal{W}^{\mathrm{OV}}(t)$ with varying width $\omega_t^{\mathrm{OV}} := |\mathcal{W}^{\mathrm{OV}}(t)|$. Again, we center these (overlapping) windows around their window location t. To obtain the varying windows $\mathcal{W}^{\mathrm{OV}}(t)$, we first determine the centers of the segments $\{\tau_j + 1, \ldots, \tau_{j+1}\}$, $j = 0, \ldots, m$, of the $(m+1)$-segmentation $\boldsymbol{\tau}_{0:(m+1)}$. The *segment centers* are defined as

$$\gamma_j := \frac{\tau_j + 1 + \tau_{j+1}}{2}, \quad j = 0, \ldots, m.$$

Then, we linearly interpolate the corresponding segment lengths $\lambda_j := (\tau_{j+1} - \tau_j)$ between the segment centers γ_j, $j = 0, \ldots, m$. Hence, the *interpolated segment lengths* are given by

$$\varsigma(t) := \frac{\gamma_{j+1} - t}{\gamma_{j+1} - \gamma_j}\lambda_j + \frac{t - \gamma_j}{\gamma_{j+1} - \gamma_j}\lambda_{j+1}, \quad \text{for } t \in [\gamma_j, \gamma_{j+1}],\ j = 0, \ldots, m.$$

Based on $\varsigma(t)$ we compute the corresponding *window width parameters*

$$\delta_t^{\mathrm{OV}} := \left\lfloor \frac{\varsigma(t) - 1}{2} \right\rfloor, \quad \text{for } t = \lceil \gamma_0 \rceil, \ldots, \lfloor \gamma_{m+1} \rfloor,$$

where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ round a number to its next smaller and bigger integer, respectively. For the *border case*, that is the start ($t = 1, \ldots, \lfloor \gamma_0 \rfloor$) and the end of the time series ($t = \lceil \gamma_{m+1} \rceil, \ldots, N$), we again define δ_t^{OV} such that the corresponding moving windows (with maximum possible

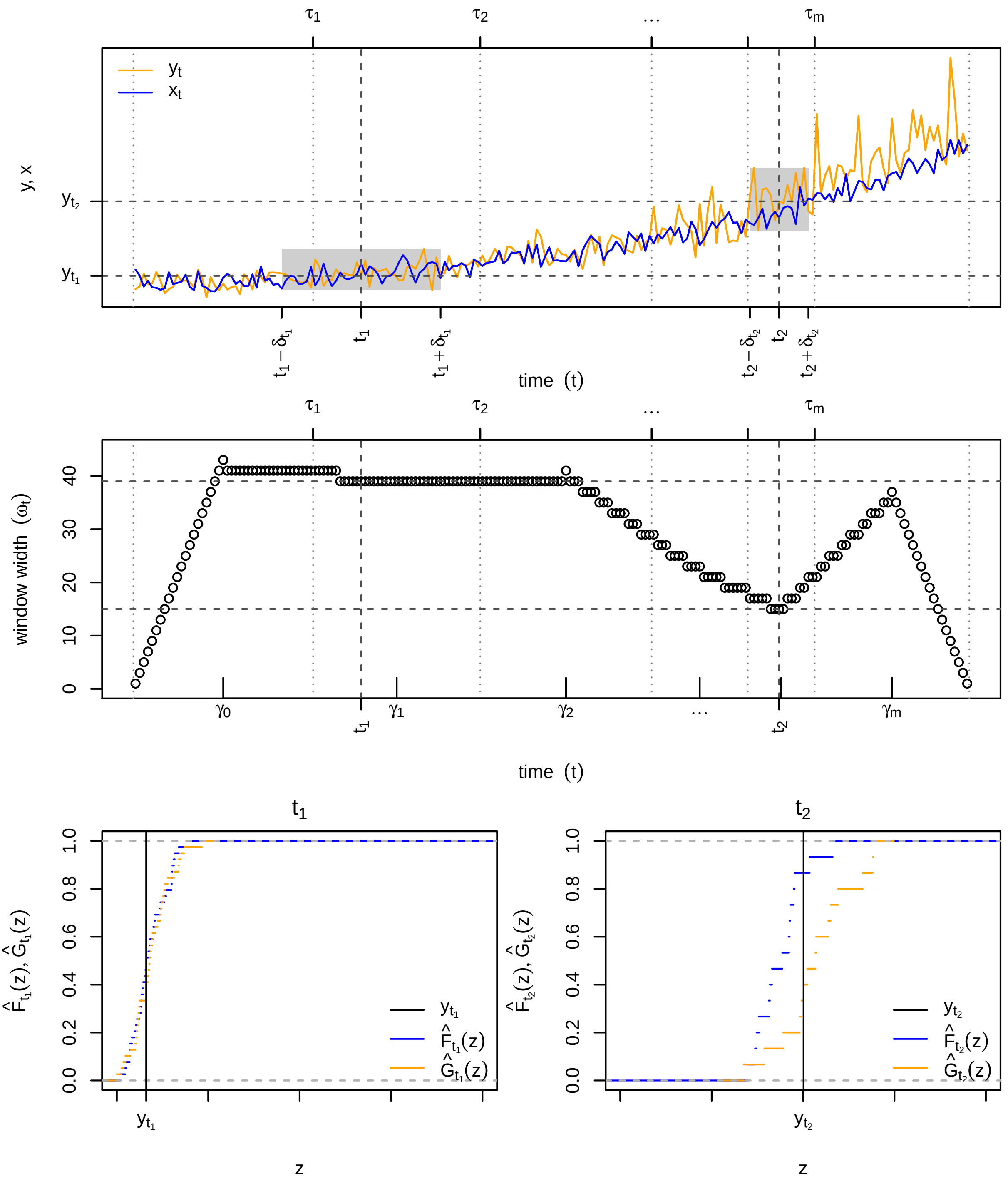


Figure 5.3: Illustration of moving window methodology for overlapping windows with varying width (OV): Realizations y_t (orange) and x_t (blue), $t = 1, \ldots, N$ (of phenomenon Y_t and corresponding model/forecast X_t), detected changepoints $\boldsymbol{\tau}_{1:m}$, and moving windows $\mathcal{W}^{\mathrm{OV}}(t_1)$ and $\mathcal{W}^{\mathrm{OV}}(t_2)$ (gray) for two selected time instances t_1 and t_2 (upper panel). Segment centers (γ_j, $j = 0, \ldots, m$) and selected window width (middle panel). Empirical CDFs for phenomenon ($\widehat{G}_t$) and model/forecast ($\widehat{F}_t$) based on window $\mathcal{W}^{\mathrm{OV}}(t)$, and observation ($y_t$) in the window location (t), for the time instances t_1 (lower left panel) and t_2 (lower right panel).

width) completely lie in $\{1, \ldots, N\}$. To sum up, the (symmetric) *overlapping windows with varying width (OV-windows)* are defined by

$$\mathcal{W}^{\mathrm{OV}}(t) := \left\{t - \delta_t^{\mathrm{OV}}, \ldots, t, \ldots, t + \delta_t^{\mathrm{OV}}\right\}, \tag{5.16}$$

where

$$\delta_t^{\mathrm{OV}} := \begin{cases} t-1, & \text{for } t = 1, \ldots, \lfloor \gamma_0 \rfloor, \\ \lfloor (\varsigma(t) - 1)/2 \rfloor, & \text{for } t = \lceil \gamma_0 \rceil, \ldots, \lfloor \gamma_{m+1} \rfloor, \\ N - t, & \text{for } t = \lceil \gamma_{m+1} \rceil, \ldots, N. \end{cases}$$

The corresponding varying window widths are obtained as $\omega_t^{\mathrm{OV}} = 2\delta_t^{\mathrm{OV}} + 1$. Moving scores and divergences obtained based on OV-windows will also be referred to as *OV-scores* and *OV-divergences*, respectively.

Figure 5.3 is the analog of Figure 5.2 for the OV-method. Its middle panel shows, how the OV-method interpolates the window widths between the segment centers γ_j, $j = 0, \ldots, m$. The resulting moving windows $\mathcal{W}^{\mathrm{OV}}(t_1)$ and $\mathcal{W}^{\mathrm{OV}}(t_2)$ for two selected time instances t_1 and t_2 are illustrated in the upper panel. We observe, that both windows are symmetric/centered around their locations t_1 and t_2, however they have different widths. The lower panel again shows for both OV-windows (located at t_1 and t_2), how the empirical CDFs for both the phenomenon ($\widehat{G}_t$) and the model/forecast ($\widehat{F}_t$) differ from each other and how much they deviate from the observation (y_t) in the window location (t). Comparison of the empirical CDFs for t_1 (left lower panel) and t_2 (right lower panel) shows that the empirical CDFs for t_2 are comparatively coarser. This is due to the fact that the window $\mathcal{W}^{\mathrm{OV}}(t_1)$ is wider than $\mathcal{W}^{\mathrm{OV}}(t_2)$ (i.e. $\omega_{t_1}^{\mathrm{OV}} > \omega_{t_2}^{\mathrm{OV}}$). Again, we expect a smaller moving CRPS for t_1 compared to t_2, since $\widehat{F}_{t_1}$ is much steeper and more centered around the corresponding observation. Furthermore, we observe a much smaller moving IQ distance for t_1 than for t_2, as the empirical CDFs $\widehat{F}_t$ and $\widehat{G}_t$ are much more similar for t_1.

To close our introduction of the OV-window selection approach, we discuss if the utilization of moving divergence averages (5.12) based on OV-windows and score divergences violates the propriety property. As already for the OF-windows, we learn from Equation (5.13) that the metric (5.12) is also not proper for OV-windows. Since the OV-windows $\mathcal{W}^{\mathrm{OV}}(t)$, $t = 1, \ldots, N$, are overlapping and their width varies continuously over time, the propriety property is violated much more in comparison to the case of the OF-windows.

Disjoint windows with varying width (DV) For the third approach, we consider moving time windows $\mathcal{W}^{\mathrm{DV}}(t)$ with varying width $\omega_t^{\mathrm{DV}} := \left|\mathcal{W}^{\mathrm{DV}}(t)\right|$. This time, we do not center these windows around their window location t. Instead, we consider the (disjoint) windows which are given by the segments $\{\tau_j + 1, \ldots, \tau_{j+1}\}$, $j = 0, \ldots, m$, of the $(m+1)$-segmentation $\boldsymbol{\tau}_{0:(m+1)}$. Then, the *disjoint windows with varying width (DV-windows)* are defined by

$$\mathcal{W}^{\mathrm{DV}}(t) := \{\tau_j + 1, \ldots, \tau_{j+1}\}, \quad \text{for } t = \tau_j + 1, \ldots, \tau_{j+1}, \, j = 0, \ldots, m. \tag{5.17}$$

The (varying) window widths, which equal the corresponding segment lengths are given by

$$\omega_t^{\mathrm{DV}} := (\tau_{j+1} - \tau_j), \quad \text{for } t = \tau_j + 1, \ldots, \tau_{j+1}, \, j = 0, \ldots, m.$$

Moving scores and divergences obtained based on DV-windows will also be referred to as *DV-scores* and *DV-divergences*, respectively.

Figure 5.4 is the analog of Figures 5.2 and 5.3 for the DV-method. Its middle panel shows, that the window width ω_t^{DV} of the disjoint windows selected by the DV-method is constant for

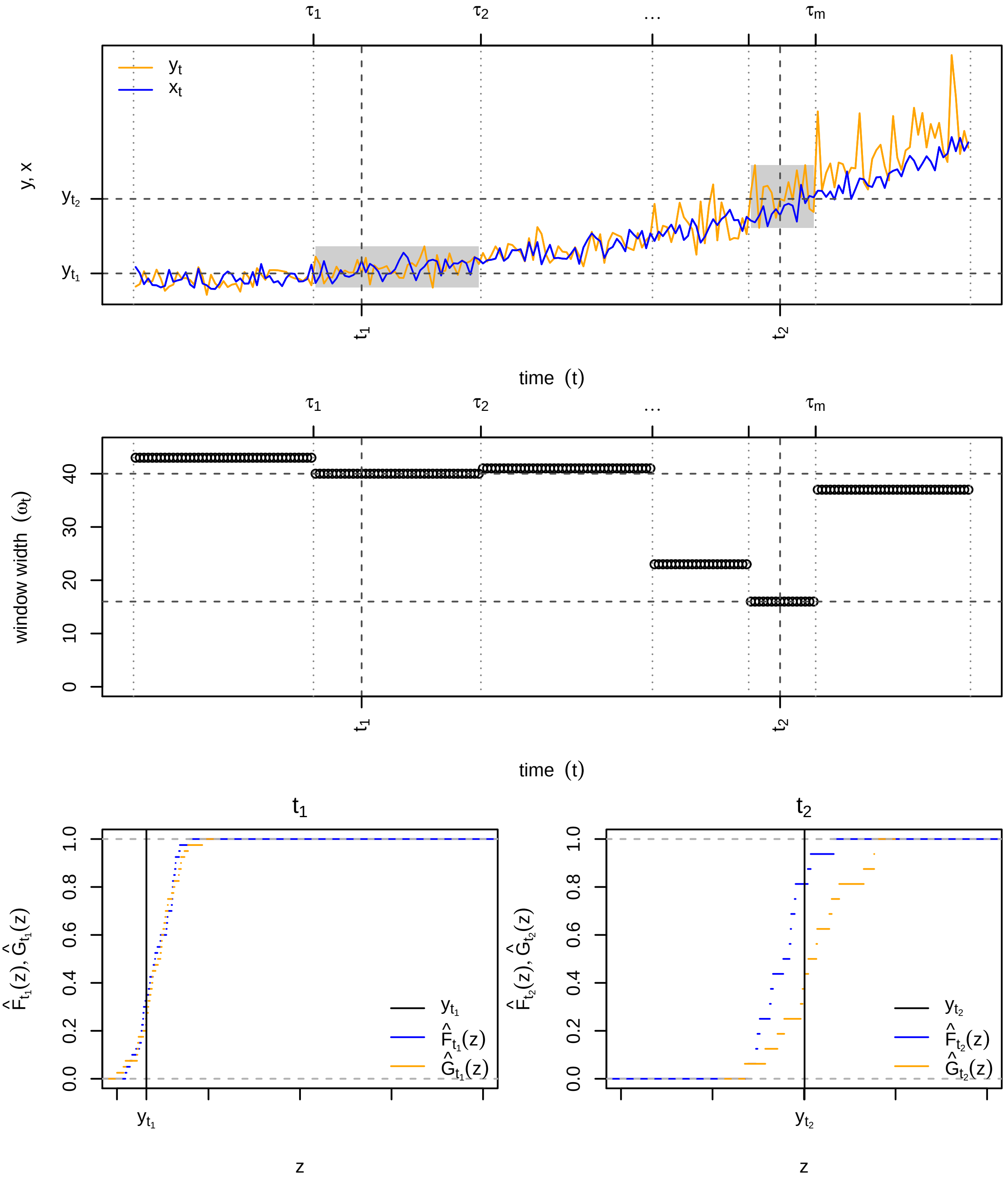


Figure 5.4: Illustration of moving window methodology for disjoint windows with varying width (DV): Realizations y_t (orange) and x_t (blue), $t = 1, \ldots, N$ (of phenomenon Y_t and corresponding model/forecast X_t), detected changepoints $\boldsymbol{\tau}_{1:m}$, and moving windows $\mathcal{W}^{\mathrm{DV}}(t_1)$ and $\mathcal{W}^{\mathrm{DV}}(t_2)$ (gray) for two selected time instances t_1 and t_2 (upper panel). Selected window width (middle panel). Empirical CDFs for phenomenon $(\widehat{G}_t)$ and model/forecast $(\widehat{F}_t)$ based on window $\mathcal{W}^{\mathrm{DV}}(t)$, and observation (y_t) in the window location (t), for the time instances t_1 (lower left panel) and t_2 (lower right panel).

each window. The moving windows $\mathcal{W}^{\mathrm{DV}}(t_1)$ and $\mathcal{W}^{\mathrm{DV}}(t_2)$ for two selected time instances t_1 and t_2 are illustrated in the upper panel. This time, both windows are not symmetric/centered around their locations t_1 and t_2. Moreover, they have different widths and do not contain any changepoint. The lower panel again shows for both DV-windows (located at t_1 and t_2), how the empirical CDFs for both the phenomenon ($\widehat{G}_t$) and the model/forecast ($\widehat{F}_t$) differ from each other and how much they deviate from the observation (y_t) in the window location (t). Comparing the left and the right side of the lower panel, we obtain very similar results as for the OV-windows (Figure 5.3).

To close our introduction of the DV-window selection approach, we discuss if the utilization of moving divergence averages (5.12) based on DV-windows and score divergences violates the propriety property or not. To answer this question, we again consider Equation (5.13), which is equivalent to the metric (5.12). For the DV-windows it holds that for all time instances $t = 1, \ldots, N$ all scores $s(\boldsymbol{x}_{\mathcal{W}(k)}, y_t)$, $k \in \{j : t \in \mathcal{W}(j)\}$, corresponding to the observation y_t equal one and the same score $s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t)$. As a consequence, (5.13) again simplifies to

$$\frac{1}{N} \sum_{t=1}^{N} \Omega(t) s(\boldsymbol{x}_{\mathcal{W}(t)}, y_t),$$

with weights

$$\Omega(t) \coloneqq \sum_{k : t \in \mathcal{W}(k)} \frac{1}{|\mathcal{W}(k)|}.$$

Note that the formula for the weights $\Omega(t)$ equals Equation (5.15). This time we have exact (and not only approximate) weights. From the Definition (5.17) of the DV-windows it follows directly that the weights $\Omega(t)$ all equal 1. Hence, utilization of moving divergence averages (5.12) based on DV-windows preserves propriety.

5.4 Organization of simulation studies

To assess the properties and judge the applicability of moving scores and divergences (see Section 5.3.2) based on different moving window selection approaches (see Section 5.3.3), we conduct a simulation study. To this end, we first consider a

changepoint scenario (C, see Section 5.5),

where the characteristics (mean and variance) of the phenomenon of interest change at certain (unknown) time instances. Moreover, as one of the main application areas of the introduced validation techniques are the validation of weather forecasts and climate models, and both weather and climate are non-stationary phenomena subject to trends and periodicity, we consider a

trend scenario (T, see Section 5.6)

and a

periodicity scenario (P, see Section 5.7).

General setup of the simulation studies For all three scenarios we consider a *phenomenon* given by

$$Y_t, \quad t = 1, \ldots, N, \tag{5.18}$$

and *five models*

$$X_t^k, \quad t = 1, \ldots, N, \tag{5.19}$$

for that phenomenon, numbered by $k = 1, \ldots, 5$. We assume that Y_t, X_t^k, $k = 1, \ldots, 5$, are normally distributed for all $t = 1, \ldots, N$, that is

$$Y_t \sim \mathcal{N}\left(\mu_{0,t}, \sigma_{0,t}^2\right)$$

and

$$X_t^k \sim \mathcal{N}\left(\mu_{k,t}, \sigma_{k,t}^2\right), \quad k = 1, \ldots, 5.$$

Since we simulate the phenomenon for the simulation studies, we also refer to Y_t, $t = 1, \ldots, N$, as *data generating process*. Depending on the scenario (changepoint (C), trend (T), periodicity (P)) both the means $\mu_{k,t}$ ($k = 0, \ldots, 5$) and the standard deviations $\sigma_{k,t}$ ($k = 0, \ldots, 5$) vary with time ($t = 1, \ldots, N$). The results of the simulation studies presented in Sections 5.5–5.7 are based on 10,000 replications of the data generating process (5.18) and each of the five models (5.19).

Methods For all 10,000 replications, we compute time series ($t = 1, \ldots, N$) of moving scores (5.9) and divergences (5.10) for each of the five models (5.19). We repeat this procedure for different window selection approaches (see Section 5.3.3), that is

- overlapping windows with fixed width (OF, see Equation (5.14)),
- overlapping windows with varying width (OV, see Equation (5.16)),
- and disjoint windows with varying width (DV, see Equation (5.17)).

Moreover, for comparison, we compute point-wise (PW) scores (5.5) and divergences (5.6). In case of the changepoint (C) and the trend scenario (T), we also consider ST-scores (5.3) and ST-divergences (5.4), which assume stationarity of the time series. We compare these time series of empirical scores and divergences to the corresponding theoretical (Theo.) scores (5.7) and divergences (5.8), which can be obtained, since we know the distributions of the data generating process (5.18) and each of the five models (5.19) for all time instances $t = 1, \dots, N$.

To get an idea of the temporal evolution of the (moving) scores and divergences, we average the obtained score/divergence time series over all 10,000 replications (see Figures 5.8–5.14, 5.18–5.24, 5.28–5.34). For this purpose we consider Continuous Ranked Probability Scores (CRPS, see Equations (2.24) and (2.32)) and Integrated Quadratic (IQ) distances (see Equations (2.36) and (2.42)). To provide an overall ranking of the different models (distinguishing between different approaches for the computation of scores and divergences) we further average over all time instances $t = 1, \dots, N$ (see Tables 5.2, 5.5 and 5.8) and also look at Squared Error (SE) scores (see Equations (2.23) and (2.31)) and Mean Value (MV) divergences (see Equations (2.35) and (2.41)). Note that utilization of moving divergence averages is not necessarily a proper evaluation method (that is in case of OF- and OV-windows; see discussion in Sections 5.3.2 and 5.3.3).

5.5 Simulation study: Changepoint scenario

To describe the changepoint scenario (C) we consider time varying means and time varying standard deviations which change after certain changepoints $\boldsymbol{\tau}_{1:m} \in \mathbb{R}^m$ (see Section 2.7). As usual we denote the corresponding $(m+1)$-segmentation by $\boldsymbol{\tau}_{0:(m+1)}$. Hence, for the changepoint scenario (C), the time varying means are defined as

$$M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\mu}) := \mu_j, \text{ for } t \in \{\tau_j + 1, \dots, \tau_{j+1}\},$$

where the parameter vector $\boldsymbol{\mu} = (\mu_0, \dots, \mu_m) \in \mathbb{R}^{m+1}$ consists of the mean values corresponding to the $m+1$ segments of the $(m+1)$-segmentation $\boldsymbol{\tau}_{0:(m+1)}$. The time varying standard deviations are defined as

$$S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}) := \sigma_j, \text{ for } t \in \{\tau_j + 1, \dots, \tau_{j+1}\},$$

where the parameter vector $\boldsymbol{\sigma} = (\sigma_0, \dots, \sigma_m) \in \mathbb{R}^{m+1}$ consists of the standard deviations corresponding to the $m + 1$ segments of the $(m + 1)$-segmentation $\boldsymbol{\tau}_{0:(m+1)}$.

Changepoint scenario (C) The subsequent equations define the *changepoint scenario (C)*, where we consider time series of length $N = 200$ with $m = 2$ changepoints and corresponding segmentation $\boldsymbol{\tau}_{0:3} = (0, 80, 130, 200)$. The data generating process (C0) and the five models (C1)–(C5) are given by

$$\begin{array}{lll}
\text{(C0)} & Y_t \sim \mathcal{N}\Big(M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:3}, \boldsymbol{\mu}^0), \big[S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}^0)\big]^2\Big) & \text{(data generating process),} \\
\hline
\text{(C1)} & X_t^1 \sim \mathcal{N}\Big(M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:3}, \boldsymbol{\mu}^1), \big[S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}^1)\big]^2\Big) & \text{(true model),} \\
\text{(C2)} & X_t^2 \sim \mathcal{N}\Big(M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:3}, \boldsymbol{\mu}^2), \big[S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}^2)\big]^2\Big) & \text{(constant mean),} \\
\text{(C3)} & X_t^3 \sim \mathcal{N}\Big(M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:3}, \boldsymbol{\mu}^3), \big[S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}^3)\big]^2\Big) & \text{(constant variance),} \\
\text{(C4)} & X_t^4 \sim \mathcal{N}\Big(M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:3}, \boldsymbol{\mu}^4), \big[S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}^4)\big]^2\Big) & \text{(constant mean and variance),} \\
\text{(C5)} & X_t^5 \sim \mathcal{N}\Big(M_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:3}, \boldsymbol{\mu}^5), \big[S_{\mathrm{C}}(t; \boldsymbol{\tau}_{0:(m+1)}, \boldsymbol{\sigma}^5)\big]^2\Big) & \text{(wrong mean and variance),}
\end{array}$$

with parameters

(C0)	$\boldsymbol{\mu}^0 = (0, 1, 0)$,	$\boldsymbol{\sigma}^0 = (0.9, 0.9, 0.3)$,	(data generating process),
(C1)	$\boldsymbol{\mu}^1 = \boldsymbol{\mu}^0$,	$\boldsymbol{\sigma}^1 = \boldsymbol{\sigma}^0$,	(true model),
(C2)	$\boldsymbol{\mu}^2 = (0.25, 0.25, 0.25)$,	$\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^0$,	(constant mean),
(C3)	$\boldsymbol{\mu}^3 = \boldsymbol{\mu}^0$,	$\boldsymbol{\sigma}^3 = (0.6, 0.6, 0.6)$,	(constant variance),
(C4)	$\boldsymbol{\mu}^4 = \boldsymbol{\mu}^2$,	$\boldsymbol{\sigma}^4 = \boldsymbol{\sigma}^3$,	(constant mean and variance),
(C5)	$\boldsymbol{\mu}^5 = (0.1, 0.9, 0.1)$,	$\boldsymbol{\sigma}^5 = \boldsymbol{\sigma}^3$,	(wrong mean and variance).

Whereas the true model (C1) is equivalent to the data generating process (C0) in terms of its parametrization, the other Models (C2)–(C5) differ from the data generating process (C0) in at least one parameter.

Illustration of the changepoint scenario (C) The differences between the data generating process (C0) and the models (C1)–(C5) in the temporal evolution of the mean and the standard deviation are visualized in Figure 5.5. In comparison to the data generating process (C0) where the mean $\mu_{0,t}$ changes after two and the variance $\sigma^2_{0,t}$ after one changepoint(s),

- the constant mean $\mu_{2,t}$ of Model (C2) differs from $\mu_{0,t}$ for all $t = 0, \ldots, N$,
- the constant variance $\sigma^2_{3,t}$ of Model (C3) differs from $\sigma^2_{0,t}$ for all $t = 0, \ldots, N$,
- the constant mean $\mu_{4,t}$ and the constant variance $\sigma^2_{4,t}$ of Model (C4) differ from $\mu_{0,t}$ and $\sigma^2_{0,t}$ for all $t = 0, \ldots, N$,
- the mean $\mu_{5,t}$ of Model (C5) deviates slightly from $\mu_{0,t}$ for all three segments defined by $\boldsymbol{\tau}_{0:3} = (0, 80, 130, 200)$ and the constant variance $\sigma^2_{5,t}$ differs from $\sigma^2_{0,t}$ for all $t = 0, \ldots, N$.

Hence, Model (C2) does not consider that the mean of the data generating process changes in two points and rather assumes a constant mean. It captures the change in the variance. Model (C3) models the change in the mean correctly, however it neglects the change in the variance (it assumes a constant variance). Models (C4) and (C5) are both misspecified in terms of mean and variance, where the misspecification of the mean is less severe for Model (C5). The true model (C1) is specified correctly. A comparative model evaluation should consider Model (C1) best.

Illustration of changepoint analysis and selection of moving windows Figure 5.6 illustrates one replication of the data generating process (C0) and the corresponding selection of moving windows according to Section 5.3.3. We observe that $m = 2$ changepoints were detected, which results in a 3-segmentation. Apparently, the changepoint locations of the data generating process were detected more or less accurately. Also the mean $\widehat{\mu}_t$ and variance estimates $\widehat{\sigma}^2_t$ corresponding to the different segments reflect the temporal evolution of their true counterparts defined by the data generating process (C0).

The segmentation into three segments is also reflected in the moving windows selected by the OV and DV approach. For the border case (time instances close to the start and the end of the observation period), the window widths ω_t^{OF} and ω_t^{OV} decrease towards 1 when approaching the start ($t = 1$) and the end ($t = N$) of the time series. For all other time instances t (apart from the border case), the window width ω_t^{OF} is constant and equals 69, and the window width ω_t^{OV} varies between 47 and 81. The window width ω_t^{DV} takes only the three different values 82, 48 and 70. Hence, in most cases the window width differs considerably from the window width $\omega_t^{\text{PW}} = 1$ used in the case of a point-wise (PW) evaluation.

With the bottom panel of Figure 5.6 we illustrate the (approximate) weights $\Omega(t)$ (see Equation (5.15)) corresponding to each observation y_t, $t = 1, \ldots, N$, in an evaluation based on moving divergence averages (5.12). As already discussed, the propriety of the metric (5.12) is given only if the weights $\Omega(t)$ are constant. From the figure we observe, that the DV-windows yield propriety, while the OF- and OV-windows do not. Especially in the beginning and in the end of the considered time interval (border case), the weights $\Omega(t)$ deviate from 1 for OF- and OV-windows. In general the metric based on OF-windows approximates propriety better than that based on OV-windows.

As Figure 5.6 is based on one replication only, we are further interested to see how many changepoints were detected for all 10,000 replications. Moreover, we are interested to know if the selected window widths vary a lot. Figure 5.7 summarizes for all 10,000 replications of the data generating process (C0) how many changepoints were detected and which window widths were selected predominantly by the OF approach. We observe that almost always (in more than

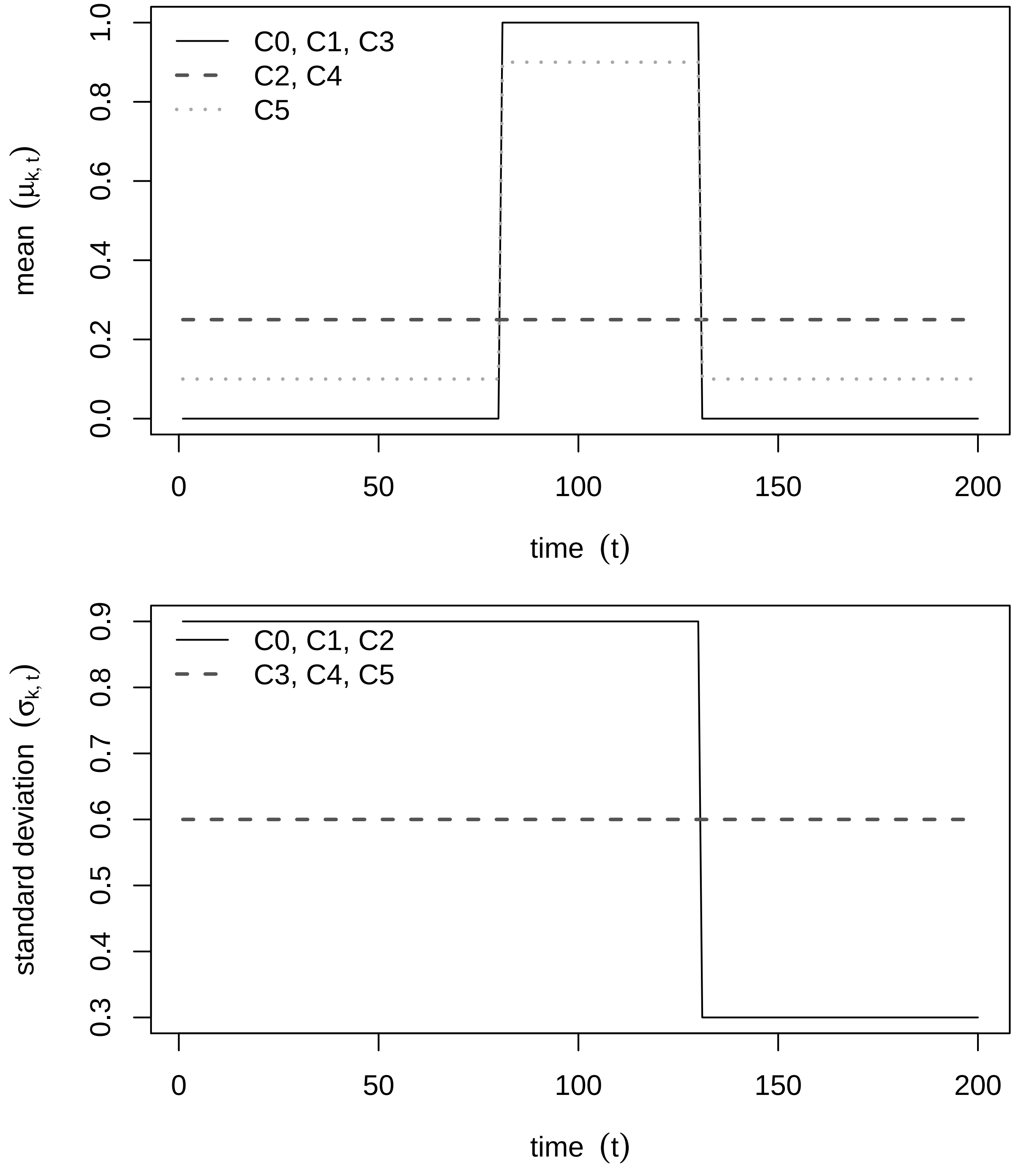


Figure 5.5: **Changepoint scenario (C):** Temporal evolution of mean $\mu_{k,t}$ (upper panel) and standard deviation $\sigma_{k,t}$ (lower panel) for the data generating process (C0) ($k = 0$) and the Models (C1)–(C5) ($k = 1, \dots, 5$).

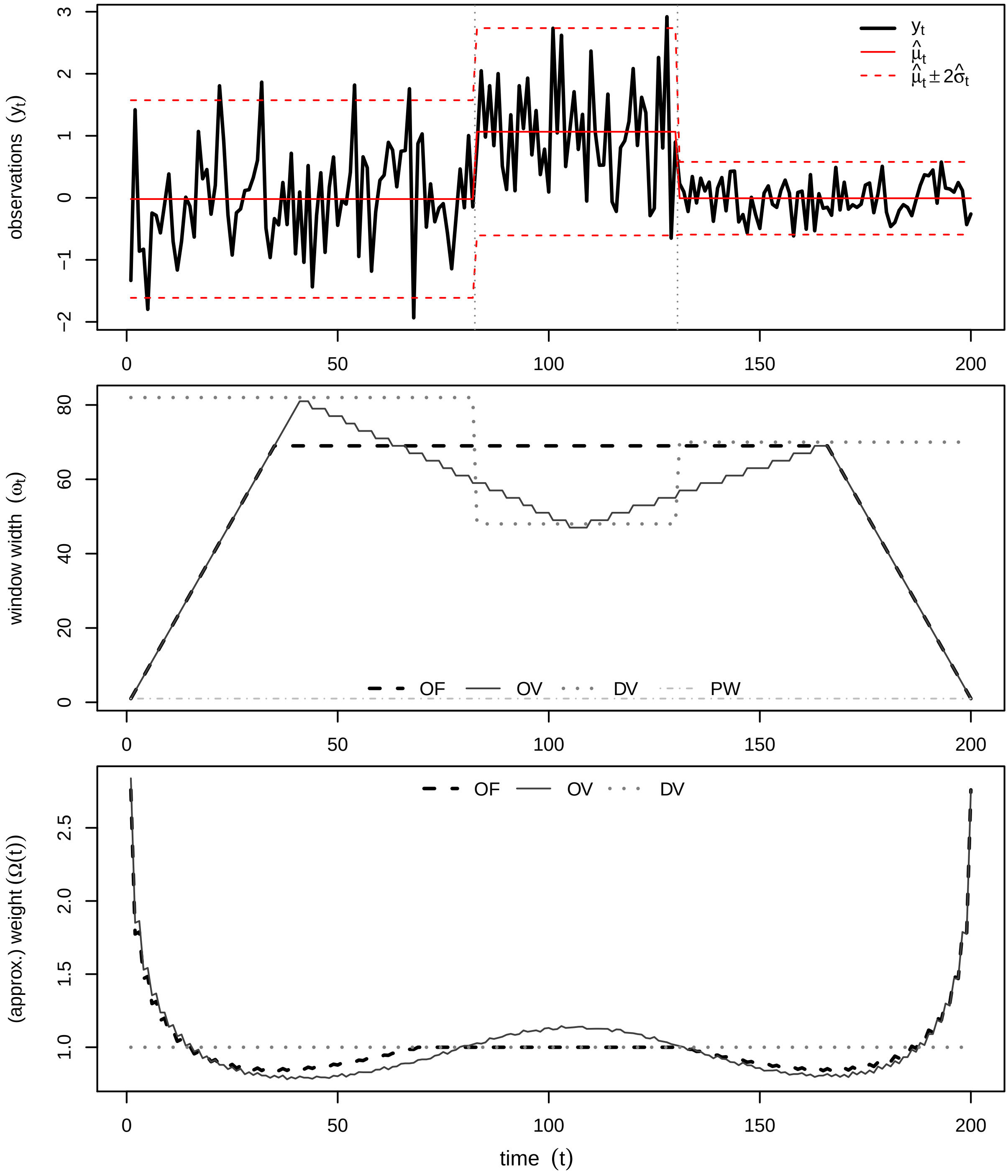


Figure 5.6: **Changepoint scenario (C):** One replication of the data generating process (C0) (observations y_t, $t = 1, \ldots, N$) and changepoint analysis. The detected segmentation (gray dotted vertical lines) and the mean and standard deviation estimates $\widehat{\mu}_t$ and $\widehat{\sigma}_t$ for each segment (red) are indicated (upper panel). The corresponding moving window widths ω_t chosen by the OF (dashed), OV (solid) and DV (dotted) selection approach, as well as the window width $\omega_t^{\mathrm{PW}} = 1$ used in the case of a point-wise (PW) evaluation (dash-dotted) are compared in the middle panel. The resulting (approximate) weights $\Omega(t)$ for each time instance t (see Equation (5.15)) are illustrated in the bottom panel.

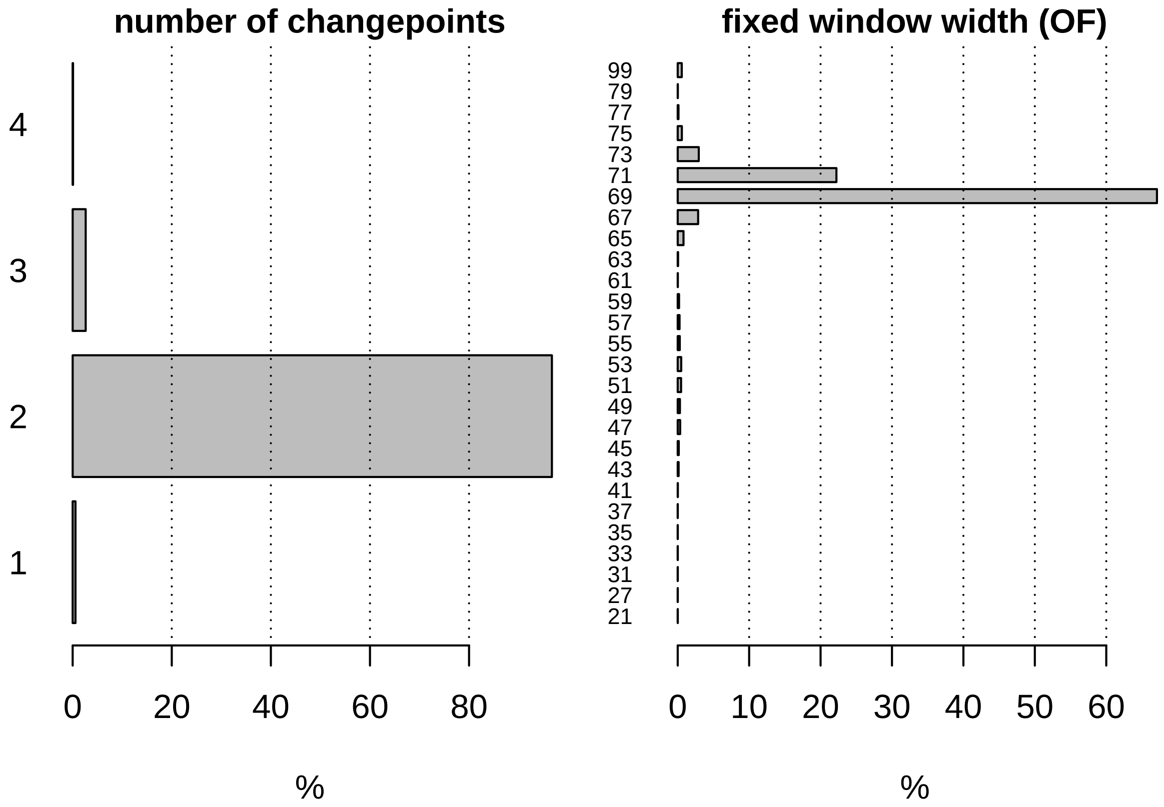


Figure 5.7: **Changepoint scenario (C):** Share (in %) of all 10,000 replications, where the given number of changepoints was detected (left panel), and where the given fixed window width ω^{OF} was selected by the OF approach (right panel).

95% of all cases) the correct number of changepoints (2) was detected and that in most cases the fixed OF-window width was either 69 or 71. Hence, in most cases a 3-segmentation close to the true segmentation was identified.

Results (model-wise comparison of different evaluation approaches) Figures 5.8–5.12 illustrate the temporal evolution of the empirical (moving) CRPS and IQ distances and their deviation from their theoretical counterparts, for Models (C1)–(C5), respectively. We observe the following:

(C) For all models (Figures 5.8–5.12):

- The PW-scores and divergences deviate considerably from the theoretical scores and divergences, respectively.
- For the start and the end of the time series (border case), the OF- and OV-scores and divergences tend towards the PW-scores and divergences, respectively.
- For most models (and most time instances) the deviation of the ST-scores and divergences from the theoretical scores and divergences, respectively, exceeds that of the moving (OF-, OV- and DV-) scores and divergences. Exceptions are addressed below.
- The simultaneous change of mean and variance affects the OF- and OV-scores and divergences more than the sole change of the mean.

- In most cases, the deviation of the moving (OF-, OV- and DV-) scores and divergences from the theoretical scores and divergences, respectively, is most pronounced close to the changepoints, where the deviation is the smallest for DV-windows.

(C1) True model (Figure 5.8):

- The theoretical CRPS remain (approximately) constant as long as the variance does not change. Their temporal evolution is not affected by a sole change in the mean. (The theoretical IQ distances are—by definition—constantly 0.)
- While the DV-scores approximate the theoretical scores best, the OV-scores yield an improvement compared to the OF-scores.
- Here, the ST-divergence approximates the theoretical divergences best. Despite the presence of non-stationarity this behavior can be observed for the true model. This is because the true model specifies the characteristics of the data generating process correctly for each single time instance. Construction of an empirical CDF based on the sample given by all realizations of the corresponding time series then (approximately) yields the same distribution which we obtain in case of the data generating process. Hence, the ST-divergence gets close to the theoretical divergences which are all 0. All other (moving) divergences also come very close to the theoretical divergences.

(C2) Constant mean (Figure 5.9):

- The theoretical scores and divergences jump in both changepoints.
- All in all, the DV-scores and divergences approximate their theoretical counterparts best.
- Again, the OV-scores and divergences yield a slight improvement compared to the OF-scores and divergences, respectively.

(C3) Constant variance (Figure 5.10):

- The theoretical scores and divergences jump only in the second changepoint (change of mean and variance).
- While the DV-scores approximate their theoretical counterparts best, no such general statement can be made for the moving divergences. The DV-divergences are better for the start and the end of the time series (border case), while in between, the OF- and OV-divergences approximate their theoretical counterparts slightly better.

(C4) Constant mean and variance (Figure 5.11):

- While for most time instances all moving scores approximate their theoretical counterpart very well, the DV-score performs better for the start and the end of the time series (border case).
- The DV-divergences approximate their theoretical counterparts comparatively best.

(C5) Wrong mean and variance (Figure 5.12):

- As the model misspecification does not differ much from model (C3), we observe a similar temporal evolution of the moving scores and divergences.

All in all, we conclude that the moving scores and divergences based on DV-windows are most suitable for model evaluation under the presence of changepoints.

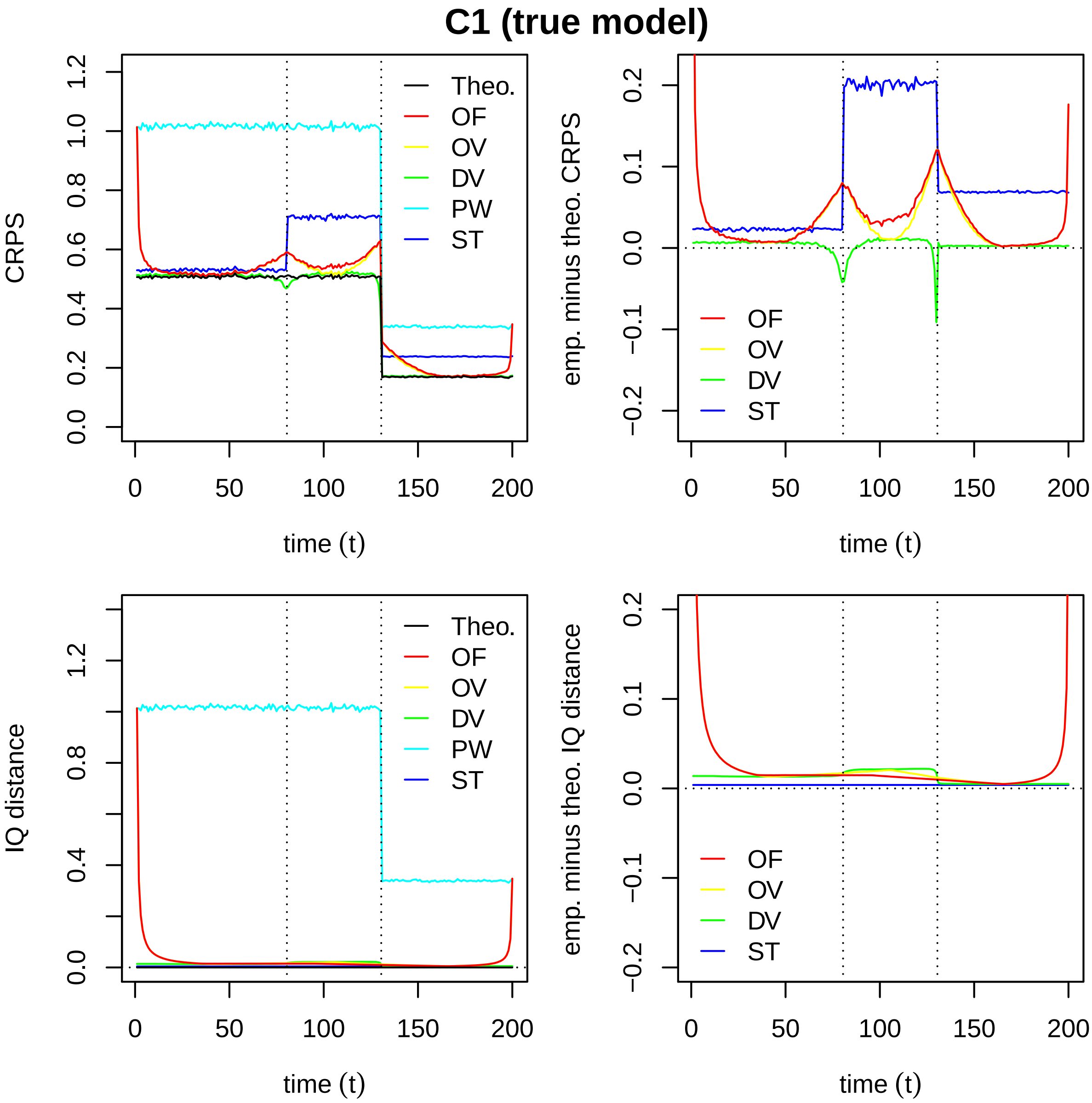


Figure 5.8: **Changepoint scenario (C1/true model):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

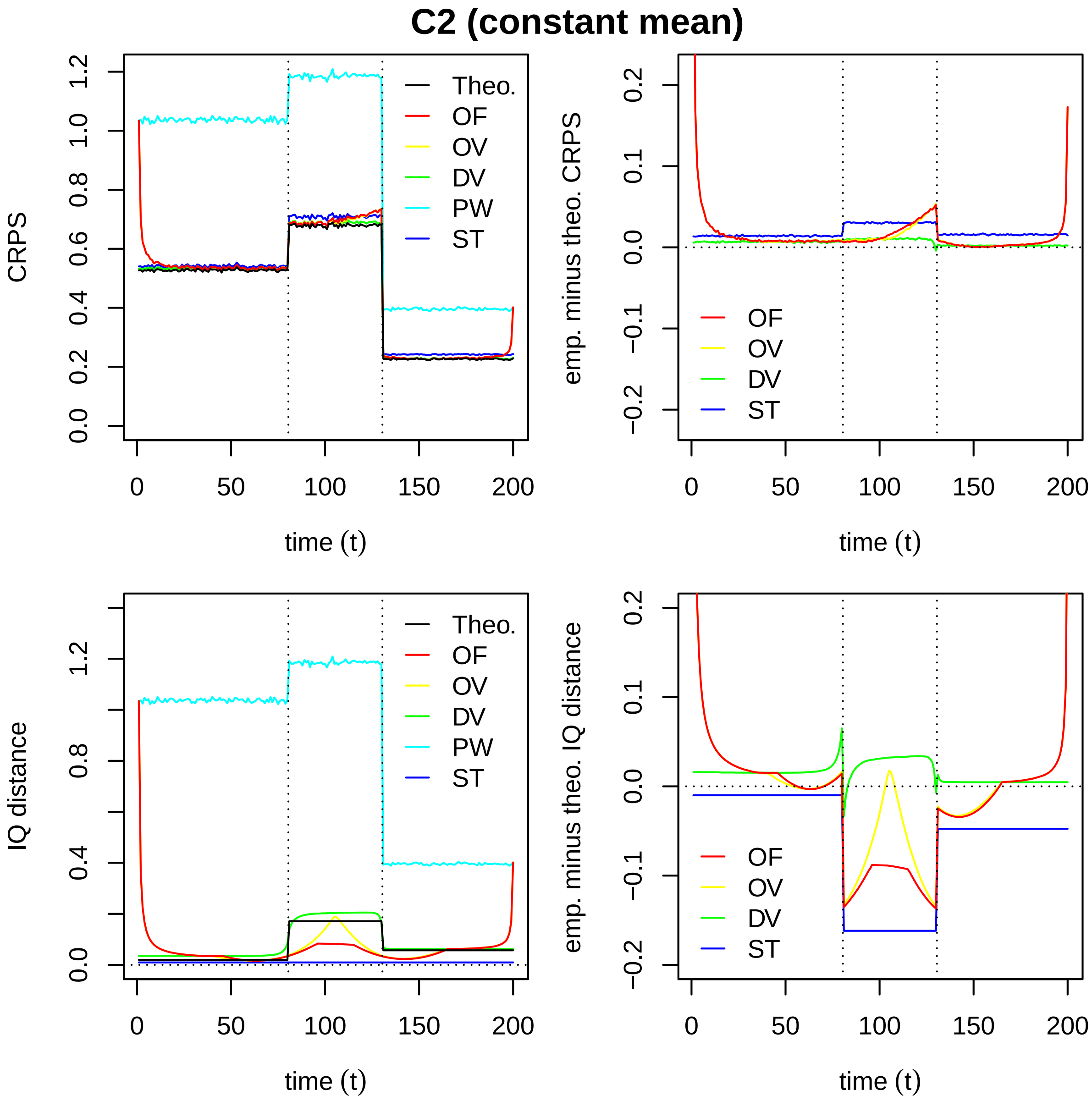


Figure 5.9: **Changepoint scenario (C2/constant mean):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

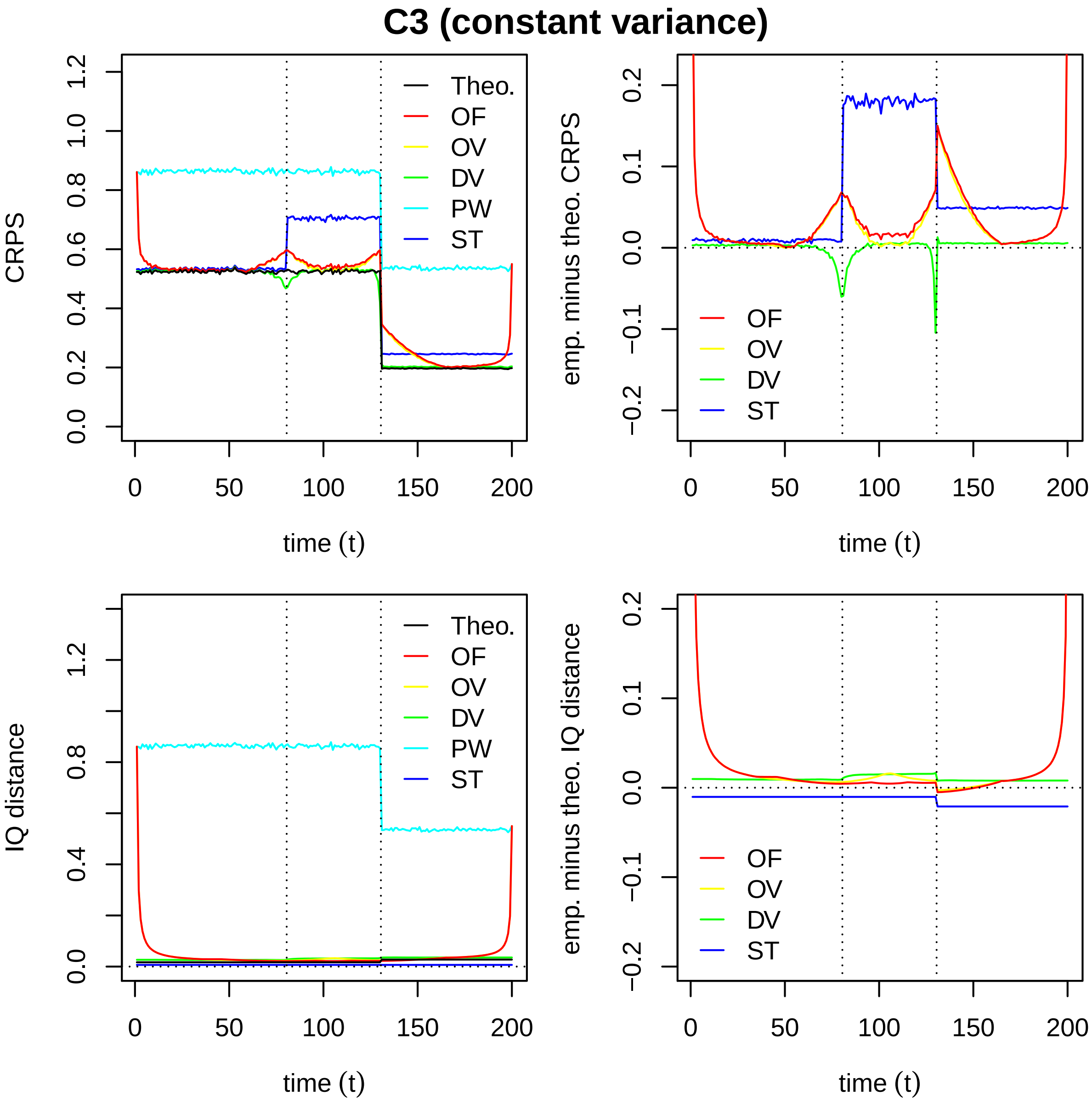


Figure 5.10: **Changepoint scenario (C3/constant variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

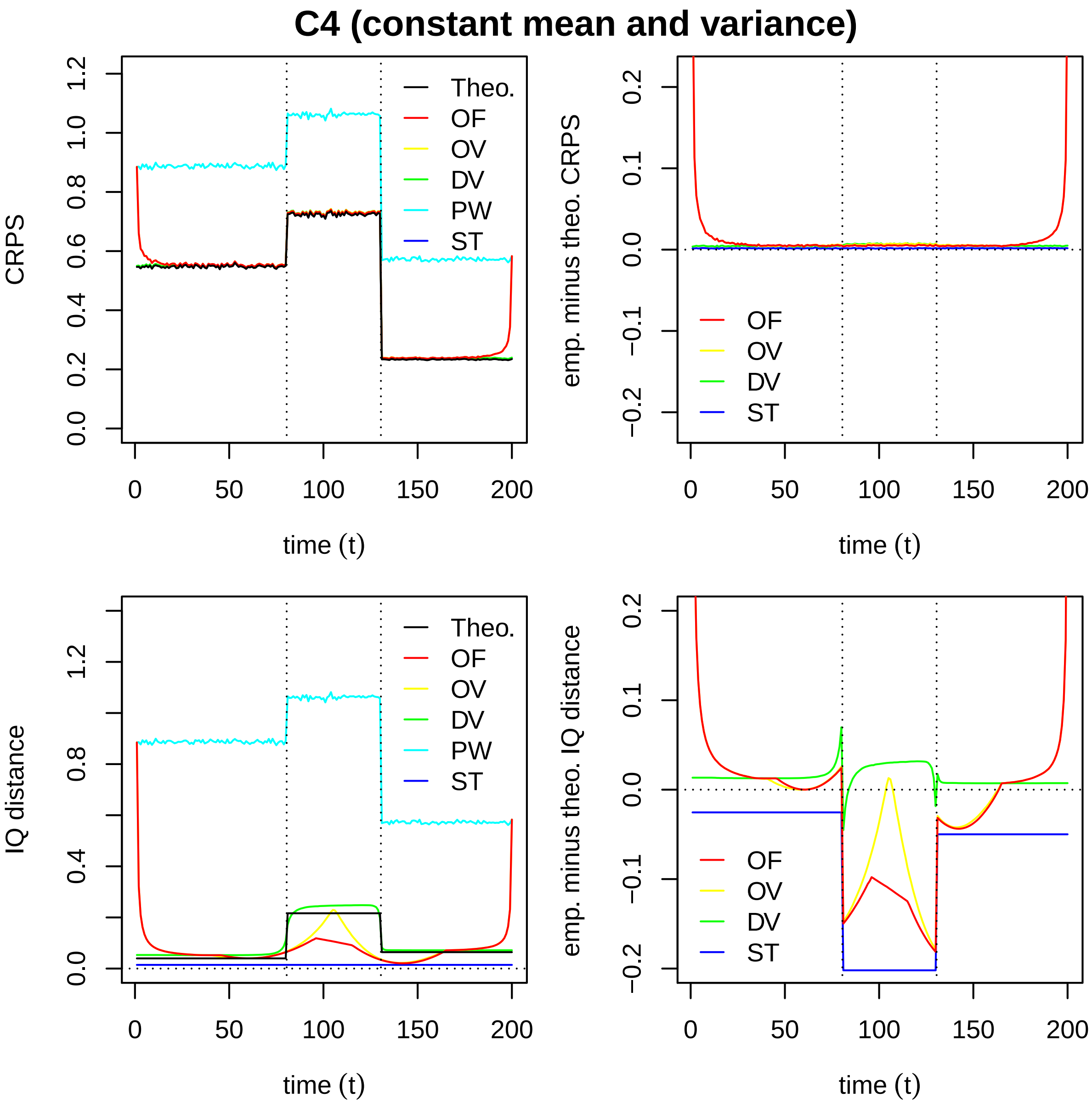


Figure 5.11: **Changepoint scenario (C4/constant mean and variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

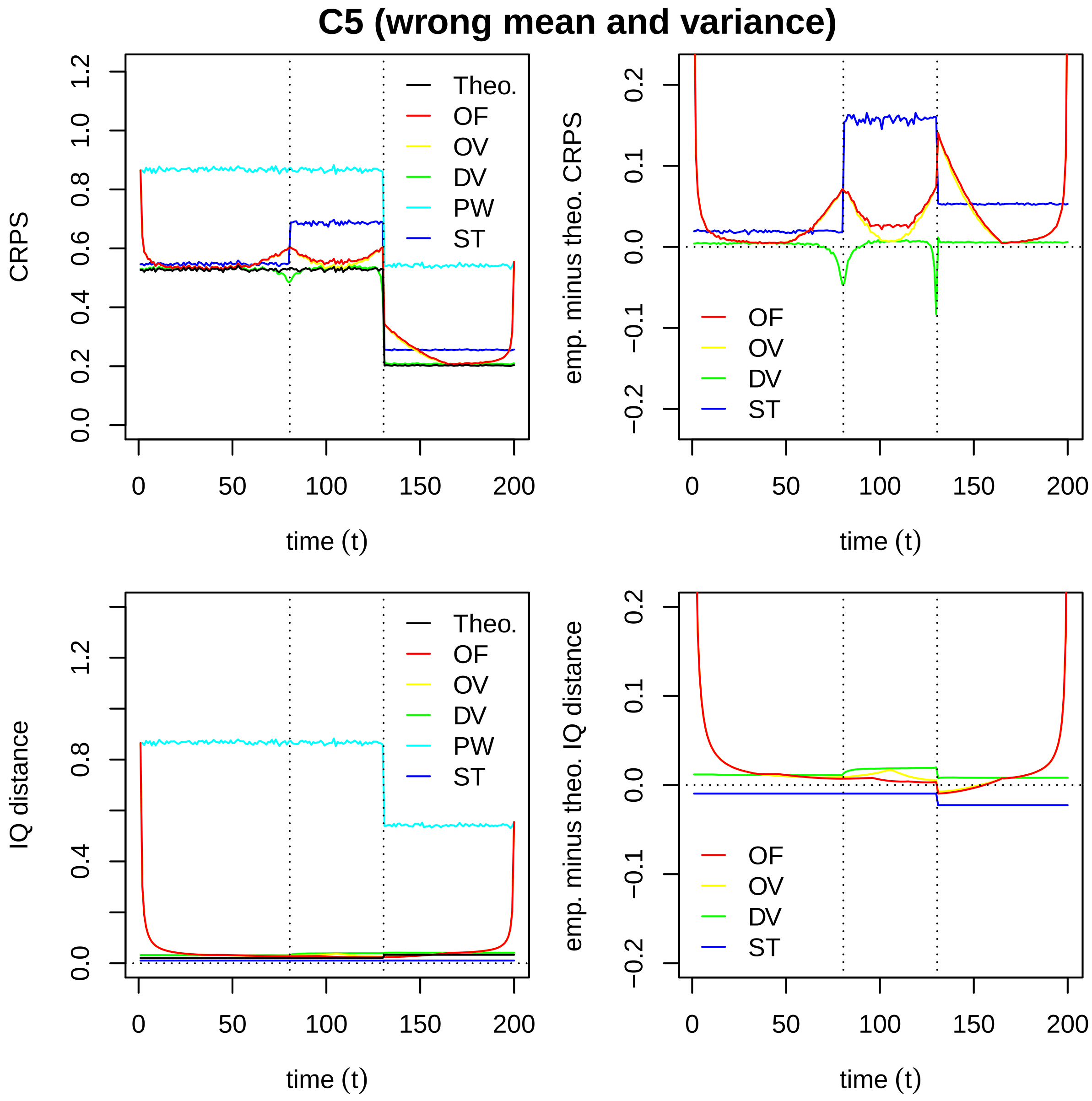


Figure 5.12: **Changepoint scenario (C5/wrong mean and variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

Results (model comparison based on different evaluation approaches) Figures 5.13 and 5.14 compare Models (C1)–(C5) based on (time series of) theoretical/empirical CRPS and IQ distances, respectively. As before, we consider theoretical (Theo.), moving OF-, moving OV-, moving DV-, point-wise (PW) and ST-scores/divergences. For the scores (Figure 5.13) we observe the following:

- Theoretical (Theo.) scores:
  - The true model (C1) is strictly the best. The differences between the five models can be explained by their different misspecifications of mean and variance.
  - For the first and last segment of the time series, the scores of all models move closely together, as their means deviate not or not much from the data generating process.
  - For the middle segment of the time series, the scores of Models (C2) and (C4) jump up, as their means deviate more from the mean of the data generating process.
  - After the variance of the data generating process drops for the last segment of the time series the scores drop as well.
  - Model (C3) performs worse than (C1) and Model (C4) performs worse than (C2), due to a misspecification of the variance.
  - Model (C5) performs worse than (C3), due to its misspecification of the mean.
  - The misspecification of Model (C4) is worst for all segments of the time series.
- Moving scores with OF-windows:
  - For the start and the end of the time series (border case), the scores of all models increase together.
  - For the first and the last segment of the time series, the models are ranked falsely, as we approach the changepoints.
  - For the middle segment of the time series, approaching the second changepoint, the true model (C1) gets ranked falsely.
- Moving scores with OV-windows:
  - The scores evolve similar to those for the OF-windows. Our observations are the same.
- Moving scores with DV-windows:
  - For most time instances the models are ranked appropriately.
  - Wrong rankings occur only very close to the changepoints.
- Point-wise (PW) scores:
  - For the first and the middle segment of the time series, Models (C1) and (C2) get a far too bad evaluation.
  - For the last segment of the time series, Models (C3) and (C5) are evaluated too bad.
- Scores under the assumption of stationarity (ST):
  - For all segments of the time series the scores for the different models do not differ very much from one another.
  - The observed rankings are wrong for all segments of the time series.

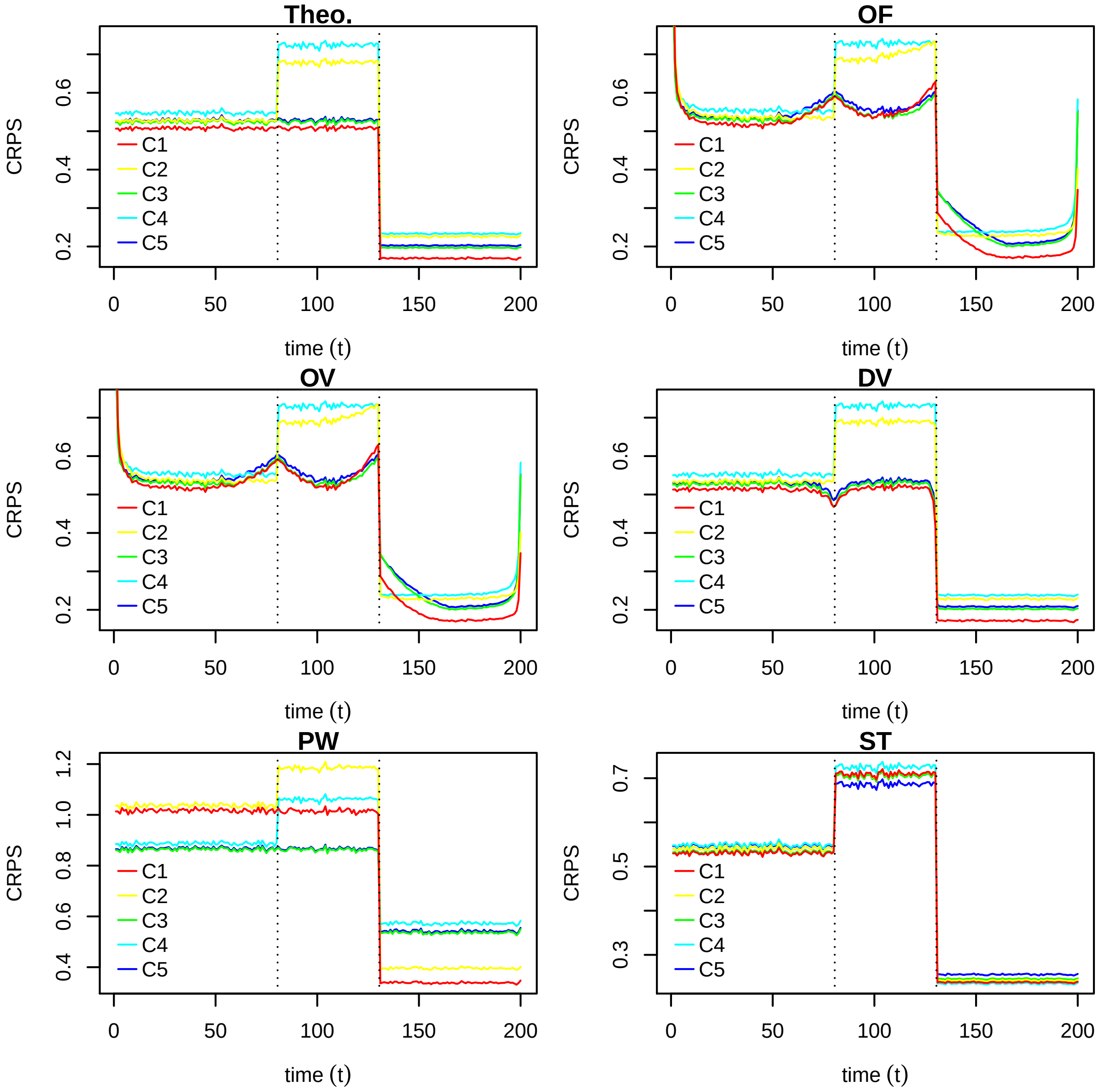


Figure 5.13: **Changepoint scenario (C):** Comparison of Models (C1)–(C5) based on (time series of) theoretical/empirical CRPS. Comparison based on theoretical scores (upper left panel), moving scores computed using the OF approach (upper right panel), moving scores computed using the OV approach (middle left panel), moving scores computed using the DV approach (middle right panel), point-wise (PW) scores (lower left panel) and based on ST-scores under the assumption of stationarity (lower right panel).

For the divergences (Figure 5.14) we observe the following:

- Theoretical (Theo.) divergences:
  - The true model (C1) is strictly the best. The differences between the five models can be explained by their different misspecifications of mean and variance.
  - For the middle segment of the time series, the divergences of Models (C2) and (C4) jump up, as their means deviate more from the mean of the data generating process.
  - After the variance of the data generating process drops for the last segment of the time series, the divergences of Models (C3) and (C5) jump up a little bit, which is due to their misspecification of the variance.
  - Model (C3) performs worse than (C1) and Model (C4) performs worse than (C2), due to a misspecification of the variance.
  - Model (C5) performs worse than (C3), due to its misspecification of the mean.
  - The misspecification of Model (C4) is worst for all segments of the time series.
  - The models are ranked in the same way as for the CRPS (see above).
- Moving divergences with OF-windows:
  - For the start and the end of the time series (border case), the divergences of all models increase together.
  - For the first segment of the time series, the ranking of Model (C2) is mostly false.
  - For the last segment of the time series, the divergences of Models (C2) and (C4) are partly to low.
  - The models are ranked correctly for the middle segment of the time series.
- Moving divergences with OV-windows:
  - For the middle segment of the time series, the divergences evolve different compared to those for the OF-windows.
  - In terms of model ranking, we make the same observations as for the OF-windows.
- Moving divergences with DV-windows:
  - For most time instances the models are ranked appropriately.
  - Only for the first segment of the time series, where the Models (C2), (C3) and (C5) perform very similar in terms of theoretical divergences, Model (C2) is ranked falsely.
- Point-wise (PW) divergences:
  - The point-wise IQ distances are equal to the point-wise CRPS. Hence, we make the same observations as above.
- Divergences under the assumption of stationarity (ST):
  - The ST-divergences do not differentiate between the different segments.
  - Model (C5) is ranked wrong for the middle and the last segment of the time series.

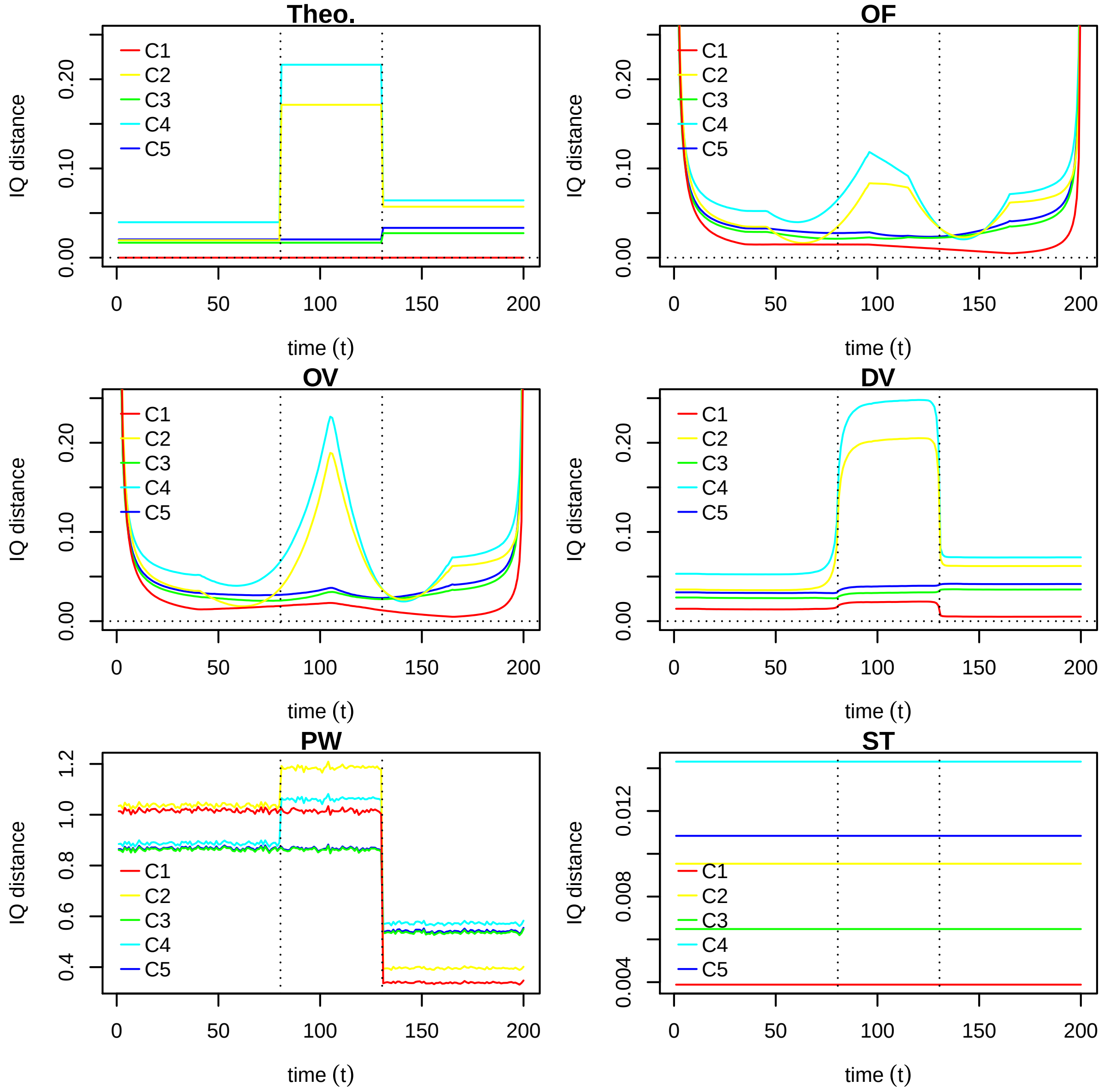


Figure 5.14: **Changepoint scenario (C):** Comparison of Models (C1)–(C5) based on (time series of) theoretical/empirical IQ distances. Comparison based on theoretical divergences (upper left panel), moving divergences computed using the OF approach (upper right panel), moving divergences computed using the OV approach (middle left panel), moving divergences computed using the DV approach (middle right panel), point-wise (PW) divergences (lower left panel) and based on ST-divergences under the assumption of stationarity (lower right panel).

Table 5.2: **Changepoint scenario (C):** Comparison/ranking of Models (C1)–(C5) based on average (empirical/theoretical) SE scores, CRPS, MV divergences and IQ distances. Average scores/divergences (left) and corresponding model rankings (right) are provided, distinguishing between theoretical scores/divergences (Th) and empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST). The rankings based on the theoretical scores/divergences (Th) are considered as the true model rankings, which are used to judge the rankings given by the empirical scores/divergences. Note that moving divergence averages based on OF- and OV-windows do not warrant propriety.

		average scores/divergences						model rank					
		Th	OF	OV	DV	PW	ST	Th	OF	OV	DV	PW	ST
SE score	(C1)	0.558	0.637	0.627	0.559	1.117	0.748	1	2	2	2	4	3
	(C2)	0.746	0.764	0.765	0.754	1.305	0.748	4	5	5	5	5	3
	(C3)	0.558	0.632	0.622	0.556	0.919	0.747	1	1	1	1	1	1
	(C4)	0.746	0.759	0.759	0.751	1.107	0.747	4	4	4	4	3	1
	(C5)	0.568	0.648	0.638	0.570	0.929	0.750	3	3	3	3	2	5
CRPS	(C1)	0.389	0.425	0.422	0.392	0.779	0.473	1	1	1	1	3	1
	(C2)	0.460	0.476	0.476	0.466	0.850	0.479	4	4	4	4	5	3
	(C3)	0.410	0.441	0.439	0.411	0.749	0.475	2	2	2	2	1	2
	(C4)	0.482	0.494	0.495	0.487	0.821	0.483	5	5	5	5	4	5
	(C5)	0.414	0.449	0.447	0.417	0.753	0.480	3	3	3	3	2	4
MV div.	(C1)	0.000	0.036	0.038	0.018	1.117	0.005	1	2	2	2	4	3
	(C2)	0.188	0.111	0.136	0.214	1.305	0.005	4	5	5	5	5	3
	(C3)	0.000	0.032	0.033	0.015	0.919	0.004	1	1	1	1	1	1
	(C4)	0.188	0.106	0.131	0.210	1.107	0.004	4	4	4	4	3	1
	(C5)	0.010	0.037	0.039	0.029	0.929	0.007	3	3	3	3	2	5
IQ dist.	(C1)	0.000	0.027	0.028	0.012	0.779	0.004	1	1	1	1	3	1
	(C2)	0.071	0.060	0.068	0.086	0.850	0.009	4	4	4	4	5	3
	(C3)	0.020	0.042	0.044	0.031	0.749	0.006	2	2	2	2	1	2
	(C4)	0.092	0.075	0.085	0.107	0.821	0.014	5	5	5	5	4	5
	(C5)	0.025	0.047	0.048	0.037	0.753	0.011	3	3	3	3	2	4

Model ranking based on average scores and divergences To have an overview which evaluation approaches rank the models (C1)–(C5) correctly and which do not, we provide Table 5.2. Besides the CRPS and the IQ distance, we now also consider SE scores and MV divergences. The table provides the average scores and divergences (averages over all time instances and all 10,000 replications of the simulation study) for the different considered evaluation approaches (OF, OV, DV, PW and ST) and compares them to the corresponding theoretical (Th) scores and divergences, respectively. Note that utilization of moving divergence averages is not necessarily a proper evaluation method (that is in case of OF- and OV-windows; see discussion in Sections 5.3.2 and 5.3.3). Based on the average scores/divergences we rank the different models, distinguishing between the theoretical scores/divergences and the five different evaluation approaches. Hence, for each approach (column) and separately for each score/divergence type (SE score, CRPS, MV divergence, IQ distance), the model with the smallest average score/divergence is ranked best (1) and the model with the highest average score/divergence is ranked worst (5). Accordingly, the models with intermediate scores/divergences are ranked 2–4. If the score/divergence averages of two or more models are equal, all of these models get the same (minimum) rank. The resulting model rankings are compared in the right half of the table.

To decide if an evaluation approach is good or not, we consider the following two criteria. For good evaluation approaches

(E1) the true model (C1) should be ranked lowest (best), and

(E2) the ranking should be identical to that based on the theoretical (Th) scores/divergences.

All in all, we observe that (E1) and (E2) are fulfilled for the CRPS and the IQ distances calculated based on the OF, OV and DV approach.

Now, we discuss the results from Table 5.2 in more detail, distinguishing between the different types of scores and divergences:

- SE scores:
  - The theoretical scores can not differentiate between Models (C1) and (C3), and also not between (C2) and (C4). This is due to the fact, that SE scores neglect misspecifications of the variance.
  - The moving (OF-, OV- and DV-) scores falsely rank (C3) better than (C1), and (C4) better than (C2).
  - The rankings based on the PW- and ST-scores are both wrong.
  - The ST-scores are unable to differentiate between Models (C1) and (C2), and also between (C3) and (C4).
- CRPS:
  - The scores are able to differentiate between all five models.
  - The moving (OF-, OV- and DV-) scores rank all five models correctly.
  - The PW-scores yield a faulty model ranking.
  - The ST-scores rank Models (C2) and (C5) wrong.
- MV divergences:
  - We obtain the same (faulty) model rankings as for the SE scores.
- IQ distances:
  - Same as for the CRPS, the moving (OF, OV and DV) IQ distances rank all five models correctly.
  - The rankings based on the PW- and ST-divergences (same as for the CRPS) are both wrong.

Further analysis of the model rankings Whereas in Table 5.2 we investigated model rankings which were obtained by averaging over all 10,000 replications of the simulation study (we refer to them as *average based model rankings*), we now look into all 10,000 replications separately, to see if the individual model rankings correspond to the *average based model rankings* found in Table 5.2. For this we consider the models pair-wise and calculate the share of all 10,000 replications for which the average (moving) score/divergence of the first model is smaller (better) than the average (moving) score/divergence of the second model. This time the (moving) scores/divergences are averaged only temporally, that is over all available time instances.

Table 5.3: **Changepoint scenario (C):** Pair-wise comparison of Models (C1)–(C5), distinguishing between theoretical CRPS (Th), moving CRPS computed using the OF, OV and DV approach, point-wise CRPS (PW) and CRPS under the assumption of stationarity (ST). Each table entry equals the share of all 10,000 replications of the simulation study for which the average (moving) score of the model indicated in the corresponding row is smaller (better) than the average (moving) score of the model indicated in the corresponding column of the table.

			Th					**OF**		
	(C1)	(C2)	(C3)	(C4)	(C5)	(C1)	(C2)	(C3)	(C4)	(C5)
(C1)	–	1.00	1.00	1.00	1.00	–	1.00	0.99	1.00	1.00
(C2)	0.00	–	0.00	1.00	0.00	0.00	–	0.01	0.99	0.02
(C3)	0.00	1.00	–	1.00	0.86	0.01	0.99	–	1.00	0.96
(C4)	0.00	0.00	0.00	–	0.00	0.00	0.01	0.00	–	0.00
(C5)	0.00	1.00	0.14	1.00	–	0.00	0.98	0.04	1.00	–
			PW					**OV**		
	(C1)	(C2)	(C3)	(C4)	(C5)	(C1)	(C2)	(C3)	(C4)	(C5)
(C1)	–	0.99	0.04	0.90	0.08	–	1.00	0.99	1.00	1.00
(C2)	0.00	–	0.00	0.05	0.00	0.00	–	0.02	0.99	0.02
(C3)	0.96	1.00	–	1.00	0.74	0.01	0.98	–	1.00	0.95
(C4)	0.10	0.95	0.00	–	0.00	0.00	0.01	0.00	–	0.00
(C5)	0.92	1.00	0.26	1.00	–	0.00	0.98	0.05	1.00	–
			ST					**DV**		
	(C1)	(C2)	(C3)	(C4)	(C5)	(C1)	(C2)	(C3)	(C4)	(C5)
(C1)	–	0.92	0.87	0.98	0.91	–	1.00	1.00	1.00	1.00
(C2)	0.08	–	0.17	0.99	0.65	0.00	–	0.01	0.99	0.01
(C3)	0.13	0.83	–	0.98	0.91	0.00	0.99	–	1.00	0.87
(C4)	0.02	0.01	0.02	–	0.16	0.00	0.01	0.00	–	0.00
(C5)	0.09	0.35	0.09	0.84	–	0.00	0.99	0.13	1.00	–

The results of these computations (shares) are summarized in Tables 5.3 and 5.4. Table 5.3 is based on CRPS, Table 5.4 on IQ distances. Both tables again distinguish between theoretical scores/divergences (Th) and empirical scores/divergences calculated using different approaches (OF, OV, DV, PW and ST). Each block of Tables 5.3 and 5.4 corresponds to one particular approach. It is best to read each block row-wise. Each row provides the shares telling how often the model indicated to the left (model under consideration) had a better average score/divergence than the other models (indicated above). A share equal to 0 means that the model under consideration is never considered better than the other model, a share equal to 1 means that it is always considered better than the other one, a share in $(0, 0.5)$ means that the model under consideration is considered worse than the other one for the majority of all replications and a share in $(0.5, 1)$ means that it is considered better than the other one for the majority of all replications.

First, we are interested to know, if the pair-wise and replication-wise consideration in Tables 5.3 and 5.4 supports the model rankings found in Table 5.2. If we rank model A better than model B if the share corresponding to the model pair (A, B) is greater than 0.5, then we obtain exactly the same rankings as in Table 5.2.

Second, we are interested to see, how certain the different evaluation approaches are about the relative pair-wise rankings they indicate. Table entries close to 0 or 1 support certainty,

Table 5.4: **Changepoint scenario (C):** Pair-wise comparison of Models (C1)–(C5), distinguishing between theoretical IQ distances (Th), moving IQ distances computed using the OF, OV and DV approach, point-wise IQ distances (PW) and IQ distances under the assumption of stationarity (ST). Each table entry equals the share of all 10,000 replications of the simulation study for which the average (moving) divergence of the model indicated in the corresponding row is smaller (better) than the average (moving) divergence of the model indicated in the corresponding column of the table.

	Th					**OF**				
	(C1)	(C2)	(C3)	(C4)	(C5)	(C1)	(C2)	(C3)	(C4)	(C5)
(C1)	–	1.00	1.00	1.00	1.00	–	0.99	0.99	1.00	0.99
(C2)	0.00	–	0.00	1.00	0.00	0.01	–	0.11	0.98	0.13
(C3)	0.00	1.00	–	1.00	1.00	0.01	0.89	–	0.99	0.82
(C4)	0.00	0.00	0.00	–	0.00	0.00	0.02	0.01	–	0.00
(C5)	0.00	1.00	0.00	1.00	–	0.01	0.87	0.18	1.00	–
	PW					**OV**				
	(C1)	(C2)	(C3)	(C4)	(C5)	(C1)	(C2)	(C3)	(C4)	(C5)
(C1)	–	0.99	0.04	0.90	0.08	–	1.00	0.99	1.00	0.99
(C2)	0.00	–	0.00	0.05	0.00	0.00	–	0.07	0.99	0.07
(C3)	0.96	1.00	–	1.00	0.74	0.01	0.93	–	0.99	0.82
(C4)	0.10	0.95	0.00	–	0.00	0.00	0.01	0.01	–	0.00
(C5)	0.92	1.00	0.26	1.00	–	0.01	0.93	0.18	1.00	–
	ST					**DV**				
	(C1)	(C2)	(C3)	(C4)	(C5)	(C1)	(C2)	(C3)	(C4)	(C5)
(C1)	–	0.92	0.87	0.98	0.91	–	1.00	1.00	1.00	1.00
(C2)	0.08	–	0.17	0.99	0.65	0.00	–	0.01	0.99	0.01
(C3)	0.13	0.83	–	0.98	0.91	0.00	0.99	–	1.00	0.87
(C4)	0.02	0.01	0.02	–	0.16	0.00	0.01	0.00	–	0.00
(C5)	0.09	0.35	0.09	0.84	–	0.00	0.99	0.13	1.00	–

entries close to 0.5 indicate that the evaluation approach under consideration has difficulties to decide in favor of one of the corresponding models. In most cases we observe shares close to 0 or 1. We point out cases, where this does not hold. The ST-scores and divergences for instance have difficulties with the relative ranking of models (C2) and (C5). The theoretical CRPS mostly allow for unambiguous decisions, only the decision between models (C3) and (C5) is not always the same. The DV-scores seem to mirror this behavior best. In contrast, the theoretical IQ distances always allow for unambiguous decisions. Generally, the shares obtained for OF-, OV- and DV-scores and divergences come very close to their theoretical counterparts.

All in all, we conclude that the average based model rankings found in Table 5.2 (for CRPS and IQ distances) and the differentiation between the different models in these rankings are justified.

5.6 Simulation study: Trend scenario

For the trend scenario (T) we consider trends in the mean and standard deviation of a time series. To model these trends, we consider the function

$$h_{\mathrm{T}}(t;\boldsymbol{\theta}) \coloneqq \theta_0 + \theta_1 t \exp\left(\theta_2 t\right), \tag{5.20}$$

with *intercept parameter* θ_0, *linear trend parameter* θ_1 and *non-linear trend parameter* θ_2, where $\boldsymbol{\theta} \coloneqq (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3$. Based on (5.20) we model the trend in the mean by

$$M_{\mathrm{T}}(t;\boldsymbol{\mu}) \coloneqq h_{\mathrm{T}}(t;\boldsymbol{\mu}), \text{ for all } t = 1, \ldots, N,$$

with parameter vector $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2) \in \mathbb{R}^3$. Hence, the mean grows exponentially, if $\mu_1, \mu_2 > 0$. The trend in the standard deviation is modeled as

$$S_{\mathrm{T}}(t;\boldsymbol{\sigma}) \coloneqq h_{\mathrm{T}}(t;\boldsymbol{\sigma}), \text{ for all } t = 1, \ldots, N,$$

with parameter vector $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \sigma_2) \in \mathbb{R}^3$. The standard deviation increases exponentially, if $\sigma_1, \sigma_2 > 0$.

Trend scenario (T) The subsequent equations define the *trend scenario (T)*, where we consider time series of length $N = 200$. The data generating process (T0) and the five models (T1)–(T5) are given by

(T0)	$Y_t \sim \mathcal{N}\left(M_{\mathrm{T}}(t;\boldsymbol{\mu}^0), \left[S_{\mathrm{T}}(t;\boldsymbol{\sigma}^0)\right]^2\right)$	(data generating process),
(T1)	$X_t^1 \sim \mathcal{N}\left(M_{\mathrm{T}}(t;\boldsymbol{\mu}^1), \left[S_{\mathrm{T}}(t;\boldsymbol{\sigma}^1)\right]^2\right)$	(true model),
(T2)	$X_t^2 \sim \mathcal{N}\left(M_{\mathrm{T}}(t;\boldsymbol{\mu}^2), \left[S_{\mathrm{T}}(t;\boldsymbol{\sigma}^2)\right]^2\right)$	(wrong mean),
(T3)	$X_t^3 \sim \mathcal{N}\left(M_{\mathrm{T}}(t;\boldsymbol{\mu}^3), \left[S_{\mathrm{T}}(t;\boldsymbol{\sigma}^3)\right]^2\right)$	(wrong variance),
(T4)	$X_t^4 \sim \mathcal{N}\left(M_{\mathrm{T}}(t;\boldsymbol{\mu}^4), \left[S_{\mathrm{T}}(t;\boldsymbol{\sigma}^4)\right]^2\right)$	(wrong mean and linear variance),
(T5)	$X_t^5 \sim \mathcal{N}\left(M_{\mathrm{T}}(t;\boldsymbol{\mu}^5), \left[S_{\mathrm{T}}(t;\boldsymbol{\sigma}^5)\right]^2\right)$	(wrong mean and constant variance),

with parameters

(T0)	$\boldsymbol{\mu}^0 = (0, 1/3, 2)/N$,	$\boldsymbol{\sigma}^0 = (20, 0.05, 2)/N$,	(data generating process),
(T1)	$\boldsymbol{\mu}^1 = \boldsymbol{\mu}^0$,	$\boldsymbol{\sigma}^1 = \boldsymbol{\sigma}^0$,	(true model),
(T2)	$\boldsymbol{\mu}^2 = (0, 1/3, 1.9)/N$,	$\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^0$,	(wrong mean),
(T3)	$\boldsymbol{\mu}^3 = \boldsymbol{\mu}^0$,	$\boldsymbol{\sigma}^3 = (20, 0.0375, 1.5)/N$,	(wrong variance),
(T4)	$\boldsymbol{\mu}^4 = \boldsymbol{\mu}^2$,	$\boldsymbol{\sigma}^4 = (20, 0.05, 0)/N$,	(wrong mean and linear variance),
(T5)	$\boldsymbol{\mu}^5 = \boldsymbol{\mu}^2$,	$\boldsymbol{\sigma}^5 = (20, 0, 0)/N$,	(wrong mean and constant variance).

Whereas the true model (T1) is equivalent to the data generating process (T0) in terms of its parametrization, the other Models (T2)–(T5) differ from the data generating process (T0) in at least one parameter.

Illustration of the trend scenario (T) The differences between the data generating process (T0) and the models (T1)–(T5) in the temporal evolution of the mean and the standard deviation are visualized in Figure 5.15. For the data generating process (T0) we see, that the mean $\mu_{0,t}$ and the variance $\sigma^2_{0,t}$ both grow exponentially. In comparison to the data generating process (T0)

- the mean $\mu_{2,t}$ of Model (T2) increases slower than $\mu_{0,t}$,
- the variance $\sigma^2_{3,t}$ of Model (T3) increases slower than $\sigma^2_{0,t}$,
- the mean $\mu_{4,t}$ of Model (T4) increases slower than $\mu_{0,t}$ and the variance $\sigma^2_{4,t}$ grows only linearly,
- the mean $\mu_{5,t}$ of Model (T5) increases slower than $\mu_{0,t}$ and the variance $\sigma^2_{5,t}$ remains constant.

Hence, the mean of Model (T2) is always below the mean observed for the data generating process (T0), where the variance is modeled correctly. Model (T3) models the exponentially increasing mean correctly, however, the variance it assumes is always to low. Both, Models (T4) and (T5) are wrong in terms of mean and variance, where the misspecification of the variance is less severe for Model (T4). The parameters of the true model (T1) coincide with those of the data generating process (T0). Thus, a ranking of all Models (T1)–(T5) should rank Model (T1) best.

Illustration of changepoint analysis and selection of moving windows Figure 5.16 illustrates one replication of the data generating process (T0) and the corresponding selection of moving windows according to Section 5.3.3. We observe that 7 irregularly spaced changepoints were detected. Obviously, the exponential growth of the mean and the variance play a role for the detection of the changepoints.

The segmentation into eight segments of different length is also reflected in the moving windows selected by the OV and DV approach. The minimum detected segment length is 13 and the maximum detected segment length is 42. Hence, the window width ω_t^{DV} takes values between 13 and 42. Apart from the border case (time instances t close to the start and the end of the time series), the window width ω_t^{OF} is constant and equals 21, and the window width ω_t^{OV} varies between 13 and 41. The difference of OF-, OV- and DV-windows in comparison to a point-wise (PW) evaluation ($\omega_t^{\mathrm{PW}} = 1$) become clear from the figure.

With the bottom panel of Figure 5.16 we illustrate the (approximate) weights $\Omega(t)$ (see Equation (5.15)) corresponding to each observation y_t, $t = 1, \dots, N$, in an evaluation based on moving divergence averages (5.12). As already discussed, the propriety of the metric (5.12) is given only if the weights $\Omega(t)$ are constant. From the figure we observe, that the DV-windows yield propriety, while the OF- and OV-windows do not. For the OF-windows, the weights $\Omega(t)$ deviate from 1 only in the beginning and in the end of the considered time interval (border case). We clearly see, that the metric based on OV-windows approximates propriety much worse than that based on OF-windows.

As Figure 5.16 is based on one replication only, we are further interested to see how many changepoints were detected for all 10,000 replications. Moreover, we are interested to know if the selected window widths vary a lot. Figure 5.17 summarizes for all 10,000 replications of the data generating process (T0) how many changepoints were detected and which window widths were selected predominantly by the OF approach. We observe that in most cases ($> 90\%$) five or six changepoints were detected. Moreover, we observe that in most cases the fixed OF-window width

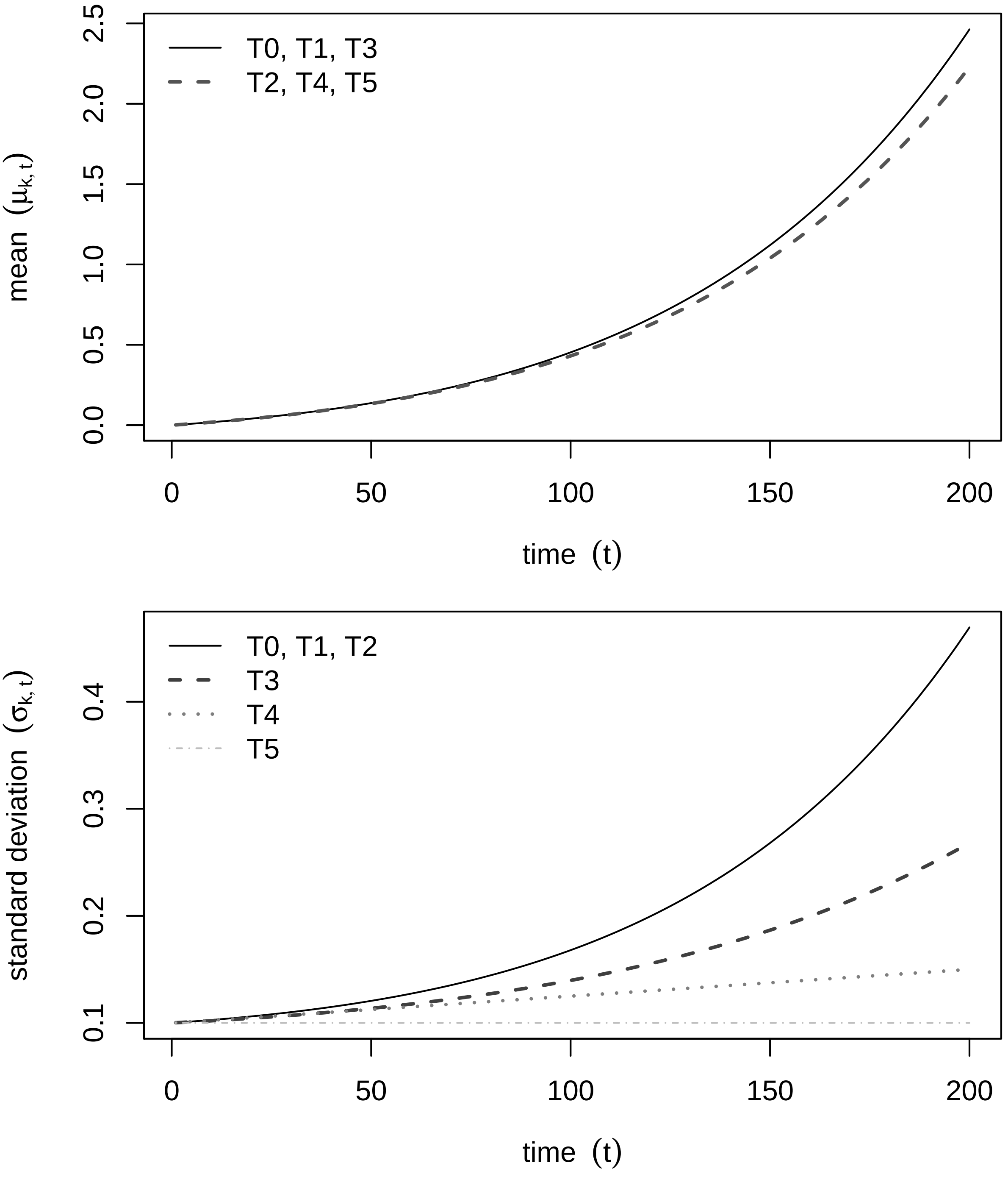


Figure 5.15: **Trend scenario (T):** Temporal evolution of mean $\mu_{k,t}$ (upper panel) and standard deviation $\sigma_{k,t}$ (lower panel) for the data generating process (T0) $(k = 0)$ and the models (T1)–(T5) $(k = 1, \ldots, 5)$.

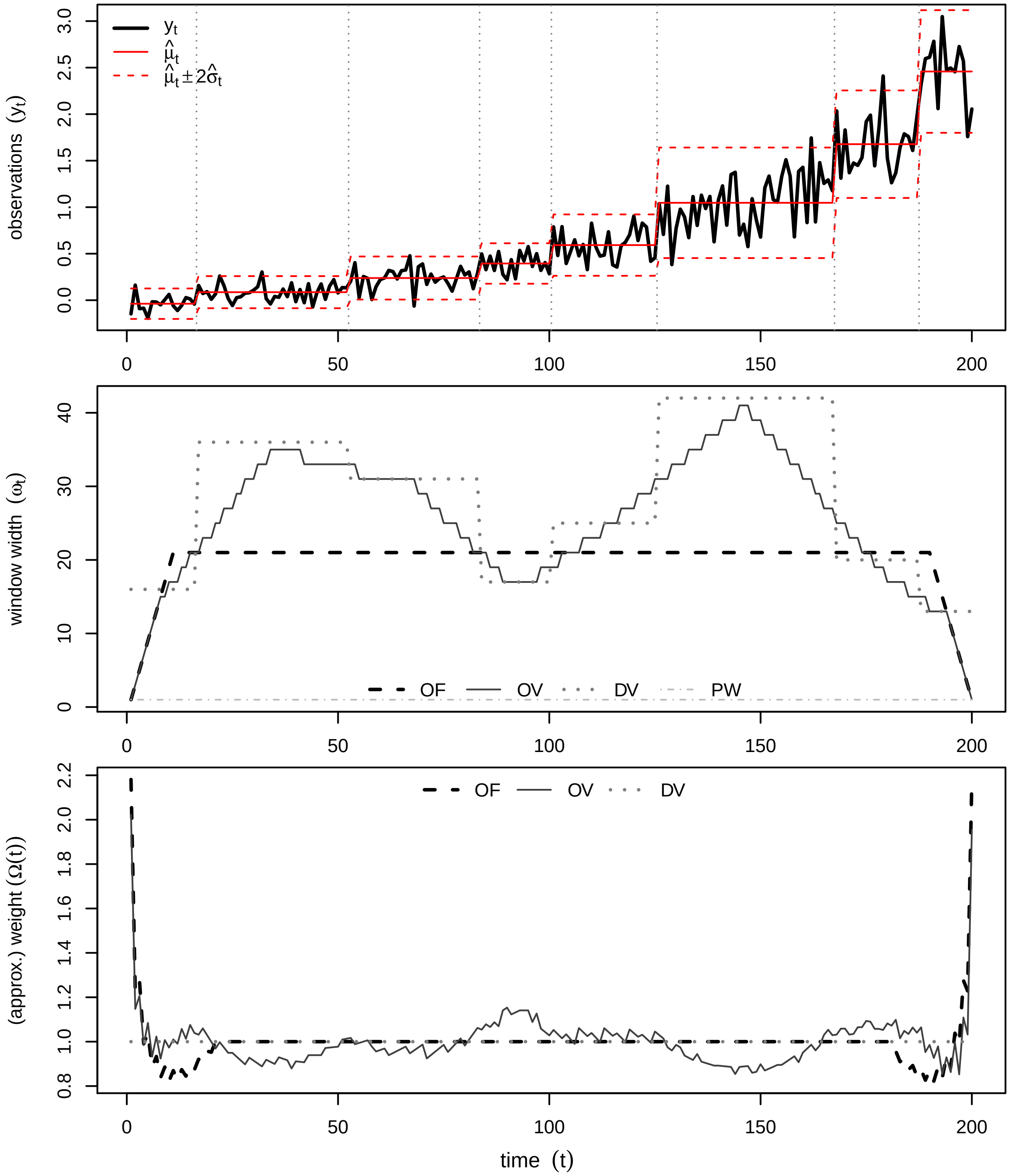


Figure 5.16: **Trend scenario (T):** One replication of the data generating process (T0) (observations y_t, $t = 1, \ldots, N$) and changepoint analysis. The detected segmentation (gray dotted vertical lines) and the mean and standard deviation estimates $\widehat{\mu}_t$ and $\widehat{\sigma}_t$ for each segment (red) are indicated (upper panel). The corresponding moving window widths ω_t chosen by the OF (dashed), OV (solid) and DV (dotted) selection approach, as well as the window width $\omega_t^{\mathrm{PW}} = 1$ used in the case of a point-wise (PW) evaluation (dash-dotted) are compared in the middle panel. The resulting (approximate) weights $\Omega(t)$ for each time instance t (see Equation (5.15)) are illustrated in the bottom panel.

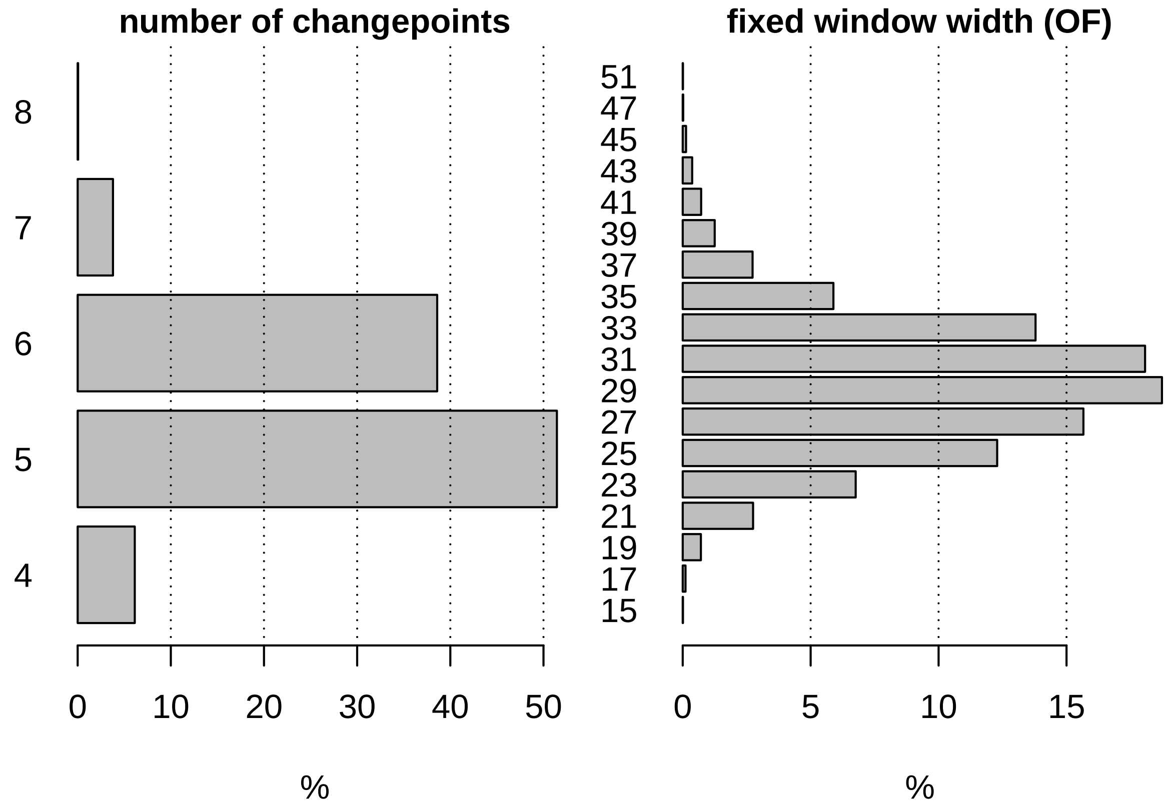


Figure 5.17: **Trend scenario (T):** Share (in %) of all 10,000 replications, where the given number of changepoints was detected (left panel), and where the given fixed window width ω^{OF} was selected by the OF approach (right panel).

was between 23 and 35. While there seems to be not much variation in the detected number of changepoints, the window widths vary considerably.

Results (model-wise comparison of different evaluation approaches) Figures 5.18–5.22 illustrate the temporal evolution of the empirical (moving) CRPS and IQ distances and their deviation from their theoretical counterparts, for Models (T1)–(T5), respectively. We observe the following:

(T) For all models (Figures 5.18–5.22):

- The PW-scores and divergences deviate very much from the theoretical scores and divergences, respectively.
- The ST-scores deviate considerably from the theoretical scores.
- For the start and the end of the time series (border case), the OF- and OV-scores and divergences tend towards the PW-scores and divergences, respectively.
- The deviation of the different scores and divergences from the theoretical scores and divergences, respectively, increases with increasing trend steepness.

(T1) True model (Figure 5.18):

- The theoretical CRPS increases with increasing variance. (The theoretical IQ distances are—by definition—constantly 0.)

- While the DV-scores approximate the theoretical scores best, the OF- and OV-scores perform similar.
- Here, the ST-divergence approximates the theoretical divergences best. Despite the presence of non-stationarity this behavior can be observed for the true model. This is because the true model specifies the characteristics of the data generating process correctly for each single time instance. Construction of an empirical CDF based on the sample given by all realizations of the corresponding time series then (approximately) yields the same distribution which we obtain in case of the data generating process. Hence, the ST-divergence gets close to the theoretical divergences which are all 0. All other (moving) divergences also come close to the theoretical divergences.

(T2) Wrong mean (Figure 5.19):

- The theoretical scores and divergences increase both with increasing trend steepness (mean and variance).
- The DV-scores approximate their theoretical counterpart best. OF- and OV-scores come close to the DV-scores.
- For most time instances, the ST-divergences approximate the theoretical divergences best. However, for steep trends, the ST-divergences become smaller than the theoretical divergences. That is, they evaluate Model (T2) too good.
- OF- and OV-divergences perform similar to each other.

(T3) Wrong variance (Figure 5.20):

- The theoretical scores and divergences increase both with increasing trend steepness (mean and variance).
- For most time instances, the different moving (OF-, OV- and DV-) scores approximate their theoretical counterparts similarly well. For steep trends, these scores evaluate Model (T3) too good (the DV-scores deviate most).
- The different moving (OF-, OV- and DV-) divergences perform similar compared to each other. Again, the ST-divergences become smaller than the theoretical divergences.

(T4) Wrong mean and linear variance (Figure 5.21):

- For steep trends, the DV-scores deviate most from the theoretical scores.
- The negative deviation from the theoretical divergences is worst for the ST-divergences. The moving (OF-, OV- and DV-) divergences deviate much less.

(T5) Wrong mean and constant variance (Figure 5.22):

- The results are similar as for model (T4). However, the negative deviation from the theoretical scores and divergences is more pronounced.

From Figures 5.18–5.22 we conclude that the moving scores and divergences based on OF-, OV- and DV-windows all approximate their theoretical counterparts reasonably well. If they provide the correct model rankings is investigated subsequently.

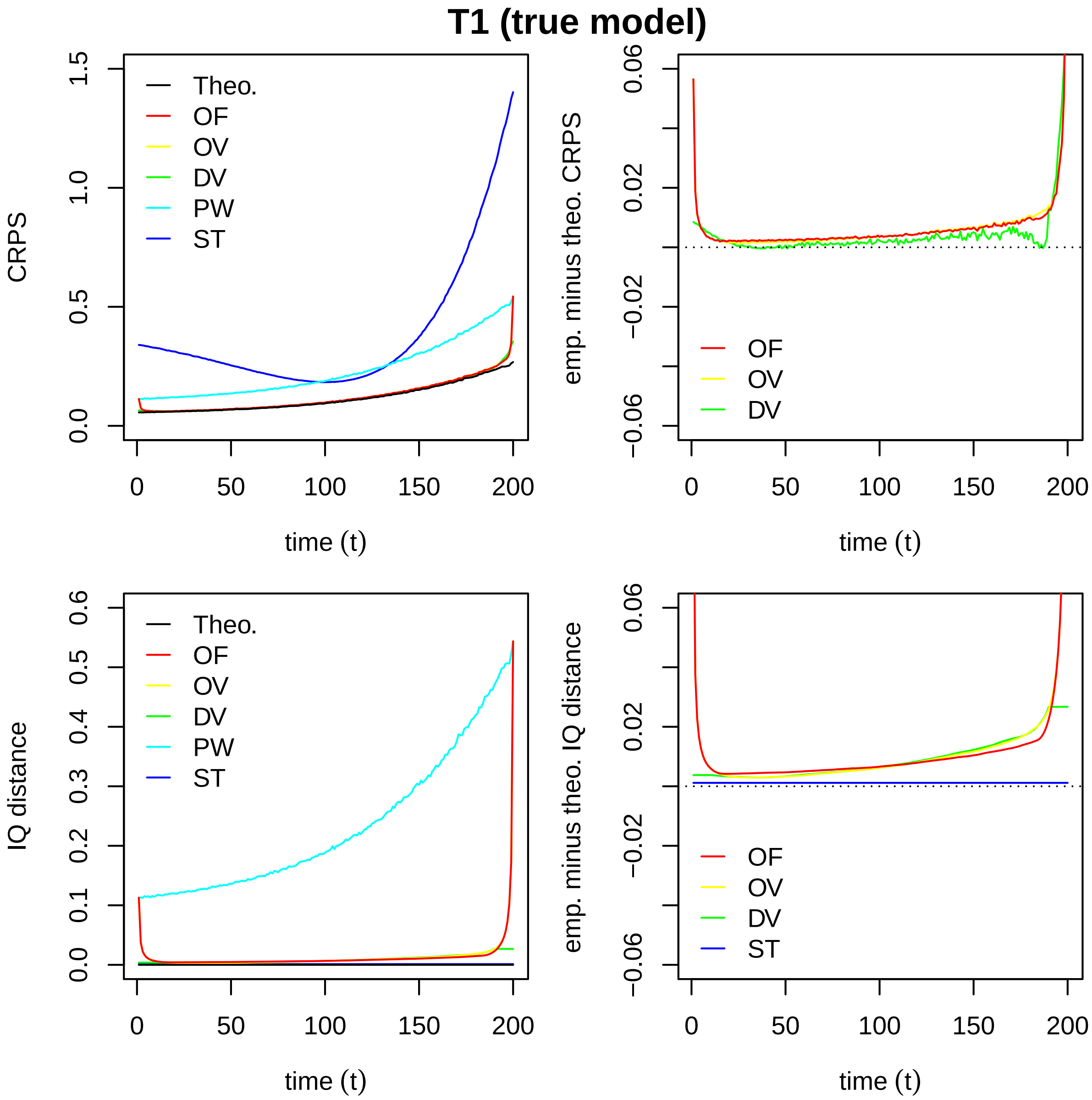


Figure 5.18: **Trend scenario (T1/true model):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

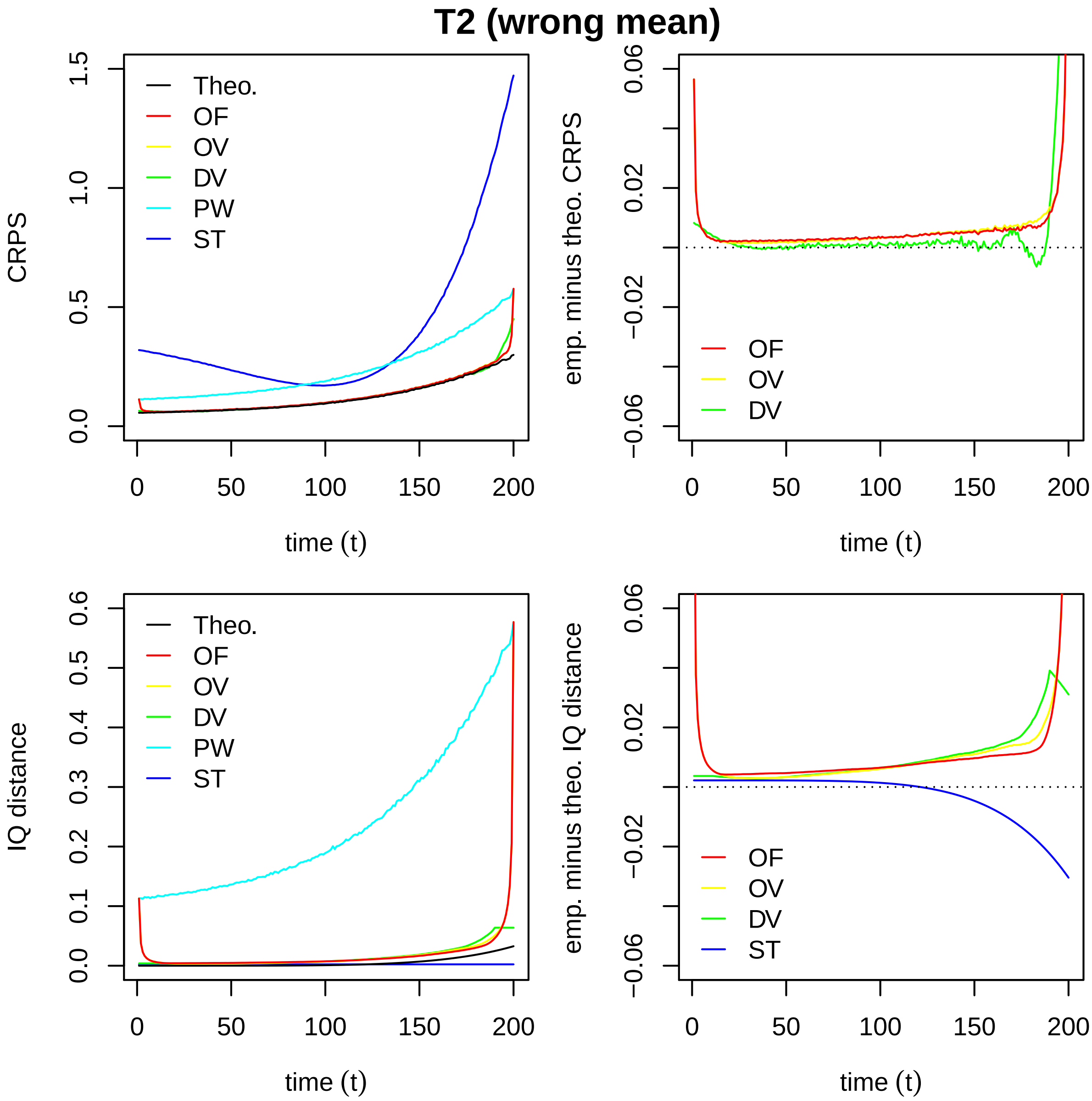


Figure 5.19: **Trend scenario (T2/wrong mean):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

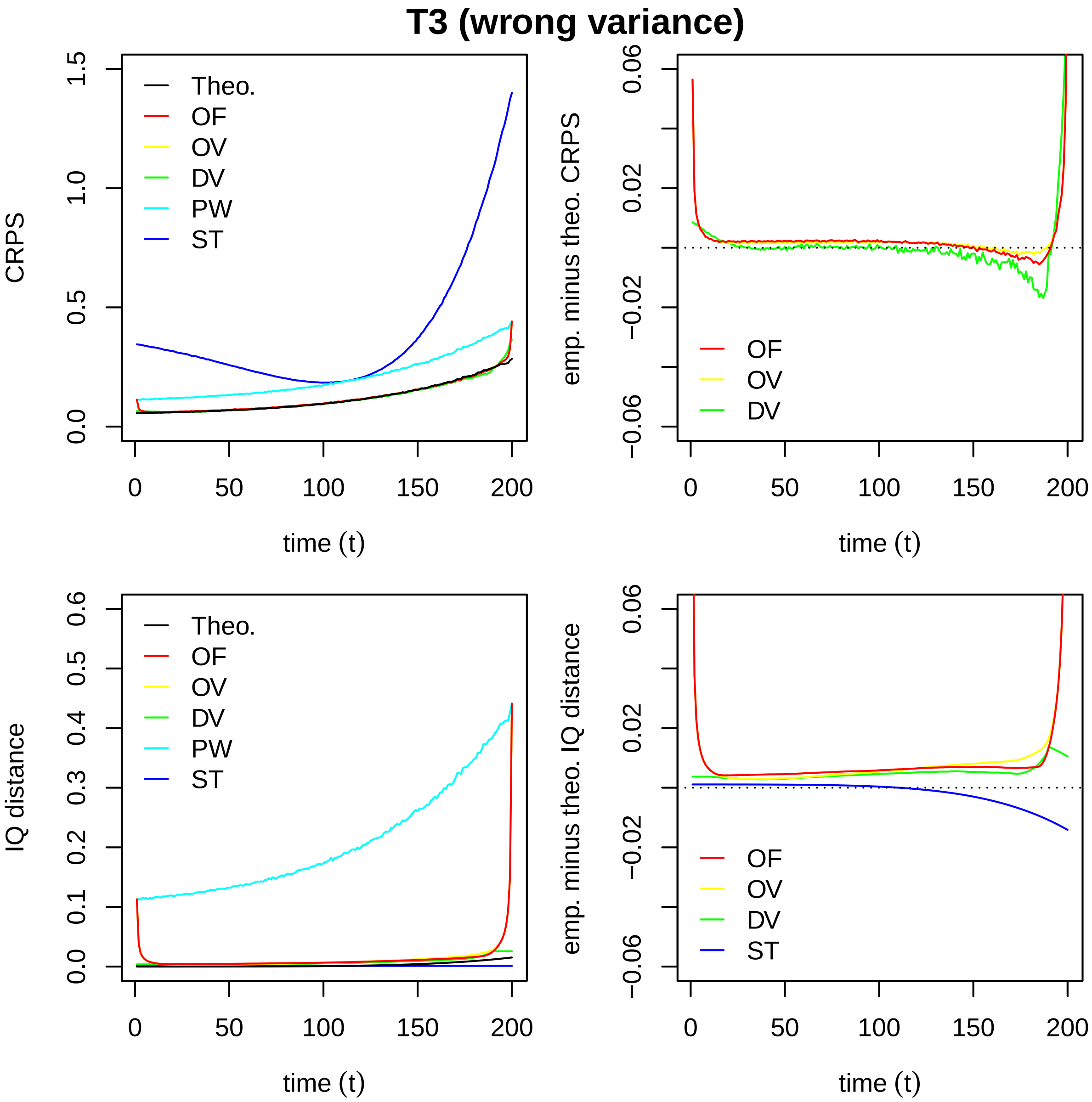


Figure 5.20: **Trend scenario (T3/wrong variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

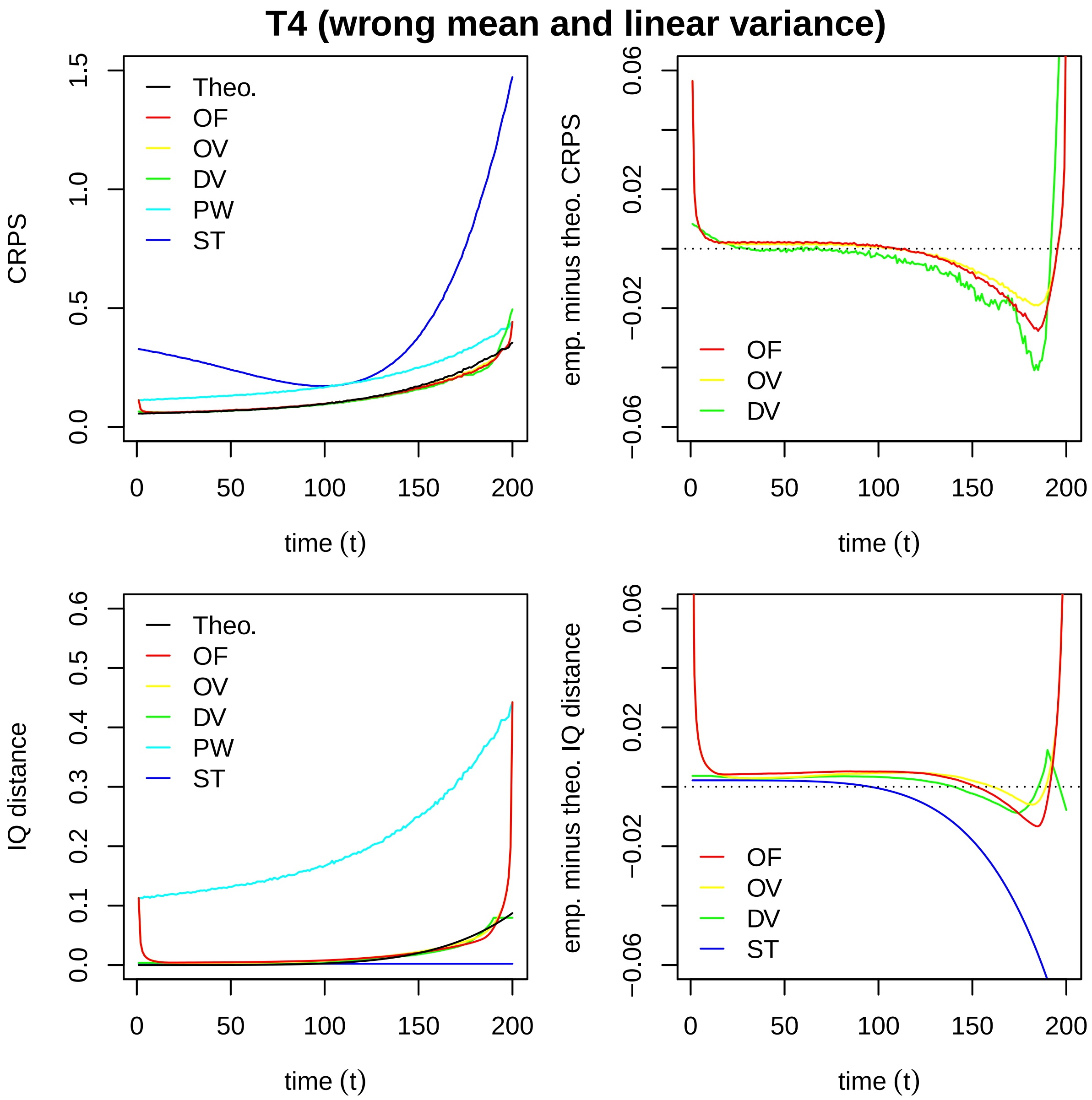


Figure 5.21: **Trend scenario (T4/wrong mean and linear variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

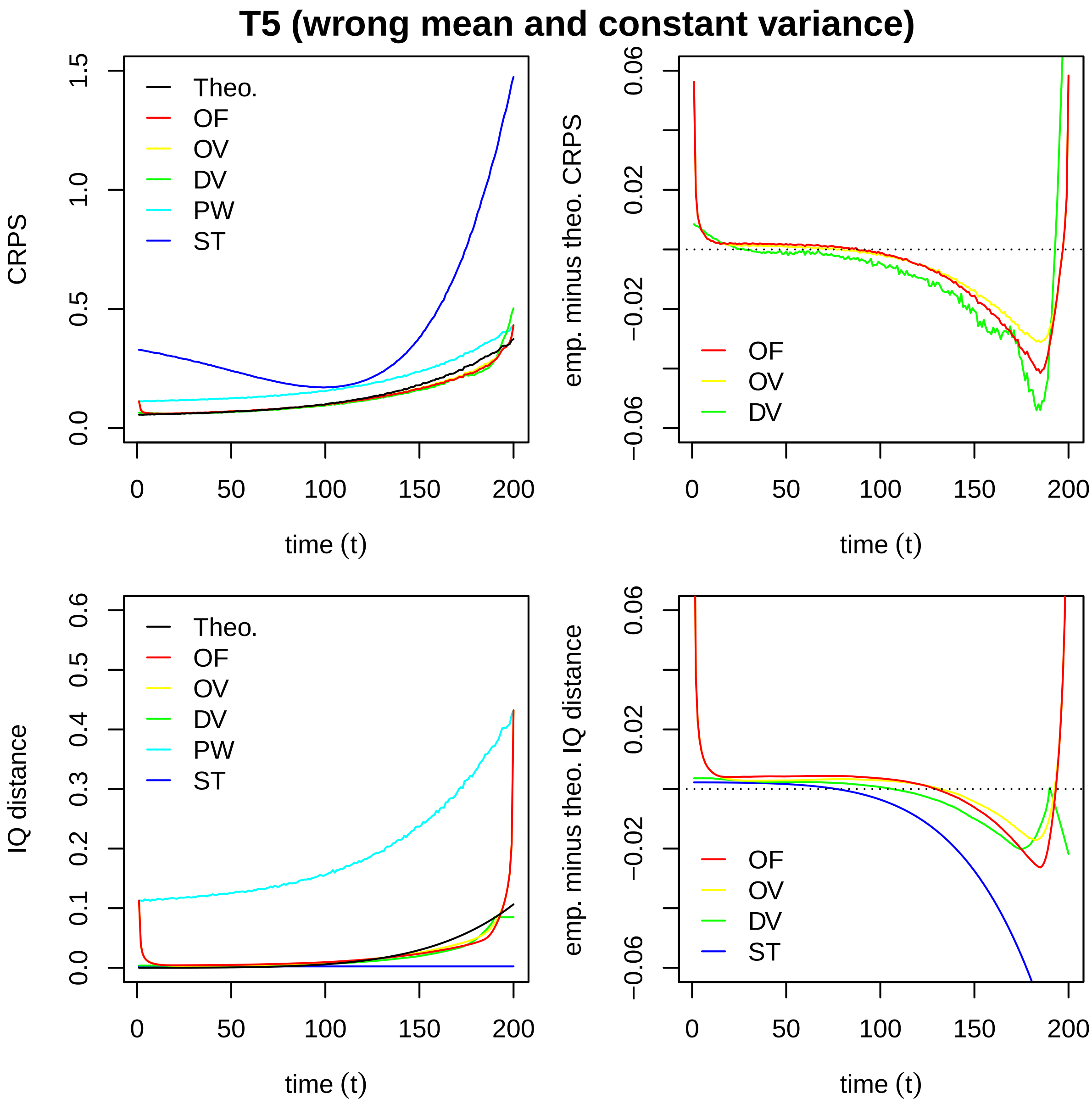


Figure 5.22: **Trend scenario (T5/wrong mean and constant variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

Results (model comparison based on different evaluation approaches) Figures 5.23 and 5.24 compare Models (T1)–(T5) based on (time series of) theoretical/empirical CRPS and IQ distances, respectively. As before, we consider theoretical (Theo.), moving OF-, moving OV-, moving DV-, point-wise (PW) and ST-scores/divergences. For the scores (Figure 5.23) we observe the following:

- Theoretical (Theo.) scores:
  - The true model (T1) is strictly the best. The differences between the five models can be explained by their different misspecifications of mean and variance.
  - For the flat part of the time series (flat trend), the scores of all models move closely together, as their means deviate not or not much from the data generating process.
  - As the trend steepness (mean an variance) increases, the scores of the different models deviate more and more from each other.
  - Model (T2) performs worse than (T1), due to its misspecification of the mean.
  - Model (T3) performs worse than (T1) and Models (T4) and (T5) perform worse than (T2), due to their misspecification of the variance.
  - The misspecification of Model (T5) is worst.
- Moving scores with OF-windows:
  - The differentiation among the different models is more difficult compared to the theoretical scores.
  - With one exception (Model (T3)) the models get ranked correctly for most time instances.
  - For steep trends Model (T3) is evaluated better than the true model (T1).
- Moving scores with OV-windows:
  - The scores evolve similar to those for the OF-windows. Our observations are the same.
- Moving scores with DV-windows:
  - Compared to OF- and OV-scores the differentiation among the different models is even more difficult.
  - Again, Model (T3) is evaluated better than the true model (T1).
- Point-wise (PW) scores:
  - Models (T1) and (T2) are evaluated far too bad.
  - Models (T4) and (T5) are evaluated too good.
- Scores under the assumption of stationarity (ST):
  - For all time instances, the scores for the different models do not differ very much from one another.
  - The observed rankings are wrong for the whole time period.

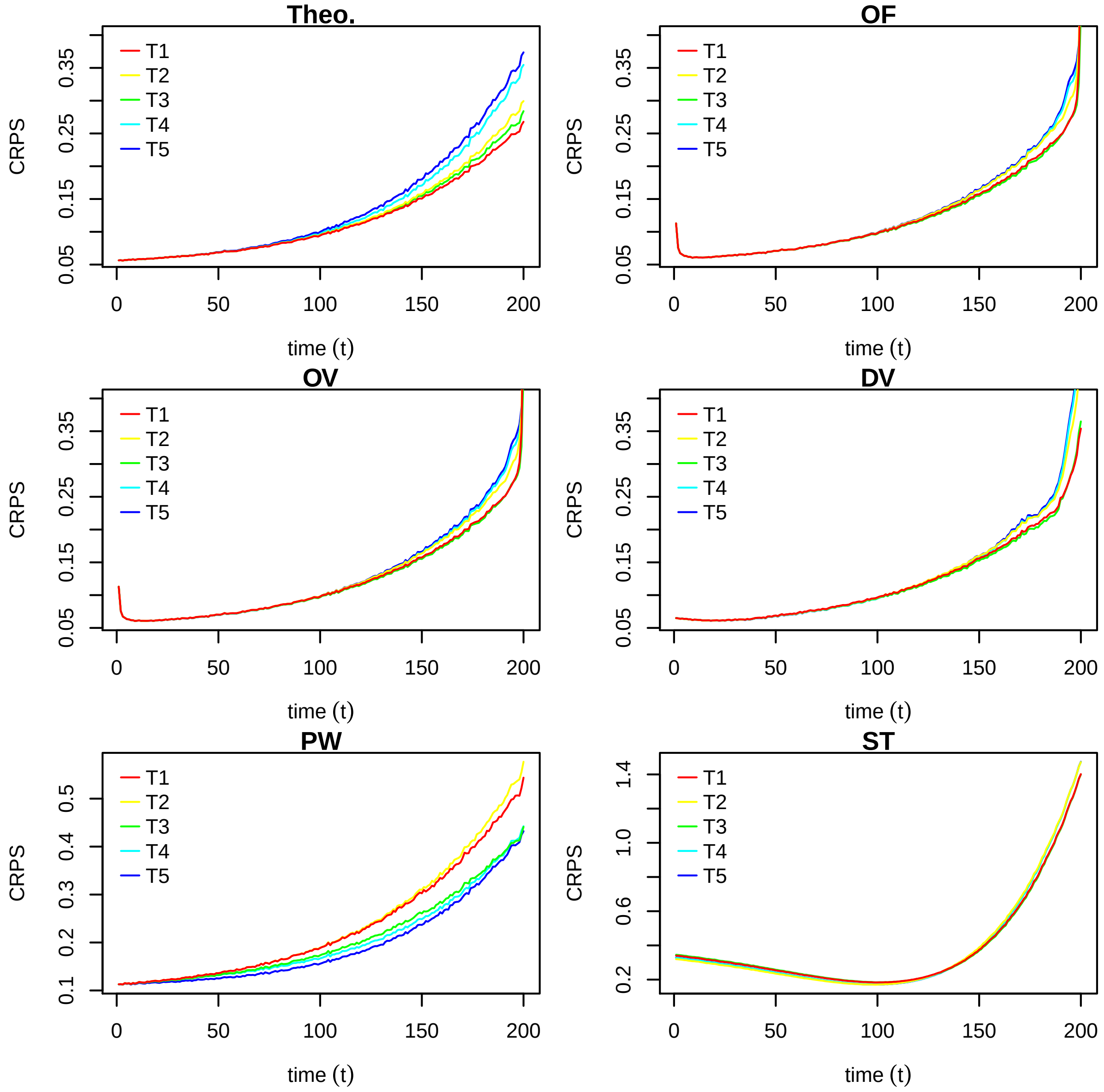


Figure 5.23: **Trend scenario (T):** Comparison of Models (T1)–(T5) based on (time series of) theoretical/empirical CRPS. Comparison based on theoretical scores (upper left panel), moving scores computed using the OF approach (upper right panel), moving scores computed using the OV approach (middle left panel), moving scores computed using the DV approach (middle right panel), point-wise (PW) scores (lower left panel) and based on ST-scores under the assumption of stationarity (lower right panel).

For the divergences (Figure 5.24) we observe the following:

- Theoretical (Theo.) divergences:
  - The true model (T1) is strictly the best. The differences between the five models can be explained by their different misspecifications of mean and variance.
  - For the flat part of the time series (flat trend), the divergences of all models move closely together, as their means deviate not or not much from the data generating process.
  - As the trend steepness (mean an variance) increases, the divergences of the different models deviate more and more from each other.
  - Model (T2) performs worse than (T1), due to its misspecification of the mean.
  - Model (T3) performs worse than (T1) and Models (T4) and (T5) perform worse than (T2), due to their misspecification of the variance.
  - The misspecification of Model (T5) is worst.
  - The models are ranked in the same way as for the CRPS (see above).
- Moving divergences with OF-windows:
  - The differentiation among the different models is more difficult compared to the theoretical divergences.
  - The divergences for Model (T3) can hardly be differentiated from those for the true model (T1). Other than that, the models get ranked correctly for most time instances.
- Moving divergences with OV-windows:
  - The divergences evolve similar to those for the OF-windows. Our observations are the same.
- Moving divergences with DV-windows:
  - Compared to OF- and OV-divergences we obtain more spurious model rankings.
  - For many time instances, Model (T3) is evaluated better than the true model (T1).
  - The flat ends of the DV-divergence curves occur due to the restrictions on the minimum size (11) of the DV-windows.
- Point-wise (PW) divergences:
  - The point-wise IQ distances are equal to the point-wise CRPS. Hence, we make the same observations as above.
- Divergences under the assumption of stationarity (ST):
  - The ST-divergences do not differentiate between different time instances.
  - Models (T2) and (T3) are ranked too bad and too good, respectively.

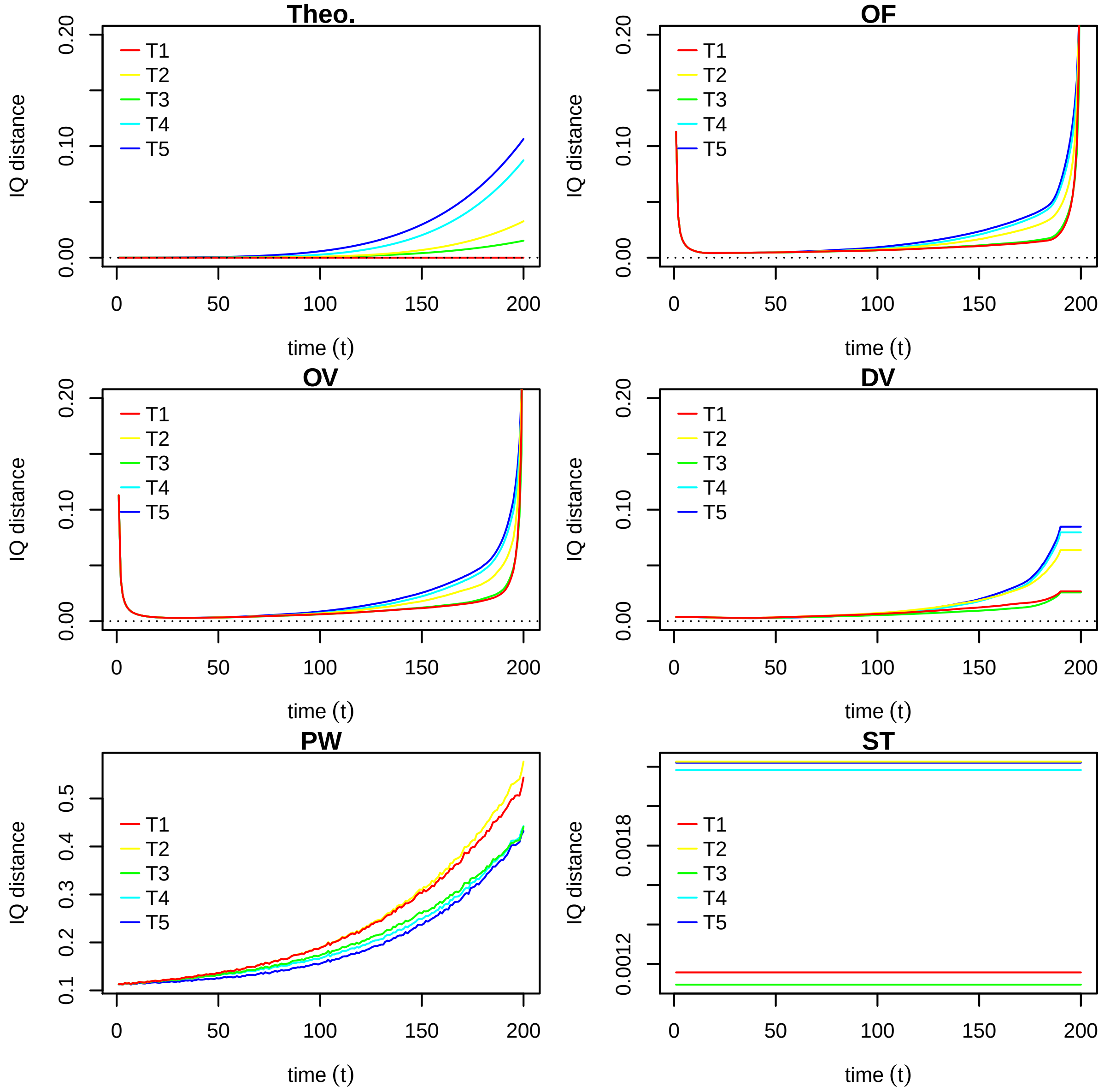


Figure 5.24: **Trend scenario (T):** Comparison of Models (T1)–(T5) based on (time series of) theoretical/empirical IQ distances. Comparison based on theoretical divergences (upper left panel), moving divergences computed using the OF approach (upper right panel), moving divergences computed using the OV approach (middle left panel), moving divergences computed using the DV approach (middle right panel), point-wise (PW) divergences (lower left panel) and based on ST-divergences under the assumption of stationarity (lower right panel).

Table 5.5: **Trend scenario (T):** Comparison/ranking of Models (T1)–(T5) based on average moving SE scores, average moving CRPS, average moving MV divergences and average moving IQ distances. Average scores/divergences (left) and model rankings (right) are provided, distinguishing between theoretical (Th) scores/divergences (true model ranking), moving scores/divergences computed using the OF approach, moving scores/divergences computed using the OV approach, moving scores/divergences computed using the DV approach, point-wise (PW) scores/divergences and scores/divergences under the assumption of stationarity (ST). Note that moving divergence averages based on OF- and OV-windows do not warrant propriety.

		average scores/divergences						model rank					
		Th	OF	OV	DV	PW	ST	Th	OF	OV	DV	PW	ST
SE score	(T1)	0.053	0.057	0.057	0.056	0.106	0.515	1	2	2	2	4	2
	(T2)	0.059	0.063	0.063	0.064	0.112	0.518	3	5	5	5	5	5
	(T3)	0.053	0.054	0.054	0.054	0.079	0.515	1	1	1	1	3	1
	(T4)	0.059	0.059	0.060	0.062	0.075	0.518	3	4	4	4	2	4
	(T5)	0.059	0.059	0.059	0.062	0.069	0.518	3	3	3	3	1	3
CRPS	(T1)	0.116	0.124	0.124	0.120	0.232	0.387	1	2	2	2	4	2
	(T2)	0.121	0.128	0.128	0.126	0.237	0.388	3	3	3	3	5	5
	(T3)	0.119	0.122	0.122	0.119	0.206	0.386	2	1	1	1	3	1
	(T4)	0.131	0.128	0.129	0.127	0.201	0.388	4	4	4	4	2	3
	(T5)	0.136	0.130	0.130	0.128	0.193	0.388	5	5	5	5	1	4
MV div.	(T1)	0.000	0.008	0.008	0.004	0.106	0.001	1	2	2	2	4	2
	(T2)	0.007	0.015	0.015	0.012	0.112	0.003	3	5	5	5	5	5
	(T3)	0.000	0.006	0.006	0.003	0.079	0.000	1	1	1	1	3	1
	(T4)	0.007	0.011	0.012	0.011	0.075	0.003	3	4	4	4	2	4
	(T5)	0.007	0.011	0.011	0.010	0.069	0.003	3	3	3	3	1	3
IQ dist.	(T1)	0.000	0.014	0.014	0.009	0.232	0.001	1	2	2	2	4	2
	(T2)	0.005	0.019	0.019	0.015	0.237	0.002	3	3	3	3	5	5
	(T3)	0.003	0.014	0.014	0.008	0.206	0.001	2	1	1	1	3	1
	(T4)	0.015	0.021	0.021	0.016	0.201	0.002	4	4	4	4	2	3
	(T5)	0.020	0.023	0.023	0.017	0.193	0.002	5	5	5	5	1	4

Model ranking based on average scores and divergences In order to see, if any evaluation approach ranks the models (T1)–(T5) correctly, we provide Table 5.5. Besides the CRPS and the IQ distance, we now also consider SE scores and MV divergences. The table provides the average scores and divergences (averages over all time instances and all 10,000 replications of the simulation study) for the different considered evaluation approaches (OF, OV, DV, PW and ST) and compares them to the corresponding theoretical (Th) scores and divergences, respectively. Note that utilization of moving divergence averages is not necessarily a proper evaluation method (that is in case of OF- and OV-windows; see discussion in Sections 5.3.2 and 5.3.3). Based on the average scores/divergences we rank the different models, distinguishing between the theoretical scores/divergences and the five different evaluation approaches. Hence, for each approach (column) and separately for each score/divergence type (SE score, CRPS, MV divergence, IQ distance), the model with the smallest average score/divergence is ranked best (1) and the model with the highest average score/divergence is ranked worst (5). Accordingly, the models with intermediate scores/divergences are ranked 2–4. If the score/divergence averages of two or more models are equal, all of these models get the same (minimum) rank. The resulting model rankings are compared in the right half of the table.

To decide if an evaluation approach is good or not, we again consider the following two criteria. For good evaluation approaches

(E1) the true model (T1) should be ranked lowest (best), and

(E2) the ranking should be identical to that based on the theoretical (Th) scores/divergences.

All in all, we observe that (E1) and (E2) are satisfied best for the CRPS and the IQ distances calculated based on the OF, OV and DV approach.

Now, we discuss the results from Table 5.5 in more detail, distinguishing between the different types of scores and divergences:

- SE scores:
    - The theoretical scores can not differentiate between Models (T1) and (T3), and also not between (T2), (T4) and (T5). This is due to the fact, that SE scores neglect misspecifications of the variance.
    - The moving (OF-, OV- and DV-) scores falsely rank (T3) better than (T1). Moreover, they rank (T4) and (T5) better than (T2).
    - The rankings based on the PW- and ST-scores are both wrong.
    - The ST-scores provide the same ranking as the OF-, OV- and DV-scores.
- CRPS:
    - The scores are able to differentiate between all five models.
    - The moving (OF-, OV- and DV-) scores falsely rank Model (T3) better than (T1). All other models are ranked correctly.
    - The ranking provided by the PW-scores is completely wrong.
    - The ST-scores rank Model (T2) too bad and Model (T3) too good.
- MV divergences:
    - We obtain the same (faulty) model rankings as for the SE scores.
- IQ distances:
    - Same as for the CRPS, the moving (OF, OV and DV) IQ distances rank Model (T3) better than (T1). All other models are ranked correctly.
    - The rankings based on the PW- and ST-divergences (same as for the CRPS) are both wrong.

Further analysis of the model rankings Whereas in Table 5.5 we investigated model rankings which were obtained by averaging over all 10,000 replications of the simulation study (we refer to them as *average based model rankings*), we now look into all 10,000 replications separately, to see if the individual model rankings correspond to the *average based model rankings* found in Table 5.5. For this we consider the models pair-wise and calculate the share of all 10,000 replications for which the average (moving) score/divergence of the first model is smaller (better) than the average (moving) score/divergence of the second model. This time the (moving) scores/divergences are averaged only temporally, that is over all available time instances.

Table 5.6: **Trend scenario (T):** Pair-wise comparison of Models (T1)–(T5), distinguishing between theoretical CRPS (Th), moving CRPS computed using the OF, OV and DV approach, point-wise CRPS (PW) and CRPS under the assumption of stationarity (ST). Each table entry equals the share of all 10,000 replications of the simulation study for which the average (moving) score of the model indicated in the corresponding row is smaller (better) than the average (moving) score of the model indicated in the corresponding column of the table.

	Th					**OF**				
	(T1)	(T2)	(T3)	(T4)	(T5)	(T1)	(T2)	(T3)	(T4)	(T5)
(T1)	–	0.94	0.98	1.00	1.00	–	0.84	0.15	0.83	0.87
(T2)	0.06	–	0.26	1.00	1.00	0.16	–	0.09	0.53	0.65
(T3)	0.02	0.74	–	1.00	1.00	0.85	0.91	–	0.92	0.95
(T4)	0.00	0.00	0.00	–	1.00	0.17	0.47	0.08	–	0.94
(T5)	0.00	0.00	0.00	0.00	–	0.13	0.35	0.05	0.06	–
	PW					**OV**				
	(T1)	(T2)	(T3)	(T4)	(T5)	(T1)	(T2)	(T3)	(T4)	(T5)
(T1)	–	0.83	0.00	0.00	0.00	–	0.85	0.19	0.85	0.89
(T2)	0.17	–	0.00	0.00	0.00	0.15	–	0.09	0.60	0.72
(T3)	1.00	1.00	–	0.22	0.02	0.81	0.91	–	0.93	0.96
(T4)	1.00	1.00	0.78	–	0.00	0.15	0.40	0.07	–	0.97
(T5)	1.00	1.00	0.98	1.00	–	0.11	0.28	0.04	0.03	–
	ST					**DV**				
	(T1)	(T2)	(T3)	(T4)	(T5)	(T1)	(T2)	(T3)	(T4)	(T5)
(T1)	–	0.88	0.42	0.91	0.92	–	0.92	0.19	0.92	0.94
(T2)	0.12	–	0.12	0.51	0.54	0.08	–	0.05	0.63	0.71
(T3)	0.58	0.88	–	0.93	0.93	0.81	0.95	–	0.97	0.98
(T4)	0.09	0.49	0.07	–	0.64	0.08	0.37	0.03	–	0.92
(T5)	0.08	0.46	0.07	0.36	–	0.06	0.29	0.02	0.08	–

The results of these computations (shares) are summarized in Tables 5.6 and 5.7. Table 5.6 is based on CRPS, Table 5.7 on IQ distances. Both tables again distinguish between theoretical scores/divergences (Th) and empirical scores/divergences calculated using different approaches (OF, OV, DV, PW and ST). Each block of Tables 5.6 and 5.7 corresponds to one particular approach. It is best to read each block row-wise. Each row provides the shares telling how often the model indicated to the left (model under consideration) had a better average score/divergence than the other models (indicated above). A share equal to 0 means that the model under consideration is never considered better than the other model, a share equal to 1 means that it is always considered better than the other one, a share in $(0, 0.5)$ means that the model under consideration is considered worse than the other one for the majority of all replications and a share in $(0.5, 1)$ means that it is considered better than the other one for the majority of all replications.

First, we are interested to know, if the pair-wise and replication-wise consideration in Tables 5.6 and 5.7 supports the model rankings found in Table 5.5. If we rank model A better than model B if the share corresponding to the model pair (A, B) is greater than 0.5, then we mostly obtain the same rankings as in Table 5.5. The ST-scores and divergences are the only exception. Here, Model (T2) is ranked better than Models (T4) and (T5), where the average based model ranking suggests that Model (T2) is ranked worse than (T4) and (T5).

Table 5.7: **Trend scenario (T):** Pair-wise comparison of Models (T1)–(T5), distinguishing between theoretical IQ distances (Th), moving IQ distances computed using the OF, OV and DV approach, point-wise IQ distances (PW) and IQ distances under the assumption of stationarity (ST). Each table entry equals the share of all 10,000 replications of the simulation study for which the average (moving) divergence of the model indicated in the corresponding row is smaller (better) than the average (moving) divergence of the model indicated in the corresponding column of the table.

			Th					**OF**		
	(T1)	(T2)	(T3)	(T4)	(T5)	(T1)	(T2)	(T3)	(T4)	(T5)
(T1)	–	1.00	1.00	1.00	1.00	–	0.85	0.45	0.91	0.95
(T2)	0.00	–	0.00	1.00	1.00	0.15	–	0.14	0.77	0.85
(T3)	0.00	1.00	–	1.00	1.00	0.55	0.86	–	0.94	0.97
(T4)	0.00	0.00	0.00	–	1.00	0.09	0.23	0.06	–	0.99
(T5)	0.00	0.00	0.00	0.00	–	0.05	0.15	0.03	0.01	–
			PW					**OV**		
	(T1)	(T2)	(T3)	(T4)	(T5)	(T1)	(T2)	(T3)	(T4)	(T5)
(T1)	–	0.83	0.00	0.00	0.00	–	0.86	0.44	0.91	0.95
(T2)	0.17	–	0.00	0.00	0.00	0.14	–	0.14	0.78	0.88
(T3)	1.00	1.00	–	0.22	0.02	0.56	0.86	–	0.95	0.97
(T4)	1.00	1.00	0.78	–	0.00	0.09	0.22	0.05	–	0.99
(T5)	1.00	1.00	0.98	1.00	–	0.05	0.12	0.03	0.01	–
			ST					**DV**		
	(T1)	(T2)	(T3)	(T4)	(T5)	(T1)	(T2)	(T3)	(T4)	(T5)
(T1)	–	0.88	0.42	0.91	0.92	–	0.92	0.19	0.92	0.94
(T2)	0.12	–	0.12	0.51	0.54	0.08	–	0.05	0.63	0.71
(T3)	0.58	0.88	–	0.93	0.93	0.81	0.95	–	0.97	0.98
(T4)	0.09	0.49	0.07	–	0.64	0.08	0.37	0.03	–	0.92
(T5)	0.08	0.46	0.07	0.36	–	0.06	0.29	0.02	0.08	–

Second, we are interested to see, how certain the different evaluation approaches are about the relative pair-wise rankings they indicate. Table entries close to 0 or 1 support certainty, entries close to 0.5 indicate that the evaluation approach under consideration has difficulties to decide in favor of one of the corresponding models. Whereas the theoretical CRPS not always allow for unambiguous decisions (Models (T1)–(T3)), the theoretical IQ distances always allow for unambiguous decisions. For the empirical scores/divergences we observe shares smaller than 0.25 or greater than 0.75, in most cases. We point out cases, where this does not hold. The ST-scores and divergences are worst in ranking the different models unambiguously. They have problems with the relative ranking for the model pairs (T1,T3), (T2,T4), (T2,T5) and (T4,T5). The shares obtained for OF-, OV- and DV-scores (CRPS) yield uncertainty for the relative ranking of (T2) with respect to (T4) and (T5) and vice versa. For OF- and OV-divergences only the distinction between Models (T1) and (T3) is problematic.

All in all, we find that the average based model rankings found in Table 5.5 (for CRPS and IQ distances) and the differentiation between the different models in these rankings are justified in most cases. Cases where the relative ranking of models can not be trusted are addressed in the two paragraphs above.

5.7 Simulation study: Periodicity scenario

For the periodicity scenario (P) we consider time series with periodically varying mean and standard deviation. To model the periodicity, we consider the function

$$h_{\mathrm{P}}(t;\boldsymbol{\theta}) \coloneqq \theta_0 + \theta_1 \sin\left(2\pi t \theta_2\right), \tag{5.21}$$

with *intercept parameter* θ_0, *amplitude parameter* θ_1 and *frequency parameter* θ_2, where $\boldsymbol{\theta} \coloneqq (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3$. Based on (5.21) we model the periodicity in the mean by

$$M_{\mathrm{P}}(t;\boldsymbol{\mu}) \coloneqq h_{\mathrm{P}}(t;\boldsymbol{\mu}), \text{ for all } t = 1, \dots, N,$$

with parameter vector $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2) \in \mathbb{R}^3$. Hence, the mean oscillates periodically (following a sine curve), if it holds $\mu_1, \mu_2 \neq 0$. The length of one period is given by $1/\mu_2$. The periodicity in the standard deviation is modeled as

$$S_{\mathrm{P}}(t;\boldsymbol{\sigma}) \coloneqq \exp\left(h_{\mathrm{P}}(t;\boldsymbol{\sigma})\right), \text{ for all } t = 1, \dots, N,$$

with parameter vector $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \sigma_2) \in \mathbb{R}^3$. The standard deviation oscillates periodically (following the exponential of a sine curve), if it holds $\sigma_1, \sigma_2 \neq 0$, where the length of one period is given by $1/\sigma_2$. It is strictly positive.

Periodicity scenario (P) The subsequent equations define the *periodicity scenario (P)*, where we consider time series of length $N = 730$. The data generating process (P0) and the five models (P1)–(P5) are given by

$$\begin{array}{lll}
\text{(P0)} & Y_t \sim \mathcal{N}\left(M_{\mathrm{P}}(t;\boldsymbol{\mu}^0), \left[S_{\mathrm{P}}(t;\boldsymbol{\sigma}^0)\right]^2\right) & \text{(data generating process),} \\
\hline
\text{(P1)} & X_t^1 \sim \mathcal{N}\left(M_{\mathrm{P}}(t;\boldsymbol{\mu}^1), \left[S_{\mathrm{P}}(t;\boldsymbol{\sigma}^1)\right]^2\right) & \text{(true model),} \\
\text{(P2)} & X_t^2 \sim \mathcal{N}\left(M_{\mathrm{P}}(t;\boldsymbol{\mu}^2), \left[S_{\mathrm{P}}(t;\boldsymbol{\sigma}^2)\right]^2\right) & \text{(wrong mean),} \\
\text{(P3)} & X_t^3 \sim \mathcal{N}\left(M_{\mathrm{P}}(t;\boldsymbol{\mu}^3), \left[S_{\mathrm{P}}(t;\boldsymbol{\sigma}^3)\right]^2\right) & \text{(wrong variance),} \\
\text{(P4)} & X_t^4 \sim \mathcal{N}\left(M_{\mathrm{P}}(t;\boldsymbol{\mu}^4), \left[S_{\mathrm{P}}(t;\boldsymbol{\sigma}^4)\right]^2\right) & \text{(wrong mean and variance),} \\
\text{(P5)} & X_t^5 \sim \mathcal{N}\left(M_{\mathrm{P}}(t;\boldsymbol{\mu}^5), \left[S_{\mathrm{P}}(t;\boldsymbol{\sigma}^5)\right]^2\right) & \text{(wrong mean and constant variance),}
\end{array}$$

with parameters

$$\begin{array}{llll}
\text{(P0)} & \boldsymbol{\mu}^0 = (0, 10, 1/365), & \boldsymbol{\sigma}^0 = (0, -0.5, 1/365), & \text{(data generating process),} \\
\hline
\text{(P1)} & \boldsymbol{\mu}^1 = \boldsymbol{\mu}^0, & \boldsymbol{\sigma}^1 = \boldsymbol{\sigma}^0, & \text{(true model),} \\
\text{(P2)} & \boldsymbol{\mu}^2 = (0, 9.5, 1/365), & \boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^0, & \text{(wrong mean),} \\
\text{(P3)} & \boldsymbol{\mu}^3 = \boldsymbol{\mu}^0, & \boldsymbol{\sigma}^3 = (0, -0.25, 1/365), & \text{(wrong variance),} \\
\text{(P4)} & \boldsymbol{\mu}^4 = \boldsymbol{\mu}^2, & \boldsymbol{\sigma}^4 = \boldsymbol{\sigma}^3 & \text{(wrong mean and variance),} \\
\text{(P5)} & \boldsymbol{\mu}^5 = \boldsymbol{\mu}^2, & \boldsymbol{\sigma}^5 = (0, 0, 1/365), & \text{(wrong mean and constant variance).}
\end{array}$$

Whereas the true model (P1) is equivalent to the data generating process (P0) in terms of its parametrization, the other Models (P2)–(P5) differ from the data generating process (P0) in at least one parameter.

Illustration of the periodicity scenario (P) The differences between the data generating process (P0) and the models (P1)–(P5) in the temporal evolution of the mean and the standard deviation are visualized in Figure 5.25. For the data generating process (P0) the mean $\mu_{0,t}$ varies sinusoidal and the variance $\sigma_{0,t}^2$ varies according to the exponential of a sine curve. It can be seen, that times of high means coincide with times of low variance and vice versa. In comparison to the data generating process (P0)

- the amplitude of the mean $\mu_{2,t}$ of Model (P2) is smaller than that of $\mu_{0,t}$,
- the amplitude of the variance $\sigma_{3,t}^2$ of Model (P3) is smaller than that of $\sigma_{0,t}^2$,
- the amplitudes of the mean $\mu_{4,t}$ and the variance $\sigma_{4,t}^2$ of Model (P4) are smaller than those of $\mu_{0,t}$ and $\sigma_{0,t}^2$,
- the amplitude of the mean $\mu_{5,t}$ of Model (P5) is smaller than that of $\mu_{0,t}$ and the variance $\sigma_{5,t}^2$ is constant.

Hence, Model (P2) underestimates the magnitude of the mean oscillations observed for the data generating process (P0), where it captures the variance oscillations correctly. Model (P3) captures the mean oscillations correctly, however the oscillations in the variance are underestimated. Models (P4) and (P5) are both wrong in terms of mean and variance. Whereas Model (P4) at least models some of the variation in the variance, Model (P5) completely ignores this characteristic of the data generating process (P0). Only the true model (P1) models the data generating process (P0) correctly. Thus, a comparative evaluation of the Models (P1)–(P5) should favor Model (P1).

Illustration of changepoint analysis and selection of moving windows Figure 5.26 illustrates one replication of the data generating process (P0) and the corresponding selection of moving windows according to Section 5.3.3. Again, irregularly spaced changepoints were detected. While big segment lengths (48, 63, 75, 90) occur around the minima and maxima of each period, the remaining segments are comparatively short (11–31).

This results in the window widths illustrated in the lower panel of Figure 5.26, which are symptomatic for periodic time series. The window widths ω_t^{OV} and ω_t^{DV} of the small windows mostly vary around the window width ω_t^{OF}, which is constant (19) for time instances t apart from the border case (far enough from the start and the end of the time series). As we move closer to the period minima and maxima, the window widths ω_t^{OV} and ω_t^{DV} are considerably larger.

With the bottom panel of Figure 5.26 we illustrate the (approximate) weights $\Omega(t)$ (see Equation (5.15)) corresponding to each observation y_t, $t = 1, \dots, N$, in an evaluation based on moving divergence averages (5.12). As already discussed, the propriety of the metric (5.12) is given only if the weights $\Omega(t)$ are constant. From the figure we observe, that the DV-windows yield propriety, while the OF- and OV-windows do not. For the OF-windows, the weights $\Omega(t)$ deviate from 1 only in the beginning and in the end of the considered time interval (border case). We clearly see, that the metric based on OV-windows approximates propriety much worse than that based on OF-windows.

Figure 5.27 summarizes for all 10,000 replications of the data generating process (P0) how many changepoints were detected and which window widths were selected predominantly by the OF approach. We observe that in most cases ($> 95\%$) between 26 and 32 changepoints were detected. Moreover, we observe that in most cases the fixed OF-window width was either 15, 17,

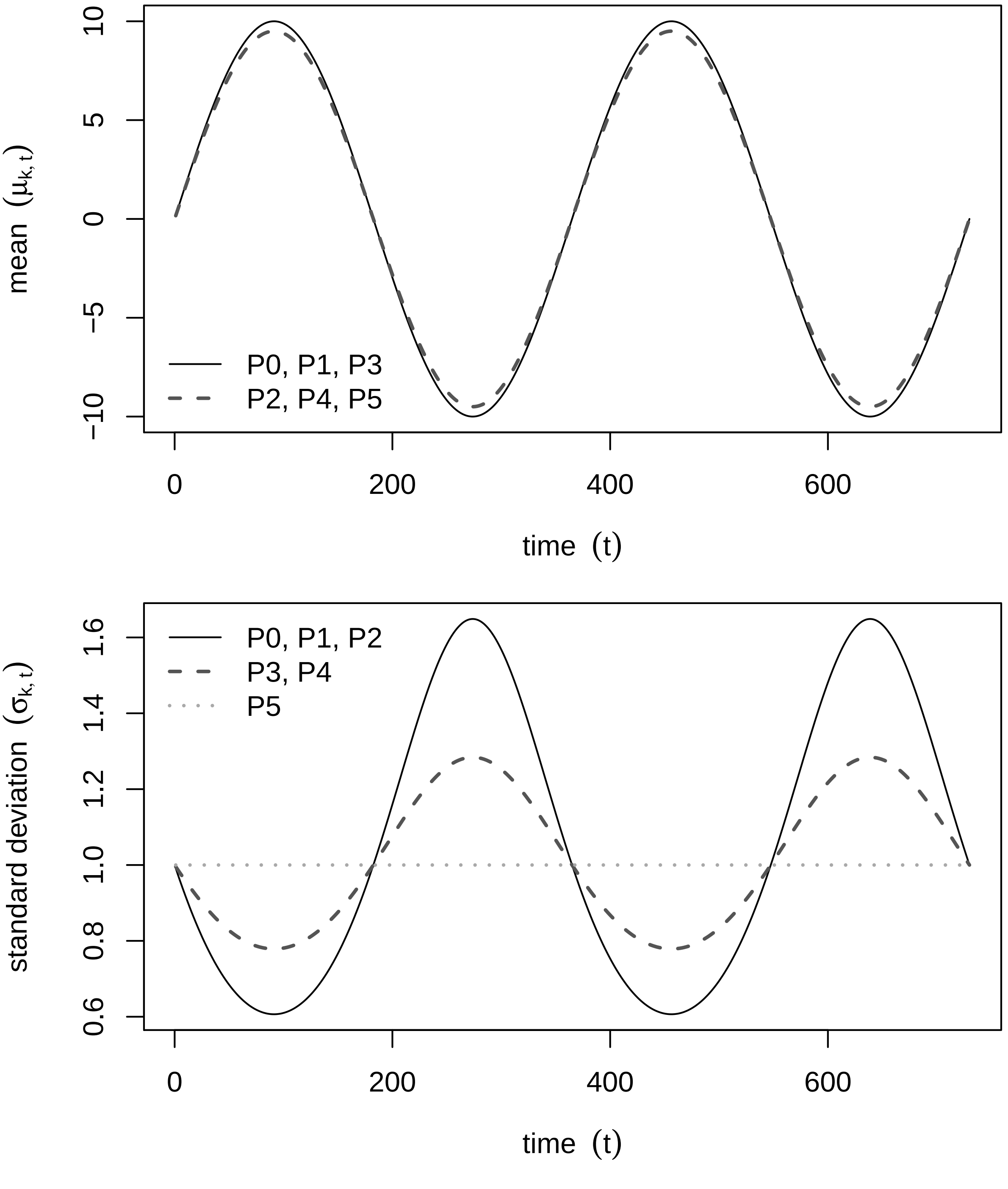


Figure 5.25: **Periodicity scenario (P):** Temporal evolution of mean $\mu_{k,t}$ (upper panel) and standard deviation $\sigma_{k,t}$ (lower panel) for the data generating process (P0) ($k = 0$) and the models (P1)–(P5) ($k = 1, \ldots, 5$).

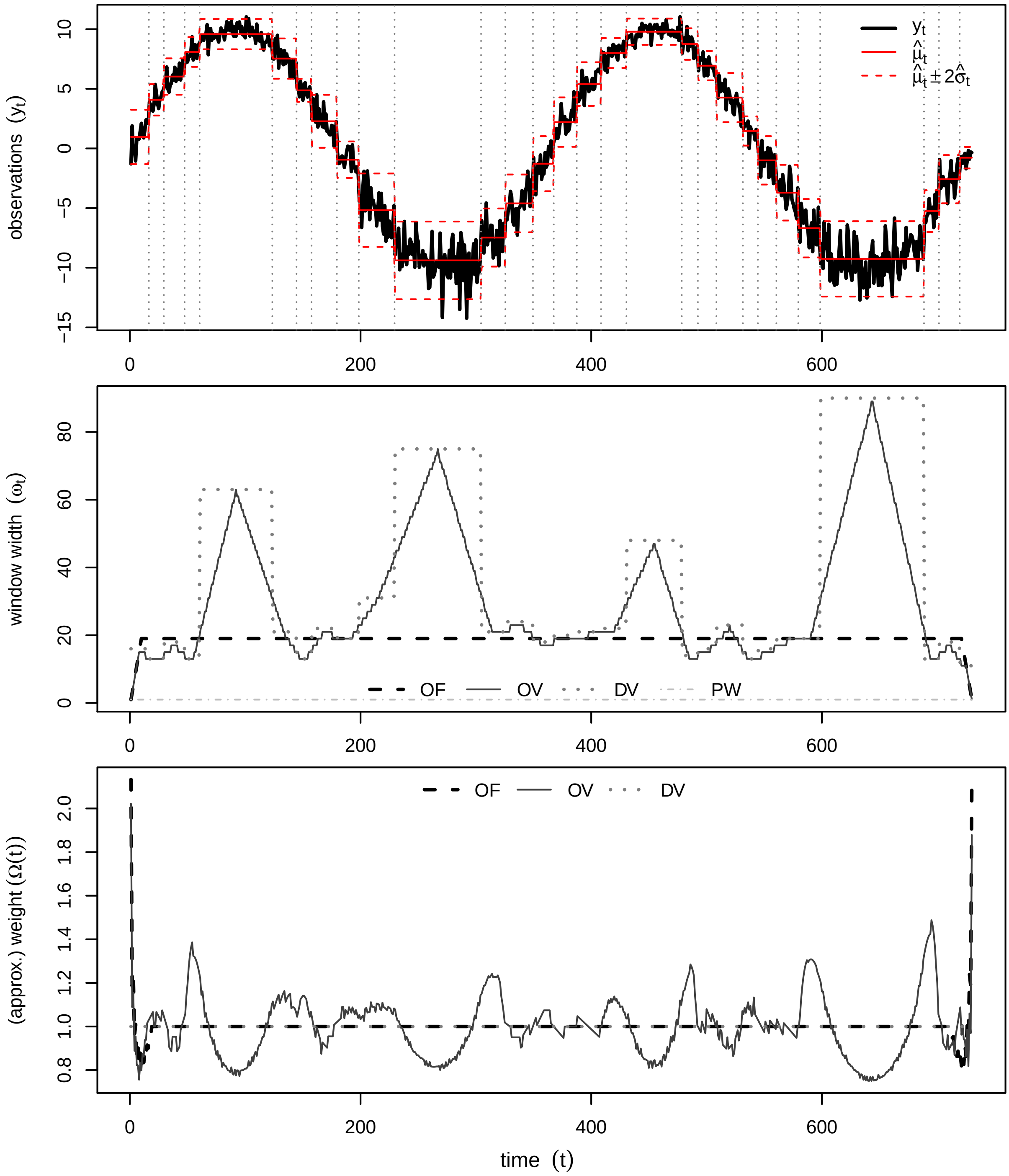


Figure 5.26: **Periodicity scenario (P):** One replication of the data generating process (P0) (observations y_t, $t = 1, \ldots, N$) and changepoint analysis. The detected segmentation (gray dotted vertical lines) and the mean and standard deviation estimates $\widehat{\mu}_t$ and $\widehat{\sigma}_t$ for each segment (red) are indicated (upper panel). The corresponding moving window widths ω_t chosen by the OF (dashed), OV (solid) and DV (dotted) selection approach, as well as the window width $\omega_t^{\mathrm{PW}} = 1$ used in the case of a point-wise (PW) evaluation (dash-dotted) are compared in the middle panel. The resulting (approximate) weights $\Omega(t)$ for each time instance t (see Equation (5.15)) are illustrated in the bottom panel.

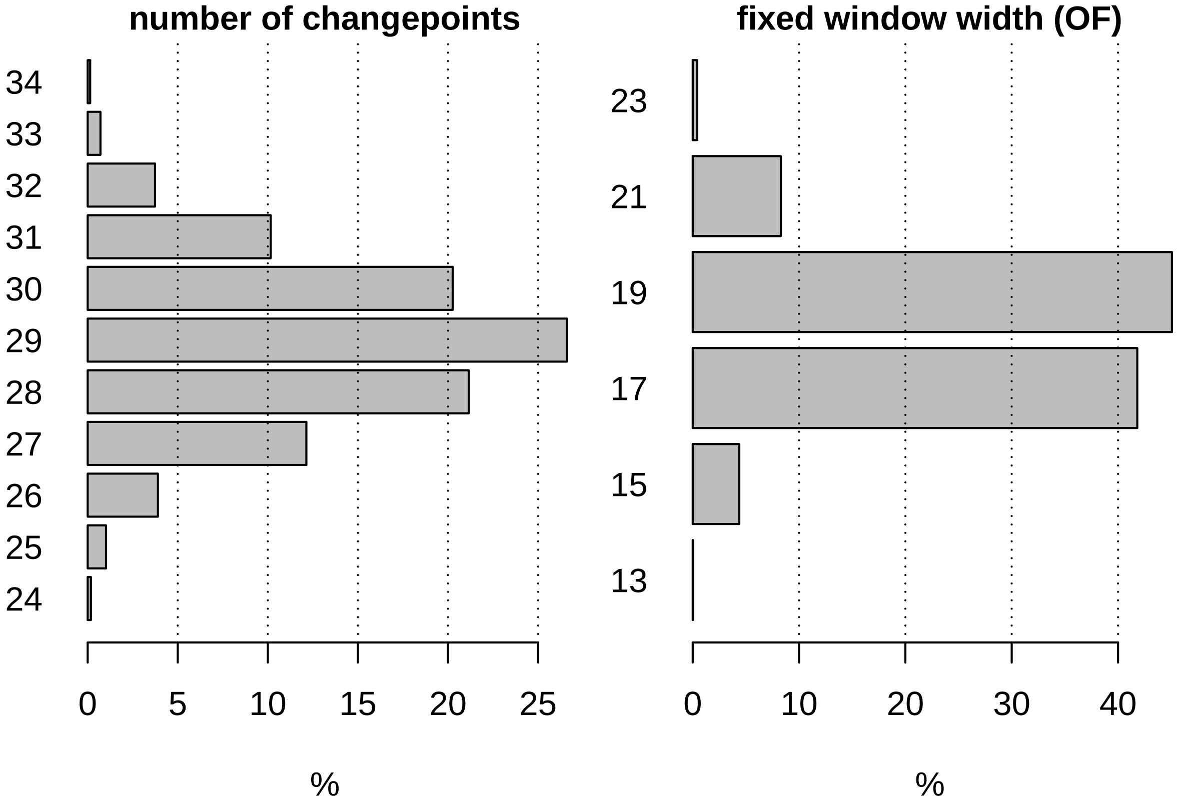


Figure 5.27: **Periodicity scenario (P):** Share (in %) of all 10,000 replications, where the given number of changepoints was detected (left panel), and where the given fixed window width ω^{OF} was selected by the OF approach (right panel).

19 or 21. Looking at Figure 5.27, there seems to be not much variation in the detected number of changepoints and the selected window widths.

Results (model-wise comparison of different evaluation approaches) Note, that for this scenario we do not compute ST-scores and divergences. This is due to the fact, that this would require repeated computation of sample CRPS (2.32) and IQ distances (2.42) based on a sample size equal to the length of the time series $N = 730$, which is computationally not feasible in a reasonable time. Since for the periodicity scenario (P) the stationarity assumption is gravely violated, we already know in advance that we can not trust in ST-scores and divergences and forgo them.

Figures 5.28–5.32 illustrate the temporal evolution of the empirical (moving) CRPS and IQ distances and their deviation from their theoretical counterparts, for Models (P1)–(P5), respectively. We observe the following:

- The theoretical CRPS increases with increasing variance of the data generating process of interest.
- For a misspecification of the mean the theoretical IQ distance increases more if the variance of the data generating process is low than if it is high.
- The PW-scores and divergences deviate very much from the theoretical scores and divergences, respectively.

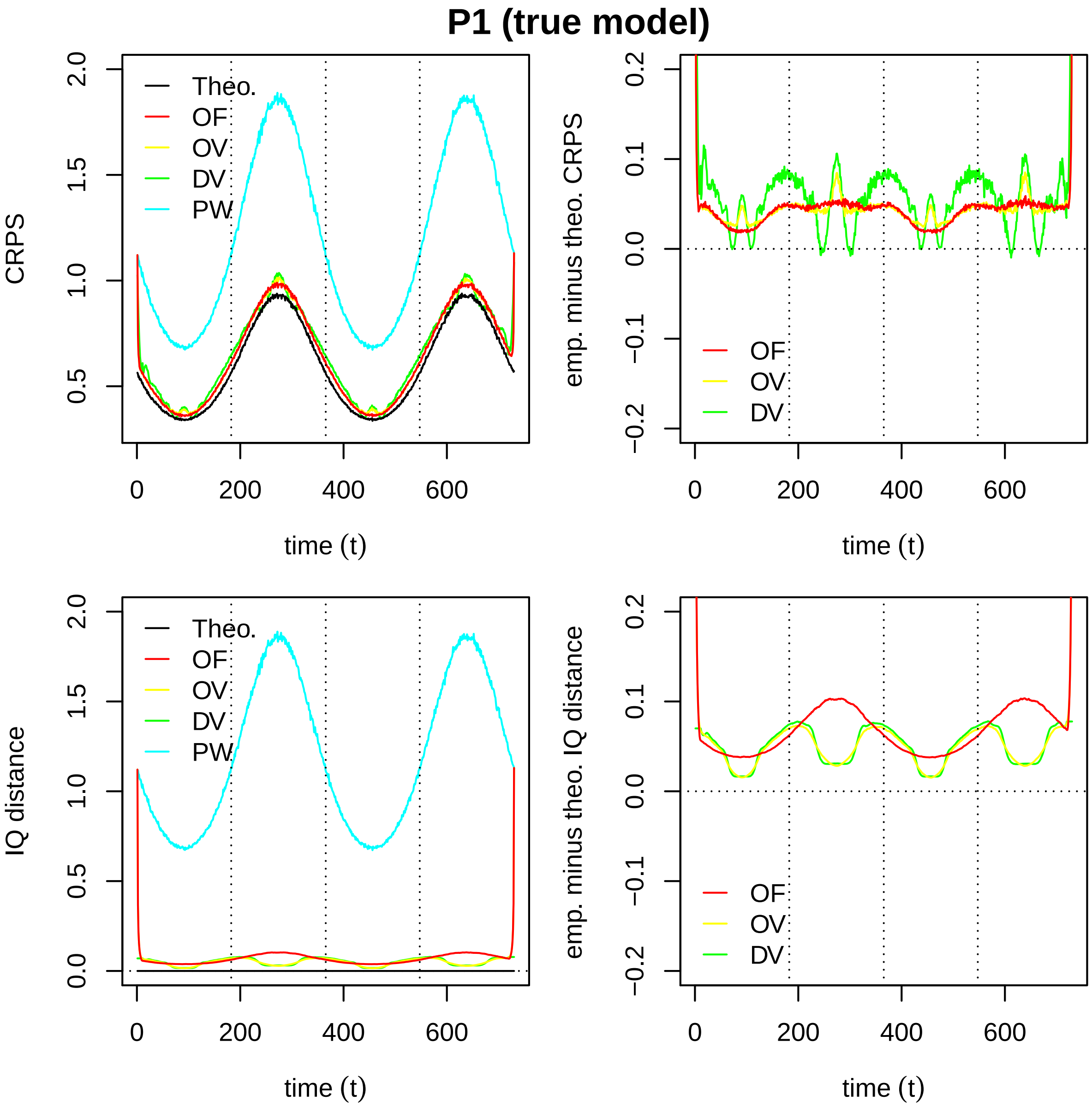


Figure 5.28: **Periodicity scenario (P1/true model):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

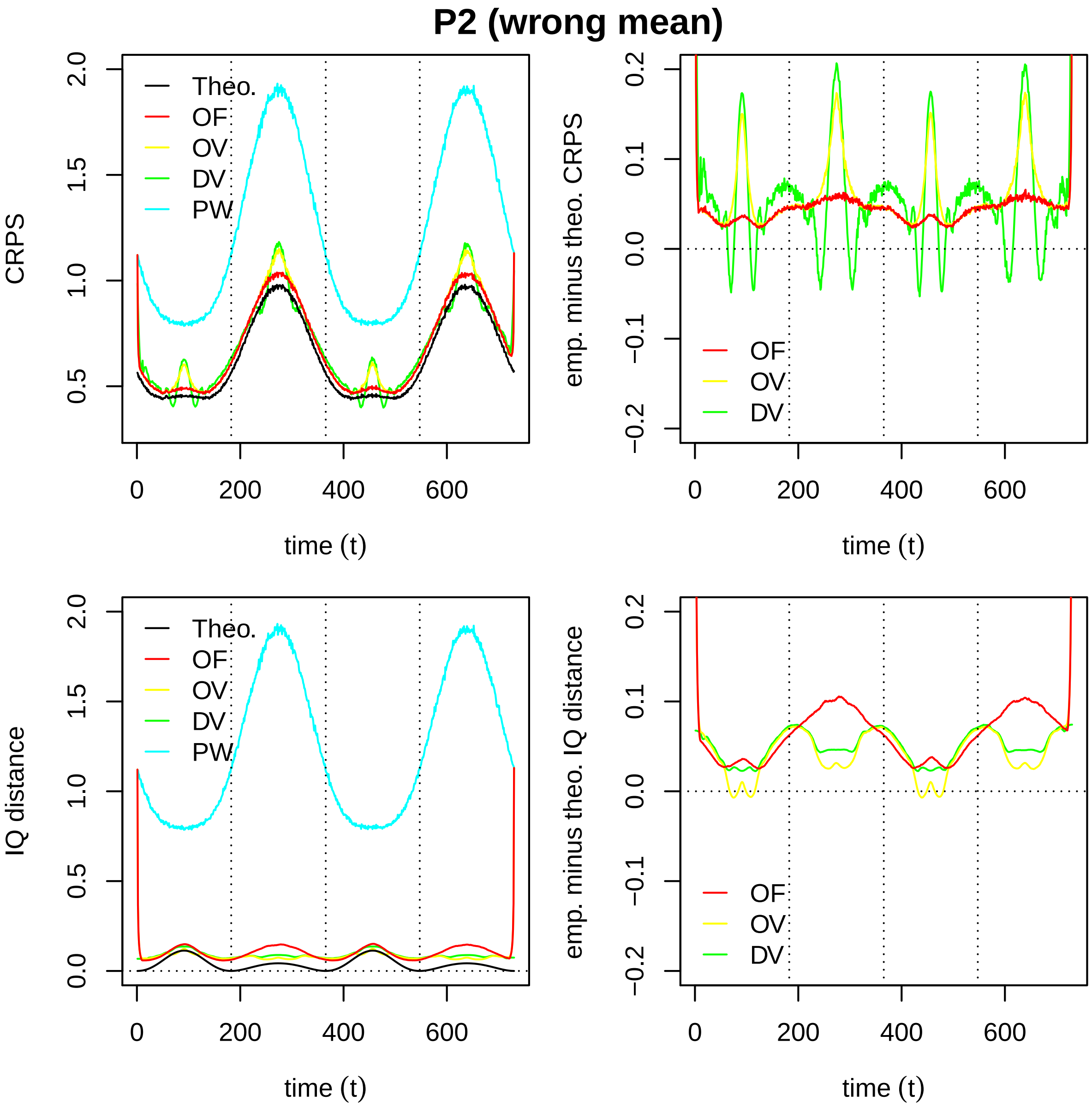


Figure 5.29: **Periodicity scenario (P2/wrong mean):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

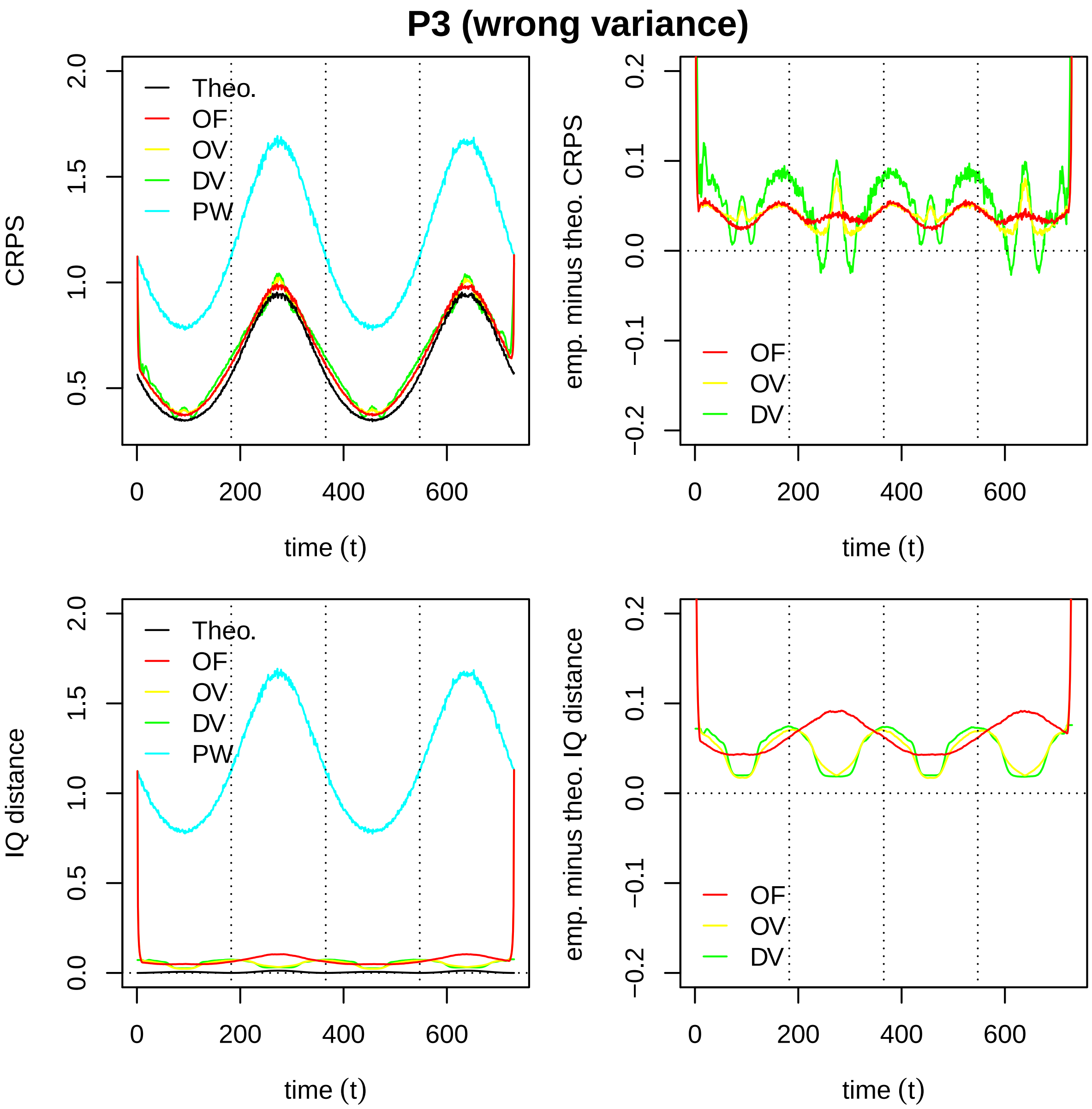


Figure 5.30: **Periodicity scenario (P3/wrong variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

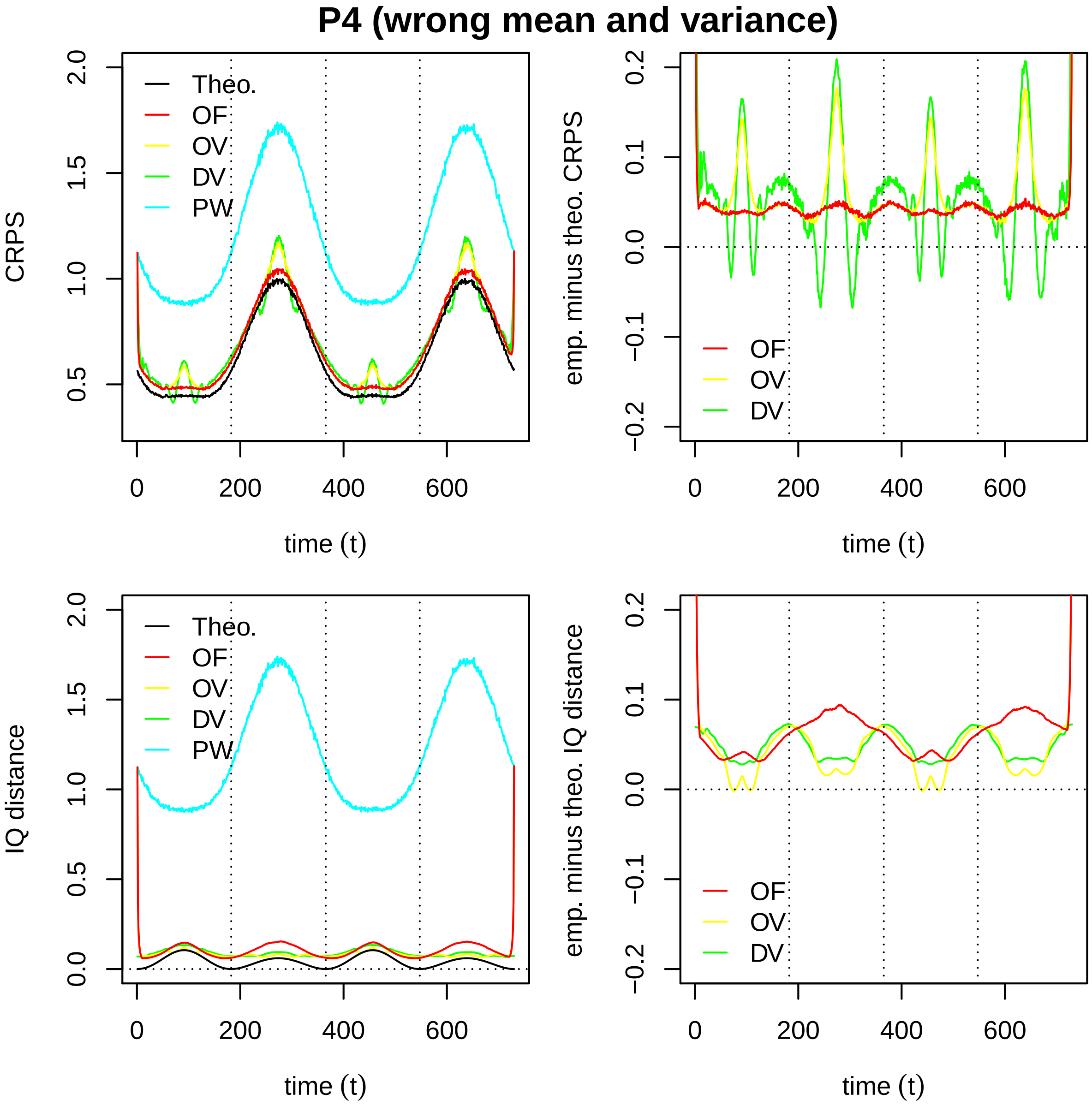


Figure 5.31: **Periodicity scenario (P4/wrong mean and variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

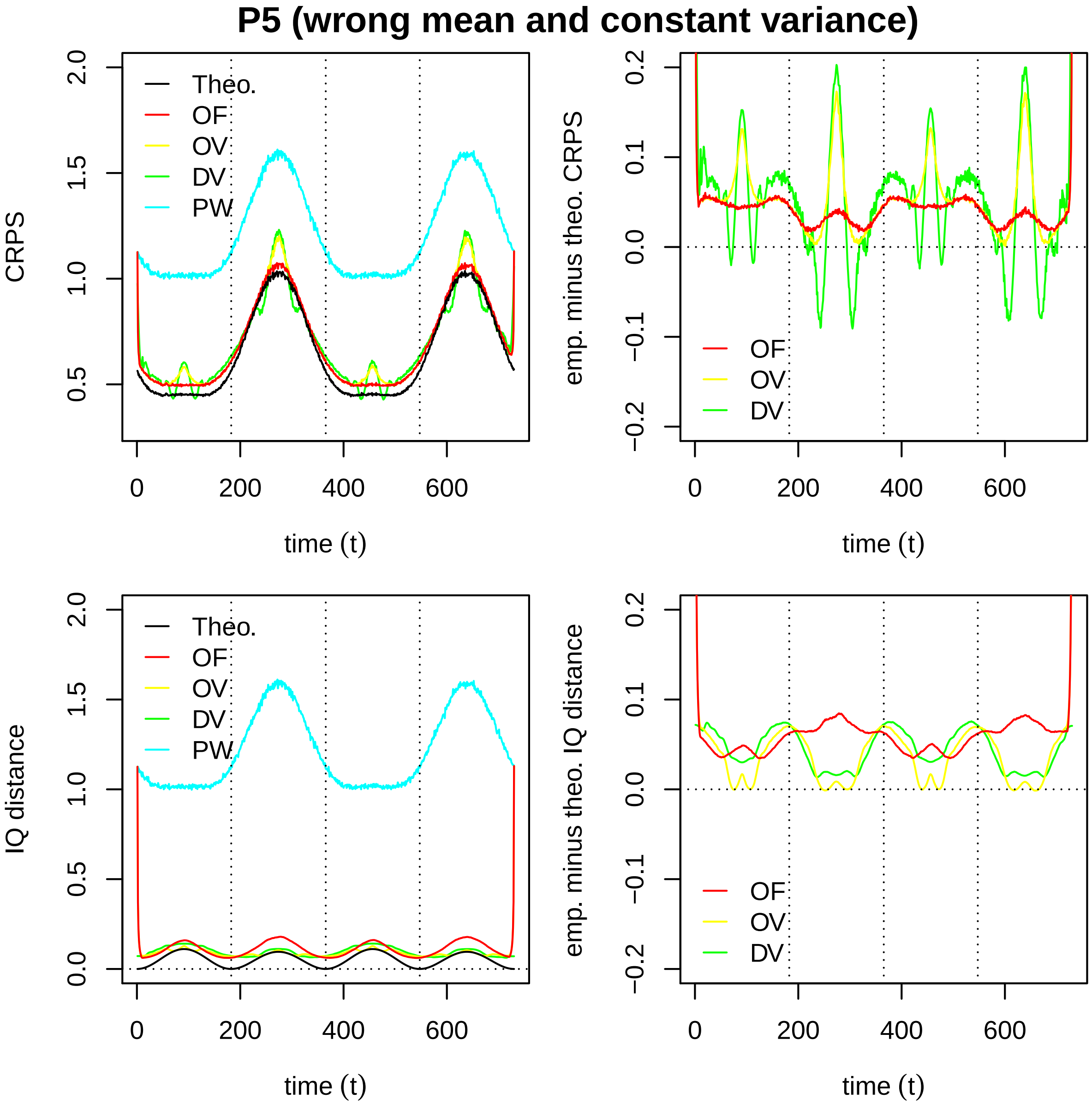


Figure 5.32: **Periodicity scenario (P5/wrong mean and constant variance):** Comparison of empirical scores/divergences computed using different approaches (OF, OV, DV, PW, ST) with theoretical (Theo.) scores/divergences. CRPS time series (upper left panel) and difference between empirical CRPS (OF, OV, DV, ST) and theoretical CRPS time series (upper right panel). IQ distance time series (lower left panel) and difference between empirical IQ distance (OF, OV, DV, ST) and theoretical IQ distance time series (lower right panel). The depicted time series are averages based on all 10,000 replications of the simulation study. The right panel does not show the (complete) time series for all different approaches, since some of them differ too much from 0 (for some time instances).

- For the start and the end of the time series (border case), the OF- and OV-scores and divergences tend towards the PW-scores and divergences, respectively.
- The deviation of the moving (OF-, OV- and DV-) scores from the theoretical scores is higher for time periods of higher variance of the data generating process and also for time periods of steep trends.
- Among the different types of moving scores, the OF-scores deviate least from the theoretical scores. The OV- and DV-scores are far more sensitive to periodic oscillations of mean and variance.
- The deviation of the moving OF-divergences from the theoretical divergences is higher for time periods of higher variance of the data generating process.
- The deviation of the moving OV- and DV-divergences from the theoretical divergences is higher for time periods of steep trends.
- Among the different types of moving divergences, the OV-divergences reflect the temporal evolution of the theoretical divergences best. The DV-divergences generally yield similar results as the OV-divergences.

All in all, we conclude from the above considerations that under the presence of periodicity we favor OF-scores and OV-divergences, which approximate their theoretical counterparts reasonably well.

Results (model comparison based on different evaluation approaches) Figures 5.33 and 5.34 compare Models (P1)–(P5) based on (time series of) theoretical/empirical CRPS and IQ distances, respectively. We consider theoretical (Theo.), moving OF-, moving OV-, moving DV- and point-wise (PW) scores/divergences. In order to discuss the results, we differentiate between four segments of equal length of the time series, as indicated with dotted vertical lines in Figures 5.33 and 5.34. For the first and third segment, the mean of the data generating process reaches its maximum, the variance its minimum. For the second and fourth segment, the mean of the data generating process reaches its minimum, the variance its maximum. In between (transition from one segment to the other) the mean and the variance of the data generating process increase or decrease, the curve corresponding to the realization is steep.

For the scores we observe the following from Figure 5.33:

- Theoretical (Theo.) scores:
  - The true model (P1) is strictly the best. The differences between the five models can be explained by their different misspecifications of mean and variance.
  - For the steep parts of the time series (transition from one segment to the other), the scores of all models move closely together, as their means and variances deviate not or not much from the data generating process.
  - Where the mean of the data generating process reaches its maximum and the variance its minimum (first and third segment), the scores of Models (P1) and (P3) are similar to each other, also the scores of Models (P2), (P4) and (P5). That is due to the fact, that the mean of Models (P1) and (P3) is the same, as well as the mean of Models (P2), (P4) and (P5).

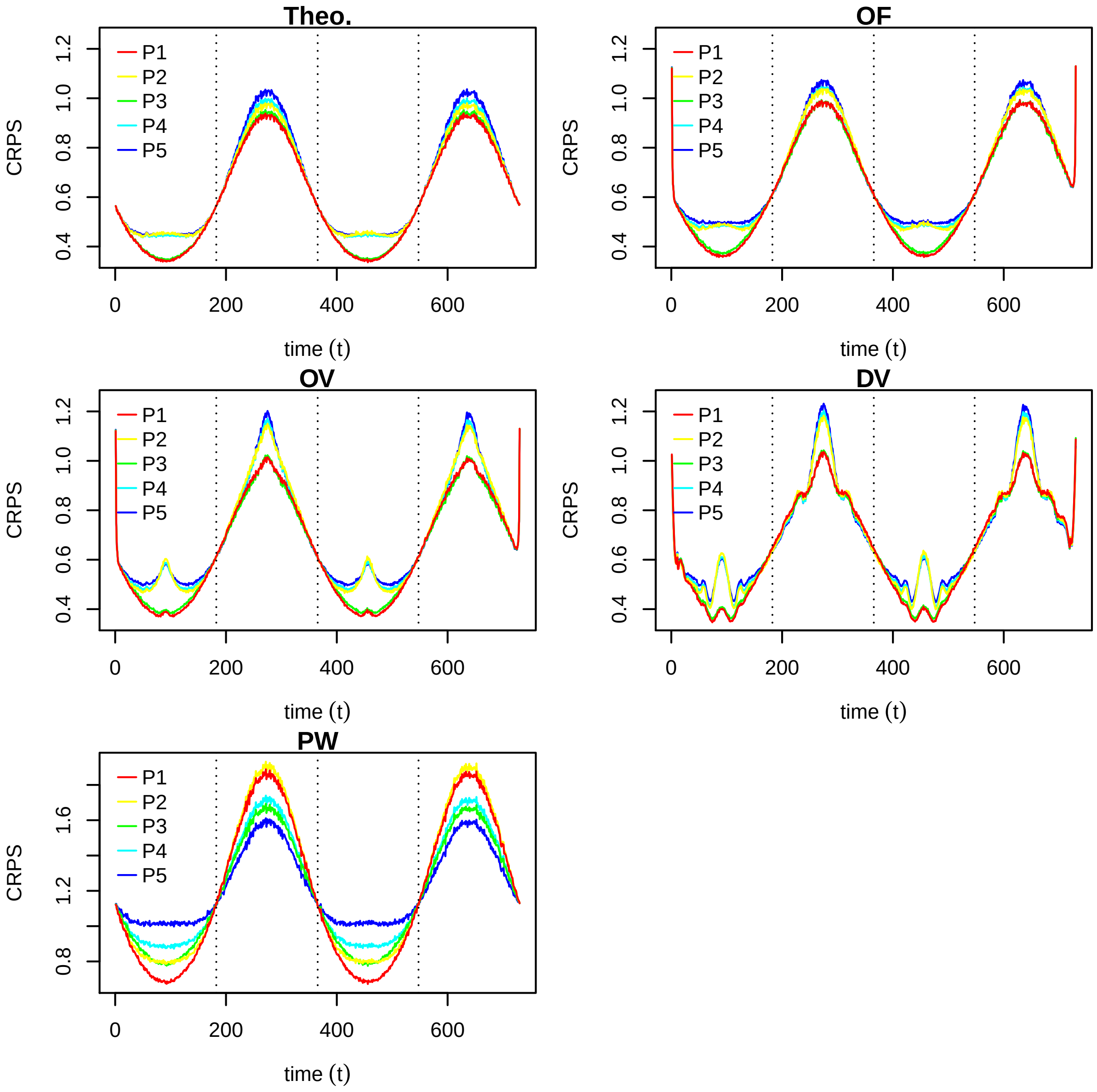


Figure 5.33: **Periodicity scenario (P):** Comparison of Models (P1)–(P5) based on (time series of) theoretical/empirical CRPS. Comparison based on theoretical scores (upper left panel), moving scores computed using the OF approach (upper right panel), moving scores computed using the OV approach (middle left panel), moving scores computed using the DV approach (middle right panel) and point-wise (PW) scores (lower left panel).

  - Where the mean of the data generating process reaches its minimum and the variance its maximum (second and fourth segment), the scores of the different models deviate most from each other.
  - Model (P2) performs worse than (P1), due to its misspecification of the mean.
  - Model (P3) performs worse than (P1) and Models (P4) and (P5) perform worse than (P2), due to their misspecification of the variance.
  - The misspecification of Model (P5) is worst.

- Moving scores with OF-windows:
  - The scores behave very similar to the theoretical scores.
  - For some time instances, the ranking of Models (P2) and (P4) is switched in comparison to the theoretical scores.
  - For the second and fourth segment the differentiation between Models (P1) and (P3) becomes difficult.

- Moving scores with OV-windows:
  - Especially in the centers of all four segments, where the mean and the variance of the data generating process reach their extremes, the scores behave different from the theoretical scores.
  - Except for Models (P2) (first and third segment) and (P3) (second and fourth segment), the models get ranked correctly for most time instances.
  - In comparison to the OF-scores, the OV-scores behave much less like the theoretical scores.

- Moving scores with DV-windows:
  - Compared to the OV-scores the similarity to the theoretical scores diminishes further.
  - Especially for time instances in the second and fourth segment wrong model rankings occur.

- Point-wise (PW) scores:
  - The model rankings are false for (almost) all time instances.
  - For the first and the third segment, Model (P2) is evaluated too good, Model (P3) too bad.
  - For the second and the fourth segment, Models (P1) and (P2) are evaluated too bad, Model (P5) is evaluated too good.

For the divergences we observe the following from Figure 5.34:

- Theoretical (Theo.) divergences:
  - The true model (P1) is strictly the best. The differences between the five models can be explained by their different misspecifications of mean and variance.
  - For the steep parts of the time series (transition from one segment to the other), the divergences of all models are close to 0, as their means and variances deviate not or not much from the data generating process.

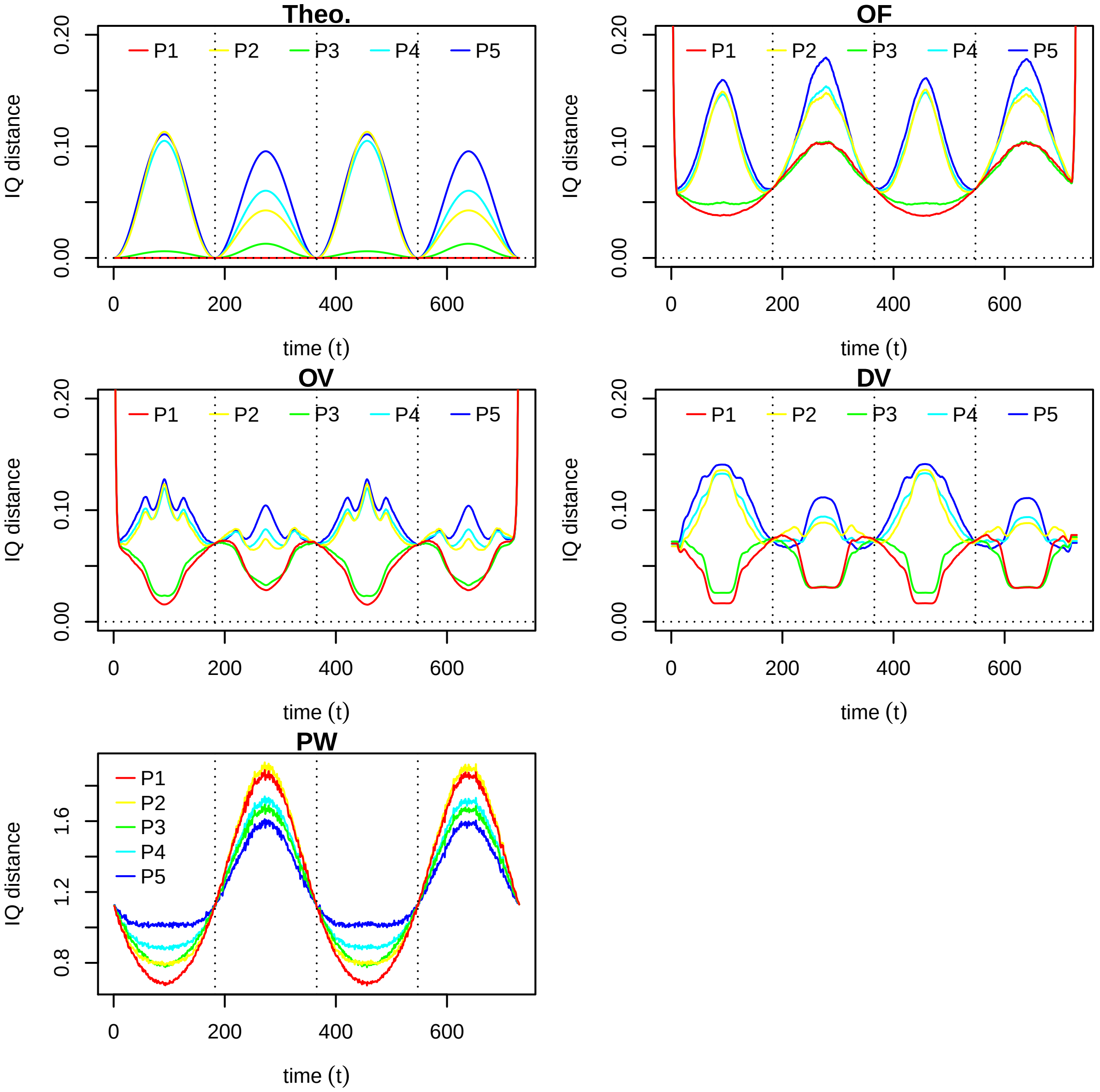


Figure 5.34: **Periodicity scenario (P):** Comparison of Models (P1)–(P5) based on (time series of) theoretical/empirical IQ distances. Comparison based on theoretical divergences (upper left panel), moving divergences computed using the OF approach (upper right panel), moving divergences computed using the OV approach (middle left panel), moving divergences computed using the DV approach (middle right panel) and point-wise (PW) divergences (lower left panel).

  - Where the mean of the data generating process reaches its maximum and the variance its minimum (first and third segment), the divergences of Models (P1) and (P3) are close to each other, also the divergences of Models (P2), (P4) and (P5). That is due to the fact, that the mean of Models (P1) and (P3) is the same, as well as the mean of Models (P2), (P4) and (P5).
  - Where the mean of the data generating process reaches its minimum and the variance its maximum (second and fourth segment), the divergences of the different models deviate most from each other.
  - Model (P2) performs worse than (P1), due to its misspecification of the mean.
  - Model (P3) performs worse than (P1) and Models (P4) and (P5) perform worse than (P2), due to their misspecification of the variance.
  - For most time instances, the misspecification of Model (P5) is worst.

- Moving divergences with OF-windows:
  - The divergences for the true model (P1) are greater than 0. For the first and the third segment, they are comparatively low. For the second and the forth segment, they are comparatively high.
  - With some exceptions, the divergences behave similar to the theoretical divergences.
  - For some time instances (first and third segment), the rankings of Models (P2), (P4) and (P5) are switched in comparison to the theoretical divergences.
  - For the second and fourth segment the differentiation between Models (P1) and (P3) is not as unambiguous as for the theoretical divergences.

- Moving divergences with OV-windows:
  - The temporal evolution of the OV-divergences differs from that of the theoretical divergences.
  - In terms of wrong model rankings the OV-divergences behave very similar to the OF-divergences.

- Moving divergences with DV-windows:
  - The (relative) temporal evolution of the DV-divergences differs very much from that of the theoretical divergences.
  - Especially for time instances in the second and fourth segment, the indicated model rankings are completely wrong.

- Point-wise (PW) divergences:
  - The point-wise IQ distances are equal to the point-wise CRPS. Hence, we make the same observations as above.

Table 5.8: **Periodicity scenario (P):** Comparison/ranking of Models (P1)–(P5) based on average moving SE scores, average moving CRPS, average moving MV divergences and average moving IQ distances. Average scores/divergences (left) and model rankings (right) are provided, distinguishing between theoretical (Th) scores/divergences (true model ranking), moving scores/divergences computed using the OF approach, moving scores/divergences computed using the OV approach, moving scores/divergences computed using the DV approach and point-wise (PW) scores/divergences. Note that moving divergence averages based on OF- and OV-windows do not warrant propriety.

		average scores/divergences					model rank				
		Th	OF	OV	DV	PW	Th	OF	OV	DV	PW
SE score	(P1)	1.266	1.341	1.361	1.427	2.531	1	2	2	2	4
	(P2)	1.391	1.485	1.589	1.574	2.656	3	5	5	5	5
	(P3)	1.266	1.330	1.357	1.424	2.328	1	1	1	1	1
	(P4)	1.391	1.474	1.585	1.571	2.453	3	4	3	3	3
	(P5)	1.391	1.471	1.586	1.572	2.390	3	3	4	4	2
CRPS	(P1)	0.600	0.643	0.645	0.657	1.200	1	1	1	1	2
	(P2)	0.638	0.683	0.699	0.694	1.238	3	3	3	3	5
	(P3)	0.605	0.646	0.646	0.659	1.178	2	2	2	2	1
	(P4)	0.641	0.684	0.700	0.695	1.214	4	4	4	4	3
	(P5)	0.652	0.694	0.706	0.702	1.216	5	5	5	5	4
MV div.	(P1)	0.000	0.149	0.101	0.087	2.531	1	2	2	2	4
	(P2)	0.125	0.273	0.221	0.234	2.656	3	5	5	5	5
	(P3)	0.000	0.137	0.097	0.084	2.328	1	1	1	1	1
	(P4)	0.125	0.261	0.218	0.231	2.453	3	4	3	3	3
	(P5)	0.125	0.258	0.218	0.231	2.390	3	3	4	4	2
IQ dist.	(P1)	0.000	0.071	0.054	0.051	1.200	1	1	1	1	2
	(P2)	0.038	0.105	0.085	0.089	1.238	3	3	3	3	5
	(P3)	0.005	0.073	0.057	0.053	1.178	2	2	2	2	1
	(P4)	0.041	0.106	0.086	0.089	1.214	4	4	4	4	3
	(P5)	0.052	0.114	0.092	0.097	1.216	5	5	5	5	4

Model ranking based on average scores and divergences In order to have an overview how the different approaches rank the models (P1)–(P5), we provide Table 5.8. Besides the CRPS and the IQ distance, we now also consider SE scores and MV divergences. The table provides the average scores and divergences (averages over all time instances and all 10,000 replications of the simulation study) for the different considered evaluation approaches (OF, OV, DV and PW) and compares them to the corresponding theoretical (Th) scores and divergences, respectively. Note that utilization of moving divergence averages is not necessarily a proper evaluation method (that is in case of OF- and OV-windows; see discussion in Sections 5.3.2 and 5.3.3). Based on the average scores/divergences we rank the different models, distinguishing between the theoretical scores/divergences and the four different evaluation approaches. Hence, for each approach (column) and separately for each score/divergence type (SE score, CRPS, MV divergence, IQ distance), the model with the smallest average score/divergence is ranked best (1) and the model with the highest average score/divergence is ranked worst (5). Accordingly, the models with intermediate scores/divergences are ranked 2–4. If the score/divergence averages of two or more models are equal, all of these models get the same (minimum) rank. The resulting model rankings are compared in the right half of the table.

To decide if an evaluation approach is good or not, we again consider the following two criteria. For good evaluation approaches

(E1) the true model (P1) should be ranked lowest (best), and

(E2) the ranking should be identical to that based on the theoretical (Th) scores/divergences.

All in all, we observe that (E1) and (E2) are fulfilled for the CRPS and the IQ distances calculated based on the OF, OV and DV approach.

Now, we discuss the results from Table 5.8 in more detail, distinguishing between the different types of scores and divergences:

- SE scores:
  - The theoretical scores can not differentiate between Models (P1) and (P3), and also not between (P2), (P4) and (P5). This is due to the fact, that SE scores neglect misspecifications of the variance.
  - The moving (OF-, OV- and DV-) scores falsely rank (P3) better than (P1), and (P5) better than (P2).
  - The OF-score and the OV- and DV-score based rankings differ from each other.
  - The ranking based on the PW-scores is wrong.
- CRPS:
  - The scores are able to differentiate between all five models.
  - The moving (OF-, OV- and DV-) scores rank all five models correctly.
  - The PW-scores yield a faulty model ranking. (P3) is ranked too good and (P2) too bad.
- MV divergences:
  - We obtain the same (faulty) model rankings as for the SE scores.
- IQ distances:
  - Same as for the CRPS, the moving (OF, OV and DV) IQ distances rank all five models correctly.
  - The ranking based on the PW-divergences (same as for the CRPS) is wrong.

Further analysis of the model rankings Whereas in Table 5.8 we investigated model rankings which were obtained by averaging over all 10,000 replications of the simulation study (we refer to them as *average based model rankings*), we now look into all 10,000 replications separately, to see if the individual model rankings correspond to the *average based model rankings* found in Table 5.8. For this we consider the models pair-wise and calculate the share of all 10,000 replications for which the average (moving) score/divergence of the first model is smaller (better) than the average (moving) score/divergence of the second model. This time the (moving) scores/divergences are averaged only temporally, that is over all available time instances. The results of these computations (shares) are summarized in Tables 5.9 and 5.10. Table 5.9 is

Table 5.9: **Periodicity scenario (P):** Pair-wise comparison of Models (P1)–(P5), distinguishing between theoretical CRPS (Th), moving CRPS computed using the OF, OV and DV approach and point-wise CRPS (PW). Each table entry equals the share of all 10,000 replications of the simulation study for which the average (moving) score of the model indicated in the corresponding row is smaller (better) than the average (moving) score of the model indicated in the corresponding column of the table.

	Th					**OF**				
	(P1)	(P2)	(P3)	(P4)	(P5)	(P1)	(P2)	(P3)	(P4)	(P5)
(P1)	–	1.00	1.00	1.00	1.00	–	1.00	0.88	1.00	1.00
(P2)	0.00	–	0.00	0.95	1.00	0.00	–	0.00	0.71	0.98
(P3)	0.00	1.00	–	1.00	1.00	0.12	1.00	–	1.00	1.00
(P4)	0.00	0.05	0.00	–	1.00	0.00	0.29	0.00	–	1.00
(P5)	0.00	0.00	0.00	0.00	–	0.00	0.02	0.00	0.00	–
	PW					**OV**				
	(P1)	(P2)	(P3)	(P4)	(P5)	(P1)	(P2)	(P3)	(P4)	(P5)
(P1)	–	1.00	0.00	0.85	0.84	–	1.00	0.79	1.00	1.00
(P2)	0.00	–	0.00	0.00	0.02	0.00	–	0.00	0.56	0.93
(P3)	1.00	1.00	–	1.00	1.00	0.21	1.00	–	1.00	1.00
(P4)	0.15	1.00	0.00	–	0.65	0.00	0.44	0.00	–	1.00
(P5)	0.16	0.98	0.00	0.35	–	0.00	0.07	0.00	0.00	–
						DV				
						(P1)	(P2)	(P3)	(P4)	(P5)
(P1)						–	1.00	0.84	1.00	1.00
(P2)						0.00	–	0.00	0.63	0.96
(P3)						0.16	1.00	–	1.00	1.00
(P4)						0.00	0.37	0.00	–	1.00
(P5)						0.00	0.04	0.00	0.00	–

based on CRPS, Table 5.10 on IQ distances. Both tables again distinguish between theoretical scores/divergences (Th) and empirical scores/divergences calculated using different approaches (OF, OV, DV and PW). Each block of Tables 5.9 and 5.10 corresponds to one particular approach. It is best to read each block row-wise. Each row provides the shares telling how often the model indicated to the left (model under consideration) had a better average score/divergence than the other models (indicated above). A share equal to 0 means that the model under consideration is never considered better than the other model, a share equal to 1 means that it is always considered better than the other one, a share in $(0, 0.5)$ means that the model under consideration is considered worse than the other one for the majority of all replications and a share in $(0.5, 1)$ means that it is considered better than the other one for the majority of all replications.

First, we are interested to know, if the pair-wise and replication-wise consideration in Tables 5.9 and 5.10 supports the model rankings found in Table 5.8. If we rank model A better than model B if the share corresponding to the model pair (A, B) is greater than 0.5, then we obtain exactly the same rankings as in Table 5.8.

Second, we are interested to see, how certain the different evaluation approaches are about the relative pair-wise rankings they indicate. Table entries close to 0 or 1 support certainty, entries close to 0.5 indicate that the evaluation approach under consideration has difficulties to

Table 5.10: **Periodicity scenario (P):** Pair-wise comparison of Models (P1)–(P5), distinguishing between theoretical IQ distances (Th), moving IQ distances computed using the OF, OV and DV approach and point-wise IQ distances (PW). Each table entry equals the share of all 10,000 replications of the simulation study for which the average (moving) divergence of the model indicated in the corresponding row is smaller (better) than the average (moving) divergence of the model indicated in the corresponding column of the table.

	Th					**OF**				
	(P1)	(P2)	(P3)	(P4)	(P5)	(P1)	(P2)	(P3)	(P4)	(P5)
(P1)	–	1.00	1.00	1.00	1.00	–	1.00	0.84	1.00	1.00
(P2)	0.00	–	0.00	1.00	1.00	0.00	–	0.00	0.66	0.98
(P3)	0.00	1.00	–	1.00	1.00	0.16	1.00	–	1.00	1.00
(P4)	0.00	0.00	0.00	–	1.00	0.00	0.34	0.00	–	1.00
(P5)	0.00	0.00	0.00	0.00	–	0.00	0.02	0.00	0.00	–
	PW					**OV**				
	(P1)	(P2)	(P3)	(P4)	(P5)	(P1)	(P2)	(P3)	(P4)	(P5)
(P1)	–	1.00	0.00	0.85	0.84	–	1.00	0.93	1.00	1.00
(P2)	0.00	–	0.00	0.00	0.02	0.00	–	0.00	0.75	0.98
(P3)	1.00	1.00	–	1.00	1.00	0.07	1.00	–	1.00	1.00
(P4)	0.15	1.00	0.00	–	0.65	0.00	0.25	0.00	–	1.00
(P5)	0.16	0.98	0.00	0.35	–	0.00	0.02	0.00	0.00	–
						DV				
						(P1)	(P2)	(P3)	(P4)	(P5)
(P1)						–	1.00	0.84	1.00	1.00
(P2)						0.00	–	0.00	0.63	0.96
(P3)						0.16	1.00	–	1.00	1.00
(P4)						0.00	0.37	0.00	–	1.00
(P5)						0.00	0.04	0.00	0.00	–

decide in favor of one of the corresponding models. In most cases we observe shares close to 0 or 1. We point out cases, where this does not hold. The PW-scores and divergences for instance have difficulties with the relative ranking of models (P4) and (P5). The theoretical CRPS mostly allow for unambiguous decisions, only the decision between models (P2) and (P4) is not always the same. Also based on OF-, OV- and DV-scores and divergences comparison in that particular case does not always seem to yield the same result. In contrast, the theoretical IQ distances always allow for unambiguous decisions.

All in all, we conclude that the average based model rankings found in Table 5.8 (for CRPS and IQ distances) and the differentiation between the different models in these rankings are justified in almost all cases.

5.8 Summary of simulation study results

Table 5.11 summarizes the results of our simulation studies conducted in Sections 5.5–5.7. Amongst others, we found that depending on the phenomenon of interest, certain types of moving windows might be more adequate than others. Moreover, in contrast to PW- or ST-scores and divergences, moving (OF, OV, DV) scores and divergences are (in most cases) able to yield an adequate (true) model ranking. One exception where the moving scores and divergences yield a faulty model ranking is the trend scenario, where only the relative ranking of the models (T1) and (T3) is inappropriate (∗). Note, that for the periodicity scenario no ST-scores and divergences were considered, due to computational infeasibility.

Table 5.11: Summary of simulation study results. The table summarizes for the changepoint scenario (C), the trend scenario (T) and the periodicity scenario (P), for which parts of the time series the evaluation based on moving scores and divergences is most difficult (critical parts), which empirical evaluation approaches are most reliable, for which evaluation approaches the true model is ranked best, for which evaluation approaches the model ranking is identical to the ranking based on theoretical scores/divergences (consistency of model ranking) and for which evaluation approaches the model rankings are unreliable (together with a count of model pairs for which the relative ranking is unreliable). The relative ranking of two models is considered unreliable, if it differs for more than 25% of all replications of the simulation study.

	Changepoint scenario	**Trend scenario**	**Periodicity scenario**
		critical parts	
	close to changepoints	steep parts	steep parts and parts with high variance
		most reliable empirical evaluation approaches	
	DV-scores and DV-divergences	OF- and OV-scores and divergences	OF-scores and OV-divergences
		true model ranked best	
SE scores	–	–	–
CRPS	OF, OV, DV, ST	–	OF, OV, DV
MV div.	–	–	–
IQ dist.	OF, OV, DV, ST	–	OF, OV, DV
		consistency of model ranking	
SE scores	–	–	–
CRPS	OF, OV, DV	(∗)	OF, OV, DV
MV div.	–	–	–
IQ dist.	OF, OV, DV	(∗)	OF, OV, DV
		unreliable evaluation approaches (# model pairs with unreliable relative ranking)	
CRPS	ST(1)	OF(2), OV(2), DV(2), ST(4)	OF(1), OV(1), DV(1), PW(1)
IQ dist.	ST(1)	OF(1), OV(1), DV(2), ST(4)	OF(1), OV(1), DV(1), PW(1)

5.9 Case study: Evaluation of Regional Climate Models

As an application of the presented methodology we evaluate four selected Regional Climate Models (RCMs), which are part of the multi-model RCM ensemble compiled by the *ENSEMBLES project* (see van der Linden and Mitchell, 2009). Specifically, we consider the RCMs of

- the Danish Meteorological Institute (*DMI*),
- the Royal Netherlands Meteorological Institute (*KNMI*),
- the Max-Planck-Institute for Meteorology (*MPI*), and
- the Swedish Meteorological and Hydrological Institute (*SMHI*),

driven by

- the *ERA-40* reanalysis data, and
- Global Climate Model simulations from the *ECHAM5* model,

respectively. More details are provided in Section 3.4. As reference data (observations) for the evaluation we use the gridded E-OBS data set (version 13.1, Haylock et al., 2008). See Section 3.3 for more details. In particular, the evaluation is based on the *daily mean temperature* output from the DMI, KNMI, MPI and SMHI model for the years 1961–1990. The RCM simulations as well as the E-OBS data are all available on the same spatial grid (0.22° rotated pole grid, North Pole at $39.25N$, $162W$) covering Europe. We restrict the analysis to an area covering most of Europe as depicted in Figure 3.4.

5.9.1 Overall model assessment

For the final model evaluation we calculate

- moving Continuous Ranked Probability Scores (CRPS, see Equations (5.9) and (2.32)), and
- moving Integrated Quadratic (IQ) distances (see Equations (5.10) and (2.42)),

based on

- overlapping windows with fixed width (OF, see Equation (5.14)),
- overlapping windows with varying width (OV, see Equation (5.16)),
- and disjoint windows with varying width (DV, see Equation (5.17)).

Moreover, we compute point-wise (PW) scores (5.5) and divergences (5.6) for comparison. Recall, that point-wise CRPS and point-wise IQ distances are both equivalent to the absolute error (2.28) and thus yield the same results. Hence, we subsequently do not differentiate further between point-wise CRPS and point-wise IQ distances. All computations are then conducted for all eight models under consideration and all 10937 (land) pixels of the study region, based on the period 1961–1990.

Table 5.12 provides overall (spatial and temporal) averages of moving CRPS, moving IQ distances and PW-scores/divergences. Note that utilization of moving divergence averages is not necessarily a proper evaluation method (that is in case of OF- and OV-windows; see discussion

Table 5.12: Overall (spatial and temporal) averages of moving CRPS, moving IQ distances and PW-scores/divergences for the DMI, KNMI, MPI, SMHI models driven by ECHAM5 (top) and ERA-40 (bottom), respectively. The relative score/divergence based rank of each model (distinguishing between ECHAM5 and ERA-40 boundary conditions) is given in brackets. Note that moving divergence averages based on OF- and OV-windows do not warrant propriety.

ECHAM5		DMI		KNMI		MPI		SMHI	
	OF	2.73	(4th)	2.42	(1st)	2.49	(3rd)	2.46	(2nd)
CRPS	OV	2.70	(4th)	2.41	(1st)	2.48	(3rd)	2.45	(2nd)
	DV	2.55	(4th)	2.26	(1st)	2.32	(3rd)	2.29	(2nd)
	PW	4.57	(4th)	4.17	(3rd)	4.07	(2nd)	3.99	(1st)
	OF	0.96	(4th)	0.66	(1st)	0.72	(3rd)	0.69	(2nd)
IQ	OV	0.92	(4th)	0.64	(1st)	0.70	(3rd)	0.67	(2nd)
	DV	1.04	(4th)	0.75	(1st)	0.81	(3rd)	0.77	(2nd)
ERA-40		DMI		KNMI		MPI		SMHI	
	OF	1.95	(1st)	1.98	(2nd)	2.07	(4th)	1.99	(3rd)
CRPS	OV	1.94	(1st)	1.96	(2nd)	2.05	(4th)	1.98	(3rd)
	DV	1.77	(1st)	1.79	(2nd)	1.89	(4th)	1.81	(3rd)
	PW	2.02	(2nd)	2.11	(3rd)	2.24	(4th)	1.99	(1st)
	OF	0.25	(1st)	0.27	(2nd)	0.36	(4th)	0.28	(3rd)
IQ	OV	0.24	(1st)	0.25	(2nd)	0.34	(4th)	0.27	(3rd)
	DV	0.26	(1st)	0.27	(2nd)	0.38	(4th)	0.29	(3rd)

in Sections 5.3.2 and 5.3.3). As expected, the ERA-40 driven models achieve better average scores/divergences as the GCM driven models. While the CRPS and the IQ divergences yield identical model rankings for all three different types of moving windows (OF, OV, DV), the PW-scores/divergences yield a different ranking. As we learned in our simulation study (Section 5.4), point-wise (PW) scores/divergences might lead to a spurious model ranking, since they do not account for higher order structures of the observed/modeled phenomenon. The results of Table 5.12 corroborate these findings. Moving scores and divergences provide a more holistic model assessment and hence should be preferred over a point-wise evaluation (that is, also over the traditional evaluation measures introduced in Section 2.6). It seems, that for the application at hand the choice of the moving window selection approach is not crucial.

5.9.2 Temporal evaluation

To assess the model accuracy over time we consider spatial averages (over the whole study area, that is over all 10937 pixels) of the (moving) scores and divergences obtained in Section 5.9.1. To illustrate the obtained averaged time series (see Figure 5.35), we further smooth them and consider only monthly averages of these daily time series. Figure 5.35 compares the resulting time series for the moving CRPS with OV-windows, the point-wise (PW) scores/divergences and the moving IQ distances with OV-windows. The corresponding figures for OF- and DV-windows can be found in Appendix B, as there is not much difference to the results for the OV-windows.

Looking at Figure 5.35, we observe obvious differences between the three metrics behind the displayed time series. While all three metrics are influenced by the seasonal oscillations of the phenomenon of interest, the moving IQ distances indicate a clear difference between the two

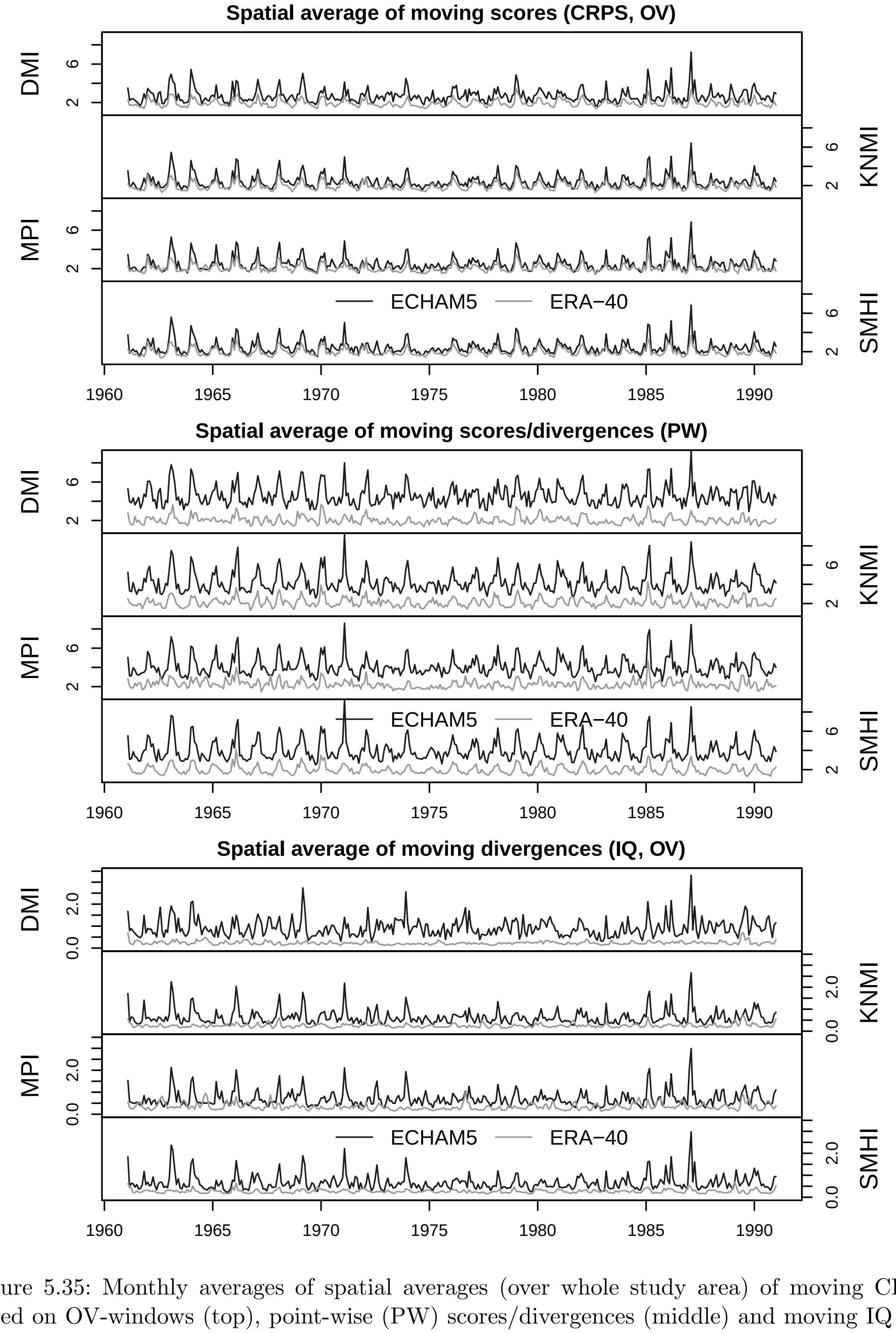


Figure 5.35: Monthly averages of spatial averages (over whole study area) of moving CRPS based on OV-windows (top), point-wise (PW) scores/divergences (middle) and moving IQ distances based on OV-windows (bottom), for the DMI, KNMI, MPI and SMHI models driven by ECHAM5 (black) and ERA-40 (gray).

different drivers (ECHAM5 and ERA-40). The moving IQ distance time series for the ERA-40 driven models oscillate much less in comparison to those for the ECHAM5 driven models. From that we conclude that the driving GCM (here ECHAM5) is not able to accurately capture the (regional) seasonal temperature dynamics.

While for the moving CRPS the differences between the different models and the different boundary conditions (ECHAM5 and ERA-40) appear to be rather small, the point-wise evaluation indicates big differences for the two different boundary conditions. This is due to the fact, that the ECHAM5 boundary conditions themselves are simulations from a Global Climate Model (GCM), while the ERA-40 boundary conditions are a reanalysis of climate observations. Hence, on a daily basis (point-wise comparison), the ERA-40 boundary conditions result in much more accurate RCM simulations. However, as we are also interested in the long term model accuracy (accurate representation of higher order characteristics of the phenomenon of interest), we prefer metrics which are able to simultaneously evaluate these long term properties.

Focusing on the moving CRPS and IQ distance time series, we observe time periods of generally weaker model performance (e.g. 1985–1987), compared to periods with much better model performance (e.g. 1974–1977). Comparison with the average temperatures illustrated in Figure 3.3 indicates a possible dependency of the model performance on how cold/mild the winters were. It seems that the models capture the temperature conditions during mild winters much better. Moreover, we find that the discrepancy distinguishing between the two different boundary conditions is more obvious for the DMI model compared to the other models. While the ERA-40 driven DMI model apparently performs best in comparison to the KNMI, MPI and SMHI model, the ECHAM5 driven DMI model performs rather worst (see also Table 5.12). Having a closer look at the moving score/divergence time series for the MPI, we observe short time periods where the ECHAM5 driven model seems to perform better than the ERA-40 driven model (e.g. 1964). This observation is more distinct considering the moving IQ distances.

To see if the above findings hold for different locations of the study area (Europe), we have to consider the moving scores/divergences for different locations separately instead of considering only a spatial average. We conduct such a spatial evaluation in the subsequent section.

5.9.3 Spatial evaluation

To be able to judge the model performance in different locations across Europe we look at the temporal averages of the (moving) score and divergence time series obtained in Section 5.9.1. Note that utilization of moving divergence averages is not necessarily a proper evaluation method (that is in case of OF- and OV-windows; see discussion in Sections 5.3.2 and 5.3.3). The temporal averages corresponding to the moving CRPS with OV-windows, the point-wise (PW) scores/divergences and the moving IQ distances with OV-windows are depicted in Figure 5.36. The corresponding figures for OF- and DV-windows are provided in Appendix B.

Again the distinction between the ERA-40 and the ECHAM5 boundary conditions is more pronounced for the point-wise scores/divergences and again we emphasize that a point-wise evaluation does not account for higher order structures of the phenomenon of interest. In particular a point-wise evaluation does not take the uncertainty of the phenomenon of interest into account, which may also vary across space. Moreover, the simulations of the ECHAM5 driven RCMs are generally much more uncertain, as already discussed in Section 5.9.2.

Among the ECHAM5 driven models, the DMI model performs worst (especially for the continental climate in eastern Europe). The MPI model (ECHAM5) has comparatively more problems in modeling the mean temperature in the north-east of Europe, where many lakes dominate the landscape. Moreover, Figure 5.36 indicates, that all models have difficulties to model the tem-

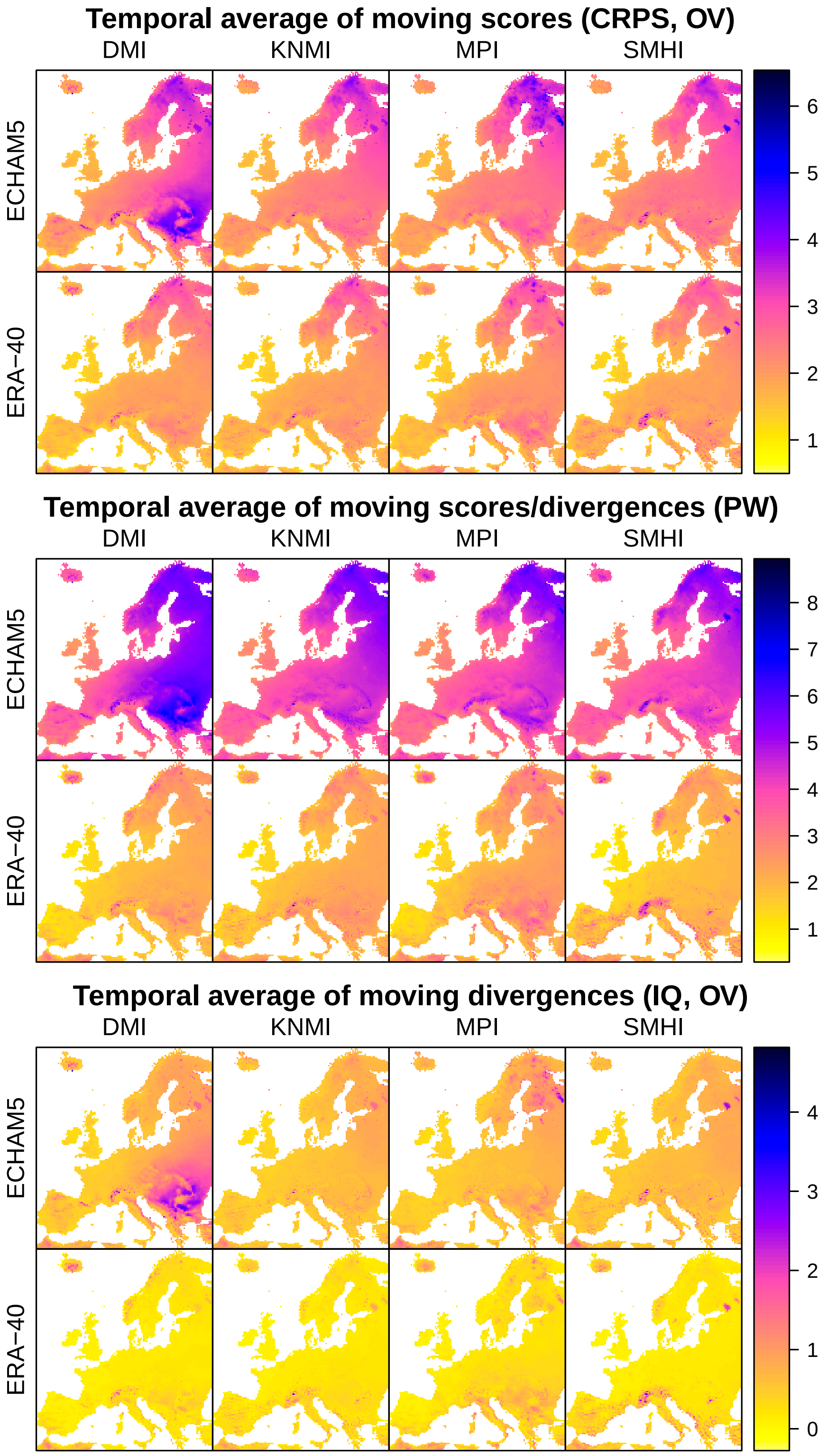

Temporal average of moving scores (CRPS, OV)
DMI
KNMI
MPI
SMHI
ECHAM5
ERA−40
6
5
4
3
2
1
Temporal average of moving scores/divergences (PW)
DMI
KNMI
MPI
SMHI
ECHAM5
ERA−40
8
7
6
5
4
3
2
1
Temporal average of moving divergences (IQ, OV)
DMI
KNMI
MPI
SMHI
ECHAM5
ERA−40
4
3
2
1
0

Figure 5.36: Maps (Europe) of temporal averages (1961–1990) of moving CRPS based on OV-windows (top), point-wise (PW) scores/divergences (middle) and moving IQ distances based on OV-windows (bottom), for the DMI, KNMI, MPI and SMHI models (ECHAM5/ERA-40). Note that moving divergence averages based on OV-windows do not warrant propriety.

perature over large bodies of water (e.g. lakes Ladoga and Onega in Russia). This observation is identified less well by the point-wise scores/divergences. The DMI model with ERA-40 boundary conditions apparently performs best for large water bodies. Further, we observe that most models have problems for mountainous areas (e.g. Alps, Pyrenees, Carpathians).

To see which model performs best in which area of Europe we provide maps (Figure 5.37) which show for each pixel of the study area which model has the lowest (best) average moving CRPS/IQ distance (distinguishing between OF-, OV- and DV-windows) and the lowest (best) average point-wise score/divergence. Again we differentiate between the two different boundary conditions (ECHAM5/ERA-40).

Whereas the moving CRPS and IQ distances indicate similar spatial patterns and mostly agree on the best model, the point-wise evaluation seems to favor the SMHI model for large parts of Europe, where the moving score/divergence approaches prefer a different model. We identify large connected areas where either the KNMI, the MPI or the SMHI model perform best for ECHAM5 boundary conditions. The DMI model seems to perform better in regions whose temperatures are more influenced by the sea. For ERA-40 boundary conditions, we identify large areas where the SMHI or the DMI model perform best. The MPI model is preferred for a large share of the Iberian Peninsula.

5.9.4 Comparison of results to an evaluation based on linear trends

Having conducted an extensive evaluation and model comparison based on our novel methodology, we now want to compare the results to one of the evaluation approaches undertaken in the literature (see discussion in Section 5.2). For this, we take the approach followed by Lorenz and Jacob (2010). Basically, we compare linear trends in the RCM output to those present in the E-OBS reference data set.

In order to compute linear trends for temperature time series, we first aggregate the daily time series to yearly time series y_t, $t = 1961, \dots, 1990$, by computing annual means. We then assume a linear regression model

$$y_t = \alpha + \beta t + \varepsilon_t, \quad t = 1961, \dots, 1990,$$

with intercept parameter α, linear trend parameter β and residuals ε_t, $t = 1961, \dots, 1990$. Based on least-squares estimation (minimization of the sum $\sum_{t=1961}^{1990}(y_t - \alpha - \beta t)^2$), we obtain estimates of the linear trend parameter $\widehat{\beta}$.

Conducting these computations separately for all $n = 10937$ pixels/locations $i = 1, \dots, n$ of our evaluation region (compare Figure 3.4), for the RCM output and the E-OBS reference data set, respectively, we obtain $(2 \cdot 4 + 1 = 9)$ spatial grids of linear trend estimates $\widehat{\beta}$. In the following we index these estimates with the corresponding pixel number $i = 1, \dots, n$. Further, we distinguish between estimates $\widehat{\beta}_i^{\text{mod}}$ corresponding to one of the RCMs and estimates $\widehat{\beta}_i^{\text{ref}}$ corresponding to the reference data set. In order to be able to compare and rank the different RCMs (for the two different boundary conditions), we finally compute absolute trend errors of the form

$$\left|\widehat{\beta}_i^{\text{mod}} - \widehat{\beta}_i^{\text{ref}}\right|, \quad i = 1, \dots, n.$$

Distinguishing between ECHAM5 and ERA-40 boundary conditions, the maps provided in Figure 5.38 visualize for each pixel of the study area, which of the four RCMs has the smallest absolute trend error, that is which model is considered best in modeling the linear temperature trend observed in the E-OBS data set. Whereas for the ECHAM5 driven models, the DMI model is preferred most over Europe, for the ERA-40 driven models the SMHI model

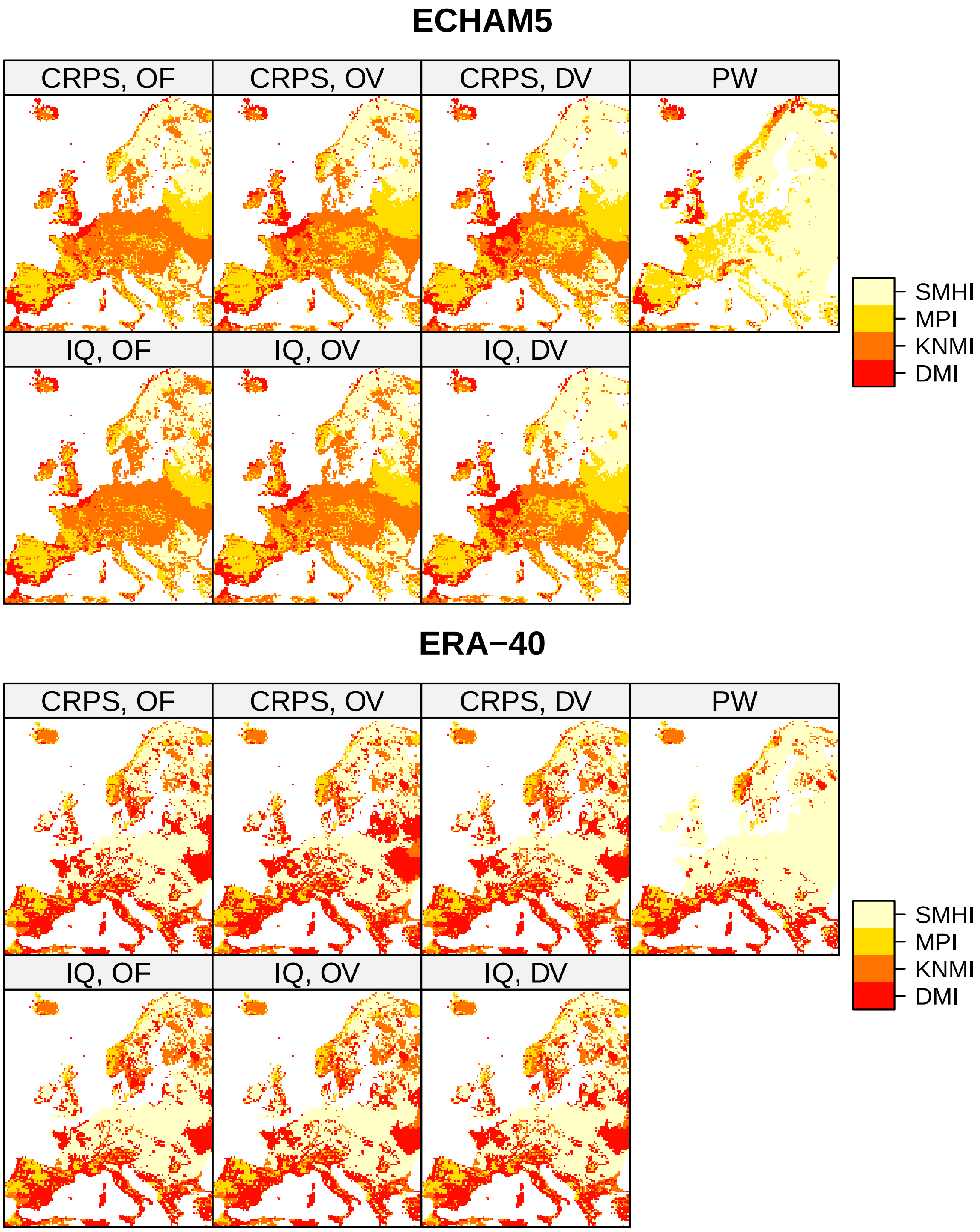


Figure 5.37: Maps showing for which areas in Europe which model (DMI, KNMI, MPI or SMHI) has the lowest (best) average (moving) score/divergence (reference period 1961–1990). Besides rankings based on moving CRPS and moving IQ distances (based on OF-, OV- and DV-windows), rankings for point-wise (PW) scores/divergences are displayed. Note that moving divergence averages based on OF- and OV-windows do not warrant propriety. As before we distinguish between ECHAM5 (top) and ERA-40 (bottom) boundary conditions for the four different models.

performs best in the north-east, the KNMI model in central and north-west Europe and the DMI model in eastern Europe. Comparison with Figure 5.37 shows, that the approach followed here and our novel approach based on moving windows yield completely different results. This is obviously due to the fact, that the approach followed in this section focuses only on one specific model characteristic, whereas our new approach intends to give a holistic picture of the model performance.

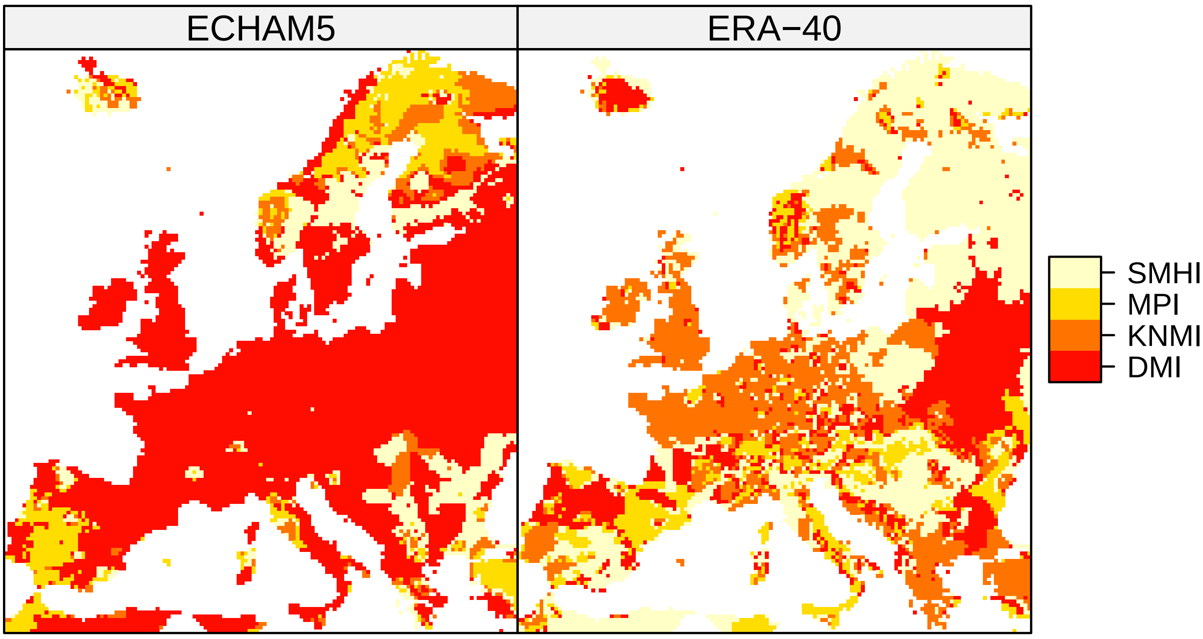


Figure 5.38: Maps showing for which areas in Europe which model (DMI, KNMI, MPI or SMHI; ECHAM5/ERA-40) has the smallest trend error.

To further summarize the results of the linear trend based evaluation and to provide a ranking of the models based on this approach, we consider spatial averages

$$\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{\beta}_i^{\text{mod}} - \widehat{\beta}_i^{\text{ref}}\right|,$$

of the absolute trend errors. The results are provided by Table 5.13. In accordance with Figure 5.38 the DMI model is considered best by far for ECHAM5 boundary conditions. For ERA-40 boundary conditions the models are ranked in the order SMHI, KNMI, DMI and MPI, where the SMHI model is considered best. Again, these results differ considerably from those for our novel evaluation approach summarized in Table 5.12. Direct comparison of the DMI model for the two different boundary conditions this time even suggests that the model driven by ECHAM5 performs better (in terms of linear trends).

Table 5.13: Average (spatial) absolute trend errors (times 10, i.e. decadal trend).

	DMI	KNMI	MPI	SMHI
ECHAM5	0.14	0.28	0.28	0.25
ERA-40	0.15	0.14	0.20	0.13

5.10 Conclusions and outlook

We propose novel methodology for the evaluation of (time series) models/forecasts under the presence of non-stationarity. Our novel approach utilizes proper scores and divergences on small moving time windows (where stationarity is assumed) in order to provide a fair and holistic comparison/assessment of models/forecasts. The moving window approach is illustrated in Figures 5.2–5.4. Three simulation studies explore the moving score/divergence technique under the presence of changepoints, trends and periodicity. A case study, evaluating/comparing Regional Climate Models, illustrates the utility of the novel technique for practical applications. We summarize our results and conclusions in the following.

Methodology

1. Being based on proper scores/divergences, our novel approach allows for a *fair* comparative model/forecast evaluation.
2. Propriety of moving divergence averages is not warranted for all considered types of moving windows. While DV-windows yield propriety of moving divergence averages, OF- and OV-windows do not. However for long time series, moving divergence averages based on OF-windows can at least be considered as approximately proper. The approximation of propriety for OV-windows is much worse than that for OF-windows.
3. Evaluation based on moving windows instead of a point-wise comparison allows to account for *higher order structures* of the modeled/forecast phenomenon.
4. Compared to most other evaluation approaches taken in practice our evaluation method does not require
   (a) a separate consideration of *different seasons*,
   (b) to make a (subjective) decision on the importance/weight of *different features of the model/forecast*.
5. Utilization of a changepoint analysis for the *choice of the moving windows* is meaningful, since a changepoint analysis segments a time series into stationary segments.

Simulation studies

5. Depending on the phenomenon of interest, certain *types of moving windows* might be more adequate than others. Under the presence of
   (a) fixed changepoints: Disjoint windows with varying width (DV).
   (b) trends or periodicity: Overlapping windows with fixed (OF) or varying width (OV).
6. For the application of the moving score/divergence methodology to *continuous outcomes*
   (a) the utilization of the Continuous Ranked Probability Score (CRPS) is preferred over the Squared Error (SE) score,
   (b) the utilization of the Integrated Quadratic (IQ) distance is preferred over the Mean Value (MV) divergence.
7. Amongst others, the simulation studies showed that moving scores/divergences

(a) are able to *approximate their theoretical counterparts* considerably well,

(b) are *better suited than point-wise (PW) scores/divergences or scores/divergences under a stationarity assumption (ST)* to yield an adequate (true) model/forecast ranking,

(c) in most cases yield the *correct model ranking* (see the trend scenario (Section 5.6) for an exception).

Application

8. The novel technique has been applied successfully for the evaluation of Regional Climate Models (RCMs). The case study showed

   (a) *identical overall model rankings* based on moving CRPS and moving IQ divergences for different window selection approaches,

   (b) that (as expected) the reanalysis (ERA-40) driven RCMs obtain better moving scores/ divergences as the corresponding GCM (ECHAM5) driven RCMs,

   (c) that moving scores/divergences allow to assess the *temporal evolution* of model/forecast performance,

   (d) that spatial maps of average moving scores/divergences allow to identify *areas of bad model/forecast performance* (e.g. water bodies, mountain ranges),

   (e) that evaluation based on only one specific model/forecast characteristic does not mirror the overall model/forecast performance and may result in an inadequate model ranking.

Future research directions Besides application of the novel evaluation methodology to compare climate models, applications in other areas might be of interest for future research. Given the context of this thesis and the development of drought indices in Chapter 4, it is of particular interest to us to investigate if an evaluation of such drought indices using the novel evaluation approach is feasible or if the presented methodology has to be adapted. Moreover, an extension of the work presented here in Chapter 5 to categorical outcomes is of interest. For example, one might want to be able to judge models which differentiate only if a certain (extreme) event occurs or not. Further extensions of our approach might concern the judgment of model accuracy in a spatial context, where we can think of the moving window as a spatial neighborhood considering either a fixed number of spatial neighbors or all locations within a certain radius.

Appendix

A List of recurring acronyms

AE	absolute error
AIC	Akaike information criterion
ARMA	autoregressive moving-average (model)
BIC	Bayesian information criterion
CDF	cumulative distribution function
COR	Pearson correlation
CRPS	Continuous Ranked Probability Score
CSI	critical success index
C-vine	canonical vine
DMI	Danish Meteorological Institute
DV	disjoint windows with varying width
ECMWF	European Centre for Medium-Range Weather Forecasts
E-OBS	ENSEMBLES daily gridded observational dataset
ERA-20C	ECMWF Atmospheric Reanalysis of the 20th Century
FAR	false alarm ratio
GCM	Global Climate Model
i.i.d.	independent and identically distributed
IPCC	Intergovernmental Panel on Climate Change
IQ	Integrated Quadratic (distance)
JDI	Joint Deficit Index
KNMI	Royal Netherlands Meteorological Institute
MAE	Mean Absolute Error

MdAE	Median Absolute Error
MPI	Max-Planck-Institute for Meteorology
MSDI	Multivariate Standardized Drought Index
MSE	Mean Square Error
MV	Mean Value (divergence)
OF	overlapping windows with fixed width
OP	Optimal Partitioning
OV	overlapping windows with varying width
PCC	pair-copula construction
PDF	probability density function
PDSI	Palmer Drought Severity Index
PELT	Pruned Exact Linear Time
PET	potential evapotranspiration
PIT	probability integral transformation
POD	probability of detection
PRE	precipitation
PW	point-wise
RCM	Regional Climate Model
RMSE	Root Mean Square Error
R-vine	regular vine
SDAT	Standardized Drought Analysis Toolbox
SE	squared error
SI	Standardized (Drought) Index
SMHI	Swedish Meteorological and Hydrological Institute
SPI	Standardized Precipitation Index
SPEI	Standardized Precipitation Evapotranspiration Index
ST	stationarity (assumption)

B Evaluation of Regional Climate Models: Further results

Temporal evaluation Figures B.1 and B.2 complement the results of Section 5.9.2 displayed in Figure 5.35. The figures illustrate monthly averages of spatial averages (over the whole study area) of the moving scores and divergences obtained in Section 5.9.1. While Figure B.1 shows the results based on OF-windows, Figure B.2 shows the results based on DV-windows.

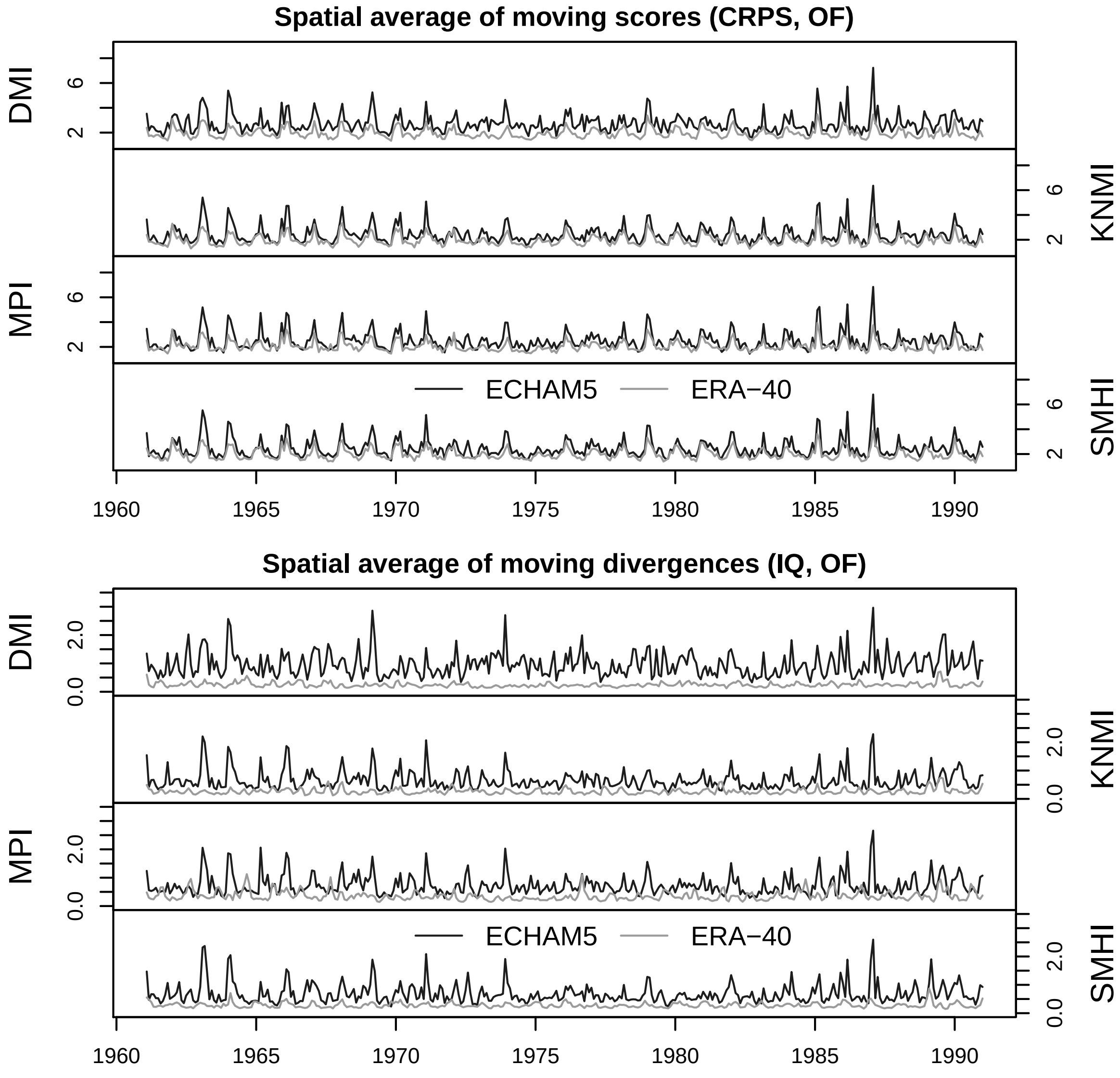


Figure B.1: Monthly averages of spatial averages (over whole study area) of moving CRPS (top) and moving IQ distances (bottom) based on OF-windows, for the DMI, KNMI, MPI and SMHI models driven by ECHAM5 (black) and ERA-40 (gray).

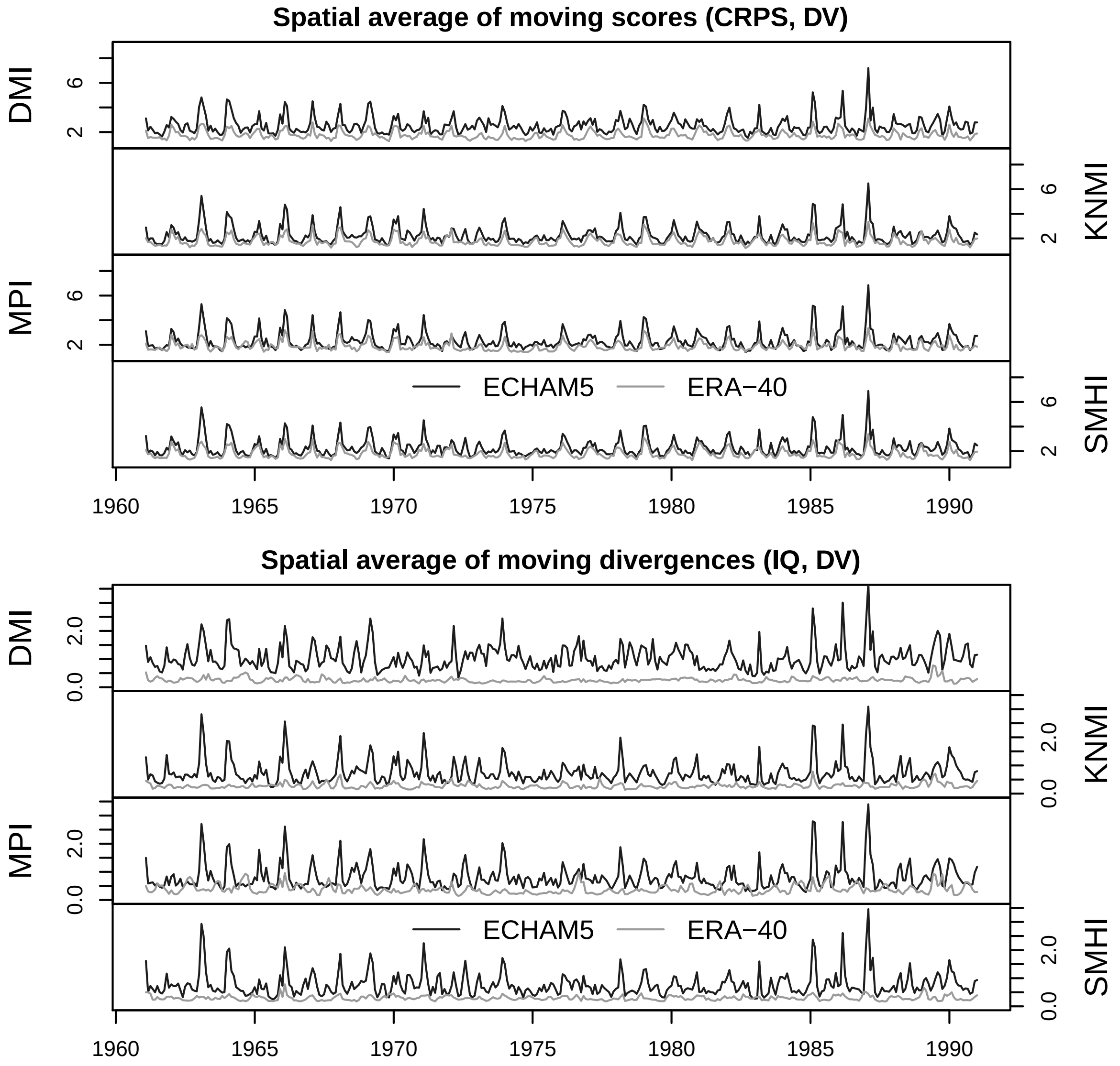
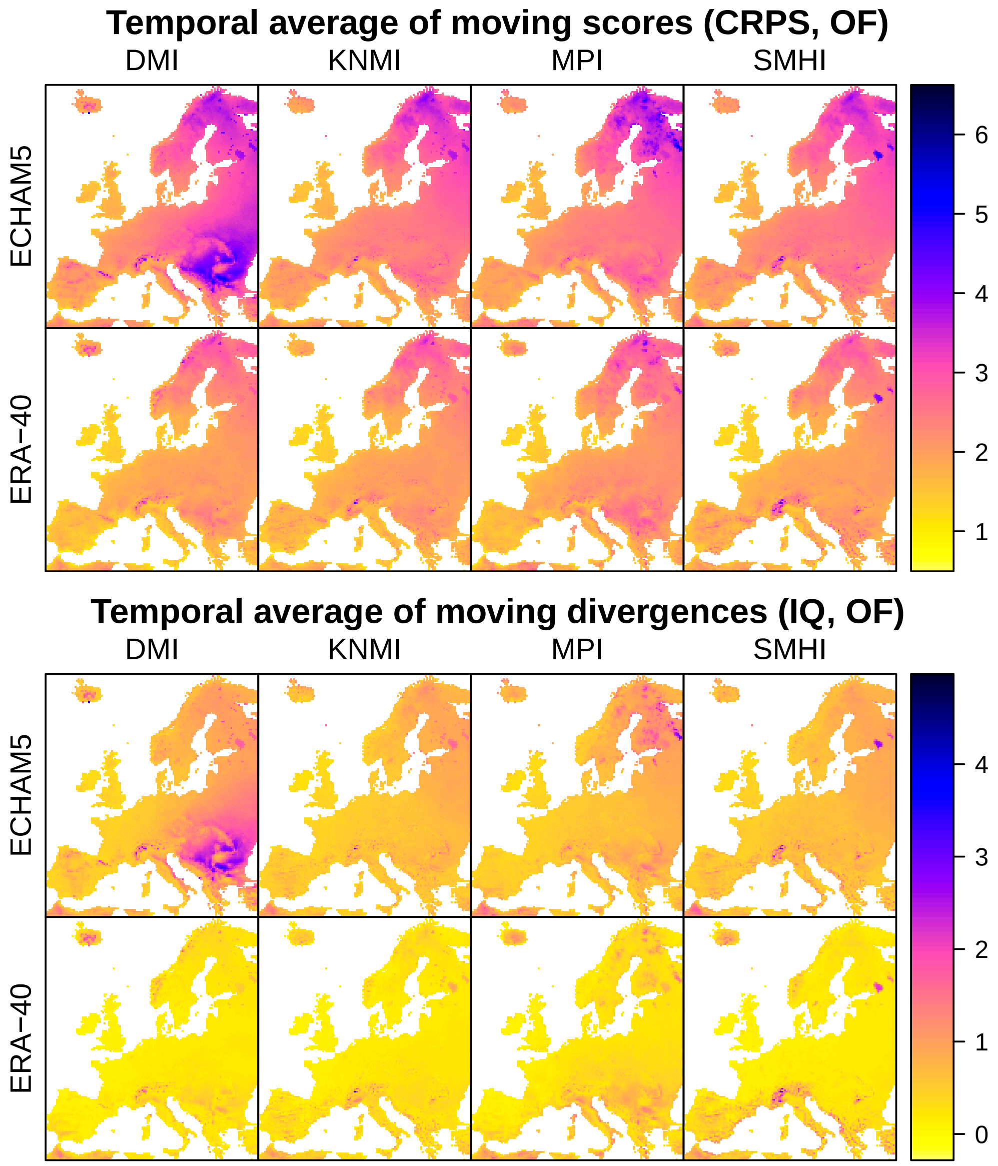


Figure B.2: Monthly averages of spatial averages (over whole study area) of moving CRPS (top) and moving IQ distances (bottom) based on DV-windows, for the DMI, KNMI, MPI and SMHI models driven by ECHAM5 (black) and ERA-40 (gray).

Spatial evaluation Figures B.3 and B.4 complement the results of Section 5.9.3 displayed in Figure 5.36. The figures illustrate temporal averages (over the period 1961–1990) of the moving scores and divergences obtained in Section 5.9.1. While Figure B.3 shows the results based on OF-windows, Figure B.4 shows the results based on DV-windows.



Temporal average of moving scores (CRPS, OF)

Temporal average of moving divergences (IQ, OF)

Figure B.3: Maps (Europe) of temporal averages (1961–1990) of moving CRPS (top) and moving IQ distances (bottom) based on OF-windows, for the DMI, KNMI, MPI and SMHI models (ECHAM5/ERA-40). Note that moving divergence averages based on OF-windows do not warrant propriety.

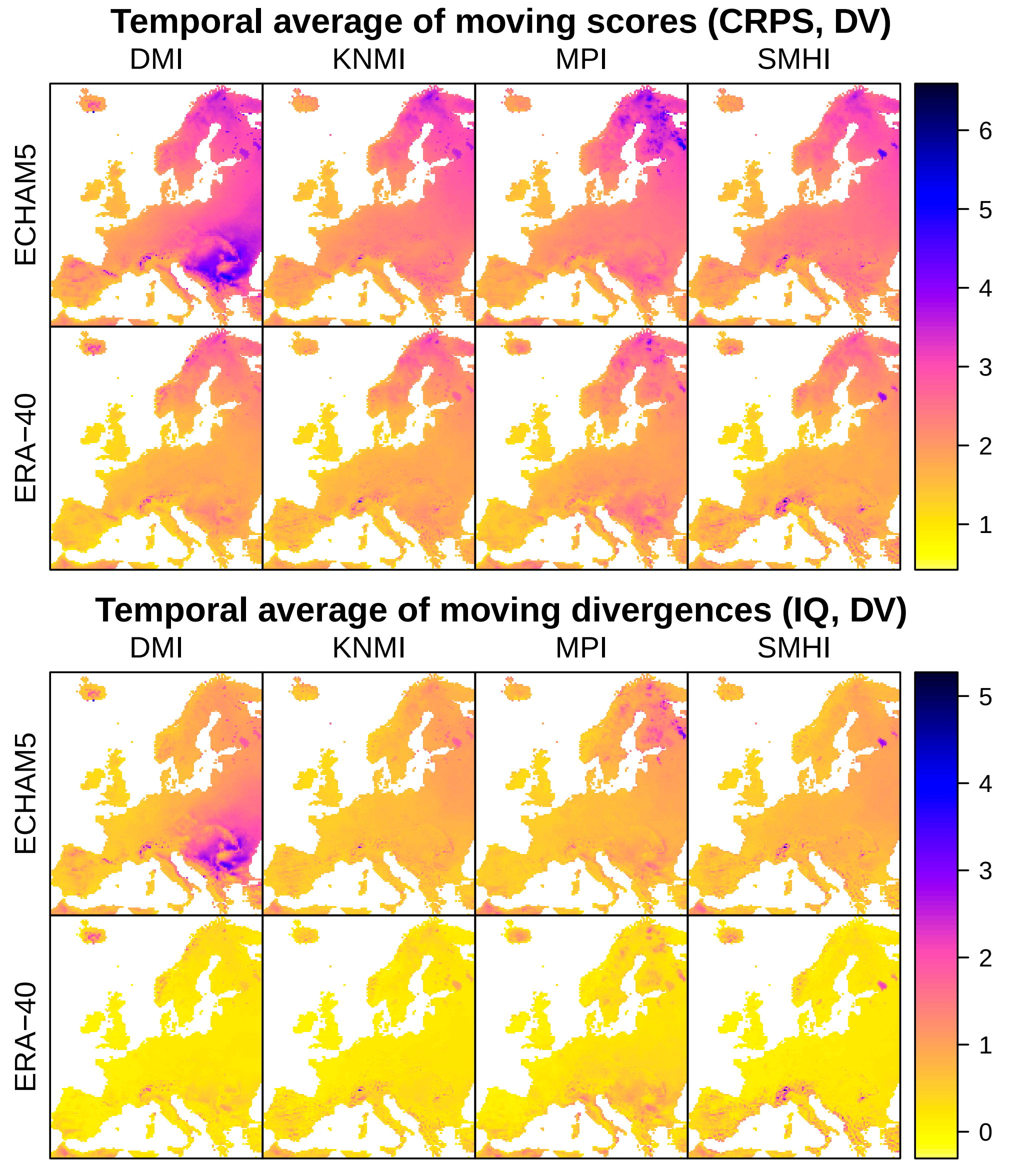

Temporal average of moving scores (CRPS, DV)
DMI
KNMI
MPI
SMHI
ECHAM5
ERA−40
6
5
4
3
2
1
Temporal average of moving divergences (IQ, DV)
DMI
KNMI
MPI
SMHI
ECHAM5
ERA−40
5
4
3
2
1
0

Figure B.4: Maps (Europe) of temporal averages (1961–1990) of moving CRPS (top) and moving IQ distances (bottom) based on DV-windows, for the DMI, KNMI, MPI and SMHI models (ECHAM5/ERA-40).

Bibliography

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

AghaKouchak, A. (2014). A baseline probabilistic drought forecasting framework using standardized soil moisture index: application to the 2012 united states drought. *Hydrology and Earth System Sciences 18*(7), 2485–2492.

AghaKouchak, A., A. Farahmand, F. S. Melton, J. Teixeira, M. C. Anderson, B. D. Wardlow, and C. R. Hain (2015). Remote sensing of drought: Progress, challenges and opportunities. *Reviews of Geophysics 53*(2), 452–480.

Arkansas Soybean Promotion Board (2011). Checkoff At Work: Soybean Statistics; Production. `http://www.themiraclebean.com/soybean-statistics`. [Online; accessed February 1, 2016].

Bachmair, S., K. Stahl, K. Collins, J. Hannaford, M. Acreman, M. Svoboda, C. Knutson, K. H. Smith, N. Wall, B. Fuchs, N. D. Crossman, and I. C. Overton (2016). Drought indicators revisited: the need for a wider consideration of environment and society. *Wiley Interdisciplinary Reviews: Water 3*(4), 516–536.

Barbe, P., C. Genest, K. Ghoudi, and B. Rémillard (1996). On Kendall's process. *Journal of Multivariate Analysis 58*(2), 197–229.

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

Bedford, T. and R. M. Cooke (2002). Vines - a new graphical model for dependent random variables. *The Annals of Statistics 30*(4), 1031–1068.

Bloomfield, J. P. and B. P. Marchant (2013). Analysis of groundwater drought building on the standardised precipitation index approach. *Hydrology and Earth System Sciences 17*(12), 4769–4787.

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

Brechmann, E. C. (2014). Hierarchical Kendall copulas: Properties and inference. *Canadian Journal of Statistics 42*(1), 78–108.

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

Deutscher Wetterdienst (2015). WebWerdis (Weather Request and Distribution System), https://werdis.dwd.de/werdis/start_js_JSP.do. Station data for Regensburg.

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.

Embrechts, P., F. Lindskog, and A. McNeil (2003). Modelling dependence with copulas and applications to risk management. In S. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384. Elsevier.

Erhardt, T. M. and C. Czado (2016). Standardized drought indices: A novel uni- and multivariate approach. In revision for Journal of the Royal Statistical Society: Series C (Applied Statistics).

Eum, H.-I., P. Gachon, R. Laprise, and T. Ouarda (2012). Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme. *Climate Dynamics 38*(7-8), 1433–1457.

European Centre for Medium-Range Weather Forecasts (2014). ERA-20C Project (ECMWF Atmospheric Reanalysis of the 20th Century). http://dx.doi.org/10.5065/D6VQ30QG.

Fang, H.-B., K.-T. Fang, and S. Kotz (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis 82*(1), 1–16.

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

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen (2013). Evaluation of climate models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Chapter 9, pp. 741–866. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Frahm, G., M. Junker, and A. Szimayer (2003). Elliptical copulas: applicability and limitations. *Statistics & Probability Letters 63*(3), 275–286.

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.

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

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

Gringorten, I. I. (1963). A plotting rule for extreme probability paper. *Journal of Geophysical Research 68*(3), 813–814.

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.

Hao, Z. and V. P. Singh (2015). Drought characterization from a multivariate perspective: A review. *Journal of Hydrology 527*, 668–678.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New (2008). A european daily high-resolution gridded dataset of surface temperature and precipitation. *Journal of Geophysical Research: Atmospheres 113.*

Hyndman, R. J. and A. B. Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting 22*(4), 679–688.

IPCC (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* [Core Writing Team, R. K. Pachauri and L. A. Meyer (Eds.)]. IPCC, Geneva, Switzerland.

Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters 12*(2), 105–108.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis 46*(2), 262–282.

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

Joe, H. (2001). *Multivariate models and dependence concepts* (1st ed.), Volume 73 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC.

Joe, H. (2014). *Dependence Modeling with Copulas.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

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

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

Killick, R., P. Fearnhead, and I. A. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association 107*(500), 1590–1598.

Killick, R., C. F. H. Nam, J. A. D. Aston, and I. A. Eckley (2012). changepoint.info: The changepoint repository. `http://changepoint.info`. [Online; accessed September 27, 2016].

Kjellström, E., F. Boberg, M. Castro, J. H. Christensen, G. Nikulin, and E. Sánchez (2010). Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. *Climate Research 44*(2-3), 135–150.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association 53*(284), 814–861.

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

Landgren, O. A., J. E. Haugen, and E. J. Førland (2014). Evaluation of regional climate model temperature and precipitation outputs over scandinavia. *Climate Research 60*, 249–264.

Lorenz, P. and D. Jacob (2010). Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40. *Climate Research 44*(2-3), 167–177.

Ma, M., L. Ren, F. Yuan, S. Jiang, Y. Liu, H. Kong, and L. Gong (2014). A new standardized palmer drought index for hydro-meteorological use. *Hydrological Processes 28*(23), 5645–5661.

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.

National Agricultural Statistics Service, United States Department of Agriculture (2015). Quick Stats, `http://quickstats.nass.usda.gov/`.

Nelsen, R. B. (2006). *An Introduction to Copulas* (2nd ed.). Springer Series in Statistics. New York: Springer.

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

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

Schepsmeier, U. (2015). Efficient information based goodness-of-fit tests for vine copula models with fixed margins: A comprehensive review. *Journal of Multivariate Analysis 138*, 34–52.

Scott, A. J. and M. Knott (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics 30*(3), 507–512.

Shukla, S. and A. W. Wood (2008). Use of a standardized runoff index for characterizing hydrologic drought. *Geophysical Research Letters 35*(2), L02405.

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.

Staudinger, M., K. Stahl, and J. Seibert (2014). A drought index accounting for snow. *Water Resources Research 50*(10), 7861–7872.

Svoboda, M., D. LeComte, M. H. 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.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification 1*(1), 522–534.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review 38*(1), 55–94.

van der Linden, P. and J. F. B. Mitchell (Eds.) (2009). *ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project*, FitzRoy Road, Exeter EX1 3PB, UK. Met Office Hadley Centre.

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.

Vicente-Serrano, S. M., J. I. López-Moreno, S. Beguería, J. Lorenzo-Lacruz, C. Azorin-Molina, and E. Morán-Tejeda (2012). Accurate computation of a streamflow drought index. *Journal of Hydrologic Engineering 17*(2), 318–332.

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

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences.* Academic Press.

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