Ruprecht-Karls-Universität Heidelberg
Fakultät für Mathematik und Informatik

Proper divergence functions for comparing and combining climate model outputs for extreme temperature indices

Diplomarbeit
von
Nadine Gissibl

Betreuer: Dr. Thordis L. Thorarinsdottir
Prof. Dr. Tilmann Gneiting
Dr. Jana Sillmann

Juli 2012

Abstract

Divergence functions measure the difference between two probability distributions F and G. They are widely used to assess the quality of probabilistic forecasts, where F represents a predictive distribution function and G represents the empirical distribution function of the events that materialize. Scoring rules are the analogue for the evaluation of a probabilistic forecast F for a single event y. A scoring rule is proper if the forecaster optimizes the expected score when y is drawn from the distribution F. This property encourages honesty and careful assessments in the prediction process. Here, divergence functions with a similar property are proposed. The Cramér-von Mises distance, the divergence function associated with the continuous ranked probability score, is particularly suitable for prediction assessment. In a case study, fifteen climate model projections for four indices for climate extremes over Europe from 1961 to 1990 are compared with corresponding re-analysis and/or observation based data sets to assess the skill of the models in simulating climate extremes. The quality of the simulations depends on the region, the season, and the index under consideration. For instance, the climate models are usually better in simulating the monthly and the yearly maximum temperature than the monthly and the yearly minimum temperature. Furthermore, weighting methods based on proper divergence functions to combine climate models are proposed. The weighted model combinations usually perform significantly better than any single climate model.

Zusammenfassung

Eine Divergenz-Funktion misst den Unterschied zweier Wahrscheinlichkeitsverteilungen F und G. Sie werden verbreitet dazu eingesetzt, die Qualität probabilistischer Vorhersagen zu beurteilen, wobei F die Vorhersageverteilung und G die empirische Verteilungsfunktion der eintretenden Ereignisse darstellt. Scoring-Funktionen sind das Analogon für die Bewertung einer probabilistischen Vorhersage F für ein einzelnes Ereignis y. Eine Scoring-Funktion heißt korrekt, falls der Prognostiker den erwarteten Score optimiert, wenn y der Verteilung F folgt. Diese Eigenschaft fördert die Aufrichtigkeit und sorgfältige Beurteilung beim Aufstellen einer Prognose. Wir stellen Divergenz-Funktionen mit einer vergleichbaren Eigenschaft vor. Die Cramér-von Mises Distanz, die mit dem continuous ranked probability score assoziierte Divergenz-Funktion, ist für die Beurteilung von Prognosen besonders geeignet. In einer Fallstudie werden Klimaprojektionen von fünfzehn Klimamodellen für vier Klimaextrem-Indizes von 1961 bis 1990 in Europa mit ensprechenden, auf Reanalysen und/oder Beobachtungen basierenden Datensätzen verglichen, um die Fähigkeiten und Schwächen der Modelle bei der Simulation von Klimaextremen zu beurteilen. Die Qualität der Simulationen ist von der Region, der Jahreszeit und dem Index abhängig. Beispielsweise können Klimamodelle meist die monatliche und jährliche Maximaltemperatur besser als die Minimaltemperatur simulieren. Außerdem führen wir Gewichtungsmethoden zur Kombination von Klimamodellen, die auf korrekten Divergenz-Funktionen basieren, ein. Die gewichteten Kombinationen der Klimamodelle schneiden meist wesentlich besser ab als jedes Klimamodell für sich.

Contents

1 Introduction

In recent years, a number of severe extreme climate events with dramatic social, economic, and environmental impact have occured. For example, in summer 2003, an exceptional heat wave hit Europe. It was by far the hottest European summer over the past 500 years (Luterbacher et al., 2004). As a result of the heat wave, more than 70,000 Europeans died (Robine et al., 2008). The financial loss due to crop failure was around US$12.3 billion. Forest fires in Portugal alone resulted in an additional US$1.6 billion in damage. The heat wave also led to unprecedented melting of the Alpine glaciers (Schär and Jendritzky, 2004). These huge amounts of damage to humans and their surroundings make it necessary to learn more about the nature of such extreme climate events and to study them in the context of natural climate variability and anthropogenic climate change. For further examples of extreme climate events, see e.g. Trenberth and Shea (2006) and Ulbrich et al. (2003).

The knowledge of significant future changes in the frequency and intensity of extreme events would provide guidance for reducing human and financial losses. Long data records are thus investigated for the analysis of rare and extreme events. Coupled atmosphere-ocean general circulation models (AOGCMs) are able to provide such long climate simulations, as they are appropriate tools to simulate past, present, and future climate states (Sillmann and Roeckner, 2008). However, AOGCMs may not always realistically represent climate extremes. Therefore, the ability of AOGCMs to simulate climate extremes has to be assessed.

One method to characterize extreme events is based on indices for climate extremes defined by an international committee to assess extremes in temperature and precipitation (Folland et al., 1999; Karl et al., 1999; Nicholls and Murray, 1999). These indices derived from daily maximum and minimum temperature and precipitation data describe a particular characteristic of an extreme event, such as its frequency, amplitude, or persistence. For instance, one may use the monthly maximum temperature to identify heat waves as in the example above. All indices consider only moderate extreme events on an annual or seasonal basis. Thus, the daily data series do not have to be very long and the indices can be readily calculated from observations, re-analysis, and model based data. This allows comparisons between extreme index data of climate models based on hindcast simulations and the corresponding re-analysis or observation based data set. To assess climate models with respect to

their ability to simulate climate extremes under present climate, we propose scoring methods that apply to the indices for climate extremes. Then, these scoring methods can also be utilized to estimate model biases to correct future simulations.

The Intergovernmental Panel on Climate Change (IPCC) glossary definition of climate which is the most common definition is as follows:

> "Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years. The classical period for averaging these variables is 30 years, as defined by the World Meteorological Organization. The relevant quantities are most often surface variables such as temperature, precipitation, and wind. Climate in a wider sense is the state, including a statistical description, of the climate system" (IPCC, 2007, p. 942).

Consequently, to evaluate a climate model in simulating climate extremes, the entire distribution of our "weather" data should be considered. We therefore propose our scoring methods within the framework of probabilistic forecasts, that is, forecasts taking the form of probability distributions. This will firstly lead us to scoring rules which assess the performance of a probabilistic forecast F for a single event y that materializes. A scoring rule is proper if the forecaster optimizes the expected score when y is drawn from the distribution F. This property is essential in scientific and operational forecast evaluation (Gneiting and Raftery, 2007). The description of climate above, however, invites a comparison of the predictive distribution with the empirical distribution function of the events that materialize over the time period under consideration. In order to realize this, divergence functions with a property similar to the propriety of scoring rules are proposed (Thorarinsdottir et al., 2012). For comparison, we also consider scoring methods for single-valued point forecasts, namely scoring functions (Gneiting, 2011).

Furthermore, since weighted averages of climate models may perform better than individual models, we present methods to combine AOGCMs by means of divergence function values. Here, the models are weighted according to their performance in a training period.

In Chapter 2, a theoretical framework for the evaluation of point forecasts and probabilistic forecasts is introduced. Depending on the type of forecast, we present various statistical measures to evaluate the predictive quality. Furthermore, we propose three weighting methods to combine climate model outputs. In Chapter 3, the

scoring methods and the weighting methods proposed in Chapter 2 are applied to climate model simulations for four indices for climate extremes, namely the maximum and minimum temperature value over a month and over a year. We first provide an overview of climate models, re-analyses, and observations in general and describe the indices for climate extremes in more detail. We then introduce the data used in this study and show the most important results. Finally, in Chapter 4, we summarize and discuss the results of the case study, in relation to currently employed methods for assessing the performance of climate models.

We thank Dr. Jana Sillmann (Canadian Centre for Climate Modelling and Analysis, University of Victoria, Canada) for helping with designing the research question, for providing the data, and for helpful discussions.

2 Theoretical background

2.1 Scores and divergence functions

We are interested in forecasts of a future event Y taking values in Ω, where Ω is usually $\mathbb{R}^d$ or a subset thereof. That is, the future observation Y is a random variable. To assess and compare the predictive performance of forecasters or forecasting procedures for Y, scoring methods are needed.

Here, we distinguish between two types of forecasts. Let $\mathcal{A}$ denote a σ-algebra of subsets of Ω, and let $\mathcal{P}$ denote a class of probability measures on $(\Omega, \mathcal{A})$, which is referred to as an *observation domain*. A *probabilistic forecast* for Y on Ω is any probability measure $P \in \mathcal{P}$. In the last few years, the view that forecasts should be probabilistic in nature has increasingly gained ground (Dawid, 1984; Gneiting and Raftery, 2007; Gneiting, 2008). However, single-valued point forecasts—that is, deterministic forecasts—are also used in many situations, for reasons of communication, tradition, or decision making (Gneiting, 2011). Generally, a *point forecast* may be a functional of a probabilistic forecast, such as the mean, the median, or a quantile.

Depending on the type of forecasts, Gneiting and Raftery (2007), Gneiting (2011), and Thorarinsdottir et al. (2012) propose various statistical measures to evaluate their quality which we will discuss below, essentially following these papers. We take them to be *negatively oriented*, that is, we prefer a forecast with a small score over one with a larger score. Hence, a forecaster wishes to minimize the score which indicates his loss.

2.1.1 Scoring functions

We start by considering point forecasts for the random variable Y. Let $\Omega \subseteq \mathbb{R}^d$ be the domain of both the point forecasts and the realizations of Y equipped with the corresponding Borel σ-algebra $\mathcal{A}$. For simplicity and clarity of presentation, we assume here that $d = 1$. All the results may be extended to higher dimensions. A *scoring function* is a mapping $S : \Omega \times \Omega \longrightarrow [0, \infty)$ such that $S(x, \cdot)$ is measurable with respect to $\mathcal{A}$ for each fixed value $x \in \Omega$. Then $S(x, y)$ states the forecaster's loss, when he predicts $x \in \Omega$ and $y \in \Omega$ materializes.

To evaluate the performance of a forecaster in providing several forecasts, the average over the individual scores is taken. Thus, if there are k forecast cases with corresponding point forecasts, $x_1, \ldots, x_k$, and observations, $y_1, \ldots, y_k$, we get

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} S(x_i, y_i)$$

as a summary measure of the predictive performance. In this study, the *mean absolute error* (MAE),

$$\text{MAE} = \frac{1}{k} \sum_{i=1}^{k} |x_i - y_i|,$$

and the *mean squared error* (MSE),

$$\text{MSE} = \frac{1}{k} \sum_{i=1}^{k} (x_i - y_i)^2,$$

are used. Larger forecast errors result in larger values of the MAE and the MSE. Therefore, if there are competing point forecasters, we clearly expect similar rankings under the MAE and the MSE.

The consistency of scoring functions plays an important role in making and evaluating point forecasts (Gneiting, 2011). Let $\mathcal{P}$ be a family of potential probability distributions for the random variable Y. That is, the class $\mathcal{P}$ contains probability measures on the measurable space $(\Omega, \mathcal{A})$. Here, the class $\mathcal{P}$ is often taken to be the set of all probability measures on Ω. A *statistical functional* is a function $T : \mathcal{P} \longrightarrow \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ denotes the power set of Ω.

Definition 2.1. The scoring function S is *consistent* for the functional T relative to the class $\mathcal{P}$ if

$$\mathbb{E}_P S(t, Y) \leqslant \mathbb{E}_P S(x, Y) \tag{2.1}$$

for all probability distributions $P \in \mathcal{P}$, all $t \in T(P)$, and all $x \in \Omega$. It is *strictly consistent* if it is consistent with equality in (2.1) if and only if $x \in T(P)$.

The expectations in (2.1) are well-defined, since S is a nonnegative function. They describe the expected loss under the probabilistic forecast $P \in \mathcal{P}$, if the forecaster predicts $t \in T(P)$ or $x \in \Omega$, respectively, for the random variable Y.

The following situation, for which we use the same notation as before, illustrates the critical importance of consistency. A forecaster quotes a quantile r at level

$\alpha \in (0,1)$ and y materializes. Then he will be penalized by $S(r,y)$. If

$$\mathbb{E}_P S(q,Y) \leqslant \mathbb{E}_P S(r,Y) \tag{2.2}$$

for all $r \in \Omega$ and all probability distributions $P \in \mathcal{P}$ with corresponding α-quantile q, the scoring function S is consistent for the quantile functional at level α. Inequality (2.2) then implies that the expected loss is minimized for the true α-quantile. The forecaster is thus encouraged to report his true beliefs. Consequently, scoring functions that are consistent for a quantile functional are suitable for assessing the quality of quantile forecasts. We refer to them as *proper scoring rules for quantiles*.

Let $I \subseteq \mathbb{R}$ be an interval, and let $\mathcal{B}(I)$ denote the Borel σ-algebra on I. Furthermore, let $\mathcal{P}$ denote the class of the Borel probability measures P on the measurable space $(I, \mathcal{B}(I))$ with finite first moment and strictly increasing distribution function. Then, the *asymmetric piecewise linear* scoring function,

$$S_\alpha(r,y) = (\mathbb{1}\{y \leqslant r\} - \alpha)(r-y), \tag{2.3}$$

is a proper scoring rule for quantiles. That is, it is consistent for the quantile functional at level $\alpha \in (0,1)$ relative to the class $\mathcal{P}$: Let q be the unique α-quantile of the probability measure $F \in \mathcal{P}$ that we identify with its cumulative distribution function (cdf) so that $F(q) = \alpha$. If $r < q$, where $r \in I$, then

$$\begin{aligned}
&\mathbb{E}_F S_\alpha(r,Y) - \mathbb{E}_F S_\alpha(q,Y) \\
&\qquad = \int_I [\mathbb{1}\{y \leqslant r\} - \alpha](r-y)\mathrm{d}F(y) \\
&\qquad\quad - \int_I [\mathbb{1}\{y \leqslant q\} - \alpha](q-y)\mathrm{d}F(y) \\
&\qquad = \int_I r\mathbb{1}\{y \leqslant r\}\mathrm{d}F(y) - \int_I q\mathbb{1}\{y \leqslant q\}\mathrm{d}F(y) \\
&\qquad\quad + \int_I [y\mathbb{1}\{y \leqslant q\} - y\mathbb{1}\{y \leqslant r\}]\mathrm{d}F(y) - \alpha r + \alpha q \\
&\qquad = rF(r) - qF(q) + \int_r^q y\mathrm{d}F(y) - \alpha r + \alpha q \\
&\qquad = rF(r) + \int_r^q y\mathrm{d}F(y) - \alpha r \\
&\qquad \geqslant rF(r) + r\int_r^q \mathrm{d}F(y) - \alpha r \\
&\qquad = 0.
\end{aligned}$$

Conversely, if $r > q$, then

$$\begin{aligned}
&\mathbb{E}_F S(r, Y) - \mathbb{E}_F S(q, Y) \\
&\qquad = rF(r) - \int_q^r y \mathrm{d}F(y) - \alpha r \\
&\qquad \geqslant rF(r) - r \int_q^r \mathrm{d}F(y) - \alpha r \\
&\qquad = 0.
\end{aligned}$$

If forecasts are in the form of probability distributions over future events, they can be better characterized by considering several quantiles. Scoring rules for quantiles are thus sometimes extended to functions $S : \Omega^k \times \Omega \longrightarrow [0, \infty)$, where $S(r_1, \ldots, r_k; y)$ represents the forecaster's loss when he quotes quantiles $r_1, \ldots, r_k$ and y materializes. For example, if S_i is a proper scoring rule for predicting the single quantile at level $\alpha_i \in (0, 1)$ for $i = 1, \ldots, k$, the scoring rule

$$S(r_1, \ldots, r_k; y) = \sum_{i=1}^{k} S_i(r_i, y)$$

is clearly proper for predicting the quantiles at levels $\alpha_1, \ldots, \alpha_k$. Note that, to avoid technical complications, the elements of $\mathcal{P}$ often have to meet certain conditions, as is, for instance, the case for the asymmetric piecewise linear scoring function in (2.3) (Gneiting and Raftery, 2007, p. 370).

2.1.2 Scoring rules

Here, we consider probabilistic forecasts for a random variable Y on a general sample space Ω. Let $\mathcal{P}$ be a convex class of probability measures on the observation domain $(\Omega, \mathcal{A})$. A *scoring rule* is a function $S : \mathcal{P} \times \Omega \longrightarrow \bar{\mathbb{R}}$ that is $\mathcal{P}$-quasi-integrable[1] in the second argument for each fixed probability measure $P \in \mathcal{P}$. Scoring rules measure the quality of probabilistic forecasts for a single event y. As in the case of scoring functions, the forecaster's loss is represented by $S(P, y)$ when he quotes $P \in \mathcal{P}$ and y materializes. The following property of a scoring rule S is of high importance in the theory of probabilistic forecasts.

[1] A function $f : \Omega \longrightarrow \bar{\mathbb{R}} = [-\infty, \infty]$ which is measurable with respect to $\mathcal{A}$ and is quasi-integrable with respect to all $P \in \mathcal{P}$ (Bauer, 1992, p. 74) is $\mathcal{P}$-*quasi-integrable*.

Definition 2.2. The scoring rule S is *proper* relative to $\mathcal{P}$ if

$$\mathbb{E}_Q S(Q, Y) \leqslant \mathbb{E}_Q S(P, Y) \tag{2.4}$$

for all probability measures P, $Q \in \mathcal{P}$. It is *strictly proper* if it is proper with equality in (2.4) if and only if $P = Q$.

If S is proper, the expected loss under a probability measure $Q \in \mathcal{P}$ is minimized if the forecaster quotes Q. Similar as for consistent scoring functions, the forecaster is thus encouraged to report his true beliefs if the predictive performance is measured by a proper scoring rule. Note that the proposed definition deviates from that in Gneiting and Raftery (2007) who take scoring rules to be positively oriented.

To evaluate a set of k forecast-observation-pairs in practical applications, average values of proper scoring rules S,

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} S(P_i, y_i),$$

where $P_1, \ldots, P_k$ denote the forecasts and $y_1, \ldots, y_k$ denote the corresponding observations, can be used as a measure of the predictive performance.

Assuming $\mathcal{P}$ to be the class of all Borel probability measures on $\mathbb{R}$ enables us to identify a probabilistic forecast $F \in \mathcal{P}$ with its cdf F. The *continuous ranked probability score* (CRPS) can be defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \big[F(x) - \mathbb{1}\{x \geqslant y\}\big]^2 \mathrm{d}x, \tag{2.5}$$

which we refer to as the *threshold decomposition* of the CRPS (Matheson and Winkler, 1976; Hersbach, 2000). An alternative form of the CRPS is the *quantile score representation*,

$$\mathrm{CRPS}(F, y) = 2 \int_0^1 \big[(\mathbb{1}\{y \leqslant F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y)\big] \mathrm{d}\alpha, \tag{2.6}$$

where we write $F^{-1}(\alpha)$ for the quantile at level $\alpha \in (0, 1)$. Since the integrand in (2.6) equals twice the asymmetric piecewise linear scoring function S_α in (2.3) for the quantile forecast $F^{-1}(\alpha)$, this representation confirms that determining a predictive cdf is equivalent to determining all predictive quantiles. Laio and Tamea (2007) show the equivalence of the threshold decomposition (2.5) and the quantile score representation (2.6).

If F has finite first moment, the CRPS can be written as

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F\mathbb{E}_F|X - X'|, \tag{2.7}$$

where X and X' are independent random variables with distribution F (Gneiting and Raftery, 2007). We call (2.7) the *kernel score representation* of the CRPS (cf. Gneiting and Raftery, 2007, p. 368).

The CRPS is a proper scoring rule relative to the class $\mathcal{P}$ and a strictly proper scoring rule relative to the subclass $\mathcal{P}_1$, that consists of the Borel probability measures on $\mathbb{R}$ with finite first moment. Both the representation (2.6) and the representation (2.7) show that the CRPS is reported in the same unit as the observation. This fact simplifies the interpretation of the resulting scores. Furthermore, if F is a point measure, the CRPS reduces to the absolute error. Thus, the CRPS is a generalization of the absolute error and can be used to directly compare deterministic and probabilistic forecasts (Gneiting and Raftery, 2007; Gneiting and Ranjan, 2011).

One might consider an alternative to the CRPS naturally induced by the kernel score representation (2.7), say

$$S(P, y) = \mathbb{E}_P|X - y|,$$

where X denotes a random variable with distribution P. This score is not a proper scoring rule relative to the class $\mathcal{P}_1$: Let X and Y, Y' be independent random variables with distribution P and Q. If $Q \in \mathcal{P}_1$, let med_Q denote the median value of the probability distribution Q. Since the scoring function $S(x, y) = |x - y|$ is strictly consistent for the median functional relative to the class $\mathcal{P}_1$ (Gneiting, 2011), the inequality

$$\mathbb{E}_Q|Y' - med_Q| \leqslant \mathbb{E}_Q|Y' - y| \tag{2.8}$$

holds for all probability measures $Q \in \mathcal{P}_1$ and all real numbers y with inequality in (2.8) if and only if $y = med_Q$. If we apply the expectation, inequality (2.8) takes the form

$$\mathbb{E}_Q|Y' - med_Q| < \mathbb{E}_Q\mathbb{E}_Q|Y' - Y| \tag{2.9}$$

for all probability measures $Q \in \mathcal{P}_1$ other than the point measure at med_Q. Let us now suppose that S is proper. Then, the inequality

$$\mathbb{E}_Q S(Q, Y) = \mathbb{E}_Q\mathbb{E}_Q|Y' - Y| \leqslant \mathbb{E}_Q S(P, Y) = \mathbb{E}_Q\mathbb{E}_P|X - Y| \tag{2.10}$$

holds for all probability measures P, $Q \in \mathcal{P}_1$. If P is the point forecast med_Q, inequality (2.10) contradicts inequality (2.9).

Gneiting and Raftery (2005) note that the following alternative to the threshold decomposition of the CRPS (2.5) in terms of F^{-1},

$$S(F, y) = \int_0^1 (F^{-1}(\alpha) - y)^2 \mathrm{d}\alpha = \mathbb{E}_F(X - y)^2,$$

is also not a proper scoring rule. This can be demonstrated similar to the above. Let $\mathcal{P}_2$ denote the class of the Borel probability measures on $\mathbb{R}$ with finite second moment. Since the scoring function $S(x, y) = (x - y)^2$ is strictly consistent for the mean functional relative to the class $\mathcal{P}_2$ (Gneiting, 2011), we can use μ_Q instead of med_Q, where μ_Q denotes the mean of the probability distribution $Q \in \mathcal{P}_2$. The alternative derivation below is even more general.

For a Borel probability measure P on $\mathbb{R}$, let X be a random variable with distribution P. Let $\mathcal{P}_{2n}$, where n is a natural number, denote the class of the Borel probability measures P on $\mathbb{R}$ for which the expectation $\mathbb{E}_P X^{2n}$ is finite. Now let X be a random variable with distribution $P \in \mathcal{P}_{2n}$. Then the negatively oriented score

$$S_{2n}(P, y) = \mathbb{E}_P(X - y)^{2n}$$

is not a proper scoring rule relative to the class $\mathcal{P}_{2n}$. To show this, let us suppose that S is proper. If X and Y, Y' are independent random variables with distribution P and Q, respectively, where P, $Q \in \mathcal{P}_{2n}$, then

$$\mathbb{E}_Q \mathbb{E}_Q (Y' - Y)^{2n} \leqslant \mathbb{E}_Q \mathbb{E}_P (X - Y)^{2n}. \tag{2.11}$$

By applying the Binomial Theorem, we obtain

$$\begin{aligned}
(X - Y)^{2n} &= \sum_{k=0}^{2n} \binom{2n}{k} X^k (-Y)^{2n-k} \\
&= \sum_{k=1}^{n} \binom{2n}{2k-1} X^{2k-1} (-Y)^{2n-(2k-1)} \\
&\quad + \sum_{k=0}^{n} \binom{2n}{2k} X^{2k} (-Y)^{2n-2k} \\
&= -\sum_{k=1}^{n} \binom{2n}{2k-1} X^{2k-1} Y^{2(n-k)+1} \\
&\quad + \sum_{k=0}^{n} \binom{2n}{2k} X^{2k} Y^{2(n-k)}.
\end{aligned}$$

If $\mathbb{E}_Q Y^{2k-1} = 0$ for all $k = 1, \dots, n$ and $\mathbb{E}_P X^{2k} \leqslant \mathbb{E}_Q (Y')^{2k}$ for all $k = 1, \dots, n$ with inequality for at least one k, inequality (2.11) becomes

$$\begin{aligned}\sum_{k=0}^{n} \binom{2n}{2k} \mathbb{E}_Q (Y')^{2k} \mathbb{E}_Q Y^{2(n-k)} &\leqslant \sum_{k=0}^{n} \binom{2n}{2k} \mathbb{E}_P X^{2k} \mathbb{E}_Q Y^{2(n-k)} \\ &< \sum_{k=0}^{n} \binom{2n}{2k} \mathbb{E}_Q (Y')^{2k} \mathbb{E}_Q Y^{2(n-k)},\end{aligned}$$

which is the desired contradiction. This contradiction can be clearly illustrated by the following example.

Example 2.1. Let $X \sim \mathcal{N}(0, \frac{\sigma^2}{4})$ and $Y \sim \mathcal{N}(0, \sigma^2)$, and let k be a natural number. Then

$$\mathbb{E}\, Y^{2k-1} = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} y^{2k-1} e^{-\frac{y^2}{2\sigma^2}} \mathrm{d}y = 0,$$

as the integrands are odd functions, and

$$\begin{aligned}\mathbb{E}\, Y^{2k} &= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} y^{2k} e^{-\frac{y^2}{2\sigma^2}} \mathrm{d}y \\ &= -\frac{1}{\sqrt{2\pi}\sigma} \sigma^2 y^{2k-1} e^{-\frac{y^2}{2\sigma^2}} \Big|_{-\infty}^{\infty} \\ &\quad + \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} \sigma^2 (2k-1) y^{2k-2} e^{-\frac{y^2}{2\sigma^2}} \mathrm{d}y \\ &= \sigma^2 (2k-1) \mathbb{E} Y^{2k-2} \\ &= \sigma^{2k} \prod_{i=1}^{k} [2k - (2i-1)].\end{aligned}$$

The second equation follows from integration by parts. For the last equation, we use $\mathbb{E}\, Y^2 = \mathbb{V}\mathrm{ar}\, Y + (\mathbb{E}\, Y)^2 = \sigma^2$. Altogether, we get

$$\mathbb{E}\, Y^k = \begin{cases} 0, & \text{if } k \text{ is odd,} \\ \sigma^k \prod_{i=1}^{\frac{k}{2}} [k - (2i-1)], & \text{if } k \text{ is even,} \end{cases}$$

and, analogously for the random variable X,

$$\mathbb{E}\, X^k = \begin{cases} 0, & \text{if } k \text{ is odd,} \\ (\frac{\sigma}{2})^k \prod_{i=1}^{\frac{k}{2}} [k - (2i-1)], & \text{if } k \text{ is even.} \end{cases}$$

Thus, it holds that $\mathbb{E}\, Y^k = 0$ if k is odd, and $\mathbb{E}\, X^k \leqslant \mathbb{E}\, Y^k$ if k is even. Due to the proof above, inequality (2.11) is violated and S_{2n} is therefore not a proper scoring rule.

We have seen that the scores $S(P, y) = \mathbb{E}_P|X - y|$ and $S_2(P, y) = \mathbb{E}_P(X - y)^2$ are not proper scoring rules relative to the class $\mathcal{P}_1$ or $\mathcal{P}_2$, respectively, where X denotes a random variable with distribution P. However in general, the following theorem provides a method to construct proper scoring rules from scoring functions. Moreover, it justifies the classification of scoring functions which are consistent for the quantile functional as proper scoring rules for quantiles.

Theorem 2.1 (Gneiting, 2011)**.** Let Ω be a subset of $\mathbb{R}^d$ equipped with the corresponding Borel σ-algebra $\mathcal{A}$, and let S be a scoring function on $\Omega \times \Omega$. Furthermore, let $\mathcal{P}$ denote a convex class of probability measures on $(\Omega, \mathcal{A})$, and let T be a statistical functional. Suppose that the scoring function S is consistent for the functional T relative to the class $\mathcal{P}$. For each $P \in \mathcal{P}$, let $t_P \in T(P)$. Then the function

$$R : \mathcal{P} \times \Omega \longrightarrow [0, \infty), \quad (P, y) \mapsto R(P, y) = S(t_P, y),$$

is a proper scoring rule.

Proof. Let Y be a random variable with probability distribution $Q \in \mathcal{P}$. If P, $Q \in \mathcal{P}$, then

$$\mathbb{E}_Q R(P, Y) = \mathbb{E}_Q S(t_P, Y) \geqslant \mathbb{E}_Q S(t_Q, Y) = \mathbb{E}_Q R(Q, Y).$$

As S is a nonnegative function, the expectations above are well-defined. □

Due to the consistency of the absolute error for the median functional relative to the class $\mathcal{P}_1$ (Gneiting, 2011), an application of Theorem 2.1 yields the proper scoring rule

$$S_{\mathrm{AE}}(P, y) = |med_P - y|. \tag{2.12}$$

As the squared error is consistent for the mean functional relative to the class $\mathcal{P}_1$ (Gneiting, 2011), the function

$$S_{\mathrm{SE}}(P, y) = (\mu_P - y)^2, \tag{2.13}$$

is also a proper scoring rule relative to the class $\mathcal{P}_1$ according by Theorem 2.1.

In the following, we will use the proper scoring rules CRPS, S_{AE}, and S_{SE}. To illustrate their properties, Figure 2.1 shows these scores as functions of the observation y,

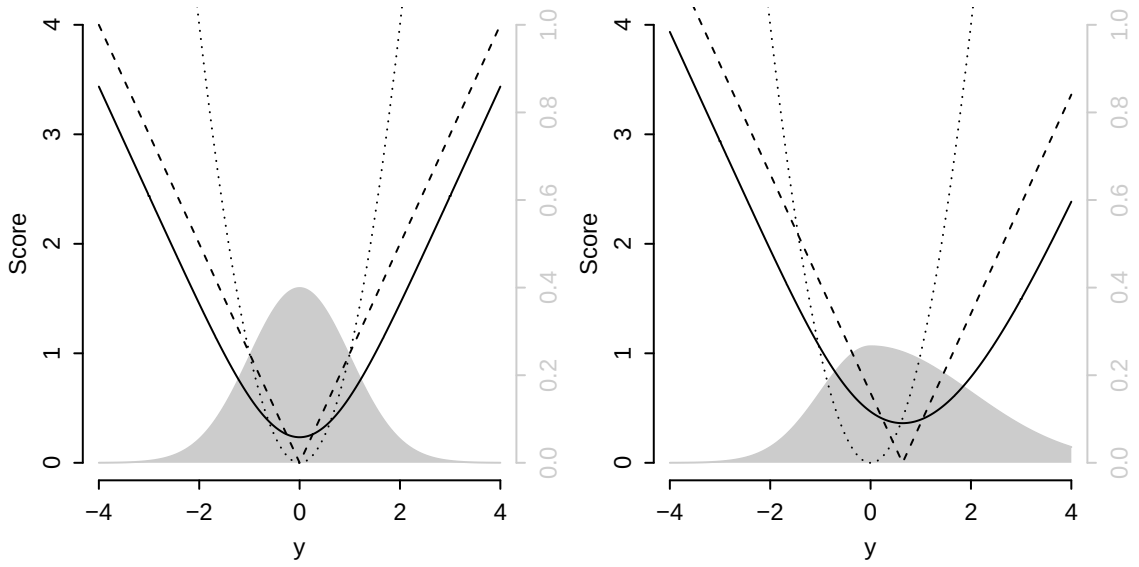

Figure 2.1: The proper scoring rules CRPS (solid line), S_{AE} (dashed line), and S_{SE} (dotted line) as functions of the observation y, when the probabilistic forecast is the standard normal distribution (left) or the two-piece normal distribution with mean $\mu = 0$ and variances $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ (right). The densities are indicated in gray.

when the probabilistic forecast is the standard normal distribution or the two-piece normal distribution with mean $\mu = 0$ and variances $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ (Gneiting and Thorarinsdottir, 2010). Due to the symmetry of the standard normal distribution, the mean and the median are the same, namely zero. The two-piece normal distribution with mean $\mu = 0$ and variances $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ has a median value of about 0.64. The functions in Figure 2.1 are axially symmetric, the CRPS and the S_{AE} with respect to the axis passing through the median value, and the S_{SE} with respect to the axis passing through the mean value. The scoring rules S_{AE} in (2.12) and S_{SE} in (2.13) are minimized for the median value or the mean value, respectively, as we would expect from their formulas for fixed probabilistic forecasts. Since the absolute error is strictly consistent for the median functional, the CRPS in (2.7) obtains its global minimum for a fixed probabilistic forecast F in $y = med_F$. Figure 2.1 also shows this characteristic. Furthermore, if the quality of the forecast is evaluated using the score S_{SE}, the forecaster's loss increases quadratically with the deviation of the observation from the predictive mean value. It increases linearly with the deviation from the predictive median value, if the scoring rule S_{AE} is used. The CRPS penalizes stronger than the S_{AE} only for small deviations of the observation from the predictive median value.

2.1.3 Divergence functions

Here, we consider probabilistic forecasts for a random variable Y on a general sample space $\Omega \subseteq \mathbb{R}$. Throughout this section, we require that all empirical measures on the real axis are included in the underlying convex class $\mathcal{P}$ of probability measures. In contrast to the scoring rules, where the score is based on the forecast distribution and a single observation, the predictive distribution is now compared to the empirical distribution of a set of observations. If $y_1, \dots, y_k$ are observations of the random variable Y with true distribution function G, the *empirical cumulative distribution function* (empirical cdf) is then defined as

$$G_k(x) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}\{x \geqslant y_i\}.$$

The empirical cdf G_k estimates the true underlying cdf G.

A *divergence function* is a function $d : \mathcal{P} \times \mathcal{P} \longrightarrow [0, \infty]$ such that $d(P, P) = 0$ for all probability measures $P \in \mathcal{P}$. The following property of a divergence function d is the analogue to the propriety of scoring rules.

Definition 2.3 (Thorarinsdottir et al., 2012)**.** Let $Y_1, \dots, Y_k$ be random variables with a common distribution function G, and let G_k be the corresponding empirical cdf. A divergence function d is *n-proper* if

$$\mathbb{E}\, d(G, G_k) \leqslant \mathbb{E}\, d(F, G_k) \tag{2.14}$$

for $k = 1, \dots, n$ and all probability measures F, $G \in \mathcal{P}$.

Since d is a nonnegative function, the expectations in (2.14) are well-defined. If the observations are drawn from G, inequality (2.14) implies that the expected divergence is minimized for the forecast $G \in \mathcal{P}$. As in the case of proper scoring rules, the forecaster is thus encouraged to be honest and to state his true beliefs.

Divergence functions that are n-proper for all $n \in \mathbb{N}$ can be obtained from regular and proper scoring rules. A scoring rule $S : \mathcal{P} \times \Omega \longrightarrow \bar{\mathbb{R}}$ is *regular* relative to the class $\mathcal{P}$ if $\mathbb{E}_Q S(P, Y)$ is real-valued for all P, $Q \in \mathcal{P}$ except possibly that $\mathbb{E}_Q S(P, Y) = \infty$ if $P \neq Q$. A function

$$d(P, Q) = \mathbb{E}_Q S(P, Y) - \mathbb{E}_Q S(Q, Y),$$

where P, $Q \in \mathcal{P}$ and S is a regular and proper scoring rule, is clearly a divergence function. We then refer to d as a *score divergence*. The following theorem shows that score divergences are n-proper for all $n \in \mathbb{N}$.

Theorem 2.2 (Thorarinsdottir et al., 2012)**.** If a divergence function d is a score divergence, then d is n-proper for all $n \in \mathbb{N}$.

Proof. Let $Y_1, \ldots, Y_k$ be random variables with a common distribution function $G \in \mathcal{P}$, and let $G_k \in \mathcal{P}$ be the corresponding empirical cdf. As d is a score divergence, there exists a regular and proper scoring rule S such that

$$d(F, G_k) = \mathbb{E}_{G_k} S(F, Y) - \mathbb{E}_{G_k} S(G_k, Y)$$

for any $k \in \mathbb{N}$, where $F \in \mathcal{P}$ and Y is a random variable with distribution G_k. Thus,

$$\begin{aligned}
\mathbb{E}\, d(F, G_k) &= \mathbb{E}\, \mathbb{E}_{G_k} S(F, Y) - \mathbb{E}\, \mathbb{E}_{G_k} S(G_k, Y) \\
&= \mathbb{E}\, \frac{1}{k} \sum_{i=1}^{k} S(F, Y_i) - \mathbb{E}\, \mathbb{E}_{G_k} S(G_k, Y) \\
&= \mathbb{E}_G S(F, Y) - \mathbb{E}\, \mathbb{E}_{G_k} S(G_k, Y) \\
&\geqslant \mathbb{E}_G S(G, Y) - \mathbb{E}\, \mathbb{E}_{G_k} S(G_k, Y) \\
&= \mathbb{E}\, \mathbb{E}_{G_k} S(G, Y) - \mathbb{E}\, \mathbb{E}_{G_k} S(G_k, Y) \\
&= \mathbb{E}\, d(G, G_k).
\end{aligned}$$

□

Let $\mathcal{P}$ consist of all Borel probability measures on $\mathbb{R}$ with finite first moment including all empirical measures on $\mathbb{R}$. The symmetric *Cramér-von Mises distance*,

$$d_{\text{C.v.M.}}(F, G_k) = \int_{-\infty}^{\infty} \big[F(x) - G_k(x)\big]^2 \mathrm{d}x,$$

is the score divergence of the CRPS (Gneiting and Raftery, 2007, p. 367). It is thus n-proper for all $n \in \mathbb{N}$ by Theorem 2.2.

If we further assume that the elements of $\mathcal{P}$ have finite second moment, the score

$$S(F, y) = [\mathbb{E}_F h(X) - h(y)]^2,$$

where X is a random variable with probability distribution F and $h : \mathbb{R} \longrightarrow \mathbb{R}$, is a regular and proper scoring rule relative to $\mathcal{P}$. According to Dawid (1998), its score

divergence, which is referred to as the *squared divergence*, is given by

$$d_S(F, G_k) = [\mathbb{E}_F h(X) - \mathbb{E}_{G_k} h(Y)]^2, \tag{2.15}$$

where Y denotes a random variable with distribution G_k. The squared divergence is also n-proper for all $n \in \mathbb{N}$ by Theorem 2.2. Let $y_1, \ldots, y_k$ denote the observations corresponding to G_k. The *squared mean value divergence*,

$$d_{\mathrm{MV}}(F, G_k) = \left[\mu_F - \frac{1}{k} \sum_{i=1}^{k} y_i \right]^2,$$

is the squared divergence in (2.15) when h is the identity function. It is the score divergence of the scoring rule S_{SE} in (2.13).

In our application, we will use the Cramér-von Mises distance $d_{\mathrm{C.v.M.}}$ and the squared mean value divergence d_{MV}. To illustrate these measures, results of a small simulation study are given in Figure 2.2. Our forecaster aims to predict a normally distributed random variable Y with mean μ_y and standard deviation σ_Y based on a training set consisting of $k = 1,000$ observations. We created such an observation set by generating $1,000$ random values from a normal distribution with mean μ_Y and standard deviation σ_Y. Figure 2.2 shows the divergence functions $d_{\mathrm{C.v.M.}}$ and d_{MV} for different values of μ_Y and σ_Y, when the forecaster quotes the standard normal distribution. In Figure 2.2 (a)–(c) the mean values are $\mu_Y \in \{-3, -2, \ldots, 2, 3\}$ with a constant value of the standard deviation σ_Y. The three figures are almost identical, even if the standard deviation is predicted correctly in (b) but not in (a) and (c). Furthermore, the divergence function d_{MV} penalizes stronger for deviations from the true mean value. Figures 2.2 (d)–(f) show the divergence functions $d_{\mathrm{C.v.M.}}$ and d_{MV} for standard deviation values of $\sigma_Y \in \{0.25, 0.5, 0.75, \ldots, 2\}$ and fixed values of μ_Y. Here, the scores are almost constant within each plot. They are considerably lower for the correct mean prediction, which is illustrated in Figure 2.2 (e). Note that the exact prediction of the true mean value is more important than the exact prediction of the true standard deviation value, particularly if the forecast verification is based on the score d_{MV}. To summarize, the simulation study shows that the squared mean value divergence d_{MV} considers only the forecast mean, as we would expect from its formula. However, we will use this divergence function, as it might be interesting for comparison.

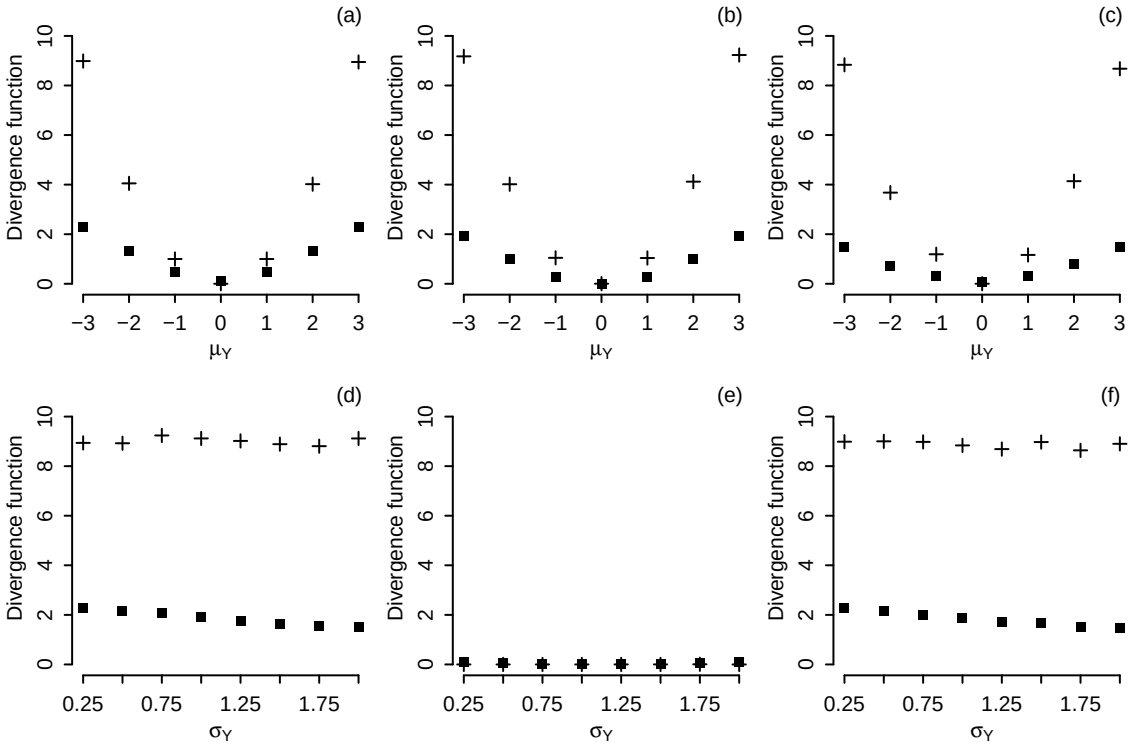

Figure 2.2: The divergence functions $d_{\text{C.v.M.}}$ (▪) and d_{MV} (+), when the observation set follows a normal distribution with mean μ_Y and standard deviation σ_Y and the probabilistic forecast is the standard normal distribution, for different values of μ_Y and σ_Y. The plots in the top row have fixed values of σ_Y as follows: (a) $\sigma_Y = 0.25$, (b) $\sigma_Y = 1$, (c) $\sigma_Y = 2$. The bottom row shows the measures for the following fixed values of μ_Y: (d) $\mu_Y = -3$, (e) $\mu_Y = 0$, (f) $\mu_Y = 3$.

2.2 Weighted combination of probabilistic forecasts

Now we consider probabilistic forecasts produced by several competing forecast procedures for a random variable Y on a finite interval $[a, b]$. The assumption $Y \in [a, b]$ may be relaxed. However, this will always hold for our application, and we thus make this assumption for clarity of the argumentation below. Let $\mathcal{P}$ denote a convex class of Borel probability measures on the corresponding Borel σ-algebra $\mathcal{B}([a, b])$ with finite second moment such that all empirical measures are included in $\mathcal{P}$. The forecast procedures may be compared and assessed by means of the abovementioned measures. We discuss combinations of competing forecasts based on the rankings achieved by the divergence functions.

Let $F^1, \ldots, F^m$ denote probabilistic forecasts for Y produced by m competing forecast procedures. The forecasts are identified with their cdfs. Furthermore, let G_k denote the empirical cdf corresponding to k observations of Y. We may combine

these forecasts by finding weights $w_1, \ldots, w_m$ such that

$$F(x) = \sum_{j=1}^{m} w_j F^j(x)$$

is a new probabilistic forecast. That is, the sum of the weights has to be equal to 1, and each weight must be nonnegative. We consider three weighting methods. The first, hereinafter referred to as the *weights$_{min}$ method*, minimizes the Cramér-von Mises distance $d_{\text{C.v.M.}}$. That is, we consider

$$\min_{w} d_{\text{C.v.M.}}(F, G_k), \tag{2.16}$$

where w denotes the vector with the weights as entries. This optimization problem is a quadratic programming problem (QP): By applying the Binomial Formula, we obtain

$$\begin{aligned} d_{\text{C.v.M.}}(F, G_k) = & \int_a^b \left[\sum_{j=1}^{m} w_j F^j(x) - G_k(x) \right]^2 \mathrm{d}x \\ = & \sum_{j=1}^{m} \sum_{l=1}^{m} w_j w_l \int_a^b F^j(x) F^l(x) \mathrm{d}x \\ & - 2 \sum_{j=1}^{m} w_j \int_a^b F^j(x) G_k(x) \mathrm{d}x \\ & + \int_a^b G_k^2(x) \mathrm{d}x. \end{aligned}$$

Thus, the optimization problem (2.16) can be formulated as

$$\min_{w} d_{\text{C.v.M.}}(F, G_k) = \min_{w} \left(\frac{1}{2} w^T Q w + q^T w \right),$$

where Q denotes the matrix with entries

$$Q_{ij} = 2 \int_a^b F^i(x) F^j(x) \mathrm{d}x$$

for $i, j = 1, \ldots, m$ and q denotes the vector with entries

$$q_j = -2 \int_a^b F^j(x) G_k(x) \mathrm{d}x$$

for $j = 1, \ldots, m$. The optimization problem (2.16) is therefore a QP under the

constraints $\sum_{i=1}^{m} w_i = 1$ and $w_j \geqslant 0$ for $j = 1, \ldots, m$. To solve this problem, we will use the interior point method described in Vanderbei (1999) as implemented in the function `ipop` of the `R` package `kernlab` (R Development Core Team, 2012).

Our second weighting method, hereinafter referred to as the *weights*$_{d_{C.v.M.}}$ *method*, is to calculate the values of the Cramér-von Mises distance $d_{\text{C.v.M.}}$ over a training set and then set

$$w_j = \frac{\frac{1}{d_{\text{C.v.M.}}(F^j, G_k)}}{\sum_{i=1}^{m} \frac{1}{d_{\text{C.v.M.}}(F^i, G_k)}}$$

for $j = 1, \ldots, m$. The conditions for the weights are clearly met. We obtain large weights for small values of the divergence function. Since we prefer a forecast procedure with a small score over one with a large score, the forecasts we consider better are given more weight.

The third weighting method, hereinafter referred to as the *weights*$_{d_{MV}}$ *method*, is similar to the second, only based on the values of the squared mean value divergence d_{MV}. Here, the weights are

$$w_j = \frac{\frac{1}{d_{\text{MV}}(F^j, G_k)}}{\sum_{i=1}^{m} \frac{1}{d_{\text{MV}}(F^i, G_k)}}$$

for $j = 1, \ldots, m$. Later, we will introduce an application of these weigthing methods in practice and will test whether they provide better forecasts than the forecast procedures themselves.

3 Case study: Verification of climate models

3.1 Climate models

Climate models are essential scientific tools for studying and simulating past, present, and future climate change. The climate system is a highly complex system that consists of several components, such as the atmosphere and the hydrosphere. Many processes and interactions in the climate system are governed by known laws of physics and chemistry and can be formulated as differential equations, such as the law of mass conversation or equations of motions. Since it is not possible to solve these equations directly, climate researchers develop computer programs on the basis of these equations. In order to be able to perform the frequently very complex, large, and processing intensive models, the simulations are only carried out at evenly distributed grid points on the earth's surface and over atmospheric layers. In the course of this, the earth is covered by an imaginary grid. Its mesh width is referred to as the spatial model resolution.

Coupled atmosphere-ocean general circulation models (AOGCMs) are the models most often used to simulate past, present, and future climate states. A general circulation model simulates the general circulation of a climate component, such as a planetary atmosphere or ocean, using three-dimensional grids. GCMs impose different boundary values. For instance, to model the atmosphere, atmospheric GCMs (AGCMs) need values of the sea surface temperature, among other things. Conversely, oceanic GCMs (OGCMs) that describe the ocean system need values of atmospheric elements, such as the air temperature. As it is not always possible to obtain the boundary conditions from observations, AGCMs and OGCMs are coupled together to form AOGCMs (Roedel and Wagner, 2011, p. 558 ff.). Further information on the climate and its modeling can be found in Von Storch et al. (1999) and Roedel and Wagner (2011).

The World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project (CMIP) studies output from AOGCMs. To this end, model outputs contributed by leading modeling centers around the world are collected and archived by the Program for Climate Model Diagnosis and Intercomparison (PCMDI). In our application, we use output data of 15 global climate models which is available in

Table 3.1: The 15 CMIP3 models evaluated in this study with originating group(s), country, and horizontal resolution in numbers of longitudes and latitudes (`http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php`).

Model	Originating Group(s) and Country	Resolution
CGCM3.1(T47)	Centre for Climate Modelling & Analysis (Canada)	96 × 48
CGCM3.1(T63)	Centre for Climate Modelling & Analysis (Canada)	128 × 64
CNRM-CM3	Méteo-France / Centre National de Recherches Météorologiques (France)	128 × 64
CSIRO-Mk3.0	CSIRO Atmospheric Research (Australia)	192 × 96
CSIRO-Mk3.5	CSIRO Atmospheric Research (Australia)	192 × 96
ECHAM5/MPI-OM	Max Planck Institute for Meteorology (Germany)	192 × 96
FGOALS-g1.0	LASG / Institute of Atmospheric Physics (China)	128 × 60
GFDL-CM2.0	US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory (USA)	144 × 90
GFDL-CM2.1	US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory (USA)	144 × 90
GISS-AOM	NASA / Goddard Institute for Space Studies (USA)	90 × 60
GISS-EH	NASA / Goddard Institute for Space Studies (USA)	72 × 46
GISS-ER	NASA / Goddard Institute for Space Studies (USA)	72 × 46
MIROC3.2(hires)	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC) (Japan)	320 × 160
MIROC3.2(medres)	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC) (Japan)	128 × 64
MRI-CGCM2.3.2	Meteorological Research Institute (Japan)	128 × 64

the data base of the third phase of the CMIP (CMIP3) (Meehl et al., 2007). This data set was also used as the base for the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC, 2007).

The 15 climate models, their originating group(s), their country of origin, and their hoizontal resolution are listed in Table 3.1. The horizontal resolution represents an important difference between the models. It is given in numbers of longitudes and latitudes, which means the following: The earth is covered by an imaginary grid which consists of 360 degrees of longitude and 180 degrees of latitude. The German model, ECHAM5/MPI-OM, for instance, has a horizontal resolution of 192 longitudes and 96 latitudes. The 360 degrees of longitudes are therefore divided in 192 longitudes and the 180 degrees of latitude in 96 latitudes, to result in a grid with $192 \times 96 = 18,432$ boxes. This corresponds to a grid mesh width of about 210 km at the equator. Here, larger values indicate finer grids. Thus, the MIROC3.2(hires) model has the best horizontal resolution, whereas the GISS-EH and GISS-ER models have the worst. A detailed description and further information on the "WCRP CMIP3 multi-model data set" can be found on the PCMDI website `http://www-pcmdi.llnl.gov`.

3.2 Re-analyses and observations

In order to assess the skill of one model in providing simulations, comparison with re-analyses and/or observational data sets is necessary (cf. Chapter 1). In a re-analysis, observational data over an extended historical period are reprocessed and assimilated into a fixed modern forecasting system. To this end, for each day of the period over which suitable observations exists, observations are combined with the results from the forecast model. The physical laws embodied in the forecast system and knowledge of the typical errors of forecasts and observations facilitate the interpretation of conflicting or indirect observations and the filling of gaps in the observational coverage. Thus, a data set which can be used for meteorological and climatological studies is produced (ECMWF, 2012). The re-analyses we use in the following are the ERA-40 re-analysis (Uppala et al., 2005) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the NCEP-1 re-analysis (Kalnay et al., 1996; Kistler et al., 2001) provided by the National Centers for Environmental Prediction and the National Center for Atmospheric Research (NCEP/NCAR). Both re-analyses have a horizontal resolution of 144 longitudes and 73 latitudes. It must be kept in mind that re-analyses may contain model errors (Uppala et al., 2005). The optimal approach would thus be to compare the GCM outputs with observations. This is difficult to achieve: Whereas model data reflect the average of the climate conditions over each grid cell, observations are taken at individual stations. It is, however, possible to convert global observational data into a grid format as we discuss in Chapter 3.4, where we apply such a data set instead of a re-analysis.

3.3 Indices for climate extremes

The Expert Team on Climate Change Detection and Indices (ETCCDI) has developed indices for extremes. As already mentioned in Chapter 1, the indices we consider here facilitate a quantitative analysis of moderate extreme events with short return periods, where each index describes a particular characteristic of an extreme event, such as its frequency, amplitude, or persistence. A set of 27 indices based on daily temperature values or daily precipitation amount were considered to be core indices (Peterson, 2005). These include e.g. the monthly maximum value of daily maximum temperature, the monthly minimum value of daily minimum temperature, the number of tropical nights, the monthly maximum 1-day precipitation, and the maximum number of consecutive dry days. Further information and an exact definition of these indices can be found on the ETCCDI website `http://cccma.seos.uvic.ca/ETCCDI`.

Table 3.2: Definition of the monthly maximum value of daily maximum temperature (mtxmax) and the monthly minimum value of daily minimum temperature (mtnmin) (`http://cccma.seos.uvic.ca/ETCCDI/list_27_indices.shtml`).

Index	Definition
Mtxmax	Let tx_i be the daily maximum temperature (in °C) on day i, $i = 1, \ldots, n$, in a month with n days. The maximum daily maximum temperature in this month is then $\mathrm{mtxmax}=\max_i(\{\mathrm{tx}_1,\ldots,\mathrm{tx}_n\})$.
Mtnmin	Let tn_i be the daily minimum temperature (in °C) on day i, $i = 1, \ldots, n$, in a month with n days. The minimum daily minimum temperature in this month is then $\mathrm{mtnmin}=\min_i(\{\mathrm{tn}_1,\ldots,\mathrm{tn}_n\})$.

The Hadley Centre for Climate Prediction and Research provides a global land-based climate extreme data set referred to as HadEX (`http://www.metoffice.gov.uk/hadobs/hadex`). The HadEX data set contains the 27 indices described in Peterson (2005) on an annual basis for the time period 1951-2003 and at grid points corresponding to a grid of 96 longitudes and 73 latitudes. The indices are based on daily data from worldwide weather observation stations. For a more detailed description of the data set and its production process, see Alexander et al. (2006).

We will concentrate on two temperature based indices, the *monthly maximum value of daily maximum temperature* (mtxmax) and the *monthly minimum value of daily minimum temperature* (mtnmin). As the HadEX data contains annual indices, we also analyze the indices on a yearly basis. Here, we obtain the *yearly maximum value of daily maximum temperature* (txmax) and the *yearly minimum value of daily minimum temperature* (tnmin). The maximum of the maximum temperature and the minimum of the minimum temperature are robust and plausible considering the relatively coarse model resolution. Furthermore, they contain useful information for climate change impact studies (Sillmann and Roeckner, 2008). The precise definition of mtxmax and mtnmin is given in Table 3.2. Txmax and tnmin are defined similarly.

3.4 Data

In our analysis, the comparison of model and observation or re-analysis based indices will be concentrated on land grid boxes in the European domain (8W-40E, 32N-72N), see Figure 3.1, during a 30-year period from 1961 to 1990. Corresponding simulations of the daily 2 m maximum or minimum temperature, signifying temperature measurements at a height of 2 m, are available from the CMIP3 data base as well as for both re-analyses[1]. In the following, we will identify climate models with

[1] CMIP3: `http://www-pcmdi.llnl.gov`,
ERA-40: `http://www.ecmwf.int/research/era/do/get/era-40`,
NCEP-1: `http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html`.

prediction models, although this is not quite correct due to the fact that outputs of a climate model are simulations of climate variables.

Using the data sets of daily temperature, the indices were calculated on the native model grid for each individual global climate model from Table 3.1 and the re-analyses. Then, both these indices and the HadEX indices were remapped for comparison via conservative remapping to a common spatial resolution. The common resolution is taken as the resolution of the ERA-40 and the NCEP-1 re-analysis, that is, the indices were remapped to a resolution of 144 longitudes and 73 latitudes.

When comparing indices based on the HadEX data, the climate models, and the re-analyses, it should be noted that the climate models and the re-analyses are based on data representative for the whole grid box area, whereas the HadEX indices were calculated for each weather observation station separately and then interpolated to the longitude-latitude grid. The methods applied to the climate models and the re-analyses thus imply a stronger smoothing of extremes than the method applied to the HadEX indices. This may result in systematic differences in the probability distributions of the data. However, we expect this effect to have less of an impact for temperature data, which is considered here, than for precipitation data (Sillmann and Roeckner, 2008).

3.5 Results

3.5.1 Model comparisons

To verify the climate model outputs against those based on the re-analyses or the observations, we apply scoring functions, proper scoring rules, and the associated divergence functions introduced in Chapter 2.1. The evaluation is made in two ways. On the one hand, we look at each grid point separately, hereinafter referred to as the *local method*, and, on the other hand, we pool data from all grid points in the study region and ignore the locality, hereinafter referred to as the *regional method*.

Let x be a single data value from a climate model, and let y be the corresponding event from a re-analysis or the HadEX data set. To compare both point forecasts, we apply the absolute error and the squared error. We repeat this for all point forecasts at any time over all available grid points in the test set. The resulting summary measure of the performance assigned to a given climate model is therefore given by the following average score,

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}S(x_{ij},y_{ij}),$$

where i is the index over the individual locations in the study region and j is the index over the available data values for each i. The number n of available values for each i depends on the time period and the season considered, as will be seen below. In the regional method, all locations are pooled together and, hence, i is equal to one. Therefore, the number n also depends on which method, the local or the regional, is used. Here, however, it is not necessary to distinguish between the regional and the local method, since both result in the same value.

The climate indices mtxmax, mtnmin, txmax, and tnmin can be considered as quantiles of a continuous quantity. Therefore, to compare the single extreme index values from a climate model to the corresponding re-analysis or observation based values, the asymmetric piecewise linear scoring function in (2.3) can also be applied. We choose the level α as follows. For mtxmax and txmax, we set $\alpha = \frac{N-\frac{1}{2}}{N}$, where N is the number of days in the current month or year, respectively. Similarly, we set $\alpha = \frac{1}{2N}$ for mtnmin and tnmin. We then assign each climate model the following average score, hereinafter referred to as S_{index} depending on the index considered,

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n} S_{\alpha_j}(x_{ij}, y_{ij}),$$

where we use the same notation as before. The local and the regional method also result in the same value for this average score.

Let F denote the empirical cdf of extreme index values from a climate model over a large time period. To compare F against the corresponding re-analysis or HadEX events y, we apply the proper scoring rules CRPS, S_{AE}, and S_{SE}. Similar as for the scoring functions above, we take the average score at each available grid point. The resulting score of a given climate model is then given by these averages,

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n} S(F_i, y_{ij}),$$

where i is the index over the individual locations in the study region and j is the index over the data values used to create F_i for each i.

As before, let F denote the empirical cdf of extreme index values from a climate model over a large time period, and let G denote the corresponding empirical cdf based on a re-analysis or the HadEX data set. To compare the two empirical cdfs, we apply the score divergences $d_{\text{C.v.M.}}$ and d_{MV}. By repeating this over all available grid points and taking the average, we assign each climate model the average divergence,

$$\frac{1}{m}\sum_{i=1}^{m} d(F_i, G_i),$$

where i is the index over the individual locations in the study region.

To summarize, the scoring functions measure how close values predicted by a model are to the re-analysis or the observation based values. The scoring rules compare the empirical cdf of a climate model against the single re-analysis or observation based values. The divergence functions assess the compatibility of the model based empirical cdfs and the re-analysis or the observation based empirical cdfs. Each score provides a ranking of the climate models. If the score assigned to a climate model is lower than the score of another model, it is considered to have a better model performance. The local method provides additionally a model ranking at each location. Thus, it can be investigated whether the model ranking is different for distinct locations and whether a model has issues with providing reliable outputs at some locations. The comparison of the different scoring methods further makes it possible to see whether the same models are considered good under all three settings.

We first calculate the averages above over all available data in our data set from 1961 to 1990. Due to the huge amount of results and their similarity, we present only the verification results for the climate model predictions of mtxmax and mtnmin when compared to the NCEP-1 re-analysis and for the predictions of txmax and tnmin when compared to the HadEX indices. Furthermore, only the most interesting results are mentioned and considered in detail. The full set of verification results averaged over the grid points is given in Appendix A. To determine the skill and the variability of the re-analyses, we also compare the two re-analysis data sets and assess the differences between the re-analyses and the HadEX data set.

Figure 3.1 provides an illustration of the model rankings that result from the calculation of the average scores. It shows the verification results for the predictions of mtnmin for both the regional and the local method, when compared to the NCEP-1 re-analysis (cf. Table A.6). The vertical lines, which are linearly scaled, visualize the individual scores, and the horizontal lines illustrate the different average scores of the individual climate models. The horizontal lines corresponding to models from the same institution (cf. Table 3.1) have the same color. Taking all verification results into account, there is, however, no climate model that is always superior to the other models of the same institution. The average scores for the NCEP-1 re-analysis and the ERA-40 re-analysis are also indicated. It is important to note that a good half of the models rank better than the ERA-40 re-analysis, in both the regional and the local analysis. In general, it makes little difference for the rankings whether the regional or the local method is considered. The individual scores also demonstrate certain properties. For example, as expected, the MAE and the MSE have similar rankings (cf. note on their definitions in Chapter 2.1.1, p. 5). Furthermore, Figure 3.1 validates that a scoring rule and its associated divergence function yield

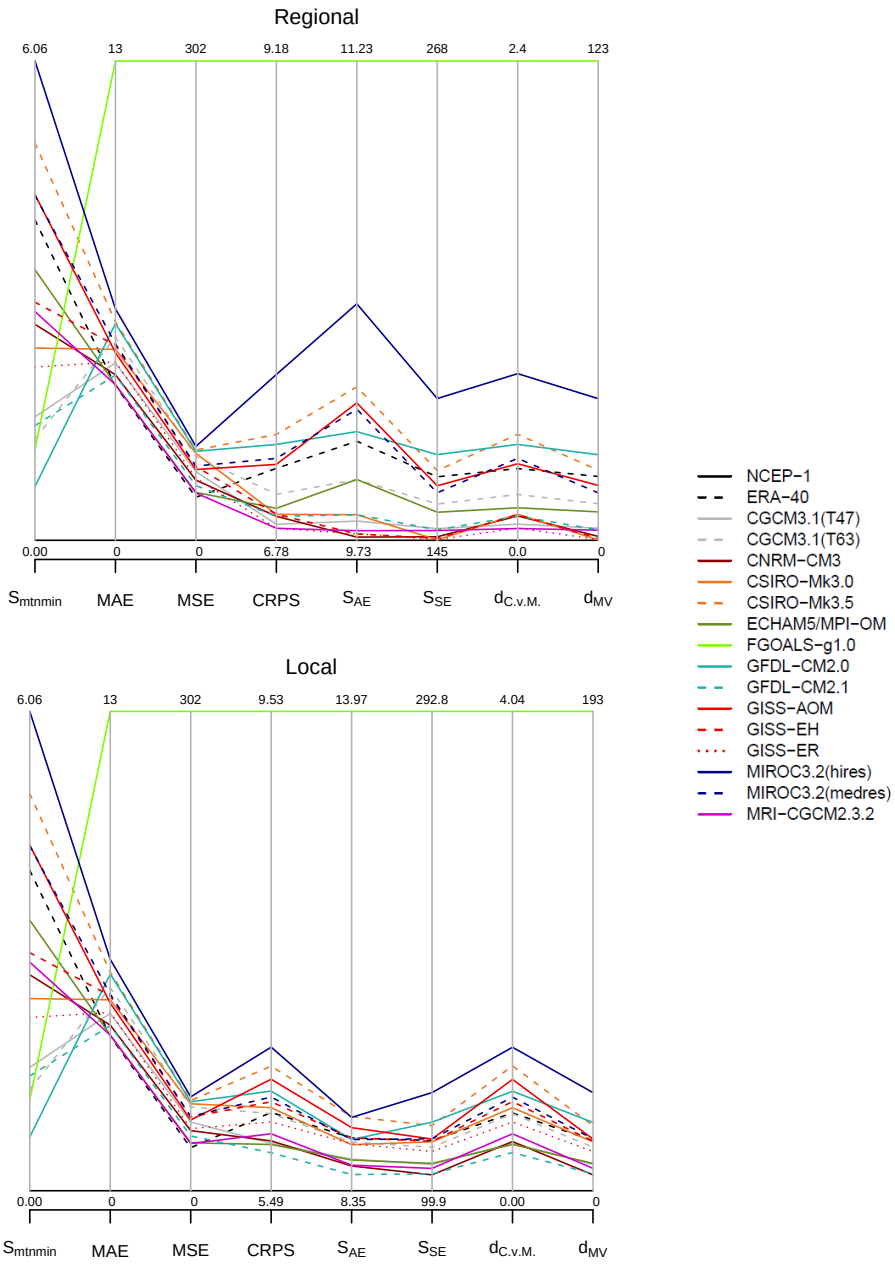

Figure 3.1: Rankings of the climate models in simulating mtnmin for the regional method (top) and the local method (bottom) when compared to the NCEP-1 re-analysis. The corresponding values of the scores and divergence functions for the NCEP-1 re-analysis and the ERA-40 re-analysis are also shown.

the same ranking. This is clear from the definition of a score divergence (see Chapter 2.1.3, p. 14) and the monotony of the expectation. Thus, the rankings of the climate models under the CRPS and the Cramér-von Mises distance $d_{\text{C.v.M.}}$ or the scoring rule S_{SE} and the squared mean value divergence d_{MV}, respectively, are identical. The rankings based on the different scores are relatively constant in general other than the ranking based on the asymmetric piecewise linear scoring function. For instance, the FGOALS-g1.0 model and the GFDL-CM2.0 model rank among the best mod-

els under the asymmetric piecewise linear scoring function, but their performances are among the worst if the verification is based on the other scores. In this case, the FGOALS-g1.0 model is by far the worst model. This might be due to its low horizontal resolution. However, this cannot be the only reason for its bad performance, since other models with low resolution, such as the CGCM3.1(T47) model, perform well. In contrast to the FGOALS-g1.0 model and the GFDL-CM2.0 model, the ERA-40 re-analysis and the MIROC3.2(medres) model rank much worse under the asymmetric piecewise linear scoring function than under the other scores.

To explain this fact, we look at the asymmetric piecewise linear scoring function more closely. If x is a prediction for the quantile at level $\alpha \in (0, 1)$ and y is the "true value", the asymmetric piecewise linear scoring function is given by

$$\begin{aligned} S_\alpha(x, y) &= (\mathbb{1}\{y \leqslant x\} - \alpha)(x - y) \\ &= \text{weight}(x, y, \alpha)|x - y|, \end{aligned}$$

where $\text{weight}(x, y, \alpha) := |\mathbb{1}\{y \leqslant x\} - \alpha|$. Thus, this proper scoring rule for quantiles is a weighted version of the absolute error. If α is small, as is, for instance, the case for mtnmin, the score is larger for an overestimating prediction than for an underestimating prediction with the same deviation from the "true value". Conversely, it is larger for an underestimating prediction than for an overestimating prediction for quantiles at large levels, such as mtxmax. As the lower the score, the better the prediction model, the asymmetric piecewise linear scoring function prefers overestimating climate models for maximum based extreme indices, and underestimating models for minimum based indices.

To confirm this in the situation of Figure 3.1, we consider the mtnmin biases from the NCEP-1 re-analysis for all climate models and grid boxes, and additionally the same for the ERA-40 re-analysis. Figure 3.2 shows the corresponding box plots. If the difference between a prediction of the NCEP-1 re-analysis and a climate model or the ERA-40 re-analysis, respectively, for mtnmin is negative, the value of the NCEP-1 re-analysis is overestimated. If it is positive, the value is underestimated. Thus, the FGOALS-g1.0 model and the GFDL-CM2.0 model usually underestimate the true mtnmin values, whereas the ERA-40 re-analysis and the MIROC3.2(medres) model usually yield an overestimate. The FGOALS-g1.0 model and the GFDL-CM2.0 model thus rank much better under the asymmetric piecewise linear scoring function than under the MAE and the MSE (see Figure 3.1), since the absolute errors caused by underestimating are given less weight. The opposite applies to the ERA-40 re-analysis and the MIROC3.2(medres) model. This seems to be a bad property. Nevertheless, there may be situations for which it is worse to overestimate the minimum temper-

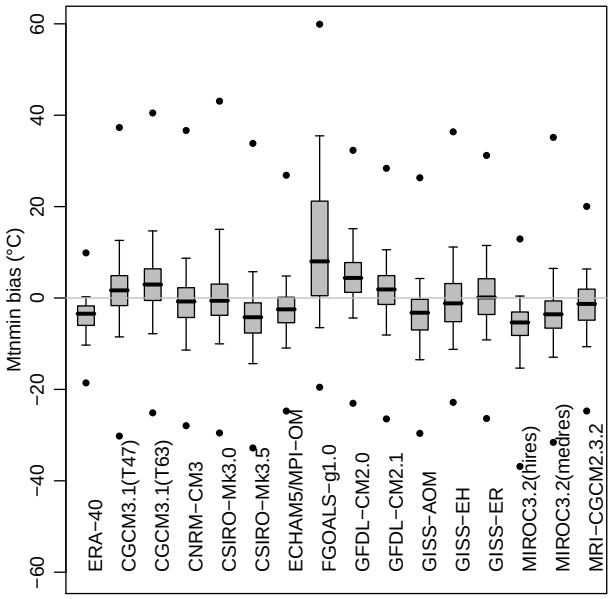

Figure 3.2: Box plots of the mtnmin biases from the NCEP-1 re-analysis for the ERA-40 re-analysis and all climate models at all grid boxes. The median is indicated with a bold line. The box marks the interquartile range; the lines mark the 5%-95% range. The points are the maximum or the minimum bias, respectively (Knutti et al., 2010, p. 2745).

ature than to underestimate it and vice versa for the maximum temperature. For instance, if people have to be warned from a certain temperature on, it is better to warn them to soon than too late. Thus, the asymmetric piecewise linear scoring function may only be used at specific locations or at certain times of the year, and, hence, we omit this score.

The rankings of the MAE and the MSE in Figure 3.1 are also illustrated by Figure 3.2. The larger the biases of a model, the larger its MAE and its MSE (cf. note on their definitions in Chapter 2.1.1, p. 5). Since the FGOALS-g1.0 model has the greatest biases, it is the worst model under both scores. Conversely, the ERA-40 re-analysis has the smallest biases and ranks best (see Figure 3.1).

The characteristics of each climate variable, such as surface temperature, differ between the seasons. Thus, similarly to the local and regional method, we consider both yearly predictions, as in Figure 3.1, and seasonal predictions. Since the seasonal predictions for mtxmax have the most interesting consequences in summer (June–August) and for mtnmin in winter (December–February), we will focus on these seasons. In the following, we consider in more detail the verification results of the climate models and the ERA-40 re-analysis for mtxmax and mtnmin when compared locally to the NCEP-1 re-analysis under the Cramér-von Mises distance $d_{\text{C.v.M.}}$. Table 3.3 shows the results for the predictions for mtxmax in the summer months and for mtnmin in the winter months. The results for the predictions of both indices in all months are also stated for comparison. The scores of the climate models for the

yearly predictions are smaller for mtxmax than for mtnmin. The same also applies for individual seasons. The climate models are therefore usually better in predicting mtxmax than mtnmin. When comparing the different seasons and the entire year, the performances of the climate models in simulating mtxmax are worst in the summer months. The same holds for mtnmin in the winter months. Consequently, the climate models have problems in simulating particularly low or high temperatures. The differences between the model performances are also greater when considering the most extreme temperatures, as the range of the model scores is greater for mtxmax in summer and for mtnmin in winter than that of the model scores based on all predictions. This also becomes clear if we compare the mean of the model scores to the corresponding smallest score of a single climate model. Furthermore, the ERA-40 re-analysis is considered as a prediction model and compared to the NCEP-1 re-analysis. It ranks in each case among the best models, but only ranks as the very best for mtxmax in the summer months. The largest divergence between the two re-analyses appears when simulating mtnmin in the winter months. To summarize the consequences of Table 3.3, the quality of a climate model in simulating extreme indices depends both on the season under consideration and on the index itself. The latter dependence is also shown by the fact that the rankings of the climate models are quite different depending on the index. For instance, the MIROC3.2(hires) model ranks among the best models in predicting mtxmax, whereas it performs among the worst in predicting mtnmin.

The GFDL-CM2.1 model is the best model for the winter and the yearly predictions of mtnmin averaged over all grid points under the Cramér-von Mises distance $d_{\text{C.v.M.}}$ (see Table 3.3). In order to investigate whether the performance of a climate model also depends on the region, we consider the results of the situation above at each location. Figure 3.3 illustrates the corresponding scores at each location in our test set. We see that the model has problems at the same locations in winter as over the entire year even if the performance for the latter is clearly better. The GFDL-CM2.1 model is better in simulating mtnmin in Central Europe than in Great Britian, Italy, or Norway. The worst performance takes place in Turkey and in the northern part of the East-European plain.

To summarize, the index and region under consideration influence the quality of a climate model and its simulations. Concerning the latter point, it is difficult to specify results which are universally valid. Instead, the rankings of the climate models in simulating an extreme index differ at the individual grid boxes (cf. also Chapter 3.5.2). However, all climate models have more or less problems in simulating extreme indices at grid points whose location is associated with a permanent extreme weather situation, such as high wind, heat, or cold.

Table 3.3: Verification results for the predictions of the climate models and the ERA-40 re-analysis for both mtxmax and mtnmin over the entire year when compared locally to the NCEP-1 re-analysis under the Cramér-von Mises distance $d_{\text{C.v.M.}}$. Furthermore, the results for the predictions of mtxmax in the summer months and for the predictions of mtnmin in the winter months are also shown. In each case, the best climate model is indicated in bold and the mean of all climate models is stated.

	Mtxmax		Mtnmin	
Forecast	Summer	All the year	Winter	All the year
NCEP-1	0.000	0.000	0.000	0.000
ERA-40	0.855	0.182	1.965	0.656
CGCM3.1(T47)	1.842	0.323	1.108	0.401
CGCM3.1(T63)	1.329	0.270	1.690	0.649
CNRM-CM3	1.331	0.452	2.024	0.415
CSIRO-Mk3.0	2.504	0.562	3.357	0.700
CSIRO-Mk3.5	**0.883**	0.265	3.656	1.053
ECHAM5/MPI-OM	1.001	0.176	1.287	0.390
FGOALS-g1.0	2.892	0.794	16.844	4.042
GFDL-CM2.0	1.820	0.441	2.456	0.839
GFDL-CM2.1	0.994	0.239	**0.971**	**0.322**
GISS-AOM	4.861	0.897	3.353	0.939
GISS-EH	1.813	0.405	2.034	0.752
GISS-ER	2.474	0.545	1.944	0.580
MIROC3.2(hires)	1.005	**0.156**	4.488	1.210
MIROC3.2(medres)	1.259	0.291	2.854	0.789
MRI-CGCM2.3.2	0.888	0.295	1.963	0.480
$\textbf{mean}_{\text{models}}$	1.793	0.407	3.335	0.904

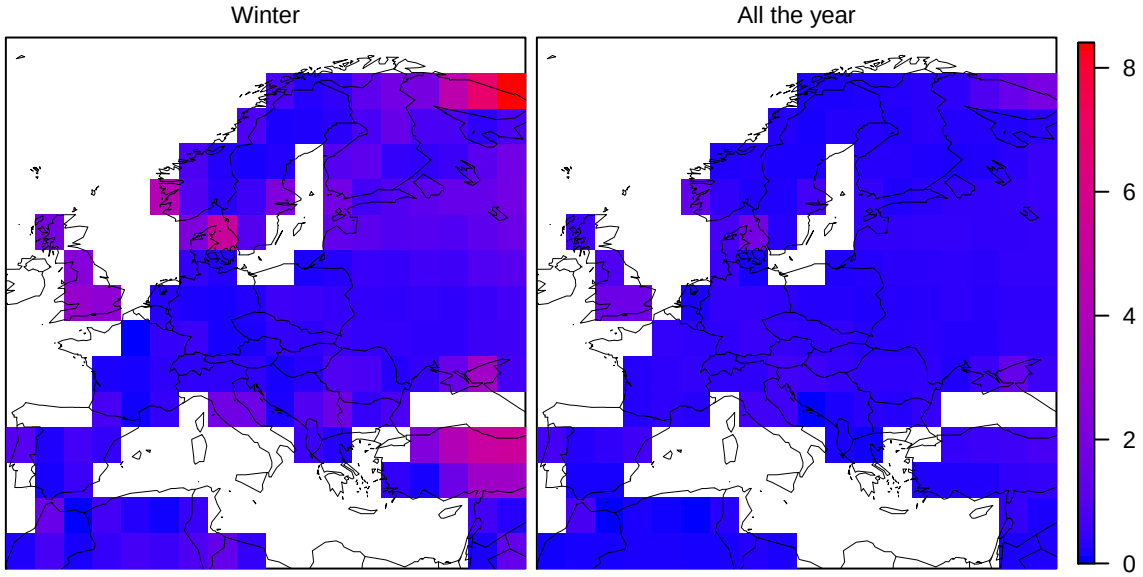

Figure 3.3: Location-specific results for the predictions of the GFDL-CM2.1 model for mtnmin in the winter months (left) and for all the year (right) at land grid points over Europe when compared to the NCEP-1 re-analysis under the Cramér-von Mises distance $d_{\text{C.v.M.}}$.

Table 3.4: Verification results for the predictions of the climate models, the ERA-40 re-analysis, and the NCEP-1 re-analysis for both txmax and tnmin when compared locally to the HadEX indices under the Cramér-von Mises distance $d_{\text{C.v.M.}}$ and the squared mean value divergence d_{MV}. In each case, the best climate model is indicated in bold and the mean of all climate models is stated.

	Txmax		Tnmin	
Forecast	$d_{\text{C.v.M.}}$	d_{MV}	$d_{\text{C.v.M.}}$	d_{MV}
HadEX	0.000	0.0	0.000	0.0
ERA-40	1.768	12.8	1.379	17.2
NCEP-1	1.620	13.9	2.303	28.4
CGCM3.1(T47)	3.648	38.0	3.993	64.4
CGCM3.1(T63)	2.897	28.4	5.109	91.8
CNRM-CM3	3.292	33.9	2.950	39.4
CSIRO-Mk3.0	4.456	51.7	4.922	89.0
CSIRO-Mk3.5	2.038	20.8	3.512	48.6
ECHAM5/MPI-OM	2.227	20.8	**0.963**	**9.9**
FGOALS-g1.0	3.660	38.2	21.552	729.0
GFDL-CM2.0	2.987	29.5	5.695	98.0
GFDL-CM2.1	2.065	18.4	2.207	31.7
GISS-AOM	7.451	93.4	2.779	38.6
GISS-EH	3.640	40.4	2.958	41.0
GISS-ER	4.515	59.4	2.957	42.3
MIROC3.2(hires)	**1.398**	**10.0**	4.155	58.8
MIROC3.2(medres)	2.875	33.8	2.764	35.6
MRI-CGCM2.3.2	2.337	18.7	1.571	17.1
$\textbf{mean}_{\text{models}}$	3.299	35.7	4.539	95.7

Using HadEX indices as the realized values gives us a feeling for how reliable the re-analyses are. Table 3.4 shows the predictive performance of the climate models, the ERA-40 re-analysis, and the NCEP-1 re-analysis for txmax and tnmin when compared locally to the HadEX indices under the divergence functions $d_{\text{C.v.M.}}$ and d_{MV}. Note that the set of grid points differs slightly from that under the monthly indices, cf. Figure 3.4. When comparing Table 3.3 and Table 3.4, the Cramér-von Mises distance $d_{\text{C.v.M.}}$ classifies the same models as good, average, or bad for mtxmax or txmax. Conversely, the rankings corresponding to mtnmin or tnmin differ considerably. For instance, the CGCM3.1(T47) model is among the best for mtnmin, but ranks quite poorly for tnmin. The performance of the GISS-AOM model is much better than the average for tnmin, whereas the model ranks significantly worse for mtnmin. Particularly because of the different rankings for mtnmin in the winter months, the NCEP-1 re-analysis might not be able to capture the indices based on minimum temperatures reliably. The same also holds for the ERA-40 re-analysis (results not shown). In general, the two re-analysis data sets perform very similarly when compared to the HadEX indices, see also Tables A.7 and A.8 in the Appendix, with the ERA-40 re-analysis performing slightly better. Furthermore, the re-analyses

are usually less skillful than the best performing models, see e.g. Table 3.4.

The rankings of the climate models for txmax and tnmin are quite similar under the divergence functions $d_{\text{C.v.M.}}$ and d_{MV} (Table 3.4). The German model, ECHAM5/MPI-OM, ranks among the best models and is the best in simulating tnmin. To gain an even deeper insight into these divergence functions, Figure 3.4 shows the verification results of this model at each location under the situation of Table 3.4. The spatial performance patterns obtained under the two divergence functions are very similar even though the magnitudes of the values differ substantially between the two measures. Thus, both divergences result in the same evaluation of the ECHAM5/MPI-OM model in simulating txmax and tnmin at each location. The scores imply that tnmin is well captured by the model in most regions. The few locations evaluated as bad are mainly located in Southern or Northern Europe. The simulation of txmax deteriorates with increasing northern latitude and only functions properly in Central Europe. Sillmann and Roeckner (2008) achieve the same results even if the comparison is carried out on the basis of 53-year time series (1951-2003).

The exact score values corresponding to Figure 3.4 are listed in Table 3.5 for the grid boxes that include Heidelberg (Germany), Moscow (Russia), and Oslo (Norway). The average score over all grid points is also documented for both verification methods and both indices. The scores for Moscow are far below average, whereas the model performance for Oslo is significantly worse. Moreover, the ECHAM5/MPI-OM model is much better in simulating tnmin in Heidelberg than in many another locations, but it only shows average performance in simulating txmax. In comparison to Moscow and Oslo, Heidelberg ranks between them, but closer to Moscow. Figure 3.5 illustrates the occurrence of these score values. The Cramér-von Mises distance $d_{\text{C.v.M.}}$ depends on the difference between both underlying cdfs, and the squared mean value divergence d_{MV} is based on the difference between their means (cf. their definitions in Chapter 2.1.3, p. 15 f.). The smaller these differences, the lower the values of the measures. To validate this, the empirical cdf of the ECHAM5/MPI-OM model predictions for txmax and tnmin and the corresponding empirical cdf based on the HadEX data set are plotted for the grid boxes that include the three cities above. The mean values of the model and the HadEX data set are also illustrated for each grid box. The empirical distributions for tnmin are closer together than for txmax. The same holds for the mean values. Therefore, the scores are lower for tnmin. Since the differences for both indices are smallest by far in Moscow, there the model performs relatively well. The empirical cdfs and the means corresponding to Oslo diverge significantly from each other, particularly for txmax, and, hence, induce the large scores from Table 3.5. This could be because the overall temperature range is largest in Oslo.

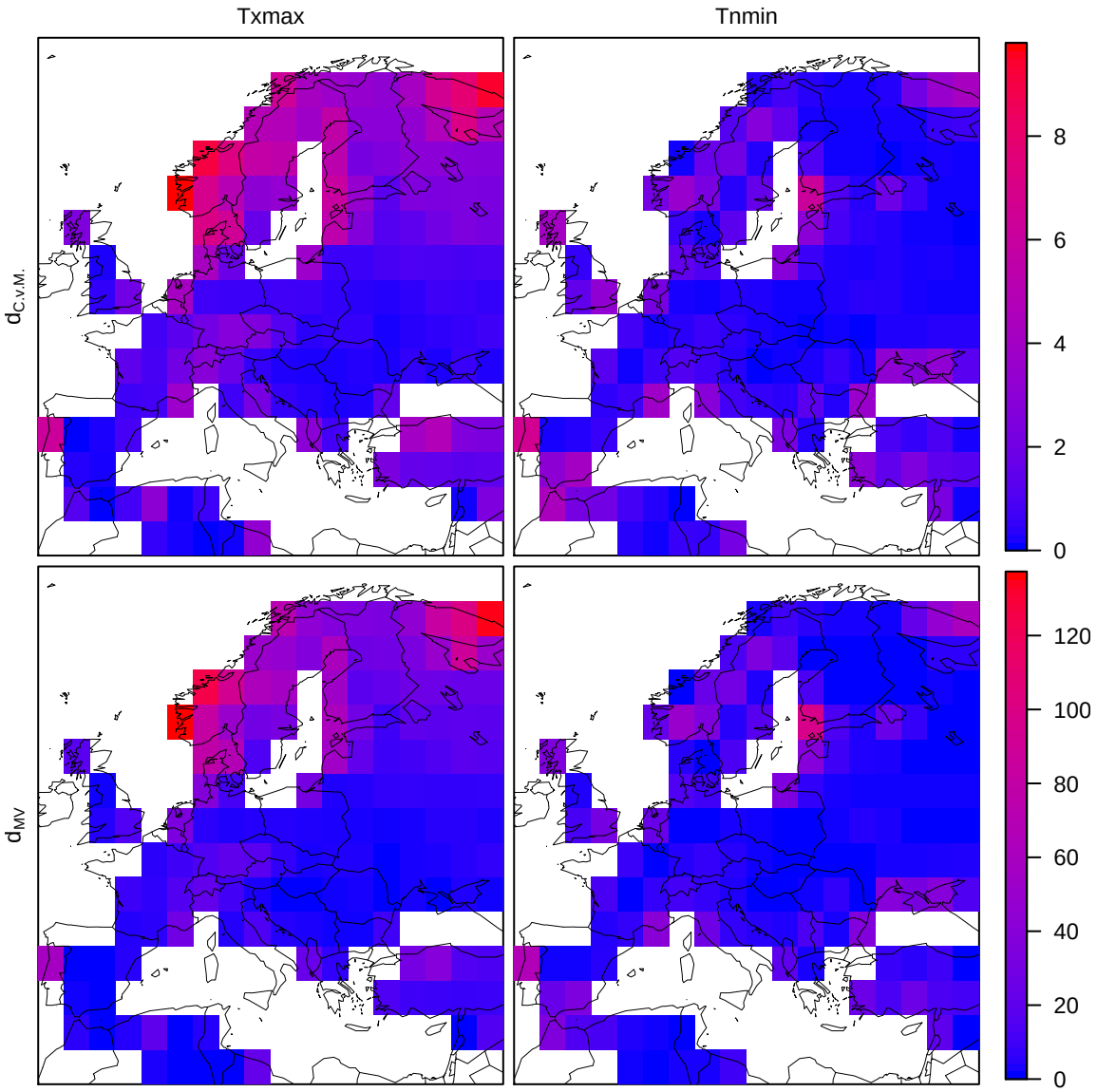

Figure 3.4: Location-specific results for the predictions of the ECHAM5/MPI-OM model for txmax (left) and tnmin (right) at land grid points over Europe when compared to the HadEX indices under the Cramér-von Mises distance $d_{\text{C.v.M.}}$ (top) and the squared mean value divergence d_{MV} (bottom).

Table 3.5: Verification results for the predictions of the ECHAM5/MPI-OM model for txmax and tnmin at the grid boxes that include Heidelberg, Moscow, and Oslo when compared to the HadEX indices under the Cramér-von Mises distance $d_{\text{C.v.M.}}$ and the squared mean value divergence d_{MV}. Furthermore, the respective mean of the scores over all grid points is stated.

	Txmax		Tnmin	
Location	$d_{\text{C.v.M.}}$	d_{MV}	$d_{\text{C.v.M.}}$	d_{MV}
Heidelberg	2.022	14.0	0.276	3.0
Moscow	0.799	5.0	0.136	1.1
Oslo	4.906	50.5	2.189	33.6
$\mathbf{mean}_{\text{loc}}$	2.227	20.8	0.963	9.9

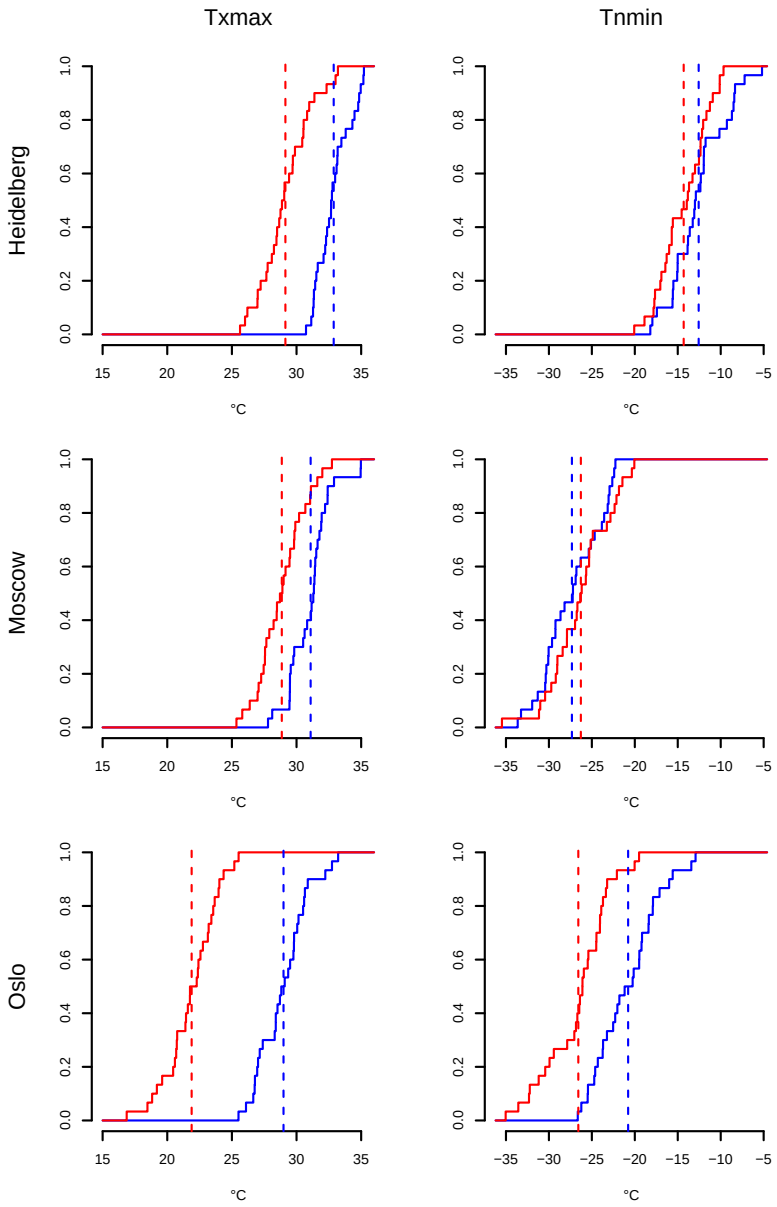

Figure 3.5: Empirical cdfs (solid lines) based on the txmax (left) and tnmin (right) predictions of the ECHAM5/MPI-OM model (red) and the corresponding HadEX values (blue) for the grid boxes that include Heidelberg, Moscow, and Oslo. The mean value of each data set is indicated with a dashed line. Note the different scales of the temperature axes for txmax and tnmin.

3.5.2 Combining model outputs

To combine the outputs of the climate models, we apply the weighting methods introduced in Chapter 2.2. The weights are usually calculated based on past data output of climate models and are then used to obtain a new cdf for future data. However, in order to evaluate the new cdf and thus to test the functionality of our approach, we divide the time period for which data is available in a training period and an out-of-sample test period.

Let F^j_{tp} and F^j_{pp} denote the empirical cdfs of extreme index values from the climate model j, $j = 1, \ldots, 15$, over the training period or the test (or prediction) period, respectively, and let G_{tp} and G_{pp} denote the corresponding empirical cdfs based on a re-analysis or the HadEX data set. The weights of the individual weighting methods are calculated from the empirical cdfs F^j_{tp} and G_{tp}. Thus, for instance, the weights$_{d_{\text{C.v.M.}}}$ method assigns model j the weight

$$w_j = \frac{\frac{1}{d_{\text{C.v.M.}}(F^j_{\text{tp}}, G_{\text{tp}})}}{\sum_{i=1}^{15} \frac{1}{d_{\text{C.v.M.}}(F^i_{\text{tp}}, G_{\text{tp}})}}.$$

The empirical cdfs from the individual climate models over the test period are then weighted to result in the new cdf

$$\sum_{j=1}^{15} w_j F^j_{\text{pp}}(x).$$

To evaluate this probabilistic forecast for the test period, it is compared against corresponding data based on the re-analyses or the observations. That is, we use the scoring rules S_{AE} and S_{SE} as well as the Cramér-von Mises distance $d_{\text{C.v.M.}}$ and assign each weighting method the corresponding average score as described in the previous section. For instance, the summary measure based on the Cramér-von Mises distance $d_{\text{C.v.M.}}$ is given by

$$\frac{1}{m} \sum_{i=1}^{m} d_{\text{C.v.M.}} \left(\sum_{j=1}^{15} w_{ij} F^j_{\text{pp,i}}, G_{\text{pp},i} \right),$$

where i is the index over the individual locations in the study region. As in the previous section, the local and the regional method are used. The index i is therefore only relevant when the local method is applied, otherwise i is equal to one.

In order to take all available data into account as test data, we perform a cross-validation study over the available data from 1961 to 1990 where we use each decade as the test period in turn. For each decade, we proceed as described above and then we average over the decadal results.

Table 3.6 provides an overview of weights obtained from the individual weighting methods. The weights are based on the mtnmin data for the years 1961 to 1980. The results for the local method are the averages of the weights over the individual grid points. The weights provide a ranking of the models over the training data. In contrast to the scores, the greater the value of the weight, the better the model. The rankings for the weights$_{d_{\text{C.v.M.}}}$ and the weights$_{d_{\text{MV}}}$ method are quite similar. Due to their definition (see Chapter 2.2, p. 19), the weights$_{d_{\text{C.v.M.}}}$ method yields the same

Table 3.6: Weights of the weighting methods $\text{weights}_{\min}$, $\text{weights}_{d_{\text{C.v.M.}}}$, and $\text{weights}_{d_{\text{MV}}}$ for mtnmin based on the comparison of all available predictions of the climate models for 1961 to 1980 against the corresponding data of the NCEP-1 re-analysis. The results for both the regional and the local method, in which the weights are averaged over the individual grid points, are shown. In each case, the largest weight is indicated in bold.

	$\text{weights}_{\min}$		$\text{weights}_{d_{\text{C.v.M.}}}$		$\text{weights}_{d_{\text{MV}}}$	
Model	Regional	Local	Regional	Local	Regional	Local
CGCM3.1(T47)	0.277	0.126	0.109	**0.109**	0.018	0.103
CGCM3.1(T63)	0.000	0.039	0.040	0.064	0.006	0.057
CNRM-CM3	0.000	0.031	0.098	0.076	0.203	0.128
CSIRO-Mk3.0	0.000	0.055	0.080	0.074	**0.320**	0.083
CSIRO-Mk3.5	0.000	0.008	0.019	0.031	0.004	0.025
ECHAM5/MPI-OM	**0.317**	**0.166**	0.065	**0.109**	0.009	0.066
FGOALS-g1.0	0.000	0.014	0.004	0.015	0.001	0.038
GFDL-CM2.0	0.098	0.075	0.020	0.046	0.003	0.024
GFDL-CM2.1	0.000	0.068	0.072	0.103	0.017	0.090
GISS-AOM	0.000	0.061	0.029	0.058	0.005	0.024
GISS-EH	0.000	0.061	0.078	0.075	0.093	0.092
GISS-ER	0.109	0.144	0.152	0.090	0.283	0.089
MIROC3.2(hires)	0.000	0.012	0.012	0.018	0.002	0.005
MIROC3.2(medres)	0.000	0.015	0.026	0.050	0.006	0.039
MRI-CGCM2.3.2	0.199	0.122	**0.196**	0.080	0.031	**0.138**

ranking of the climate models as the ranking under the Cramér-von Mises distance $d_{\text{C.v.M.}}$, and similar for the $\text{weights}_{d_{\text{MV}}}$ method. Thus, it is confirmed once again that the divergence functions $d_{\text{C.v.M.}}$ and d_{MV} yield similar model rankings for our data set. In the regional setting, the $\text{weights}_{d_{\text{C.v.M.}}}$ and the $\text{weights}_{d_{\text{MV}}}$ method assign all climate models positive weights, while the $\text{weights}_{\min}$ method does this only for a handful of models. However, when we look at the average local weights, this is no longer the case.

Table 3.7 lists the weights obtained from the local $\text{weights}_{\min}$ method at the grid boxes that include Heidelberg, Moscow, and Oslo. Here, the weights of the individual climate models differ considerably for the different grid boxes. The local results for the $\text{weights}_{\min}$ method are thus only due to averaging over the individual grid points. Similar effects can be observed for the other two weighting methods. Since the regional method ignores the locality, we expect better results from the local method in general.

We check this assumption on the basis of the yearly predictions for mtnmin when compared to the NCEP-1 re-analysis. The verification results of the weighted model combinations averaged over the decadal results are shown in Table 3.8. To test whether our weighting methods yield better predictions—in form of a cdf—than the individual climate models or the ERA-40 re-analysis, these results are also stated. The values of the average decadal scores for the individual models and the ERA-40

Table 3.7: Weights of the $\text{weights}_{\text{min}}$ method for mtnmin at the grid boxes that include Heidelberg, Moscow, and Oslo based on the comparison between all available predictions of the climate models for 1961 to 1980 in these grid boxes and the corresponding data of the NCEP-1 re-analysis. For each location, the largest weight is indicated in bold.

Model	Heidelberg	Moscow	Oslo
CGCM3.1(T47)	**0.506**	0.000	0.000
CGCM3.1(T63)	0.000	0.206	0.000
CNRM-CM3	0.000	0.000	0.000
CSIRO-Mk3.0	0.000	0.000	0.000
CSIRO-Mk3.5	0.000	0.000	0.000
ECHAM5/MPI-OM	0.263	0.000	0.037
FGOALS-g1.0	0.000	0.000	0.000
GFDL-CM2.0	0.037	0.000	0.000
GFDL-CM2.1	0.000	0.000	0.000
GISS-AOM	0.000	**0.519**	0.000
GISS-EH	0.194	0.000	0.000
GISS-ER	0.000	0.275	**0.769**
MIROC3.2(hires)	0.000	0.000	0.000
MIROC3.2(medres)	0.000	0.000	0.000
MRI-CGCM2.3.2	0.000	0.000	0.194

re-analysis are quite similar to those calculated for the entire 30-year period, see Table A.6. The results of Figure 3.1 therefore also apply here. This fact may indicate that the comparative performance of the climate models and the NCEP-1 re-analysis varies only very little across the decades of 1961 to 1990, at least on average. The weighted model combinations usually perform significantly better than any single climate model and are also competitive with the ERA-40 re-analysis. The good performance of the weighted model combinations is especially pronounced when the local method is applied for evaluation, in particular when only seasonal predictions are considered.

Table 3.9 shows the corresponding results for the summer months and the winter months only. For simplicity, we only list the individual scores of the best performing models. The $\text{weights}_{\text{min}}$ method is usually the best weighting method, and, under the Cramér-von Mises distance $d_{\text{C.v.M.}}$, this method yields an overall predictive performance which substantially outperforms all the other prediction methods. As the Cramér-von Mises distance $d_{\text{C.v.M.}}$ compares the entire cdfs, this indicates that the overall shape of the climate distribution is best captured by the $\text{weights}_{\text{min}}$ method.

In the following, since the local method also yields the more interesting results, we consider location-specific verification results corresponding to Table 3.8 and Table 3.9 under the Cramér-von Mises distance $d_{\text{C.v.M.}}$. The results of the weighted model combination based on the $\text{weights}_{\text{min}}$ method, and the ERA-40 re-analysis for the summer months, the winter months, and over the entire year are shown in Figure 3.6. In addition, the corresponding scores for the CNRM-CM3 model, the best performing model

Table 3.8: Verification results for the weighted combinations of the climate model predictions for mtnmin over the entire year averaged over decadal results when compared to the NCEP-1 re-analysis under the scoring rules S_{AE} and S_{SE} and under the Cramér-von Mises distance $d_{\mathrm{C.v.M.}}$. The results are stated for the weighting methods $\mathrm{weights}_{\mathrm{min}}$, $\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$, and $\mathrm{weights}_{d_{\mathrm{MV}}}$ as well as for the regional and the local method. Furthermore, the corresponding results for the individual models and the ERA-40 re-analysis are shown. In each case, the best climate model is indicated in bold and the mean of all climate models is stated.

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	9.73	144.9	0.000	8.34	99.7	0.000
ERA-40	10.04	161.3	0.364	8.96	120.0	0.678
$\mathrm{weights}_{\mathrm{min}}$	9.74	145.2	0.021	8.39	100.4	0.060
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	9.74	145.2	0.029	8.41	100.7	0.087
$\mathrm{weights}_{d_{\mathrm{MV}}}$	9.75	145.2	0.074	8.44	100.6	0.134
CGCM3.1(T47)	9.80	148.4	0.089	8.72	111.0	0.437
CGCM3.1(T63)	9.92	154.3	0.234	8.92	117.5	0.678
CNRM-CM3	**9.74**	146.3	0.132	8.65	**106.8**	0.450
CSIRO-Mk3.0	9.81	**145.2**	0.136	8.91	119.8	0.731
CSIRO-Mk3.5	10.21	162.9	0.536	9.23	126.1	1.081
ECHAM5/MPI-OM	9.92	152.5	0.172	8.75	111.2	0.423
FGOALS-g1.0	11.23	268.1	2.409	14.03	292.9	4.073
GFDL-CM2.0	10.07	167.1	0.486	8.97	127.7	0.873
GFDL-CM2.1	9.81	147.8	0.128	**8.55**	106.9	**0.352**
GISS-AOM	10.15	159.1	0.389	9.08	120.8	0.967
GISS-EH	9.75	145.5	0.131	8.99	119.9	0.779
GISS-ER	9.75	**145.2**	**0.062**	8.92	115.7	0.605
MIROC3.2(hires)	10.47	181.5	0.842	9.23	139.5	1.239
MIROC3.2(medres)	10.14	157.2	0.416	8.94	121.0	0.817
MRI-CGCM2.3.2	9.76	147.6	0.065	8.66	109.0	0.510
$\mathbf{mean}_{\mathrm{models}}$	10.04	161.9	0.415	9.24	129.7	0.934

in summer (see Table 3.9(a)), and for the GFDL-CM2.1 model, the best performing model in winter (see Table 3.9(b)), are illustrated. The GFDL-CM2.1 model also has the best model performance over the entire year averaged over the individual grid points under the Cramér-von Mises distance $d_{\mathrm{C.v.M.}}$ (see Table 3.8). The weighted model combinations based on the $\mathrm{weights}_{\mathrm{min}}$ method usually outperform the other methods at every location. Individual climate models show very different location-specific predictive performances. For instance, the GFDL-CM2.1 model has problems in simulating mtnmin for the summer months in Central Europe, while the performance of the CNRM-CM3 model is worse in the Mediterranean region. It is therefore not clear which of these models one should prefer. To this end, model combination seems to be a good approach and, in fact, our weighting methods improve the results significantly.

Table 3.9: Verification results for the weighted combinations of the climate model predictions for mtnmin in (a) the summer months and (b) the winter months averaged over decadal results when compared to the NCEP-1 re-analysis under the scoring rules S_{AE} and S_{SE} and under the Cramér-von Mises distance $d_{\mathrm{C.v.M.}}$. The results are stated for the weighting methods $\mathrm{weights}_{\mathrm{min}}$, $\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$, and $\mathrm{weights}_{d_{\mathrm{MV}}}$ as well as for the regional and the local method. Furthermore, the corresponding results for the best performing models and the ERA-40 re-analysis are shown. In each case, the best climate model is indicated in bold and the mean of all 15 climate models is stated.

(a) Summer

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	3.55	21.0	0.000	1.71	4.8	0.000
ERA-40	4.11	29.2	0.418	3.41	16.3	1.321
$\mathrm{weights}_{\mathrm{min}}$	3.57	21.2	0.033	1.83	5.4	0.105
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	3.56	21.5	0.059	1.85	5.6	0.130
$\mathrm{weights}_{d_{\mathrm{MV}}}$	3.61	21.3	0.077	1.84	5.5	0.139
CNRM-CM3	3.56	22.0	0.074	2.66	**10.5**	**0.685**
CSIRO-Mk3.0	**3.55**	21.3	**0.038**	**2.51**	10.7	0.709
MRI-CGCM2.3.2	3.65	**21.1**	0.100	2.96	12.9	0.948
$\mathbf{mean}_{\mathrm{models}}$	4.25	28.6	0.519	3.64	19.8	1.541

(b) Winter

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	9.13	112.9	0.000	2.94	15.2	0.000
ERA-40	9.98	135.9	0.678	5.59	46.1	2.086
$\mathrm{weights}_{\mathrm{min}}$	9.21	113.4	0.051	3.24	17.6	0.250
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	9.16	113.3	0.101	3.26	17.7	0.280
$\mathrm{weights}_{d_{\mathrm{MV}}}$	9.38	113.9	0.196	3.32	18.2	0.324
CNRM-CM3	10.45	**114.5**	0.415	5.72	46.8	2.181
GFDL-CM2.1	9.19	**114.5**	**0.111**	**4.44**	**30.4**	**1.120**
GISS-EH	**9.18**	115.2	0.201	5.66	49.1	2.208
GISS-ER	**9.18**	115.5	0.171	5.39	46.0	2.078
$\mathbf{mean}_{\mathrm{models}}$	11.17	155.9	1.191	7.10	91.6	3.489

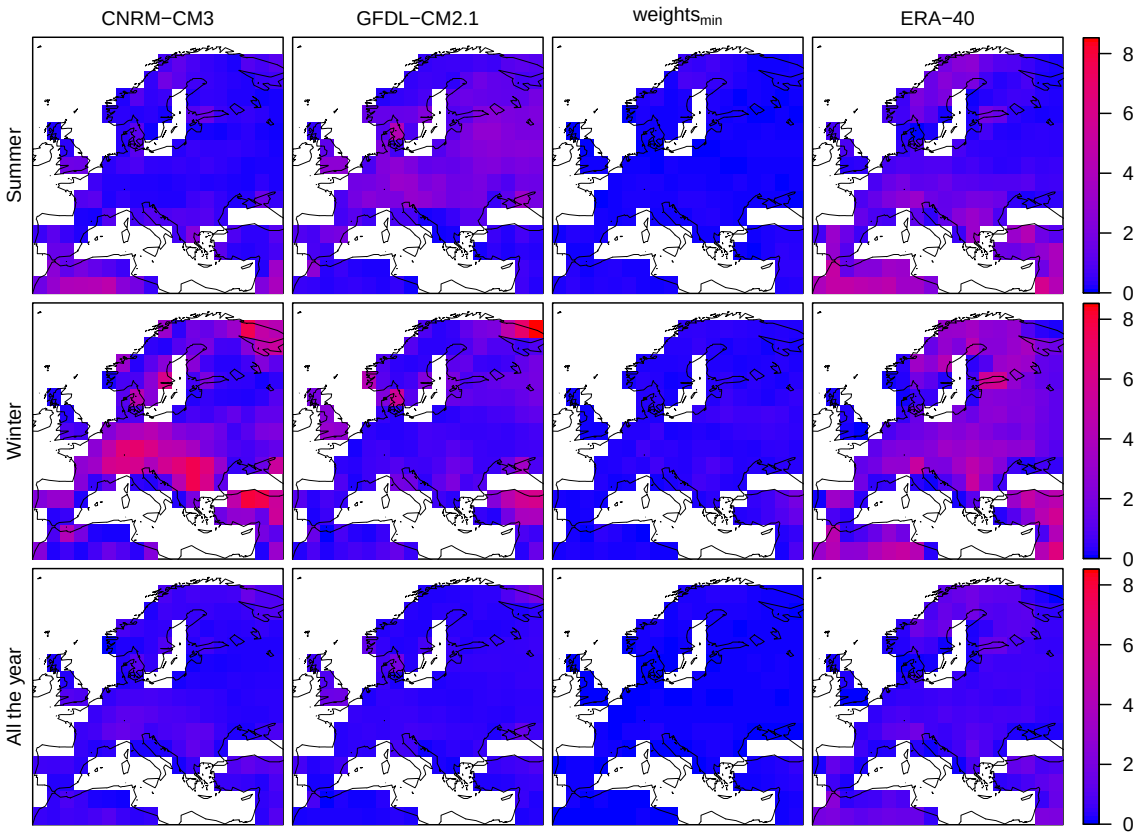

Figure 3.6: Location-specific results for the predictions of the CNRM-CM3 model, the GFDL-CM2.1 model, the weights$_{\text{min}}$ method, and the ERA-40 re-analysis for mtnmin in the summer months, the winter months, and over the entire year at land grid points over Europe averaged over the decadal results when compared to the NCEP-1 re-analysis under the Cramér-von Mises distance $d_{\text{C.v.M.}}$.

4 Discussion

We propose evaluation methods for the performance of climate models that directly assess the divergence of the predictive climate distribution from the corresponding observations and apply those to descriptive indices for climate extremes. To determine the performance of 15 global climate models with respect to their ability to simulate climate extremes, four indices, the monthly maximum value of daily maximum temperature (mtxmax), the monthly minimum value of daily minimum temperature (mtnmin), the yearly maximum value of daily maximum temperature (txmax), and the yearly minimum value of daily minimum temperature (tnmin), over Europe from 1961 to 1990 were compared with corresponding re-analysis and/or observation based data sets.

The quality of the climate models in simulating extreme indices depends on the index, the season, and the region under consideration. Different seasons and regions can only, of course, be considered for the monthly indices or the local method, respectively. The models are usually better in simulating the maximum temperature based indices than the minimum temperature based indices, and they are least skillful in the seasons in which the indices are more extreme. For instance, the model performances in simulating mtxmax are worst in the summer months. That is, the models seem to have problems with simulating particularly low or high temperatures. Individual climate models show very different location-specific predictive performances, and large divergences generally appear at grid points with a permanent extreme weather situation, such as high wind, heat, or cold, including, for instance, coastal regions, the Mediterranean region, or Scandinavia. When comparing to a re-analysis data set, the second re-analysis data set performs as well as the best climate models. However, it is only rarely better than all models, contrary to expectation, since both re-analyses represent the same observations. The same holds when both re-analyses and climate model outputs are compared to the observation based indices. Furthermore, the disparity in the verification rankings of the climate models when compared to the re-analysis on one hand and the observation based indices on the other hand may indicate that the re-analyses are not able to capture the minimum temperature based indices reliably. To fully capture the implications of our results, a more detailed analysis is needed taking the physical and numerical assumptions of each model into account.

A similar comparison has e.g. been attempted by Kiktev et al. (2003) and Sillmann and Roeckner (2008). Sillmann and Roeckner (2008) compare txmax and tnmin values of the ECHAM5/MPI-OM model for 1951 to 2003 with the corresponding HadEX values by displaying global maps of time averages and calculating time series for three European regions. Their findings are consistent with our results (cf. Chapter 3.5.1, p. 33). However, it is difficult to get an overall view of the model performance by visually comparing two maps or two functions, respectively. Since the squared mean value divergence d_{MV} measures the difference between the prediction mean and the observed mean, its application yields a summary score for the model performance with respect to the ability to simulate the mean of the indices. This is exactly what Sillmann and Roeckner (2008) examine by comparing the maps, and, due to the current definition of climate (cf. Chapter 1), this may also be important. However, to take the local performance into account, the local method should be applied.

Kiktev et al. (2003) compare model and observation based trend estimates of the indices as well as model based and gridded observed indices. For the latter, the authors use a Kolmogorov-Smirnov test. One disadvantage of this approach is the low sensitivity at the tails of the distributions. The test statistic is also not n-proper for all $n \in \mathbb{N}$, only asymptotically proper (Thorarinsdottir et al., 2012). Therefore, it is not appropriate to rank the models this way, and the Cramér-von Mises distance $d_{\mathrm{C.v.M.}}$ would be more appropriate. Furthermore, to objectively compare the similarity between the patterns of observed and model trends, the authors estimate the probability distribution function of measures of pattern similarity, including the centered pattern correlation and the uncentered pattern correlation, by a bootstrap technique. Different ensembles are then ranked based on these correlations. However, in our experience, rankings based on correlations differ substantially from those of the other scoring methods, and there is no obvious reason why the similarity is really measured.

In climate research in general, there seems to be a lack of accepted standard measures of climate model performance. Gleckler et al. (2008) use a version of the root mean squared error to assess CMIP3 20th century simulations without taking the full predictive distribution into account. Salazar et al. (2011) use a Bayesian approach to obtain full predictive distributions and compare these to observations using proper scoring rules such as the CRPS. This comparison will yield the same model rankings as the Cramér-von Mises distance $d_{\mathrm{C.v.M.}}$. However, an advantage of the divergence functions is that the resulting divergences are equal to zero if the two distributions are equal which does not necessarily hold for proper scoring rules. This simplifies comparability of different verification methods. Perkins et al. (2007) adopt the natural approach to compare climate model density distributions and the

corresponding observation based density distributions (cf. Chapter 1). A skill score that measures the similarity between two probability density functions is proposed and used for the evaluation of climate model simulations over Australia from 1961 to 2000. The metric calculates the cumulative minimum value of two distributions of each binned value and sum up the values over the number of bins used to calculate the probability density functions. It is not clear whether this comparison procedure yields a proper divergence function or not.

In addition to model verification, we propose three weighting methods based on the divergence functions $d_{\text{C.v.M.}}$ and d_{MV} to combine outputs from climate models. The weighted model combinations usually perform significantly better than any single climate model. The $\text{weights}_{\text{min}}$ method is usually the best weighting method, and, under the Cramér-von Mises distance $d_{\text{C.v.M.}}$, it substantially outperforms all the other prediction and weighting methods. A common approach is to combine the models within a Bayesian framework. However, this methodology assumes a certain parametric model for the model output which is often taken as the normal distribution for temperature, see e.g. Kallache et al. (2010).

Bibliography

Alexander, L. V., X. Zhang, T. C. Peterson, J. Caesar, B. Gleason, A. M. G. K. Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K. R. Kumar, J. Revadekar, G. Griffiths, L. Vincent, D. B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, and J. L. Vazquez-Aguirre (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research - Atmospheres 111*, D05109.

Bauer, H. (1992). *Maß- und Integrationstheorie* (2 ed.). Berlin: Walter de Gruyter.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A 147*, 278–292.

Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Department of Statistical Science, University College London.

ECMWF (2012). From weather to climate. `http://www.ecmwf.int/about/corporate_brochure/leaflets/Weather-climate-English.pdf`. Last accessed July 18, 2012.

Folland, C., C. Miller, D. Bader, M. Crowe, P. Jones, N. Plummer, M. Richman, D. Parker, J. Rogers, and P. Scholefield (1999). Workshop on indices and indicators for climate extremes, Asheville, NC, USA, 3–6 June 1997. Breakout group C: Temperature indices for climate extremes. *Climatic Change 42*, 31–43.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008). Performance metrics for climate models. *Journal of Geophysical Research 113*, D06104.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Series A 171*, 319–321.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*, 746–762.

Gneiting, T. and A. E. Raftery (2005). Strictly proper scoring rules, prediction, and estimation. Technical Report 463R, Department of Statistics, University of Washington.

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

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics 29*, 411–422.

Gneiting, T. and T. L. Thorarinsdottir (2010). Predicting inflation: Professional experts versus no-change forecasts. arXiv:1010.2318v1.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*, 559–570.

IPCC (2007). *Climate change 2007: The physical science basis. Contribution of working group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Solomon, S. et al. (eds.), Cambridge University Press.

Kallache, M., E. Maksimovich, P. Michelangeli, and P. Naveau (2010). Multimodel combination by a Bayesian hierarchical model: Assessment of ice accumulation over the oceanic Artic region. *Journal of Climate 23*, 5421–5436.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society 77*, 437–471.

Karl, T. R., N. Nicholls, and A. Ghazi (1999). CLIVAR/GCOS/WMO Workshop on indices and indicators for climate extremes. Workshop summary. *Climatic Change 42*, 3–7.

Kiktev, D., D. Sexton, L. Alexander, and C. Folland (2003). Comparison of modeled and observed trends in indices of daily climate extremes. *Journal of Climate 16*, 3560–3571.

Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, H. van den Dool, R. Jenne, and M. Fiorino (2001). The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society 82*, 247–268.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010). Challenges in combining projections from multiple climate models. *Journal of Climate 23*, 2739–2758.

Laio, F. and S. Tamea (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences 11*, 1267–1277.

Luterbacher, J., D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner (2004). European seasonal and annual temperature variability, trends, and extremes since 1500. *Science 303*, 1499–1503.

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*, 1087–1096.

Meehl, G., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. Mitchell, R. Stouffer, and K. Taylor (2007). The WCRP CMIP3 multi-model dataset: a new era in climate change research. *Bulletin of the American Meteorological Society 88*, 1383–1394.

Nicholls, N. and W. Murray (1999). Workshop on indices and indicators for climate extremes, Asheville, NC, USA, 3–6 June 1997. Breakout group B: Precipitation. *Climatic Change 42*, 23–29.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneny (2007). Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate 20*, 4356–4376.

Peterson, T. C. (2005). Climate change indices. *World Meteorological Organization Bulletin 54*, 83–86.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robine, J.-M., S. L. K. Cheung, S. L. Roy, H. V. Oyen, C. Griffiths, J.-P. Michel, and F. R. Herrmann (2008). Death toll exceeded 70,000 in Europe during the summer of 2003. *Comptes Rendus Biologies 331*, 171 – 178.

Roedel, W. and T. Wagner (2011). *Physik unserer Umwelt: Die Atmosphäre* (4 ed.). Heidelberg: Springer.

Salazar, E., B. Sansó, A. O. Finley, D. Hammerling, I. Steinsland, X. Wang, and P. Delamater (2011). Comparing and blending regional climate model predictions for the American Southwest. *Journal of Agricultural, Biological, and Environmental Statistics 16*, 586–605.

Schär, C. and G. Jendritzky (2004). Hot news from summer 2003. *Nature 432*, 559–560.

Sillmann, J. and E. Roeckner (2008). Indices for extreme events in projections of anthropogenic climate change. *Climatic Change 86*, 83–104.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl (2012). Proper divergence functions. *In preparation*.

Trenberth, K. E. and D. J. Shea (2006). Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters 33*, L12704.

Ulbrich, U., T. Brücher, A. H. Fink, G. C. Leckebusch, A. Krüger, and J. G. Pinto (2003). The Central European floods of August 2002: Part 2 -Synoptic causes and considerations with respect to climatic change. *Weather 58*, 434–442.

Uppala, S. M., P. W. Kållberg, A. J. Simmons, U. Andrae, V. D. C. Bechtold, M. Fiorino, J. K. Gibson, J. Haseler, A. Hernandez, G. A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. V. D. Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B. J. Hoskins, L. Isaksen, P. A. E. M. Janssen, R. Jenne, A. P. Mcnally, J.-F. Mahfouf, J.-J. Morcrette, N. A. Rayner, R. W. Saunders, P. Simon, A. Sterl, K. E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, and J. Woollen (2005). The ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society 131*, 2961–3012.

Vanderbei, R. J. (1999). LOQO: An interior point code for quadratic programming. *Optimization Methods and Software 11*, 451–484.

Von Storch, H., S. Güss, and M. Heimann (1999). *Das Klimasystem und seine Modellierung: Eine Einführung*. Berlin: Springer.

A Verification tables

In our case study, only a small part of the verification results could be investigated in detail. For completeness and comparison, the full set of verification results averaged over the grid boxes for the climate model predictions of mtxmax and mtnmin when compared to the NCEP-1 re-analysis and for the model predictions of txmax and tnmin when compared to the HadEX indices are presented in the following tables. Moreover, the corresponding scores for the re-analyses and, when considering the annual indices, also for the HadEX data are stated. In addition to the results for the predictions over the entire year, the results for the predictions of the monthly indices in the summer months (June–August) and in the winter months (December–February) are shown. As explained in Chapter 3.5, the comparisons were performed both regionally and locally. These results can also be found in the tables. In each case, the best climate model is indicated in bold and the mean over the average scores of the individual models is given.

A.1 Individual climate models

In this section, the verification results for the individual climate models when simulating the four extreme indices under the scoring functions MAE and MSE, under the proper scoring rules CRPS, S_{AE}, and S_{SE}, and under the divergence functions $d_{\mathrm{C.v.M.}}$ and d_{MV} are shown. All available data in our dataset for land grid boxes in the European domain from 1961 to 1990 were considered for calculation. Note that the sets of grid points for the monthly and annual indices over which is averaged differ slighty (cf. Chapter 3.5.1, p. 32).

Table A.1: Verification results for the climate predictions for **mtxmax** in the **summer** months.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	4.18	5.73	56.4	0.000	0.0
ERA-40	2.53	11.8	4.32	5.86	59.3	0.140	2.9
CGCM3.1(T47)	4.83	36.8	4.42	**5.73**	56.9	0.243	0.5
CGCM3.1(T63)	4.34	29.3	4.34	5.74	56.6	0.157	0.2
CNRM-CM3	4.10	25.6	4.43	6.34	60.5	0.254	4.0
CSIRO-Mk3.0	5.15	38.9	4.96	6.95	78.5	0.778	22.1
CSIRO-Mk3.5	3.72	22.4	4.30	5.97	57.6	0.125	1.1
ECHAM5/MPI-OM	3.70	22.0	4.35	5.94	60.9	0.166	4.5
FGOALS-g1.0	5.82	48.8	4.87	7.47	61.8	0.690	5.4
GFDL-CM2.0	5.04	37.6	4.72	6.78	70.7	0.544	14.2
GFDL-CM2.1	4.20	26.8	4.38	6.20	60.4	0.204	3.9
GISS-AOM	7.33	69.6	6.39	9.24	107.3	2.213	50.9
GISS-EH	4.81	37.6	4.37	5.74	**56.5**	0.190	**0.1**
GISS-ER	5.50	46.2	4.45	5.79	61.3	0.269	4.8
MIROC3.2(hires)	3.56	19.9	**4.24**	5.79	57.3	**0.065**	0.9
MIROC3.2(medres)	4.44	32.7	4.30	5.83	57.1	0.119	0.6
MRI-CGCM2.3.2	**3.45**	**19.0**	4.31	5.94	59.0	0.128	2.5
mean$_{\text{models}}$	4.67	34.2	4.59	6.36	64.2	0.410	7.7

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	1.47	2.07	7.2	0.000	0.0
ERA-40	2.53	11.8	2.32	3.00	15.8	0.855	8.7
CGCM3.1(T47)	4.83	36.8	3.31	4.45	29.5	1.842	22.3
CGCM3.1(T63)	4.34	29.3	2.80	3.81	21.8	1.329	14.7
CNRM-CM3	4.10	25.6	2.80	3.73	20.5	1.331	13.4
CSIRO-Mk3.0	5.15	38.9	3.97	4.85	36.2	2.504	29.1
CSIRO-Mk3.5	3.72	22.4	**2.35**	3.21	16.6	**0.883**	9.4
ECHAM5/MPI-OM	3.70	22.0	2.47	3.32	17.5	1.001	10.3
FGOALS-g1.0	5.82	48.8	4.36	5.61	43.2	2.892	36.1
GFDL-CM2.0	5.04	37.6	3.29	4.50	27.7	1.820	20.5
GFDL-CM2.1	4.20	26.8	2.46	3.43	17.1	0.994	9.9
GISS-AOM	7.33	69.6	6.33	7.23	67.9	4.861	60.8
GISS-EH	4.81	37.6	3.28	4.43	29.7	1.813	22.6
GISS-ER	5.50	46.2	3.94	5.15	38.5	2.474	31.3
MIROC3.2(hires)	3.56	19.9	2.47	3.30	16.9	1.005	9.7
MIROC3.2(medres)	4.44	32.7	2.73	3.72	23.5	1.259	16.3
MRI-CGCM2.3.2	**3.45**	**19.0**	2.36	**3.10**	**16.0**	0.888	**8.8**
mean$_{\text{models}}$	4.67	34.2	3.26	4.26	28.2	1.793	21.0

Table A.2: Verification results for the climate predictions for **mtxmax** in the **winter** months.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	4.15	6.07	56.1	0.000	0.0
ERA-40	1.57	4.2	4.21	6.08	56.2	0.060	0.1
CGCM3.1(T47)	2.90	13.5	4.22	6.08	56.3	0.074	0.2
CGCM3.1(T63)	2.86	13.6	4.25	6.11	56.9	0.099	0.8
CNRM-CM3	3.95	29.0	4.44	6.45	66.3	0.292	10.1
CSIRO-Mk3.0	2.91	16.8	**4.19**	6.09	57.4	**0.044**	1.3
CSIRO-Mk3.5	3.82	22.8	4.58	6.79	64.7	0.433	8.6
ECHAM5/MPI-OM	2.86	14.3	4.22	6.16	56.3	0.071	0.2
FGOALS-g1.0	6.29	83.6	4.83	6.19	73.6	0.679	17.5
GFDL-CM2.0	3.08	16.9	4.20	6.13	56.2	0.049	0.1
GFDL-CM2.1	3.13	16.7	4.30	6.44	57.2	0.149	1.1
GISS-AOM	3.08	17.5	4.26	6.12	56.3	0.111	0.2
GISS-EH	3.41	21.8	4.26	6.12	58.4	0.114	2.3
GISS-ER	3.38	20.5	4.28	**6.07**	58.8	0.131	2.7
MIROC3.2(hires)	**2.43**	**10.3**	4.21	6.11	56.3	0.058	0.2
MIROC3.2(medres)	2.97	16.8	4.20	6.14	**56.1**	0.056	**0.0**
MRI-CGCM2.3.2	4.05	27.7	4.56	6.26	64.9	0.415	8.8
$\textbf{mean}_{\text{models}}$	3.41	22.8	4.33	6.22	59.7	0.185	3.6

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	1.21	1.69	5.4	0.000	0.0
ERA-40	1.57	4.2	1.52	2.12	7.7	0.307	2.3
CGCM3.1(T47)	2.90	13.5	1.62	2.24	7.9	0.414	2.5
CGCM3.1(T63)	2.86	13.6	1.67	2.31	8.5	0.461	3.1
CNRM-CM3	3.95	29.0	2.66	3.65	22.2	1.449	16.8
CSIRO-Mk3.0	2.91	16.8	1.72	2.36	10.5	0.508	5.1
CSIRO-Mk3.5	3.82	22.8	2.63	3.55	17.9	1.423	12.5
ECHAM5/MPI-OM	2.86	14.3	1.63	2.24	9.2	0.414	3.8
FGOALS-g1.0	6.29	83.6	3.99	5.49	61.2	2.775	55.8
GFDL-CM2.0	3.08	16.9	1.63	2.25	9.0	0.416	3.6
GFDL-CM2.1	3.13	16.7	1.93	2.66	11.6	0.716	6.2
GISS-AOM	3.08	17.5	1.97	2.63	12.4	0.754	7.0
GISS-EH	3.41	21.8	2.15	2.92	15.9	0.938	10.5
GISS-ER	3.38	20.5	2.15	2.93	15.0	0.936	9.6
MIROC3.2(hires)	**2.43**	**10.3**	**1.53**	**2.12**	**7.6**	**0.316**	**2.2**
MIROC3.2(medres)	2.97	16.8	1.70	2.35	10.3	0.487	4.9
MRI-CGCM2.3.2	4.05	27.7	2.76	3.68	22.6	1.549	17.2
$\textbf{mean}_{\text{models}}$	3.41	22.8	2.11	2.89	16.1	0.904	10.7

Table A.3: Verification results for the climate predictions for **mtxmax** over the **entire year**.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	6.71	9.77	137.6	0.000	0.0
ERA-40	2.00	7.6	6.75	9.77	137.9	0.042	0.3
CGCM3.1(T47)	3.84	24.1	6.80	9.95	137.8	0.091	0.2
CGCM3.1(T63)	3.72	22.5	6.81	9.98	138.2	0.098	0.6
CNRM-CM3	4.13	27.8	6.91	10.07	145.4	0.200	7.9
CSIRO-Mk3.0	4.18	28.2	6.99	10.15	148.3	0.280	10.7
CSIRO-Mk3.5	3.49	19.6	6.79	9.80	140.2	0.077	2.7
ECHAM5/MPI-OM	3.18	16.8	6.75	9.81	138.1	0.039	0.5
FGOALS-g1.0	5.49	52.5	6.87	9.78	140.0	0.158	2.4
GFDL-CM2.0	4.20	28.7	6.93	10.09	145.1	0.219	7.5
GFDL-CM2.1	3.62	21.3	6.78	9.86	138.6	0.072	1.1
GISS-AOM	4.85	37.6	7.15	10.13	151.9	0.438	14.3
GISS-EH	4.00	27.1	6.79	9.90	139.4	0.077	1.8
GISS-ER	4.39	31.1	6.89	10.16	143.9	0.177	6.3
MIROC3.2(hires)	**3.09**	**15.9**	**6.71**	**9.77**	**137.6**	**0.006**	**0.0**
MIROC3.2(medres)	3.79	24.7	6.74	9.80	**137.6**	0.030	**0.0**
MRI-CGCM2.3.2	3.71	22.3	6.79	9.80	141.0	0.077	3.4
$\mathbf{mean}_{\mathrm{models}}$	3.98	26.7	6.84	9.94	141.5	0.136	4.0

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	5.21	8.03	86.0	0.000	0.0
ERA-40	2.00	7.6	5.39	8.13	89.6	0.182	3.6
CGCM3.1(T47)	3.84	24.1	5.53	8.31	90.7	0.323	4.7
CGCM3.1(T63)	3.72	22.5	5.48	8.29	89.8	0.270	3.8
CNRM-CM3	4.13	27.8	5.66	8.46	97.6	0.452	11.6
CSIRO-Mk3.0	4.18	28.2	5.77	8.50	99.3	0.562	13.3
CSIRO-Mk3.5	3.49	19.6	5.47	8.17	91.7	0.265	5.7
ECHAM5/MPI-OM	3.18	16.8	5.38	8.11	89.1	0.176	3.0
FGOALS-g1.0	5.49	52.5	6.00	8.52	103.2	0.794	17.2
GFDL-CM2.0	4.20	28.7	5.65	8.45	95.9	0.441	9.8
GFDL-CM2.1	3.62	21.3	5.45	8.22	89.6	0.239	3.5
GISS-AOM	4.85	37.6	6.10	8.47	104.0	0.897	18.0
GISS-EH	4.00	27.1	5.61	8.31	95.4	0.405	9.4
GISS-ER	4.39	31.1	5.75	8.50	99.1	0.545	13.1
MIROC3.2(hires)	**3.09**	**15.9**	**5.36**	**8.10**	**88.9**	**0.156**	**2.9**
MIROC3.2(medres)	3.79	24.7	5.50	8.33	92.7	0.291	6.6
MRI-CGCM2.3.2	3.71	22.3	5.50	8.16	92.3	0.295	6.3
$\mathbf{mean}_{\mathrm{models}}$	3.98	26.7	5.61	8.33	94.6	0.407	8.6

Table A.4: Verification results for the climate predictions for **mtnmin** in the **summer** months.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	2.56	3.55	21.0	0.000	0.0
ERA-40	2.92	12.9	2.98	4.12	29.2	0.417	8.2
CGCM3.1(T47)	3.00	13.9	2.86	4.13	25.0	0.303	3.9
CGCM3.1(T63)	3.85	21.7	3.24	4.62	31.4	0.682	10.4
CNRM-CM3	2.77	12.4	2.62	**3.55**	21.8	0.063	0.8
CSIRO-Mk3.0	2.65	11.9	**2.59**	**3.55**	21.1	**0.030**	0.1
CSIRO-Mk3.5	4.03	24.0	3.18	4.30	32.7	0.619	11.7
ECHAM5/MPI-OM	2.73	12.3	2.75	3.78	25.1	0.194	4.1
FGOALS-g1.0	4.83	35.7	3.08	3.71	22.2	0.517	1.2
GFDL-CM2.0	4.65	29.3	3.85	5.97	40.0	1.290	19.0
GFDL-CM2.1	3.38	17.1	3.08	4.63	26.7	0.518	5.7
GISS-AOM	**2.54**	**10.4**	2.66	3.67	23.0	0.097	2.0
GISS-EH	4.79	37.8	3.19	4.01	32.5	0.626	11.5
GISS-ER	3.65	21.4	2.82	3.70	23.1	0.260	2.1
MIROC3.2(hires)	5.14	31.8	4.15	5.73	47.2	1.586	26.2
MIROC3.2(medres)	4.04	23.2	3.38	4.75	35.0	0.822	14.0
MRI-CGCM2.3.2	2.91	13.4	2.66	3.64	**21.0**	0.095	**0.0**
$\textbf{mean}_{\text{models}}$	3.66	21.1	3.07	4.25	28.5	0.513	7.5

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	1.22	1.75	5.0	0.000	0.0
ERA-40	2.92	12.9	2.50	3.40	16.3	1.272	11.3
CGCM3.1(T47)	3.00	13.9	2.13	2.90	12.9	0.910	7.9
CGCM3.1(T63)	3.85	21.7	2.73	3.66	19.2	1.510	14.2
CNRM-CM3	2.77	12.4	**1.83**	2.62	**10.2**	**0.603**	**5.2**
CSIRO-Mk3.0	2.65	11.9	1.87	**2.49**	10.6	0.645	5.5
CSIRO-Mk3.5	4.03	24.0	3.02	4.02	22.6	1.796	17.6
ECHAM5/MPI-OM	2.73	12.3	2.01	2.72	11.9	0.784	6.9
FGOALS-g1.0	4.83	35.7	3.43	4.60	31.4	2.206	26.4
GFDL-CM2.0	4.65	29.3	3.60	4.65	27.7	2.376	22.7
GFDL-CM2.1	3.38	17.1	2.41	3.26	15.1	1.185	10.1
GISS-AOM	**2.54**	**10.4**	1.91	2.61	10.6	0.687	5.5
GISS-EH	4.79	37.8	3.71	4.80	36.2	2.487	31.1
GISS-ER	3.65	21.4	2.66	3.61	19.6	1.436	14.6
MIROC3.2(hires)	5.14	31.8	4.11	5.31	32.2	2.888	27.1
MIROC3.2(medres)	4.04	23.2	2.99	4.06	22.1	1.765	17.1
MRI-CGCM2.3.2	2.91	13.4	2.11	2.94	12.8	0.886	7.8
$\textbf{mean}_{\text{models}}$	3.66	21.1	2.70	3.62	19.7	1.477	14.7

Table A.5: Verification results for the climate predictions for **mtnmin** in the **winter** months.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	6.13	9.13	113.1	0.000	0.0
ERA-40	5.08	37.2	6.79	9.98	136.0	0.665	22.9
CGCM3.1(T47)	6.03	62.1	6.23	9.16	115.7	0.103	2.6
CGCM3.1(T63)	6.72	74.0	6.35	9.24	121.0	0.223	7.9
CNRM-CM3	6.55	70.0	6.52	10.46	**114.0**	0.396	**0.9**
CSIRO-Mk3.0	8.10	107.2	6.88	9.35	115.6	0.747	2.6
CSIRO-Mk3.5	8.38	101.0	7.25	11.54	138.2	1.124	25.2
ECHAM5/MPI-OM	**5.40**	**45.9**	6.38	9.49	121.8	0.256	8.7
FGOALS-g1.0	21.93	633.9	15.78	28.01	550.5	9.651	437.4
GFDL-CM2.0	7.29	84.7	6.90	9.72	140.8	0.771	27.7
GFDL-CM2.1	5.89	57.0	**6.22**	**9.15**	114.2	**0.094**	1.1
GISS-AOM	7.49	81.0	7.02	10.71	142.1	0.890	29.0
GISS-EH	6.19	62.3	6.31	9.16	114.8	0.186	1.7
GISS-ER	6.13	58.8	6.29	9.17	115.4	0.163	2.3
MIROC3.2(hires)	8.35	100.6	8.03	11.74	177.8	1.909	64.7
MIROC3.2(medres)	7.12	78.3	6.90	10.84	125.9	0.771	12.9
MRI-CGCM2.3.2	5.68	49.4	6.52	9.58	128.3	0.391	15.3
$\mathbf{mean}_{\text{models}}$	7.82	111.1	7.31	11.15	155.7	1.178	42.7

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	2.16	3.07	16.1	0.000	0.0
ERA-40	5.08	37.2	4.12	5.61	46.5	1.965	30.4
CGCM3.1(T47)	6.03	62.1	3.27	4.48	33.4	1.108	17.3
CGCM3.1(T63)	6.72	74.0	3.85	5.31	44.1	1.690	28.0
CNRM-CM3	6.55	70.0	4.18	5.73	46.2	2.024	30.1
CSIRO-Mk3.0	8.10	107.2	5.52	7.56	83.0	3.357	66.8
CSIRO-Mk3.5	8.38	101.0	5.81	7.48	75.0	3.656	58.9
ECHAM5/MPI-OM	**5.40**	**45.9**	3.45	4.73	33.9	1.287	17.7
FGOALS-g1.0	21.93	633.9	19.00	21.85	604.9	16.844	588.8
GFDL-CM2.0	7.29	84.7	4.62	6.44	61.5	2.456	45.3
GFDL-CM2.1	5.89	57.0	**3.13**	**4.36**	**30.0**	**0.971**	**13.9**
GISS-AOM	7.49	81.0	5.51	7.01	67.8	3.353	51.6
GISS-EH	6.19	62.3	4.19	5.59	48.1	2.034	32.0
GISS-ER	6.13	58.8	4.10	5.40	46.0	1.944	29.8
MIROC3.2(hires)	8.35	100.6	6.65	8.18	91.4	4.488	75.2
MIROC3.2(medres)	7.12	78.3	5.01	6.41	60.9	2.854	44.7
MRI-CGCM2.3.2	5.68	49.4	4.12	5.32	41.7	1.963	25.6
$\mathbf{mean}_{\text{models}}$	7.82	111.1	5.49	7.06	91.2	3.335	75.1

Table A.6: Verification results for the climate predictions for **mtnmin** over the **entire year**.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	6.78	9.73	145.0	0.000	0.0
ERA-40	4.20	27.1	7.14	10.04	161.3	0.360	16.4
CGCM3.1(T47)	4.82	43.3	6.86	9.79	148.0	0.081	3.1
CGCM3.1(T63)	5.52	52.8	7.01	9.92	154.3	0.230	9.4
CNRM-CM3	4.50	37.9	6.90	**9.74**	145.9	0.122	1.0
CSIRO-Mk3.0	5.19	54.8	6.91	9.81	**145.1**	0.131	**0.1**
CSIRO-Mk3.5	5.94	56.9	7.31	10.21	162.9	0.532	18.0
ECHAM5/MPI-OM	**4.23**	30.2	6.94	9.92	152.2	0.163	7.3
FGOALS-g1.0	13.03	302.2	9.18	11.23	268.1	2.405	123.2
GFDL-CM2.0	5.89	56.1	7.26	10.07	167.0	0.481	22.0
GFDL-CM2.1	4.49	34.5	6.90	9.81	147.7	0.121	2.7
GISS-AOM	5.10	44.6	7.16	10.16	159.0	0.384	14.1
GISS-EH	5.30	46.5	6.91	9.75	145.5	0.128	0.5
GISS-ER	4.85	39.2	**6.84**	9.75	145.2	0.061	0.2
MIROC3.2(hires)	6.29	59.2	7.61	10.47	181.4	0.836	36.4
MIROC3.2(medres)	5.34	46.8	7.19	10.14	157.2	0.411	12.2
MRI-CGCM2.3.2	**4.23**	**29.8**	**6.84**	9.76	147.5	**0.059**	2.6
$\textbf{mean}_{\text{models}}$	5.65	62.3	7.19	10.03	161.8	0.410	16.8

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	d_{MV}
NCEP-1	0.00	0.0	5.49	8.35	99.9	0.000	0.0
ERA-40	4.20	27.1	6.15	8.97	120.1	0.656	20.2
CGCM3.1(T47)	4.82	43.3	5.89	8.72	110.5	0.401	10.6
CGCM3.1(T63)	5.52	52.8	6.14	8.92	117.4	0.649	17.5
CNRM-CM3	4.50	37.9	5.91	8.64	**106.3**	0.415	**6.4**
CSIRO-Mk3.0	5.19	54.8	6.19	8.89	119.6	0.700	19.7
CSIRO-Mk3.5	5.94	56.9	6.54	9.22	126.1	1.053	26.2
ECHAM5/MPI-OM	**4.23**	30.2	5.88	8.71	110.9	0.390	11.0
FGOALS-g1.0	13.03	302.2	9.53	13.97	292.8	4.042	192.9
GFDL-CM2.0	5.89	56.1	6.33	8.95	127.5	0.839	27.5
GFDL-CM2.1	4.49	34.5	**5.81**	**8.54**	106.7	**0.322**	6.8
GISS-AOM	5.10	44.6	6.43	9.09	120.7	0.939	20.8
GISS-EH	5.30	46.5	6.24	8.97	119.8	0.752	19.9
GISS-ER	4.85	39.2	6.07	8.90	115.7	0.580	15.8
MIROC3.2(hires)	6.29	59.2	6.70	9.21	139.4	1.210	39.5
MIROC3.2(medres)	5.34	46.8	6.28	8.95	120.8	0.789	20.9
MRI-CGCM2.3.2	**4.23**	**29.8**	5.97	8.65	108.9	0.480	9.0
$\textbf{mean}_{\text{models}}$	5.65	62.3	6.40	9.22	129.5	0.904	29.6

Table A.7: Verification results for the climate predictions for **txmax**.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	d_{MV}
HadEX	0.00	0.0	2.55	3.61	21.0	0.000	0.0
ERA-40	3.10	13.6	3.10	4.25	30.1	0.553	9.1
NCEP-1	3.16	16.9	2.78	3.73	22.3	0.231	1.3
CGCM3.1(T47)	5.77	46.8	3.31	3.68	**21.0**	0.760	**0.0**
CGCM3.1(T63)	5.01	36.2	3.09	**3.61**	**21.0**	0.545	**0.0**
CNRM-CM3	5.23	40.9	3.53	5.13	30.6	0.987	9.7
CSIRO-Mk3.0	6.14	56.5	4.32	6.26	56.4	1.770	35.4
CSIRO-Mk3.5	4.13	29.1	2.80	3.68	**21.0**	0.253	**0.0**
ECHAM5/MPI-OM	4.09	26.7	3.22	4.17	31.4	0.678	10.4
FGOALS-g1.0	5.47	44.0	3.43	5.66	24.7	0.888	3.7
GFDL-CM2.0	5.35	40.7	3.55	5.20	38.1	1.008	17.1
GFDL-CM2.1	4.46	29.8	3.13	4.24	26.4	0.584	5.4
GISS-AOM	9.04	97.8	6.79	10.15	99.3	4.245	78.3
GISS-EH	5.75	49.1	3.18	3.62	21.7	0.636	0.7
GISS-ER	6.72	69.1	3.38	3.91	30.2	0.831	9.2
MIROC3.2(hires)	**3.07**	**14.6**	**2.60**	3.63	**21.0**	**0.056**	**0.0**
MIROC3.2(medres)	5.05	42.3	3.08	4.02	**21.0**	0.539	**0.0**
MRI-CGCM2.3.2	3.98	23.0	3.25	4.42	29.6	0.706	8.7
$\mathbf{mean}_{\mathrm{models}}$	5.28	43.1	3.51	4.76	32.9	0.966	11.9

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	d_{MV}
HadEX	0.00	0.0	0.79	1.11	2.1	0.000	0.0
ERA-40	3.10	13.6	2.55	3.21	14.9	1.768	12.8
NCEP-1	3.16	16.9	2.41	3.08	15.9	1.620	13.9
CGCM3.1(T47)	5.77	46.8	4.43	5.43	40.0	3.648	38.0
CGCM3.1(T63)	5.01	36.2	3.68	4.65	30.5	2.897	28.4
CNRM-CM3	5.23	40.9	4.08	4.96	36.0	3.292	33.9
CSIRO-Mk3.0	6.14	56.5	5.24	6.05	53.7	4.456	51.7
CSIRO-Mk3.5	4.13	29.1	2.82	3.57	22.9	2.038	20.8
ECHAM5/MPI-OM	4.09	26.7	3.01	3.85	22.9	2.227	20.8
FGOALS-g1.0	5.47	44.0	4.45	5.28	40.3	3.660	38.2
GFDL-CM2.0	5.35	40.7	3.77	4.96	31.6	2.987	29.5
GFDL-CM2.1	4.46	29.8	2.85	3.85	20.5	2.065	18.4
GISS-AOM	9.04	97.8	8.24	9.02	95.4	7.451	93.4
GISS-EH	5.75	49.1	4.43	5.54	42.5	3.640	40.4
GISS-ER	6.72	69.1	5.30	6.37	61.5	4.515	59.4
MIROC3.2(hires)	**3.07**	**14.6**	**2.18**	**2.83**	**12.0**	**1.398**	**10.0**
MIROC3.2(medres)	5.05	42.3	3.66	4.72	35.8	2.875	33.8
MRI-CGCM2.3.2	3.98	23.0	3.12	3.89	20.8	2.337	18.7
$\mathbf{mean}_{\mathrm{models}}$	5.28	43.1	4.09	5.00	37.8	3.299	35.7

Table A.8: Verification results for the climate predictions for **tnmin**.

(a) Regional

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	d_{MV}
HadEX	0.00	0.0	5.71	8.45	98.4	0.000	0.0
ERA-40	3.44	20.5	5.92	8.65	105.9	0.203	7.4
NCEP-1	4.87	34.3	5.88	8.77	102.1	0.166	3.7
CGCM3.1(T47)	7.84	89.4	6.73	9.97	128.9	1.012	30.4
CGCM3.1(T63)	8.91	119.0	7.19	10.92	151.3	1.474	52.9
CNRM-CM3	6.45	62.3	6.33	**8.46**	101.0	0.613	2.6
CSIRO-Mk3.0	8.51	113.9	7.03	8.61	118.7	1.317	20.3
CSIRO-Mk3.5	7.31	76.4	6.55	9.60	104.6	0.839	6.1
ECHAM5/MPI-OM	**4.17**	**27.7**	**5.75**	8.48	**99.5**	**0.036**	**1.1**
FGOALS-g1.0	25.31	755.1	19.35	31.63	725.8	13.639	627.4
GFDL-CM2.0	9.25	117.2	7.87	11.88	173.8	2.152	75.4
GFDL-CM2.1	5.87	53.2	6.31	9.19	116.4	0.598	18.0
GISS-AOM	6.09	57.5	6.28	9.17	110.8	0.561	12.3
GISS-EH	6.24	58.0	6.39	8.97	113.7	0.676	15.3
GISS-ER	6.16	57.1	6.34	9.07	114.3	0.630	15.9
MIROC3.2(hires)	7.03	73.6	6.98	10.16	142.7	1.268	44.3
MIROC3.2(medres)	5.86	53.0	6.30	9.11	100.2	0.584	1.7
MRI-CGCM2.3.2	4.31	29.4	5.87	8.53	104.6	0.157	6.2
$\mathbf{mean}_{\mathrm{models}}$	7.95	116.2	7.42	10.92	160.4	1.704	62.0

(b) Local

Forecast	MAE	MSE	CRPS	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	d_{MV}
HadEX	0.00	0.0	1.65	2.39	9.0	0.000	0.0
ERA-40	3.44	20.5	3.02	4.07	26.2	1.379	17.2
NCEP-1	4.87	34.3	3.95	5.10	37.4	2.303	28.4
CGCM3.1(T47)	7.84	89.4	5.64	7.19	73.4	3.993	64.4
CGCM3.1(T63)	8.91	119.0	6.75	8.48	100.8	5.109	91.8
CNRM-CM3	6.45	62.3	4.60	5.98	48.5	2.950	39.4
CSIRO-Mk3.0	8.51	113.9	6.57	8.01	98.0	4.922	89.0
CSIRO-Mk3.5	7.31	76.4	5.16	6.52	57.6	3.512	48.6
ECHAM5/MPI-OM	**4.17**	**27.7**	**2.61**	**3.54**	**18.9**	**0.963**	**9.9**
FGOALS-g1.0	25.31	755.1	23.20	25.28	738.0	21.552	729.0
GFDL-CM2.0	9.25	117.2	7.34	9.02	107.1	5.695	98.0
GFDL-CM2.1	5.87	53.2	3.85	5.34	40.7	2.207	31.7
GISS-AOM	6.09	57.5	4.42	5.47	47.6	2.779	38.6
GISS-EH	6.24	58.0	4.60	5.81	50.0	2.958	41.0
GISS-ER	6.16	57.1	4.60	5.88	51.3	2.957	42.3
MIROC3.2(hires)	7.03	73.6	5.80	6.92	67.9	4.155	58.8
MIROC3.2(medres)	5.86	53.0	4.41	5.47	44.7	2.764	35.6
MRI-CGCM2.3.2	4.31	29.4	3.22	4.10	26.2	1.571	17.1
$\mathbf{mean}_{\mathrm{models}}$	7.95	116.2	6.18	7.53	104.7	4.539	95.7

A.2 Weighted model combinations

In this section, extended verification results of the cross-validation study for the weighted model combinations are presented. The first two tables show the weights of the weighting methods $\text{weights}_{\text{min}}$, $\text{weights}_{d_{\text{C.v.M.}}}$, and $\text{weights}_{d_{\text{MV}}}$ for mtxmax and mtnmin based on the comparison between all available predictions of the climate models in the years 1961 to 1980 against the corresponding data of the NCEP-1 re-analysis. These tables also include the weights for mtxmax in the summer months and for mtnmin in the winter months. The largest weight under each verification method is indicated in bold.

The verification results for the weighted model combinations and the individual climate models averaged over decadal results under the scoring rules S_{AE} and S_{SE} and under the Cramér-von Mises distance $d_{\text{C.v.M.}}$ are stated in the further Tables. To obtain the weights for each decade, we used the remaining two decades of data as training set. To solve the minimization problem of the $\text{weights}_{\text{min}}$ method, the function `ipop` of the `R` package `kernlab` was applied (cf. Chapter 2.2, p. 19). The implemented interior point method did not converge at one grid box for the time period 1971 to 1990 and at another grid box for the time period 1961 to 1970 and 1981 to 1990 for the model predictions of mtxmax in the winter months. In these cases, we set each weight at the grid box in question as the average of the corresponding weights over the other grid boxes.

Table A.9: Weights for the predictions of the climate models for **mtxmax** in the **summer** months and over the **entire year** for 1961-1980.

(a) Regional

	$\text{weights}_{\text{min}}$		$\text{weights}_{d_{\text{C.v.M.}}}$		$\text{weights}_{d_{\text{MV}}}$	
Model	Summer	Year	Summer	Year	Summer	Year
CGCM3.1(T47)	0.000	0.000	0.047	0.034	0.068	0.080
CGCM3.1(T63)	0.000	0.000	0.069	0.030	0.118	0.015
CNRM-CM3	0.000	0.000	0.051	0.015	0.017	0.001
CSIRO-Mk3.0	0.000	0.000	0.018	0.012	0.003	0.001
CSIRO-Mk3.5	0.000	0.000	0.090	0.038	0.039	0.002
ECHAM5/MPI-OM	0.112	0.035	0.086	0.093	0.017	0.014
FGOALS-g1.0	0.000	0.000	0.017	0.021	0.008	0.003
GFDL-CM2.0	0.000	0.000	0.031	0.018	0.006	0.001
GFDL-CM2.1	0.000	0.000	0.080	0.054	0.022	0.007
GISS-AOM	0.000	0.000	0.006	0.008	0.001	0.000
GISS-EH	0.000	0.000	0.055	0.046	**0.537**	0.006
GISS-ER	0.000	0.000	0.053	0.020	0.023	0.001
MIROC3.2(hires)	**0.531**	**0.756**	**0.154**	**0.463**	0.039	0.043
MIROC3.2(medres)	0.000	0.038	0.106	0.105	0.065	**0.826**
MRI-CGCM2.3.2	0.357	0.170	0.135	0.044	0.036	0.002

(b) Local

	$\text{weights}_{\text{min}}$		$\text{weights}_{d_{\text{C.v.M.}}}$		$\text{weights}_{d_{\text{MV}}}$	
Model	Summer	Year	Summer	Year	Summer	Year
CGCM3.1(T47)	0.038	0.027	0.051	0.058	0.060	0.081
CGCM3.1(T63)	0.045	0.043	0.062	0.056	0.052	0.091
CNRM-CM3	0.053	0.035	0.061	0.036	0.058	0.021
CSIRO-Mk3.0	0.016	0.024	0.029	0.043	0.014	0.020
CSIRO-Mk3.5	0.130	0.149	**0.114**	0.072	0.102	0.079
ECHAM5/MPI-OM	0.076	0.115	0.089	0.117	0.083	0.117
FGOALS-g1.0	0.045	0.024	0.046	0.032	0.048	0.048
GFDL-CM2.0	0.026	0.015	0.044	0.041	0.065	0.030
GFDL-CM2.1	0.017	0.099	0.064	0.098	0.093	0.091
GISS-AOM	0.015	0.028	0.021	0.027	0.012	0.029
GISS-EH	0.051	0.063	0.065	0.085	0.067	0.075
GISS-ER	0.043	0.023	0.055	0.047	0.043	0.030
MIROC3.2(hires)	**0.195**	**0.167**	0.107	**0.127**	0.093	**0.162**
MIROC3.2(medres)	0.069	0.051	0.092	0.081	**0.122**	0.066
MRI-CGCM2.3.2	0.183	0.138	0.101	0.079	0.088	0.058

Table A.10: Weights for the predictions of the climate models for **mtnmin** in the **winter** months and over the **entire year** for 1961-1980.

(a) Regional

	$\text{weights}_{\text{min}}$		$\text{weights}_{d_{\text{C.v.M.}}}$		$\text{weights}_{d_{\text{MV}}}$	
Model	Winter	Year	Winter	Year	Winter	Year
CGCM3.1(T47)	**0.589**	0.277	0.153	0.109	0.026	0.018
CGCM3.1(T63)	0.000	0.000	0.115	0.040	0.018	0.006
CNRM-CM3	0.000	0.000	0.053	0.098	**0.657**	0.203
CSIRO-Mk3.0	0.000	0.000	0.027	0.080	0.033	**0.320**
CSIRO-Mk3.5	0.000	0.000	0.017	0.019	0.004	0.004
ECHAM5/MPI-OM	0.411	**0.317**	0.080	0.065	0.012	0.009
FGOALS-g1.0	0.000	0.000	0.002	0.004	0.000	0.001
GFDL-CM2.0	0.000	0.098	0.021	0.020	0.003	0.003
GFDL-CM2.1	0.000	0.000	**0.178**	0.072	0.046	0.017
GISS-AOM	0.000	0.000	0.023	0.029	0.004	0.005
GISS-EH	0.000	0.000	0.132	0.078	0.142	0.093
GISS-ER	0.000	0.109	0.109	0.152	0.036	0.283
MIROC3.2(hires)	0.000	0.000	0.010	0.012	0.002	0.002
MIROC3.2(medres)	0.000	0.000	0.029	0.026	0.009	0.006
MRI-CGCM2.3.2	0.000	0.199	0.053	**0.196**	0.007	0.031

(b) Local

	$\text{weights}_{\text{min}}$		$\text{weights}_{d_{\text{C.v.M.}}}$		$\text{weights}_{d_{\text{MV}}}$	
Model	Winter	Year	Winter	Year	Winter	Year
CGCM3.1(T47)	0.088	0.126	0.103	**0.109**	**0.143**	0.103
CGCM3.1(T63)	0.067	0.039	0.089	0.064	0.091	0.057
CNRM-CM3	0.089	0.031	0.068	0.076	0.070	0.128
CSIRO-Mk3.0	0.023	0.055	0.037	0.074	0.032	0.083
CSIRO-Mk3.5	0.004	0.008	0.025	0.031	0.053	0.025
ECHAM5/MPI-OM	**0.134**	**0.166**	**0.111**	**0.109**	0.099	0.066
FGOALS-g1.0	0.017	0.014	0.014	0.015	0.007	0.038
GFDL-CM2.0	0.082	0.075	0.068	0.046	0.064	0.024
GFDL-CM2.1	0.041	0.068	0.095	0.103	0.088	0.090
GISS-AOM	0.091	0.061	0.079	0.058	0.089	0.024
GISS-EH	0.120	0.061	0.087	0.075	0.080	0.092
GISS-ER	0.098	0.144	0.084	0.090	0.071	0.089
MIROC3.2(hires)	0.013	0.012	0.012	0.018	0.006	0.005
MIROC3.2(medres)	0.035	0.015	0.056	0.050	0.045	0.039
MRI-CGCM2.3.2	0.097	0.122	0.070	0.080	0.063	**0.138**

Table A.11: Verification results for the weighted combinations of the climate model predictions for **mtxmax** and the corresponding results for the predictions of the individual climate models and the ERA-40 re-analysis.

(a) Summer

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\text{C.v.M.}}$
NCEP-1	5.71	56.1	0.000	1.96	6.4	0.000
ERA-40	5.85	59.3	0.154	3.01	15.7	0.952
$\text{weights}_{\text{min}}$	5.73	56.5	0.022	2.17	7.8	0.193
$\text{weights}_{d_{\text{C.v.M.}}}$	5.75	56.6	0.061	2.20	8.0	0.231
$\text{weights}_{d_{\text{MV}}}$	5.76	56.8	0.154	2.24	8.1	0.262
CGCM3.1(T47)	5.75	56.7	0.258	4.46	29.5	1.966
CGCM3.1(T63)	**5.73**	**56.4**	0.169	3.83	21.9	1.448
CNRM-CM3	6.32	60.1	0.260	3.70	20.4	1.450
CSIRO-Mk3.0	6.94	78.3	0.789	4.86	36.1	2.625
CSIRO-Mk3.5	5.99	57.4	0.140	3.24	16.8	1.017
ECHAM5/MPI-OM	5.93	60.7	0.172	3.31	17.3	1.100
FGOALS-g1.0	7.46	61.7	0.700	5.61	43.1	3.008
GFDL-CM2.0	6.79	71.0	0.566	4.54	28.2	1.962
GFDL-CM2.1	6.19	60.3	0.214	3.47	17.3	1.135
GISS-AOM	9.24	107.1	2.227	7.24	67.9	4.996
GISS-EH	5.79	56.9	0.224	4.45	30.3	1.968
GISS-ER	5.80	61.5	0.297	5.19	39.2	2.641
MIROC3.2(hires)	5.79	57.1	**0.076**	3.32	16.7	1.109
MIROC3.2(medres)	5.81	56.8	0.123	3.72	23.2	1.368
MRI-CGCM2.3.2	5.92	58.8	0.137	**3.10**	**15.8**	**0.996**
$\textbf{mean}_{\text{models}}$	6.36	64.0	0.424	4.27	28.2	1.919

(b) Winter

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	6.07	56.1	0.000	1.63	5.1	0.000
ERA-40	6.09	56.2	0.065	2.13	7.6	0.353
$\mathrm{weights}_{\mathrm{min}}$	6.08	56.2	0.030	1.76	5.7	0.115
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	6.07	56.3	0.034	1.75	5.7	0.118
$\mathrm{weights}_{d_{\mathrm{MV}}}$	6.11	56.1	0.041	1.79	5.9	0.156
CGCM3.1(T47)	6.08	56.3	0.082	2.25	8.1	0.493
CGCM3.1(T63)	6.11	57.0	0.106	2.32	8.5	0.530
CNRM-CM3	6.43	66.2	0.297	3.64	22.3	1.524
CSIRO-Mk3.0	6.08	57.3	**0.050**	2.40	10.7	0.588
CSIRO-Mk3.5	6.80	64.7	0.438	3.51	17.8	1.494
ECHAM5/MPI-OM	6.17	56.3	0.078	2.28	9.4	0.496
FGOALS-g1.0	6.20	73.7	0.690	5.53	61.9	2.878
GFDL-CM2.0	6.13	56.2	0.056	2.31	9.1	0.492
GFDL-CM2.1	6.43	57.1	0.157	2.67	11.7	0.788
GISS-AOM	6.12	56.3	0.114	2.63	12.4	0.825
GISS-EH	6.13	58.4	0.122	2.94	16.1	1.028
GISS-ER	**6.07**	58.8	0.134	2.93	15.0	1.001
MIROC3.2(hires)	6.11	56.3	0.065	**2.13**	**7.8**	**0.389**
MIROC3.2(medres)	6.14	**56.1**	0.059	2.36	10.3	0.553
MRI-CGCM2.3.2	6.27	65.0	0.425	3.69	22.9	1.645
$\mathbf{mean}_{\mathrm{models}}$	6.22	59.7	0.192	2.91	16.3	0.981

(c) All the year

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	9.76	137.4	0.000	8.02	85.7	0.000
ERA-40	9.77	137.8	0.046	8.13	89.5	0.201
$\mathrm{weights}_{\mathrm{min}}$	9.77	137.6	0.010	8.06	86.3	0.049
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	9.79	137.7	0.014	8.10	86.8	0.075
$\mathrm{weights}_{d_{\mathrm{MV}}}$	9.78	137.6	0.017	8.11	86.4	0.103
CGCM3.1(T47)	9.95	137.7	0.094	8.30	90.6	0.348
CGCM3.1(T63)	9.99	138.1	0.102	8.29	89.8	0.297
CNRM-CM3	10.08	145.3	0.202	8.47	97.5	0.477
CSIRO-Mk3.0	10.14	148.2	0.284	8.50	99.2	0.587
CSIRO-Mk3.5	9.80	140.2	0.082	8.17	91.7	0.292
ECHAM5/MPI-OM	9.81	138.0	0.042	8.12	89.0	0.199
FGOALS-g1.0	9.78	140.0	0.162	8.52	103.3	0.824
GFDL-CM2.0	10.09	145.2	0.227	8.47	96.0	0.473
GFDL-CM2.1	9.86	138.5	0.076	8.22	89.5	0.265
GISS-AOM	10.13	151.8	0.442	8.48	103.9	0.923
GISS-EH	9.90	139.5	0.087	8.32	95.6	0.439
GISS-ER	10.16	143.9	0.183	8.50	99.2	0.575
MIROC3.2(hires)	**9.77**	**137.5**	**0.010**	**8.10**	**88.9**	**0.179**
MIROC3.2(medres)	9.79	**137.5**	0.032	8.33	92.5	0.312
MRI-CGCM2.3.2	9.80	140.8	0.081	8.16	92.2	0.320
$\mathbf{mean}_{\mathrm{models}}$	9.94	141.5	0.140	8.33	94.6	0.434

Table A.12: Verification results for the weighted combinations of the climate model predictions for **mtnmin** and the corresponding results for the predictions of the individual climate models and the ERA-40 re-analysis.

(a) Summer

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	3.55	21.0	0.000	1.71	4.8	0.000
ERA-40	4.11	29.2	0.418	3.41	16.3	1.321
$\mathrm{weights}_{\mathrm{min}}$	3.57	21.2	0.033	1.83	5.4	0.105
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	3.56	21.5	0.059	1.85	5.6	0.130
$\mathrm{weights}_{d_{\mathrm{MV}}}$	3.61	21.3	0.077	1.84	5.5	0.139
CGCM3.1(T47)	4.14	25.1	0.310	2.92	13.0	0.976
CGCM3.1(T63)	4.63	31.6	0.694	3.68	19.5	1.589
CNRM-CM3	3.56	22.0	0.074	2.66	**10.5**	**0.685**
CSIRO-Mk3.0	**3.55**	21.3	**0.038**	**2.51**	10.7	0.709
CSIRO-Mk3.5	4.29	32.9	0.630	4.01	23.0	1.872
ECHAM5/MPI-OM	3.77	25.1	0.197	2.73	12.0	0.840
FGOALS-g1.0	3.71	22.3	0.519	4.62	31.5	2.261
GFDL-CM2.0	5.98	40.0	1.294	4.67	27.8	2.433
GFDL-CM2.1	4.65	26.8	0.523	3.29	15.2	1.249
GISS-AOM	3.67	23.1	0.102	2.65	10.6	0.743
GISS-EH	4.01	32.5	0.628	4.82	36.2	2.536
GISS-ER	3.71	23.1	0.264	3.62	19.9	1.501
MIROC3.2(hires)	5.72	47.2	1.590	5.32	32.2	2.941
MIROC3.2(medres)	4.75	35.1	0.826	4.09	22.3	1.827
MRI-CGCM2.3.2	3.65	**21.1**	0.100	2.96	12.9	0.948
$\mathbf{mean}_{\mathrm{models}}$	4.25	28.6	0.519	3.64	19.8	1.541

(b) Winter

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	9.13	112.9	0.000	2.94	15.2	0.000
ERA-40	9.98	135.9	0.678	5.59	46.1	2.086
$\mathrm{weights}_{\mathrm{min}}$	9.21	113.4	0.051	3.24	17.6	0.250
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	9.16	113.3	0.101	3.26	17.7	0.280
$\mathrm{weights}_{d_{\mathrm{MV}}}$	9.38	113.9	0.196	3.32	18.2	0.324
CGCM3.1(T47)	9.24	116.1	0.125	4.63	34.7	1.299
CGCM3.1(T63)	9.26	121.2	0.235	5.37	44.6	1.836
CNRM-CM3	10.45	**114.5**	0.415	5.72	46.8	2.181
CSIRO-Mk3.0	9.35	115.6	0.752	7.51	82.9	3.487
CSIRO-Mk3.5	11.54	138.4	1.137	7.52	75.4	3.803
ECHAM5/MPI-OM	9.50	121.7	0.261	4.74	33.8	1.411
FGOALS-g1.0	28.00	550.4	9.664	21.90	605.4	17.019
GFDL-CM2.0	9.76	141.3	0.792	6.55	62.2	2.667
GFDL-CM2.1	9.19	**114.5**	**0.111**	**4.44**	**30.4**	**1.120**
GISS-AOM	10.71	142.0	0.894	7.01	67.8	3.469
GISS-EH	**9.18**	115.2	0.201	5.66	49.1	2.208
GISS-ER	**9.18**	115.5	0.171	5.39	46.0	2.078
MIROC3.2(hires)	11.70	177.7	1.917	8.17	91.3	4.625
MIROC3.2(medres)	10.90	126.0	0.784	6.46	61.5	3.008
MRI-CGCM2.3.2	9.58	128.8	0.411	5.37	42.5	2.121
$\mathbf{mean}_{\mathrm{models}}$	11.17	155.9	1.191	7.10	91.6	3.489

(c) All the year

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
NCEP-1	9.73	144.9	0.000	8.34	99.7	0.000
ERA-40	10.04	161.3	0.364	8.96	120.0	0.678
$\mathrm{weights}_{\mathrm{min}}$	9.74	145.2	0.021	8.39	100.4	0.060
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	9.74	145.2	0.029	8.41	100.7	0.087
$\mathrm{weights}_{d_{\mathrm{MV}}}$	9.75	145.2	0.074	8.44	100.6	0.134
CGCM3.1(T47)	9.80	148.4	0.089	8.72	111.0	0.437
CGCM3.1(T63)	9.92	154.3	0.234	8.92	117.5	0.678
CNRM-CM3	**9.74**	146.3	0.132	8.65	**106.8**	0.450
CSIRO-Mk3.0	9.81	**145.2**	0.136	8.91	119.8	0.731
CSIRO-Mk3.5	10.21	162.9	0.536	9.23	126.1	1.081
ECHAM5/MPI-OM	9.92	152.5	0.172	8.75	111.2	0.423
FGOALS-g1.0	11.23	268.1	2.409	14.03	292.9	4.073
GFDL-CM2.0	10.07	167.1	0.486	8.97	127.7	0.873
GFDL-CM2.1	9.81	147.8	0.128	**8.55**	106.9	**0.352**
GISS-AOM	10.15	159.1	0.389	9.08	120.8	0.967
GISS-EH	9.75	145.5	0.131	8.99	119.9	0.779
GISS-ER	9.75	**145.2**	**0.062**	8.92	115.7	0.605
MIROC3.2(hires)	10.47	181.5	0.842	9.23	139.5	1.239
MIROC3.2(medres)	10.14	157.2	0.416	8.94	121.0	0.817
MRI-CGCM2.3.2	9.76	147.6	0.065	8.66	109.0	0.510
$\mathbf{mean}_{\mathrm{models}}$	10.04	161.9	0.415	9.24	129.7	0.934

Table A.13: Verification results for the weighted combinations of the climate model predictions for **txmax** and the corresponding results for the predictions of the individual climate models and the re-analyses.

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
HadEX	3.60	21.0	0.000	1.06	1.9	0.000
ERA-40	4.26	30.1	0.557	3.23	14.8	1.862
NCEP-1	3.72	22.7	0.252	3.16	16.7	1.826
$\mathrm{weights}_{\mathrm{min}}$	3.64	21.0	0.067	1.56	3.8	0.410
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	3.68	21.7	0.233	1.94	7.7	0.711
$\mathrm{weights}_{d_{\mathrm{MV}}}$	3.72	21.1	0.517	1.86	6.5	0.658
CGCM3.1(T47)	3.73	21.1	0.768	5.44	40.4	3.794
CGCM3.1(T63)	**3.60**	**21.0**	0.550	4.61	30.6	3.035
CNRM-CM3	5.16	30.7	0.995	4.95	36.5	3.454
CSIRO-Mk3.0	6.28	56.4	1.776	6.06	54.0	4.585
CSIRO-Mk3.5	3.76	21.3	0.277	3.67	23.6	2.211
ECHAM5/MPI-OM	4.22	31.4	0.687	3.90	23.1	2.361
FGOALS-g1.0	5.64	24.7	0.894	5.29	40.5	3.778
GFDL-CM2.0	5.17	38.3	1.019	4.98	32.2	3.139
GFDL-CM2.1	4.27	26.4	0.592	3.99	21.5	2.272
GISS-AOM	10.20	99.3	4.253	9.06	95.7	7.590
GISS-EH	3.65	21.9	0.649	5.59	43.0	3.801
GISS-ER	3.90	30.4	0.842	6.48	62.6	4.723
MIROC3.2(hires)	3.63	**21.0**	**0.065**	**2.85**	**12.3**	**1.525**
MIROC3.2(medres)	4.02	21.1	0.546	4.78	36.4	3.052
MRI-CGCM2.3.2	4.42	29.6	0.712	3.89	21.0	2.459
$\mathbf{mean}_{\mathrm{models}}$	4.78	33.0	0.975	5.04	38.2	3.452

Table A.14: Verification results for the weighted combinations of the climate model predictions for **tnmin** and the corresponding results for the predictions of the individual climate models and the re-analyses.

	Regional			Local		
Forecast	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$	S_{AE}	S_{SE}	$d_{\mathrm{C.v.M.}}$
HadEX	8.39	97.5	0.000	2.13	7.8	0.000
ERA-40	8.60	105.4	0.219	4.01	25.9	1.598
NCEP-1	8.77	101.7	0.201	5.12	37.6	2.601
$\mathrm{weights}_{\mathrm{min}}$	8.45	98.2	0.044	2.56	9.8	0.359
$\mathrm{weights}_{d_{\mathrm{C.v.M.}}}$	8.45	98.4	0.092	2.59	10.6	0.407
$\mathrm{weights}_{d_{\mathrm{MV}}}$	8.50	98.7	0.236	2.64	10.6	0.435
CGCM3.1(T47)	10.07	128.4	1.057	7.34	73.9	4.452
CGCM3.1(T63)	10.94	151.6	1.524	8.53	102.3	5.458
CNRM-CM3	8.56	100.8	0.674	6.12	50.0	3.363
CSIRO-Mk3.0	8.58	118.1	1.347	8.08	97.9	5.227
CSIRO-Mk3.5	9.62	104.6	0.885	6.52	58.2	3.798
ECHAM5/MPI-OM	**8.47**	**99.0**	**0.055**	**3.50**	**18.5**	**1.194**
FGOALS-g1.0	31.76	725.0	13.680	25.29	738.4	21.900
GFDL-CM2.0	11.93	172.9	2.186	8.96	106.8	6.059
GFDL-CM2.1	9.13	115.7	0.613	5.35	40.3	2.475
GISS-AOM	9.14	110.1	0.580	5.47	47.5	3.030
GISS-EH	9.02	114.1	0.728	5.87	50.8	3.277
GISS-ER	9.04	114.2	0.665	5.97	51.3	3.269
MIROC3.2(hires)	10.08	142.1	1.283	6.89	67.6	4.430
MIROC3.2(medres)	9.10	100.0	0.608	5.38	44.7	3.030
MRI-CGCM2.3.2	8.55	104.6	0.195	4.16	26.6	1.852
$\mathbf{mean}_{\mathrm{models}}$	10.93	160.1	1.739	7.56	105.0	4.854

Erklärung

Hiermit versichere ich, dass ich meine Arbeit selbständig unter Anleitung verfasst habe, dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, und dass ich alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entlehnt sind, durch die Angabe der Quellen als Entlehnungen kenntlich gemacht habe.

Heidelberg, den 19. Juli 2012