Regional decadal climate predictions for Europe – Feasibility & Skill –

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
von der Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Marianne Uhlig M.Sc.
aus Karl-Marx-Stadt

Tag der mündlichen Prüfung:	13. Mai 2016
Referent:	Prof. Dr. Christoph Kottmeier
Korreferent:	Prof. Dr. Andreas Fink

Abstract

The climate is highly variable on all time-scales. External forcings like the addition of anthropoghenic carbon dioxid into the climate system as well as internal feedbacks within the climate system contribute this variability. To understand the variability and to predict its changes is a major scientific field of research and in the interest of many users. For the time frame up to a hundred years ahead the IPCC performs climate projections to estimate the effects of the changing forced boundary conditions due to the anthropogenic green-house gas emissions on the climate trends. For the short time-scales the numerical weather prediction is well established. Such predictions based on initial conditions have been extended to seasonal scales in the last two decades. In between these time frames emerges a new research field of the decadal predictions, which is located in the transition between the initial value and the forced boundary condition problem.

Decadal climate predictions cover the time frame of interest for many decision support activities and impacts. Only in the last few years, predictions over a few decades have become feasibly due to the availability of large observational data sets and advancements in climate modelling. To be able to develop mitigation strategies for potential negative impacts of the changing climate, as well as to take advantage of opportunities for adapting, it is neccessary to understand the mechanism behind the natural variations and how much they contribute to regional variations of the trend.

While seasonal climate predictions and climate projections have made great advancements, decadal climate prediction is still in its early stages. There are many open questions and topics especially in regard to regional applications. This works aims to close some of the knowledge gaps in regards to climate variability on a decadal time-scale and the prediction thereof on a regional scale in Europe.

The German research program MiKlip aims at the development of a decadal ensemble predictions system. A module within MiKlip is dedicated to develop a regional downscaling system for the global predictions. The regional focuses of the downscaling experiments are Europe, which is the main focus of this work, and Africa. Previous studies indicate some potential predictability for both regions but many questions have yet to be answered. With the use of the global prediction system consisting of the Max-Planck-Institute for Meteorology Earth System Model MPI-ESM and the downscaling by two regional models (RCM) COSMO-CLM and

REMO, a regional decadal climate ensemble is established. The predictions are compared to gridded observational data (E-OBS) in a comprehensive analysis.

More to the point, this work will answer three questions. First, what is the potential of a prediction in Europe on a decadal scale? To answer, the prediction potential of several variables in Europeis assessed using only observations. Decadal predictability in Europe originates from low frequency climate variations, especially in the North Atlantic. One possibility to quantify a variable's predictive potential could be to compare the distribution of a decadal forecast to a climatological one. How much and for how long the predicted probability density function (PDF) can be distinguished from the corresponding climatological PDF will then give a measure of predictability. I use the relative entropy from information theory to measure the difference between the PDF's.

Second, how should regional decadal predictions be evaluated? And, in conjunction with that, third, how can the predictions and their analysis be improved? Various analysis techniques including skill metrics are applied to determine if there is decadal variability and the representation thereof in the global and regional ensemble. Additional to basic skill the following questions are posed in this study: At which temporal and spatial scales can predictive skill be achieved? And is the skill of the global hindcasts preserved or even improved with regionalisation?

Previous studies indicate that extremes might partly show a higher predictive skill than mean precipitation or temperatures. Skilful predictions of decadal tendencies of extremes (like droughts, heat waves or storms) also exhibit a higher value to potential users than variations of mean quantities. Therefore, the decadal variations of extremes and their predictability are considered and found in the case of moderate temperature extremes and combined extremes of temperature and precipitation as well as variables that contain some memory to have the most potential to be predictable with a 10-year forecast. A positive skill for near surface temperature and precipitation was found especially for multi-year averages. This skill varies on season and region. The same can be found for the added value of downscaling global decadal climate predictions.

Though this work makes use of only one prediction system, some of the results can be generalised, especially regarding the potential predictability and the optimal length of multi-year windows to determine decadal signals. So, a few recommendations for regional decadal prediction systems can be made, as so far that annual data is too noisy to exhibit significant skill and that the size of the ensemble has a huge impact on the skill as well. It has also emerged that the linear trend in temperature in particular is a large contributor to the skill, and, when included, improved the performance of the models significantly. A first try into prediction large, extreme climate anomalies (climate events), has shown that the MiKlip model is able to reproduce 9 out of 11 anomalous hot summer events. Subsuming, a positive outlook into the future of predicting events with great impact on decadal time scales emerges from the overall skill of the MiKlip

prediction system, the potential predictability for many variables, as well as continuing advances in climate modelling.

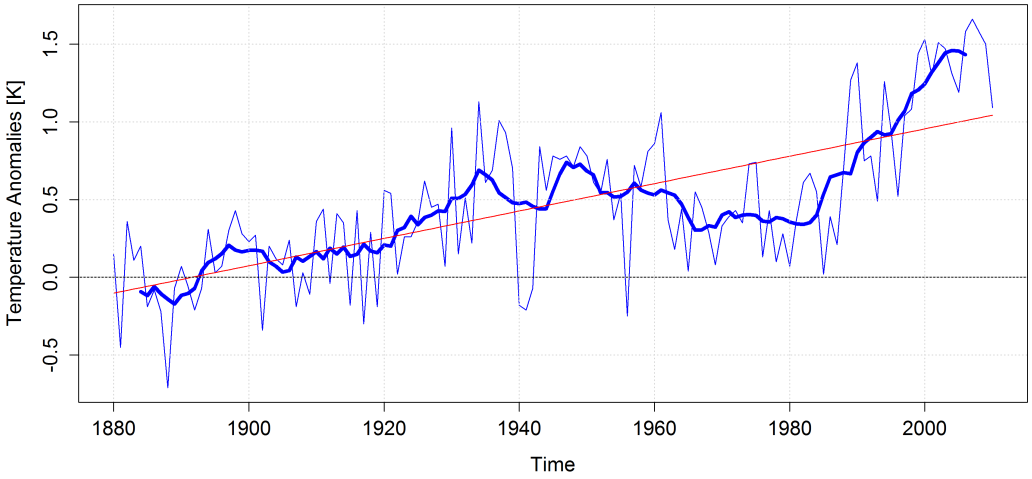
Contents

1. Introduction

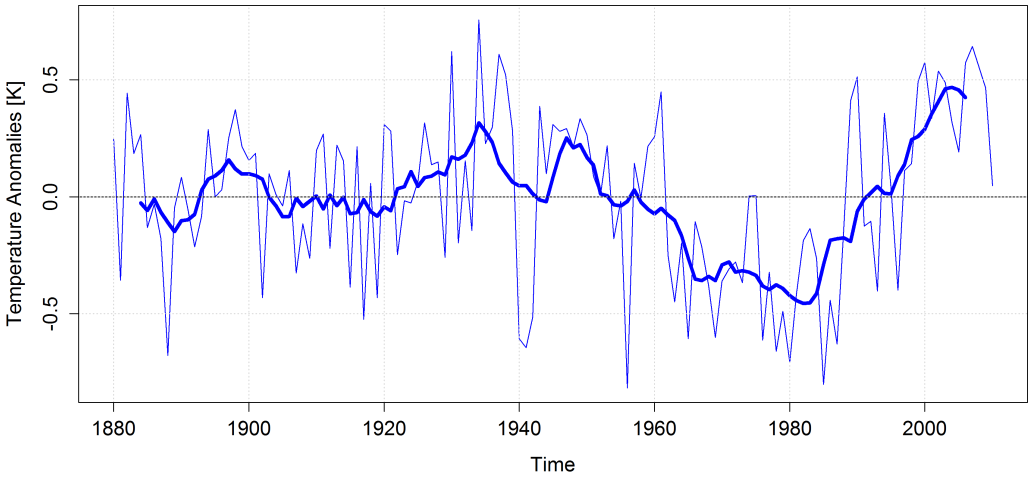
Lets look at a time series of observed temperature of the last century, like the mean surface air temperature over European land areas as in figure 1.1a. The most distinctive feature is the rate of warming. Upon removing the linear trend (red line) it becomes apparent that there is more to the story than anthropogenic climate change. Atmospheric conditions are not a linear progression due to the input of energy. The development of temperature over Europe sans linear trend (v. figure 1.1b) reveals internal features including a possible oscillation with a wave length of over 60 years. Consequently, the European (and possibly global) surface air temperature over the last one and a half century can be described as a long term warming with a strong superimposed multi-decadal variability [*Latif and Keenlyside*, 2011; *IPCC*, 2013].

While an accurate prediction of future climate for years or decades ahead might not be possible, certain aspects of the climate could be predicted. On the one hand there are for example green house gas concentrations that influence the development of the temperature trend; on the other hand there are naturally occurring external or internal variability. Internal variability stems from instabilities naturally occurring in the climate system. If such variabilities include extensive long-lived anomalies (e.g. of the upper ocean) they can influence the atmosphere locally as well as globally trough teleconnections. The El Niño-Southern Oscillation phenomena of the pacific is perhaps the most known influencing the climate not only in South America but also as far as Antarctica [*Turner*, 2004]and even Europe [*Ineson and Scaife*, 2009]. The predictability of El Niño is about 6 months [*Chen et al.*, 2004]. In the case of Europe the most prevalent long-term and potentially predictable variability is the Multidecadal Atlantic Oscillation (AMO) [*Schlesinger*, 1994; *Latif and Keenlyside*, 2011]. The AMO is a temperature oscillation of a period of about 65-70 years and is usually defined from the patterns of sea surface temperature (SST) variability in the North Atlantic once any linear trend has been removed [*Enfield et al.*, 2001]. The AMO is correlated to temperatures and rainfall over much of the Northern Hemisphere, in particular European and North American summer climate and has been brought into context with climate phenomena such as the Sahel droughts and North Atlantic hurricane activity [*Knight et al.*, 2005; *Keenlyside et al.*, 2008; *Matei et al.*, 2012].

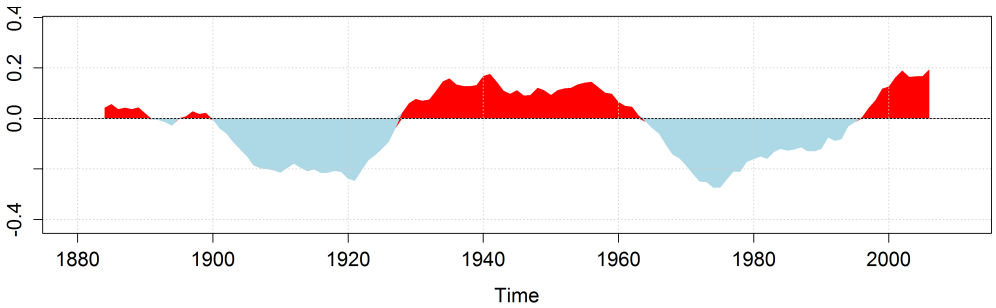
Up to date climate predictions have been mostly been considered on two time scales: seasonal and centennial. While seasonal prediction is an initial value problem, centennial climate projection are regarded as a boundary value problem, i.e. the development of the climate primarily



(a) Temperature over Europe with trend



(b) Temperature over Europe with the linear trend removed



(c) AMO Index

Figure 1.1.: Mean annual surface air temperature deviations in relation to pre-industrial values (1850–1899) over European land areas from the HAdCRUT4 data set from the UK MET Office Hadley Centre and the University of East Anglia Climate Research Unit with (1.1a) and without (1.1b) linear trend. Europe is defined as the area between 35° to 70 ° North and -25° to 30° East, plus Turkey (35° to 40° North and 30° to 45° East) [Morice et al., 2012; Met Office Hadley Centre, 2016]. The thin blue lines indicate the annual mean temperature and the thick blue lines represent the annual means filtered using a running mean of 9 years. The red line in Fig. 1.1a is the linear trend, that is removed for Fig. 1.1b

dependent on external forcings (e.g. anthropogenic green house gas emissions) [*IPCC*, 2007, 2013].

Predictions of climate for a decade ahead would be a problem of both the initial and the boundary conditions. Within a decade the global trend can be come a large fraction of explainable variance: While the linear trend in the European temperature for the last century was about 0.88 K (v. figure 1.1a) the estimated warming for the 21st century is between 4 - 6 K, making it a change in temperature of about 0.5 per decade. Also, long-term, multi-year or multi-decadal internal variability such as the AMO (v. Fig. 1.1c) comes into play on decadal time scales and require an initialization of models on the accurate phase of the variability. In the past the intra-decadal variability was about 0.28 K (Mean standard deviation of the annual detrended HadCRUT4 European temperature, as shown in Fig. 1.1b, making the expected rate of warming of the same order as the likely occurring internal variability. Further illustration is shown in Fig. 1.2. The graph shows not only the mean annual surface air over European land areas but also the AMO index [*NOAA–ESRL*, 2015; *Enfield et al.*, 2001] and two different types of forcings: the well mixed green house gases and the stratospheric aerosols (volcanoes) [*Hansen et al.*, 2005; *NASA/GSFC/HSL*, 2015]. A combination of all radiative forcings (sans the stratospheric aerosols) and the AMO index in a simplistic way to show the drivers of European temperature give the red curve. The actual European temperature and this hypothetical temperature curve are highly correlated ($r = 0.8841$).

Therein lies already the crux of decadal predictions: The AMO signal is derived from SST anomalies with a removed linear trend. While this detrending is to remove the influence of green house gas-induced warming, this signal is not linear and in some time periods might have a similar gradient as the AMO. Therefore the North Atlantic SST anomaly at the end of the twentieth century could be equally divided between the externally forced component and internally generated variability [*Ting et al.*, 2009]. The separation of signals is not always trivial.

Therefore, decadal prediction have become the focus of a number of initiatives including the international efforts like the CMIP5 experiments [*IPCC*, 2013; *Hurrell et al.*, 2011; *Taylor et al.*, 2012] and as well as national ones [*Marotzke et al.*, 2012; *Meehl et al.*, 2013]. The aim is to improve model uncertainties and expand the understanding of the mechanisms behind decadal variabilities and their regional effects.

Decadal climate predictions are usually started from a recently observed climate (initialized) in order to forecast the climate dependent on both natural (internal) variability and anthropogenic (external) climate change [*Hurrell et al.*, 2010; *Meehl et al.*, 2013]. Though recent studies have shown that initialized decadal prediction exhibit some skill above climate projections driving by external forcings and possible sources of predictability for decadal predictions have been identified decadal climate predictions as opposed to weather or seasonal/inter-annual predictions is still a young research field [*Kim et al.*, 2012; *Müller et al.*, 2012]. Additionally regions were

identified that show potential for decadal predictions [*Smith et al.*, 2007; *Keenlyside et al.*, 2008; *Pohlmann et al.*, 2009; *Smith et al.*, 2012b], e.g. the North Atlantic region.

Current research aims to develop and improve prediction systems and find the source for skill. The German Federal Ministry of Education and Research instigated the Project **MiKlip** to tackle a number of issues arising in the field of decadal climate predictions and establish a decadal climate prediction system that can be applied by an operational agency such as the German Meteorological Service [*BMBF*, 2007]. One of them is to fill the request for regional scale climate predictions for i.a. Europe as studies have shown that downscaling can add value to a prediction [*Hawkins and Sutton*, 2009; *Feser et al.*, 2011; *Kanamitsu and DeHaan*, 2011]. Before, most studies concentrated on global means or atmospheric conditions over oceans. While the reasoning of primarily oceanic variabilities being the sources of decadal predictability, supports this approach, regional information on climate is always of interest also with potential users of decadal predictions in mind. Decadal predictions are a relatively new field and aside from the efforts in MiKlip a systematic downscaling of decadal predictions has never been at-

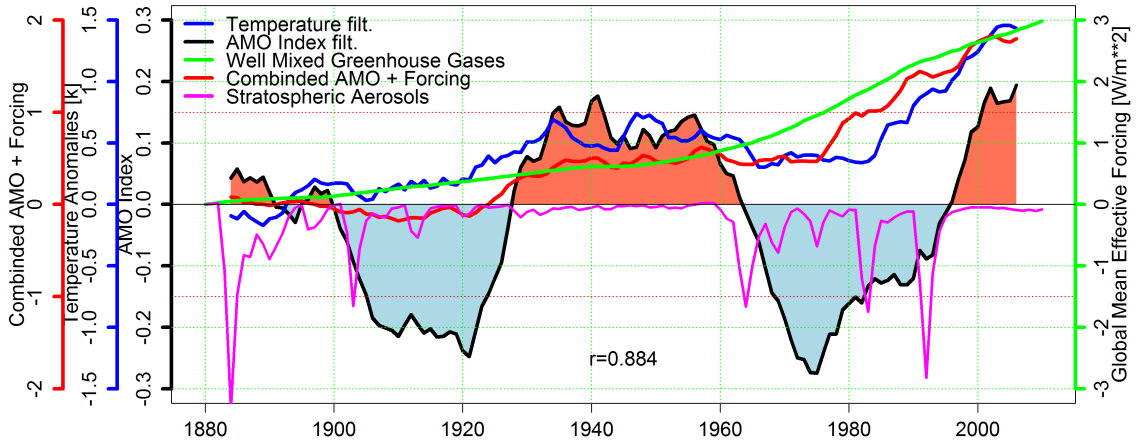


Figure 1.2.: Annual mean surface air temperature over European land areas from the from the HAdCRUT4 data set [Morice et al., 2012; Met Office Hadley Centre, 2016] filtered using a 9-year running mean (blue curve) compared to the Atlantic Multidecadal Oscillation (AMO) index [NOAA–ESRL, 2015; Enfield et al., 2001] filtered by a 9-year running mean (black curve) and two samples from the NASA/GSFC/HSL [2015] radiative forcings data set: the "well mixed green house gases" (green) and the "stratospheric aerosol forcing" (magenta) [Hansen et al., 2005]. The red curves represents the sum of all radiative forcings sans the stratospheric aerosols and the AMO index. The relationship between the forcings and internal variability in conjunction with the mean annual temperature over Europe is illustrated by a correlation coefficient of 0.8841. For further information on the data used refer to App. A.

tempted. Consequently a number of questions and aspects of regional decadal predictions are still unanswered:

1. What can potentially be predicted in Europe on a decadal scale?
2. How should regional decadal predictions be evaluated?
3. How can the predictions and their analysis be improved?

In summary, while studies have been published that indicate skilful predictions of climate up to a decade ahead as laid out in **Chapter 2**, fundamental questions remain not only pertaining the specific case of regional decadal climate predictions fore Europe. Challenges for decadal predictions include the initialization, the uncertainty of models and ensembles, their evaluation and the provision of regional information for users [*Murphy et al.*, 2010].

The first issue to deal with when starting predictions is the choice about what to predict. Usually mean values, e.g. surface temperature, contain valuable information and are also widely used for verification. However other variables may be even more valuable. One feature of future climate is the occurrence of extreme climate anomalies. But are those predictable? A measure of potential predictability would greatly aid in the decision making about employing decadal climate predictions for specific applications. The skill of climate predictions stems from good representation of slow acting parts of the climate system (e.g. the oceans) the transport of skill towards the atmosphere over land is one main challenge of regional decadal predictions. While there are known sources for predictability, particular slow ocean variability, e.g. the Atlantic Multidecadal Oscillation (AMO), the implications for atmospheric variables over land are not necessarily straight forward or easily distinguishable from other factors. For instance phases of the AMO index can coincide with the green house gas emission forcings and consequential long term climate trend. In this work, I state multi-year to decadal variability can be used to identify valuable information on predictability. Decadal variability constitutes a signal on the time scale of the decadal prediction that given a perfect model and initialization lead to a skilful prediction. Hence the first step of this comprehensive study is to link the predictability to internal variability within the variable to be predicted. An elaborate time series analysis to identify the strength and frequency of multi-year variabilities in observations is introduced in **Chapter 3** that can aid in the choice of variabilities worth to be predicted.

Not only is the potential for decadal climate prediction of interest but the skill of said predictions once a system is established. First one has to define what constitute the goodness of a prediction. *Murphy* [1993] identified three types of goodness: consistency, quality and value (utility). While consistency addresses the subjectivity of forecasters and their experience and the value the economic worth to the user of the prediction, the quality represents the correspondence between prediction and observation [*Jolliffe and Stephenson*, 2011]. While the first

is rarely addressed and will only be briefly touched upon here, the third "goodness" can be addressed via the potential of user specific variables, while a verification framework, the likes of *Goddard et al.* [2013], suffices the latter and answers the second question. The verification framework specifically developed and tested for regional decadal predictions is introduced in **Chapter 4**. The methods include continuous skill metrics to explicate predictive skill of the regional ensemble as well as the value added to global predictions.

The third issue is an expansion of the second: A deeper look into the temporal and spatial scales on which predictions can be skilful and expedient. Studies have shown that the highest skill in decadal predictions can be achieved for multi-year averages so far [*Goddard et al.*, 2013]. Does that apply to all forecast applications? And do these result suggest the aforementioned extreme climate anomalies have to be spanning several years? To investigate, different filters, both temporal and spatial, are applied in **Chapter 5** and the results are analysed pertaining to forecast quality and its significance depending on the internal organization (autocorrelation) to identify scales on which meaningful predictions can be achieved.

In a last step to round off this study, one first look into the prediction of extreme climate anomalies as a user specific application of decadal climate prediction is attempted in **Chapter 6**. Considerations into the types of events with the largest impacts on societies, economies and ecosystems lead to a preliminary definition of such events and the prediction thereof. Finally this thesis results will be summarized in **Chapter 7** and some recommendation towards regional decadal predictions systems given the experiences with the MiKlip ensembles outlined.

Addressing all these problems and advancing climate research will reduce uncertainty in decadal climate predictions systems particular when the issue of understanding of impacts inevitable comes up. That's why I will answer these questions or develop an easily applicable framework to determine the answers given any forecasting situation. I apply the theoretical methods to the hindcasts ensembles produced in MiKlip. Though the MiKlip ensembles consists of simulations of only one global model and two regional models, the findings in this thesis can be generalised. Thus not only will the MiKlip ensemble be verified but also general statements about the feasibility of regional decadal climate predictions in particular for Europe can be made.

2. Predicting the climate up to a decade ahead

2.1. The potential for Decadal Predictions

The climate system can be considered to consist of four subcomponents - atmosphere, ocean, land and cryosphere (land ice and sea ice). Each differs in its physical characteristics (e.g. heat conductivity, energy budget). Most importantly, the atmosphere and oceans as free fluids are able to transport heat and water around the planet controlled by the earth' rotation and external radiative forcing (solar heating) and produce a dynamical equilibrium and a radiative balance.

For the most part of the planet the ocean sea surface temperature varies accordingly to a white noise generator [*Hasselmann*, 1976; *Anderson et al.*, 2009]. Nevertheless there is indication of limited regions where natural variabilities exceed white noise thus providing a potentially predictable signal. Many studies have identified these variabilities to originate from the North Atlantic, the South Atlantic including the circumpolar current, the North Pacific and the Tropical Pacific [*Anderson et al.*, 2009], albeit on different time scales. On seasonal to inter-annual time scales the tropical Pacific provides much predictability as a result of the El Niño/Southern Oscillation phenomena. On times scales of multiple years (5+ yr), studies like *Boer and Lambert* [2008] have identified other regions of potential predictability, e.g. the North Atlantic and the North Pacific.

The time scales under scrutiny in most recent years have been the decades. Climate predictions up to 10 years could provide advise on climate change policies, as they could provide information on near term climate change and natural variability. Climate information users have identified this time scale as being important to infrastructure planner, water resource managers and many others [*BMBF*, 2007]. Now, decadal prediction systems similar to the seasonal systems are being developed and were included in the CMIP5 experiments [*Anderson et al.*, 2009; *Hurrell et al.*, 2011; *IPCC*, 2013].

Consequently, decadal predictions are influenced by at least 3 factors [*Meehl et al.*, 2009]:

1. adjustment to the increase of greenhouse gases that has already occurred

2. external radiative forcing (e.g. further increases of green house gases that have already occurred or recovery of the ozone hole, volcanoes) and

Table 2.1.: Correlation Coefficients between the mean annual European temperature of the HadCRUT4 data set[Met Office Hadley Centre, 2016; Morice et al., 2012] with the AMO index [NOAA–ESRL, 2015; Enfield et al., 2001] and Global Mean Effective Forcing [NASA/GSFC/HSL, 2015; Hansen et al., 2005]: Well Mixed Green House Gases (WMGHG), Stratospheric Aerosols and all forcings combined with the AMO index but without the Stratospheric Aerosols (cf. Ch. 1 & Fig. 1.2)

Correlation of	Temperature with trend	Temperature detrended
AMO	(9-year filter) 0.46	(9-year filter) 0.67
AMO	0.31	0.32
WMGHG	0.70	0.07
StratAER	0.16	0.15
Combined Forcings	0.69	0.08

3. internally generated variability (e.g. variations in oceanic surface temperature and currents).

Therein lies the challenges of decadal climate predictions, as they include two different processes: Climate change and natural variability. Firstly, the predictions require initialization of a coupled general circulation model and information of the initial state of atmosphere, ocean, cryosphere and land to predict accurate modes of natural variability as well as information of anthropogenic radiative forcing [*Hurrell et al.*, 2010]. And secondly the mechanism for natural variability are still not well understood, poorly documented and very differently represented in models [*Latif and Keenlyside*, 2011]. Additionally, most internal generated natural variability is considered high frequency noise (interannual variability). Only in some cases, a signal exceeds that noise resulting is at least potentially predictable [*Branstator and Teng*, 2010; *Teng and Branstator*, 2011].

With the ocean as the slower acting subcomponent in the climate system it is unsurprising that most predictability stems from oceanic processes. The predictability over land is small with some indicators of processes , probably linked to snow cover changes, around the Himalayan being somewhat predictable [*Boer and Lambert*, 2008], as well as some teleconncections hinting at a transport of information from the oceans towards land [*Knight et al.*, 2005; *Anderson et al.*, 2009; *Smith et al.*, 2012b]. A major source of potential predictability on the decadal scale has been identified in the Atlantic Multidecadal Oscillation (AMO) that is believed to be linked to changes in the thermohaline circulation. The AMO is a quasi periodic feature with a frequency of 60 to 70 yr. In general there is some hope for predictability in the Atlantic sector on time scales up to about 10 yr arising from knowledge of the ocean initial state [*Anderson et al.*, 2009]

or in the absence of external forcings even several decades [*Knight et al.*, 2005] and possibly influencing north hemispheric mean temperatures, especially in Europe.

The combination of natural variability and a changing climate becomes most important when considering the dangers though changing frequencies and intensities of extreme events including droughts, floods, storms, fire and heat waves [*Smith et al.*, 2012b]. High impact climate change and regional scale variability and change, e.g. large scale climate events as the "Dust Bowl" in 1930 , the Sahelian Drought in the Seventies and Eighties or , more recent, the ongoing drought in the South West of North America or decadal variability in the Atlantic hurricane activity, have been attributed to difference in land and ocean temperatures as well as among the oceans [*Hurrell et al.*, 2010].

2.2. First Decadal Hindcasts

To predict the regional scale climate variability is the task of decadal climate predictions with the need for knowledge about climate events with high impacts growing. First attempts at decadal hindcasts investigated the impact of initialization from an observed state [*Smith et al.*, 2007; *Keenlyside et al.*, 2008; *Hurrell et al.*, 2010]. They indicated advancements in skill due to initialization at a global scale and over the North Atlantic. These studies took a simple approach that was to be adapted by many those followed: First initializing a global climate model using observed anomalies and running it for 10 years while accounting for external forcings and changes therein (both natural and anthropogenic). However, the results differ for initialization technique, data and model. As an example, *Smith et al.* [2007] showed predictability of global mean temperature emerging arising from the upper ocean heat content initialization while *Keenlyside et al.* [2008] found the predictability of North Atlantic SST to be a consequence of the Atlantic meridional overturning circulation. Since the differences in results are likely also stemming from differences in climate models, *Hurrell et al.* [2010] argued for the need of a multi-model approach. Further research by *Smith et al.* [2010] managed to expand skill prediction of Atlantic hurricane frequency beyond one season.

Consequently the CMIP5 experiments did include decadal experiments for the first time [*Hurrell et al.*, 2011; *Taylor et al.*, 2012; *Kirtman et al.*, 2013]. Two sets of near-term integrations have been realised: Firstly, a set of 10-year hindcasts initialized from observed climate states near the years 1960, 1965, and every 5 yr to 2005. And secondly, by 20 years extended simulations initialized in 1960 and 1980 for 30-year hindcasts and one 30-year prediction to the year 2035 initialized in 2005. On the longer time scale the external forcing from increased green house gases should dominate the response.

Global mean temperature, the Atlantic Multidecadal Oscillation (AMO; [*Enfield et al.*, 2001]) and the Inter-decadal Pacific Oscillation (IPO; [*Mantua and Hare*, 2002]) or Pacific Decadal

Oscillation (PDO) are used to determine the ability of decadal forecast systems to predict multi-annual averages of climate variability [*Kim et al.*, 2012; *Oldenborgh et al.*, 2012; *Doblas-Reyes et al.*, 2013; *Goddard et al.*, 2013].

Oldenborgh et al. [2012] showed that the CMIP5 decadal experiments have skill in predicting the Earth's temperature at regional scales over the past 50 years - a skill that could mostly be attributed to changes in radiative forcings but also to the initialization. Especially oceanic regions, e.g. the subtropical East Pacific and the North Atlantic show skill beyond persistence for forecast years 6-9 and 2-5 respectively. The inter annual variability of the AMO is not captured well. But longer means thereof can be predicted skilfully up to 5 years. Similar results have been shown by *Pohlmann et al.* [2009], *Kim et al.* [2012] and *Matei et al.* [2012], too.

Kim et al. [2012] have demonstrated that most of the models in the CMIP5 decadal experiments overestimated trends, predicting less warming in earlier decades and too much warming at the end of the investigation period. Additionally, high prediction skill is concentrated in regions where externally forced components and low-frequency climate variability are dominant: E.g. the Indian Ocean, the North Atlantic and the Pacific Ocean. The AMO index is in general better predicted than the PDO. The poor skill of temperature predictions over the central North Pacific have among others been linked to the relatively low importance of the linear trend leading to a rapid decline in skill after a few forecast years [*Oldenborgh et al.*, 2012].

It has been stated that a skilful prediction of the multi-year variations of oscillations like the AMO could possibly lead to regional decadal forecasts beyond trend e.g. over Europe when taking teleconnections into account [*Oldenborgh et al.*, 2012; *Smith et al.*, 2012b]. However, imperfect skill and weak teleconnections in all but few regions challenge that assertion.

Pohlmann et al. [2009] stressed the importance of initialization on the skill of prediction of North Atlantic sea surface temperature (SST) but emphasize that both initial and boundary conditions must be taken into account for decadal predictions. The coupled model of the Max Planck Institute (MPI) [*Stevens et al.*, 2013] of Meteorology which consists of the atmosphere model ECHAM6 and the MPI ocean model (MPI-OM) [*Jungclaus et al.*, 2013] when initialized with oceanic synthesis fields using an anomaly coupling scheme is able to outperform a damped persistence or a trend forecast in certain regions, esp. the North Atlantic. The higher skill of SST prediction is concentrated in the first half of the forecast years, whereas with longer forecast runs the trend becomes an important contributer to skill. Errors in the initialization fields lead to decreased skill of global temperature predictions showcasing that decadal predictions have an additional source of uncertainty on uninitialized climate projections.

Caron et al. [2014] estimated the skill of Atlantic hurricane activity in the CMIP5 decadal hindcasts and found increased skill in comparison to unitialized predictions as well as simple forecasts based on climatology or persistence for predictions up to 9 years. Most of the skill stems from the ability of the prediction systems to realistically simulate the change in Atlantic

sea surface temperature. Still, *Caron et al.* [2014] urge caution to be overly optimistic, as there are still features, the prediction system are not able to capture, e.g. underestimating SST in the early 2000's and therefore underestimating hurricane activity. They state that better representation especially of ocean heat uptake and circulation in coupled climate models will further improve decadal predictions.

Since the beginning of the CMIP5 experiments decadal climate predictions have been established as a source of information on future climate evolution additional to climate projections. Several decadal prediction systems have been introduced and improved. For instance the Met Office Hadley Centre Decadal Prediction System [*Knight et al.*, 2014] and the MiKlip Decadal Prediction System [*Pohlmann et al.*, 2013] improved predictions in the tropics and the Pacific due to improvements of the model and initialization thus improving global temperature predictions.

Also, multi-model ensemble of decadal prediction experiments performed in the framework of the COMBINE project (Comprehensive Modelling of the EARTH System for Better Climate Prediction and Projection) have been examined [*Bellucci et al.*, 2014]. As with every further study the consensus is confirmed, that current initialized climate models are not only able to estimate anthropogenic warming but also parts of internal climate variability. Also, as is known in the case of seasonal forecasting multi-model ensemble generally outperform ensemble member. An examination of different initialization methods shows an improvement in the Indo-Pacific region for full field initialization over anomaly initialization.

Other studies have more focused on the handling of decadal hindcasts [*Hurrell et al.*, 2010; *Goddard et al.*, 2013]. *Hurrell et al.* [2010] among others have pointed out, that the evaluation of decadal hindcasts is challenging owed to the relatively small sample size compared to daily weather and even seasonal forecasts: observational data for initialization and evaluation of hindcasts is limited in the time it covers therefore limiting the number of possible hindcasts. Decadal hindcasts are of a probabilistic nature. Hence the reliability and the resolution apart of simple accuracy measure should be assessed. *Goddard et al.* [2013] put forward a verification framework including measures for accuracy and reliability that has become widely applied.

2.3. MiKlip ("Medium Range Climate Predictions")

Another effort to further knowledge into decadal predictions is MiKlip ("Medium Range Climate Predictions") founded by the German Ministry of Education and Research (BMBF) [*Marotzke et al.*, 2012; *Marotzke et al.*]. It also includes the up to now only effort worldwide of dynamically downscaling global decadal hindcasts [*Kottmeier and Feldmann*, 2012].

The project aims to to predict the climate for the next decade by the development of a prediction system and scientific exploration of the capabilities and perspectives of decadal predictions.

The prediction system is composed of 3 ensemble generation, the first two will be considered here.

Within the first ensemble generation 10 realization initialized every ten years between 1961 and 2010 of the of the earth system model of the Max-Planck-Institute with low resolution (MPI-ESM-LR, MPI Earth System Model Low Resolution [*Stevens et al.*, 2013], atmosphere:T63/L47, ocean: 1.5°L40) and 3 realizations every year in between were run following the CMIP5 decadal experiment design [*Taylor et al.*, 2012]. The coupled atmosphere-ocean-sea ice model incorporates the general circulation model ECHAM6 and the ocean model MPI-OM [*Jungclaus et al.*, 2013]. For initialization, ocean temperature and salinity were estimated by forcing the MPI-OM with the fluxes of momentum, heat and fresh water from the NCEP-NCAR reanalysis [**?**]. Ocean temperature and salinity anomalies then initialized 3 decadal hindcasts ensemble members of the first of January of each year [*Müller et al.*, 2012]. The perturbation used to create the ensemble members was a 1-day lag around the first of January [*Matei et al.*, 2012]. The first ensemble generation of MiKlip is called "baseline 0" and any data used from this first version of decadal prediction system will be called "b0" from here on out.

The second ensemble generation ("baseline 1", in the following "b1") made use of the same coupled earth system model MPI-ESM-LR but deployed different initialization data. Here the Ocean is initialized with temperature and salinity anomalies from the ORA-S4 reanalysis [*Balmaseda et al.*, 2013]. Additionally, the atmosphere, too, is initialized with full-field 3-d temperature, vorticity, divergence and surface pressure fields from the ERA-40 and the ERA-Interim reanalysis [*Uppala et al.*, 2005; *Dee et al.*, 2011]. An ensemble of 10 members is started around the first of January lagged each by on day as in b0 [*Pohlmann et al.*, 2013]. The second stage also includes a 5 member ensemble of a higher vertical resolution, but it will not be discussed here.

Already the first ensemble generation showed promising result. The initialized decadal hindcasts outperformed the uninitialized ones for yearly and multi year means predominately in the North Atlantic for all forecast years as have previous studies also revealed. Winter means are predicted with high skill over Northern Europe whereas for summer and autumn means the area of high skill shifts towards Central and South Europe [*Müller et al.*, 2012]. A systematic weakness of the b0 system was discovered in the Tropics, especially the Pacific.

The second ensemble generation strived to remedy that by changing the ocean initialization data. While not necessarily producing high temperature skill over the tropical oceans, the negative skill of the b0 system thereby reduced. Over Europe the differences between b0 and b1 are small in either direction. When investigating other variables such as quantity net water flux at the ocean surface (also, E-P "evaporation minus precipitation"), *Stolzenberger et al.* [2015] found that while less predictable than basic dynamical variables, the skill improved with the initialization of both the ocean and the atmosphere increasing the predictability in the inner tropics from 1 to 2 years. For later prediction years, the hindcasts have only slightly higher

skill than the uninitialized ensemble, but the members are closer to each other than to the uninitialized ensemble. Thereby indicating the influence of the initialization on the prediction even in the second half of the decade [*Stolzenberger et al.*, 2015].

Another point addressed using the MiKlip ensembles was the ensemble size by *Kadow et al.* [2015]. The decadal experiment during the first ensemble generation of the CMIP5 experiment design only included initialization every 5 years [*Taylor et al.*, 2012] which was been observed to lead to unreliable skill assessment [*Goddard et al.*, 2013]. Since these first experiments, most prediction systems are now initialized annually. Comparison the three member ensemble used by *Pohlmann et al.* [2013] to a ten member ensemble found clear improvements in the skill all over the globe. Also, under-dispersion of the original three member ensemble was reduced: The larger ensemble better representing the uncertainty of the prediction by the ensemble spread [*Kadow et al.*, 2015].

Recent studies implied that full-field initialization of ocean and temperature would improve decadal prediction [*Bellucci et al.*, 2014; *Polkova et al.*, 2014]. The third ensemble generation of MiKlip has taken that into account and switch the initialization process to a full-field of ocean and atmosphere. However these decadal hindcasts are not considered here.

Additionally, an investigation of an extended period (1901-2010) was launched. The extended period lead to larger regions with significant skill for temperature. It is believed to be a result of the larger contribution of the trend but also long term climate variability as in the North Atlantic can have wavelengths that exceed 50 years [*Müller et al.*, 2014]. One can be cautiously optimistic as the MiKlip system was able to not only outperform the unitialized climate runs during the North Atlantic warming events during the 1990 but also the 1920. This suggests that there is the possibility of predicting future large scale climate events.

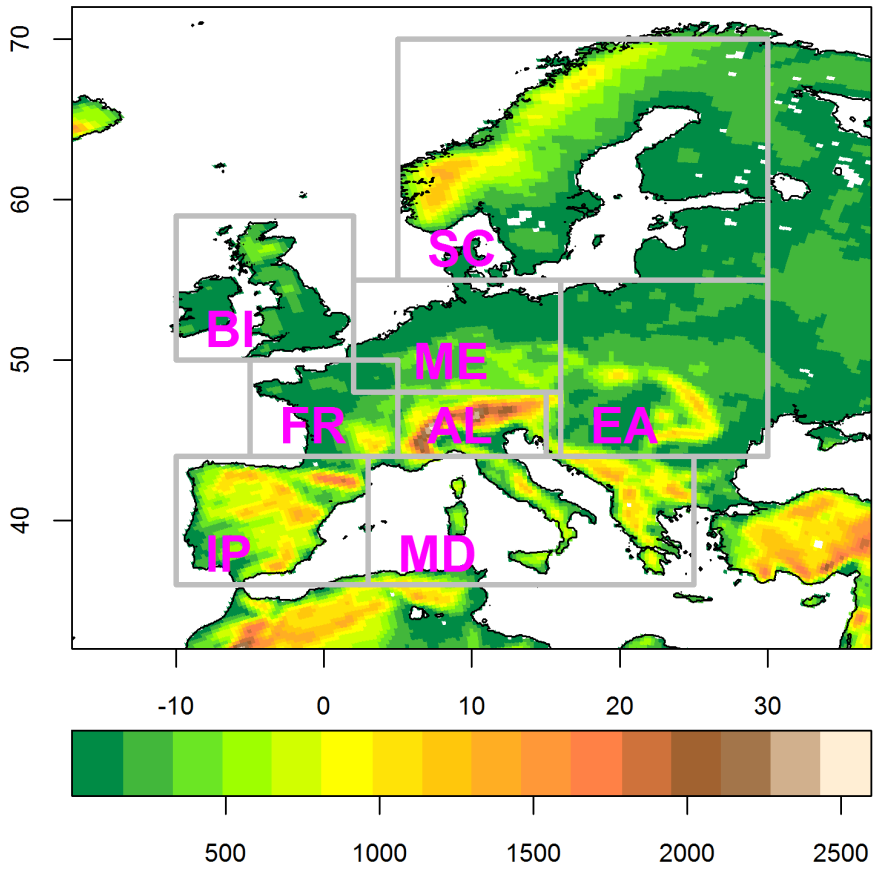


Figure 2.1.: The European sub-areas as defined by Christensen and Christensen [2007]. they will be referred to as "PRUDENCE regions" from now.

Downscaling of decadal hindcasts in MiKlip

At this point, my work pitches in: A unique part of the project MiKlip is the Regionalization of the decadal ensembles. So far it is the only attempt worldwide to do so. From the first ensemble generation 10 realization initialized every ten years between 1961 and 2010 of the MPI-ESM-LR [*Stevens et al.*, 2013] were downscaled to 0.22 degrees using the regional climate models COSMO-CLM (CCLM) [*Doms and Schättler*, 2002] (Hereafter donated as b0-EUR22). COSMO-CLM (COnsortium for Small-scale MOdeling in CLimate Mode) is a non-hydrostatic regional climate model developed from the Local Model (LM) of the German Meteorological Service.

For the decade 2001–2010 the regional model REMO [*Jacob*, 2001] was also used to downscale the 10 realisations starting in 2001 and making the last decade a regional ensemble of 20 members in total. REMO was developed at the Max Planck Institute for Meteorology (MPI-M) on the basis of the former numerical weather prediction model of the German Weather Service (EUROPA-MODELL, EM) [*Jacob*, 2001].

The downscaling of the second ensemble generation (baseline1, hereafter b1) was more extensive. A similar regional ensemble to the b0-EUR22 of 10 members initialized decadally with a resolution of 0.22 degrees was compiled (hereafter b1-EUR22) using only CCLM. Additionally 5 realizations initialized annually between 1961–2001 were downscales with the CCLM to a resolution of 0.44 degrees (b1-EUR44) and 2 realizations with REMO (v. Fig. 2.3) that will not be considered here.

The decadal predictions are compared to the E-OBS v8.0 climatology [*Haylock et al.*, 2008; *EU-FP6 project ENSEMBLES & ECA&D project* , 2016] . E-OBS is a gridded observational data set with a resolution of 25 kilometres. Several observation-based indices like the AMO-Index [*Enfield et al.*, 2001; *NOAA–ESRL*, 2015] or the radiative forcings [*Hansen et al.*, 2005; *NASA/GSFC/HSL*, 2015] are also utilized. More detailed descriptions can be found in the App. A.

The regional ensemble of MiKlip has been analysed in several studies on different aspects. *Mieruch et al.* [2014] found in regards to the b0-EUR22 ensemble that downscaling preserves overall skill. Reliability can be improved by regionalisation for summer temperatures. While *Reyers et al.* [2015] found a high correlation between the wind energy output of the dynamical downscaling as well as the statistical-dynamical downscaling with the observations as well as the abilities to reproduce realistic PDFs of the 10-m winde for most stations in Europe.

My work will expand analyses for all ensemble generations of MiKlip and will answer a few fundamental questions about decadal climate predictions on the way. For some of the following analyses the Europe has to be split into smaller regions. The PRUDENCE regions as defined

by *Christensen and Christensen* [2007] seperated Europe into 8 sub-areas that are more or less homogenous, climatologically speaking. Fig. 2.1 shows the regions.

Some analyses require the data to be bias corrected, e.g. the calculation of threshold based indices as are discussed in Sec. 3.3 [*Zhang et al.*, 2011]. *Sillmann et al.* [2014] suggest a simple mean bias correction. The method involves the removal of the bias in the mean annual cycle of the models and the use of percentile thresholds from a reference data set. First, the annual cycle of the data is estimated for each model. This annual cycle is then subtracted from the original model time series and substituted with the corresponding annual cycle estimated from the observational data set. By doing so the model distribution is shifted to match the observed distribution. In Fig. 2.2 the bias correction is demonstrated using daily maximum temperature from on grid point (of roughly the coordinates of Karlsruhe). The cold bias of CCLM is clearly corrected while the shape of the distribution stayed the same. If now a threshold as derived from the observations (e.g. a percentile) the exceedance thereof in the model simulations is due to differences in the simulated daily temperature variability.

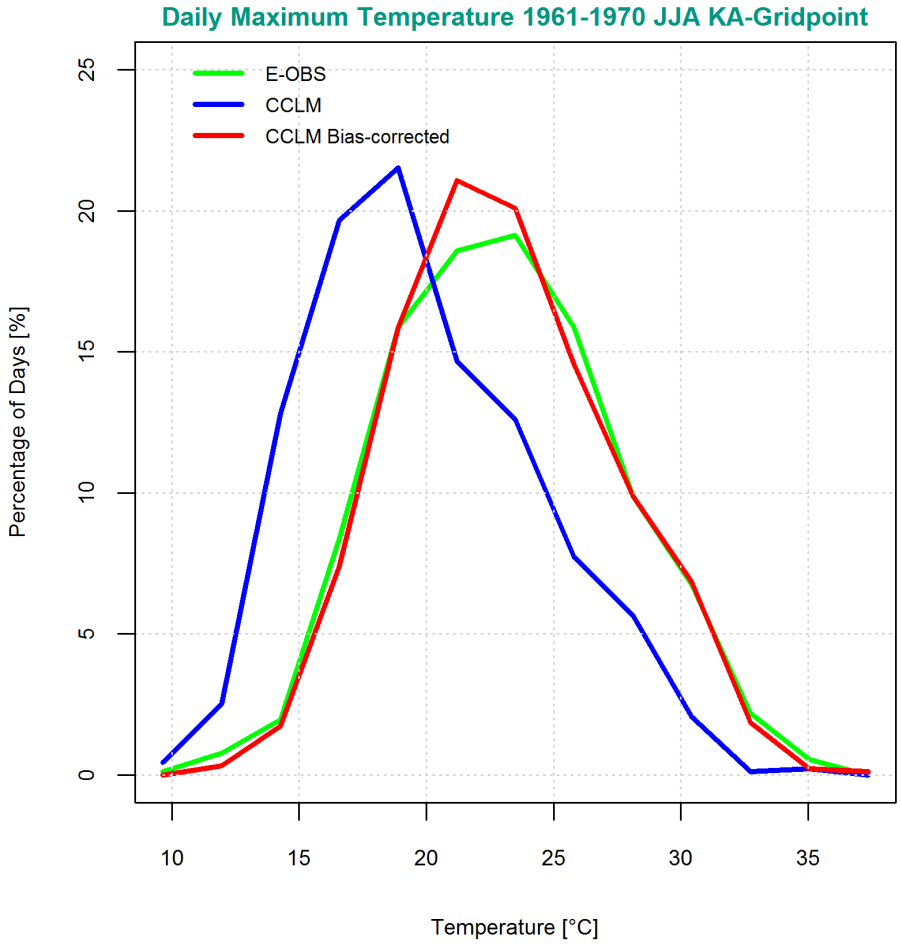
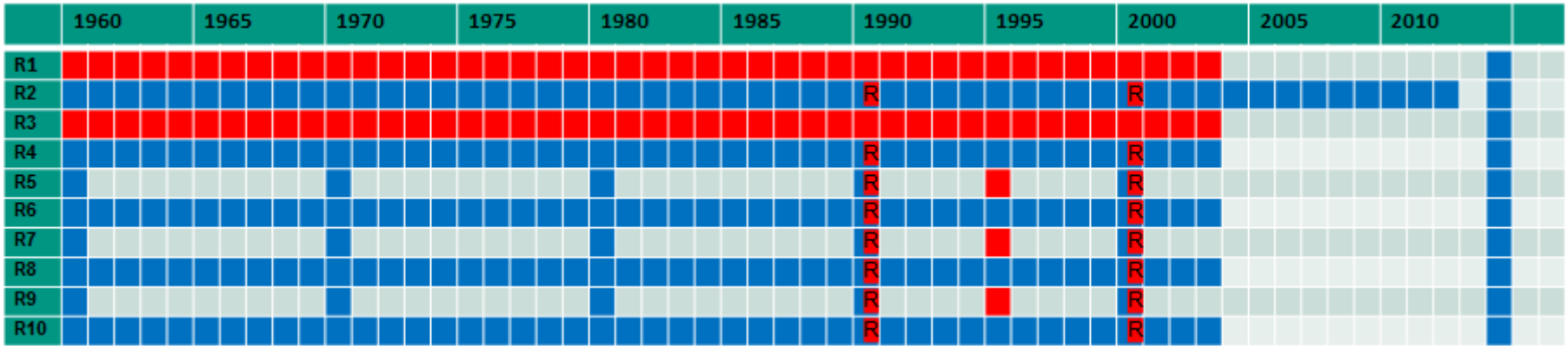
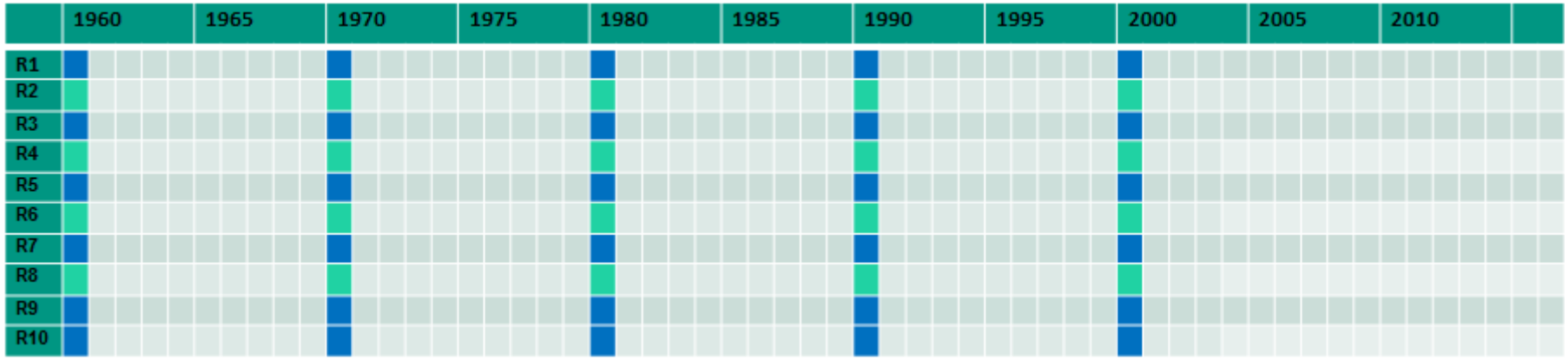


Figure 2.2.: The bias correction suggest in Sillmann et al. [2014] has the effect that while the general form of the distribution stays the mean of corrected data is shifted to match the observed one. The data used to demonstrate this effect is the daily maximum temperature of the grid point of roughly the coordinates of Karlsruhe from the b1-EUR44 CCLM ensemble for the decade 1961–1970.



(a) EUR44



(b) EUR22 & EUR0625/EUR022

Figure 2.3.: Ensemble generation in Miklip Modul C – Regionalization: The tables show the individual runs of the ensemble members (realizations R1 to R10) for three resolutions: In 2.3a 0.44 °, in 2.3b 0.22 °in blue as well as 0.0625 °and 0.22 °in green. The 0.0625 °ensemble will not be considered here. The nomenclature is to be read as follows: A decadal run of of the decade 1960 will start at the First of January 1961 and ends with the 31st of December 1970. In addition to the here shown simulations, a full ten member-ensemble for the decade 2000 (2001-2010) of REMO in the resolution 0.22 °is available.

3. Potential predictability on decadal time scales

Already, the *IPCC* [1990] stressed the importance of decadal-scale climate changes, as such might oppose or reinforce effects of long-term changes in greenhouse gases. Non-linear interactions in the Earth-ocean-atmosphere system lead to those unforced internal climate variabilities. Observations as well as simulations of coupled general circulation models (AOGCM) exhibit an inherent interannual variability including on decadal and longer timescales, e.g. the variation pattern of 30-70 years in the North Atlantic sector associated with the AMO. An in-phase prediction of such implies a potential for skilful predictions on such scales [*IPCC*, 1990; *IPCC*, 2007; *Griffies and Bryan*, 1997; *Latif et al.*, 2006].

Predictions up to 10 years of sea surface temperature of the North Atlantic show promising results [*Müller et al.*, 2012]. The combined effects of the internal and forced drivers of climate in either direction will be superimposed on the trend due to increasing green house gases [*IPCC*, 1990; *Smith et al.*, 2012b]. The interaction of atmosphere and ocean, most likely the source of decadal variability, and its ramifications on non-adjacent regions are neither properly understood nor documented [*Hurrell et al.*, 2010].

Since predictions over land are of particular interest, the question of importance is, whether and to which extend the information is transported from the ocean to the land. As knowledge about the teleconnections is scarce the more appropriate question would be:

Do decadal scale variations over the oceans provide a signal greater than the noise over the continents?

In the case of Europe, the regional climate is not always directly coupled to the variability in ocean temperature. The connections between the climate system components can be indirect and possibly weak.

How can one measure the decadal-scale signal, if neither the teleconnection nor the primary variability within the climate system is fully described? I.e. *Smith et al.* [2012b] took known climate indices, e.g. the Atlantic Multidecadal Oscillation (AMO) or the North Atlantic Oscillation (NAO), and correlated their indices with temperature and precipitation observation over land to reveal teleconnections. A method like this requires full knowledge about the climate system components on the other side of the teleconnection to atmospheric land variables. That is not necessarily always the case. Therefore a different method that does not require the teleconnection but only the variable to be predicted is needed.

Here, the term *potential predictability* will be "defined as variability [...] that exceeds the variability due to [...] stochastic processes" according to *Feng et al.* [2012]. In the case of decadal predictions the two major factors are contributing to this natural variability of atmospheric variables:

1. long term climate trend, that with appropriate emissions scenarios can be highly predictable

2. and teleconnections that couple the different parts of the climate system, both slow (i.e. ocean, soil) and fast (atmosphere).

The main focus here will be the on the second point as it can be assumed that the prediction of a climate trend is feasible [*Knight et al.*, 2014] given a reasonable accurate initialization. Also, the trend due to external forcing and anthropogenic emissions is already the main focus of climate projections [*IPCC*, 1990; *IPCC*, 2007, 2013]. Decadal predictions aim further in that they like to accurately predict internal variability of the climate system.

Therefore, a potentially predictable variable on the decadal scale includes a strong expression of the teleconnections. The variable would have to have a low Signal-To-Noise-Ratio on short time scales, that increases with longer time scales, but has also a big long term trend component and an underlying variability with a multi-year frequency. However, finding such a variable is only half of the story. Another factor playing into the ultimative choosing of variable to predict are the wishes of potential users. A descriptive and comparative test for the potential predictability is therefore needed to fully describe the full potential of the variable as well as compare it quantitatively to others to simplify the choosing process.

Ideally one could characterize the predictability that stems from slow variations in the climate system solely by finding whether or not the frequency spectrum of the predictand includes the frequency of the natural oscillation providing the predictable factor. But how does one accomplish that without knowing all essential slow climate variations and whether the their influence on the predictand is linear and direct? Finding the scale with the biggest signal-to-noise-ratio theoretically requires knowledge about the signal's frequency as well. A correlation with a green house gas radiative forcing or the Atlantic Multidecadal Oscillation index implies they are solely responsible for predictable decadal variability. But teleconnections or climate responses to forcings my not be linear and some sources not known. Together with a non-parametric test however an almost complete picture about a variable could be given.

In this work I will simplify this problem towards an easy test for decadal variability within the predictand variable. Identifying the strength of variabilities with frequencies up to 10 years allows then the comparison of variables and offers an insight into possibilities for the prediction of user relevant variables regardless the source of the predictability and the ability of the current

prediction system to actually forecast them. First, I will lay out the methodology to quantify forecast utility. This analysis is tested in a controlled experiment with predefined synthetical data and then applied to observations of air temperature and precipitation over Europe and the North Atlantic. The final aim is to make a selection of variables that are potentially predictable on the decadal time scale.

3.1. A simple test for decadal variability

Following the definition by *Feng et al.* [2012] (see above) a time series analysis of observational data can yield a quantification of potential predictability.

Assuming the probability distribution p of a forecast that evolves in time, will given a reasonable amount of time approach the an equilibrium distribution q, one can assume the climatology (or the uninitialized runs) as the equilibrium q and the decadal prediction as p. For the prediction to be potentially valuable it should hold more/different information than the climatology because of decadal variability.

In this case the terms "forecast" and "climatology" are used loosely. Instead of making an actual prediction using a GCM, a "perfect model" scenario is assumed in which prediction and observation are in perfect accordance. Defining potential predictability as solely dependable on the observations removes the uncertainty stemming from the model. A "forecast" is any fraction of the observed time series, its entirety defined as "climatology".

One possibility to test for decadal variability could be to compare the distribution of a decadal forecast, i.e. a 10 year slice of the observation, to the distribution of the entire observed time series. If the distributions differ significantly one can assume there to be a decadal component to the forecast [*Meehl et al.*, 2013]. Methods include simple test for diversity of distributions or relative entropy [*Kleeman*, 2002; *Tippett et al.*, 2004].

The χ^2-Test

The most simple distribution test is the χ^2-Test might suffice [*Wilks*, 2006]. Taking the test value. The χ^2-Test however has some deficits. For instance the test theoretically requires knowledge about the number of degrees of freedom. Ideally the distribution parameter provides the degrees of freedom. But the data's distribution is not necessarily known. Also, the χ^2-Test does not allow for empty classes into which the data is sorted. An alternative is the Wilcox-Mann-Whitney-Test [*Wilks*, 2006]. The Wilcox-Mann-Whitney-Test tests primarily for the distribution's location. The test is a cumulative test that allows empty classes. However, the data sets should be similar in size or sufficiently large. That is not given when comparing a 10 year forecast to 50 year climatology

Implementing Relative Entropy

Another way to compare distributions is relative entropy. The term originates from Information Theory and measures the information loss sustained by taking only part of all information available into consideration (i.e. assuming climatology when a forecast is available). In *Kleeman* [2002] and *Branstator and Teng* [2010] relative entropy has been implemented for ensemble simulations. They, too, assumed a perfect model. Due to the uncertainty in the value of the initial conditions, the values are given by a probability distribution p. This distribution evolves in time and will, given a reasonable amount of time, approach the equilibrium distribution q. If one assumes the long time behaviour of the system matches its equilibrium then q can be seen as the climatological distribution.

Now the question of the forecast's utility is asked: How much information is added to a particular situation by its availability? Practically, information of the past or the climatology is already at hand so forecast should add to that.

Information Theory now defines a natural measure to this known as *relative entropy* R. If a discrete set of States i are being predicted R is given by

$$R = \sum_i p_i \times ln\left(\frac{p_i}{q_i}\right), \qquad [3.1]$$

where q_i is the climatological distribution and p_i is the prediction's distribution. R is also known as the *Kullback−Leibler−Distance*. It measures the distance between the distributions and will vanish if both are identical. In the context of atmospheric models, prediction utility always declines with the length of the forecast, due to chaos, as the prediction distribution converges on the equilibrium one.

Relative entropy is always non-negative, as it is a measure for information, it is non-symmetric and non-parametric. There are no requirements for the data to be tested.

Application to synthetical data

To further illustrate the method of using potential predictability synthetically data that exhibit narrowly defined properties is tested for its potential predictability on decadal scales.

In figure 3.1 three synthetical time series are shown. The purple time series is an autoregressive process of the first order (AR(1)) with an underlying multi-decadal oscillation with a wave length of about 25 years. The yellow time series is an autoregressive process of the first order and the grey time series is white noise. All three time series are constructed to have a zero mean value (as would detrended anomalies) and roughly the same standard deviation. They represent 50 years of monthly data, i.e. 600 time steps per time series.

First the relative entropy of a forecast starting at the first vertical line in figure 3.1 is shown in Fig. 3.2 with solid, thick lines of the same colour scheme as in Fig. 3.1. For this test perfect

forecasts are assumed (i.e. error equal zero) and the relative entropy measures the information gain by the availability of the forecast in addition to the climatology. The climatology is assumed to be without trend and stationary, and the whole 50 years available are used. The relative entropy is shown for forecasts of different length from 1 year up to 49 years (abscissa). The purple line represents the forecasts of the multi-decadal oscillation. It becomes obvious that with longer forecast time the relative entropy decreases as the forecast's distribution approaches climatology. But it is not a steady decline. Local maxima can be found at about 10 and about 20 years forecast length. Incidentally, these are forecasts (in the case of the first starting date) that are still short of the first complete cycle of the multi decadal oscillation. Once the forecast length exceeds the wave length of the oscillation it will per definition include the whole climatology and the relative entropy should approach zero, because the synthetical time series is constructed to be only a multi-decadal oscillation and noise. The dotted lines show the 95th percentile of the relative entropy. *Branstator and Teng* [2010] define this as the level of prediction utility.

The yellow and the grey developments of relative entropy are (almost) always lower than the purple line, with the forecasts of the AR(1) process exceeding the white noise. The information quotient of forecasts of a the AR(1) process or the white noise and their respective climatology is lower than the multi decadal oscillation especially for long forecast periods. For short periods the relative entropy is as high as in the case of the multi decadal oscillation. However the relative

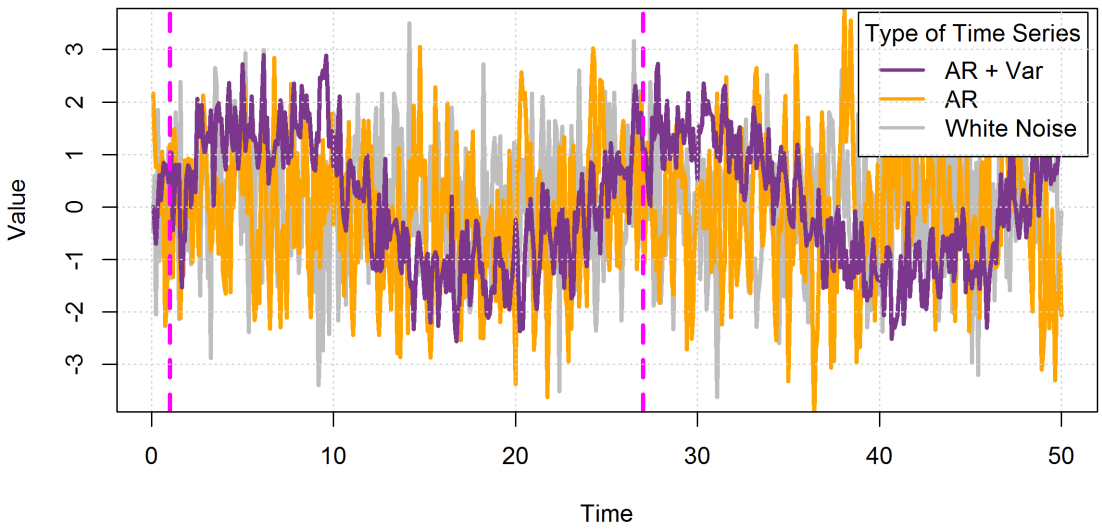


Figure 3.1.: Three different time series will be used to illustrate the concept of relative entropy. On is pure random (white noise, grey line), the second is an autoregressive process of the first order (AR(1), yellow line) and the third is an autoregressive process with a multi decadal oscillation represented by a simple sine function (purple line). The vertical lines indicate the starting dates of the forecasts to be later analysed .

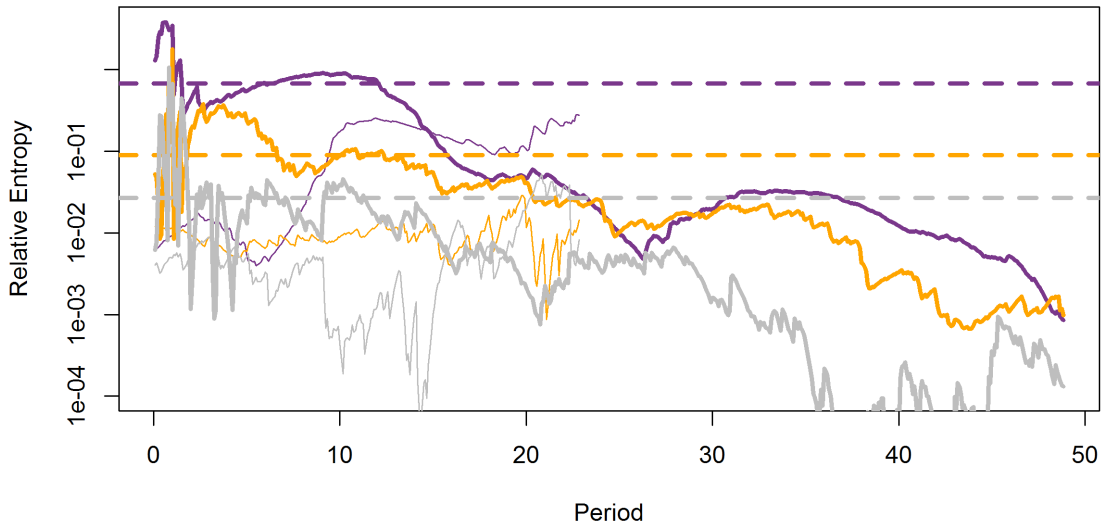


Figure 3.2.: The relative entropy of forecasts starting at the first vertical line in figure 3.1 (solid thick lines) and the second vertical line in figure 3.1 (solid thin lines) are shown. In this case, the relative entropy measures the information gain of the forecast opposed to the climatology (here the whole time series). All possible forecast lengths from 1 year up to 49 years (solid thick lines) are shown. In the case of the second starting date of figure 3.1 the possible forecast length are shortened.

entropy declines faster, since even with relatively short forecasts in the case of white noise and an AR(1) process the climatology is then already covered.

The thin solid lines in Fig. 3.2 highlight the influence different starting dates can have. Here the second vertical line in Fig. 3.1 was used as the starting date. However some characteristics do apply, too. The thin purple line of second forecast of the multi decadal oscillation, too, has a local maximum at about 10 years and another could be at about 20 (However the forecast ends here due to the late starting date, hence the second maximum is not definite). The yellow and the grey thin lines of the AR(1) and the white noise forecasts indicate even smaller relative entropy than the forecasts from the first starting date.

Relative entropy has no unit and is not a quantitative measure, but can be used in comparisons. In this hypothetical case one would argue quite rightfully for the utility of the multi decadal oscillation to be predictable on the decadal time-scale as opposed to the AR(1) process and the white noise. Both are time series that are less deterministic and therefore less predictable on a multi year scale.

But are these two starting dates representative for the whole time series? In Fig. 3.3 the average of all possible forecasts of the length 1 to 40 within the time series' 50 years are plotted in the same manner as in Fig. 3.2. I.e. the number of 1 year-long forecasts possible in this

time series is 589. The number of 40 year time slices, that can be fitted into the time series, is significantly lower: 121.The curves are smoother compared to to individual forecasts in Fig. 3.3 but show a similar pattern: the potential of the multi decadal oscillation is higher than the AR(1) process' and the white noise's forecasts. However the length of utility is for all 3 forecasts the same, as the relative entropy decreases past the 95th percentile within about 7 years. Comparing all three forecasts, the potential predictability of the multi decadal is highest for all forecast lengths.

Relative entropy has been used to determine forecast utility before [*Kleeman*, 2002; *Abramov et al.*, 2005; *Branstator and Teng*, 2010; *Meehl et al.*, 2013]. These earlier studies applied relative entropy to ensemble predictions to measure e.g. the influence of initialization, by establishing the forecast length with which the information content is similar to that of climatology essentially when i.a. the green house gas effect becomes a driving factor of the climate. These studies and the previous example of synthetical data show that relative entropy can answer two questions:

1. On which forecast lengths does a variable indicate potential predictability stemming from initialization?

2. When analysing a fixed forecast length for several variables, which shows the most potential?

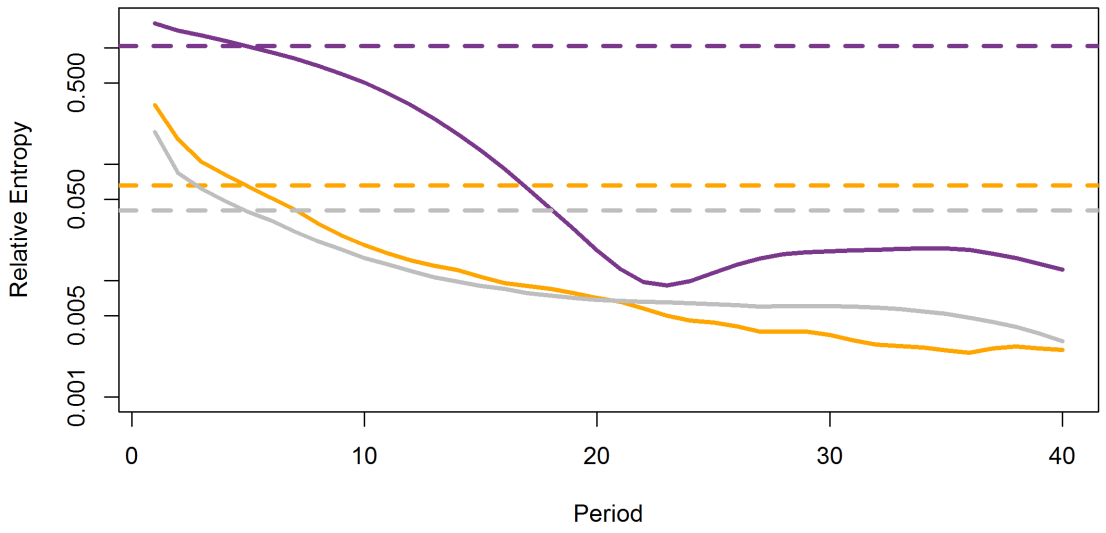


Figure 3.3.: The mean relative entropy of forecasts of the length of 1 year up to 40 years. The relative entropy is averaged over every value of all possible forecasts: in the case of 1 year forecasts there are fifty available, but there are only ten 40 year forecasts. The dotted horizontal lines indicate the 95 th percentile of the entropy distribution.

Relative entropy can be used to compare forecast lengths and variables. Together with pre-existing knowledge about teleconnections, an educated choice about variables with decadal potential is doable.

3.2. The potential predictability of exemplary meteorological time series

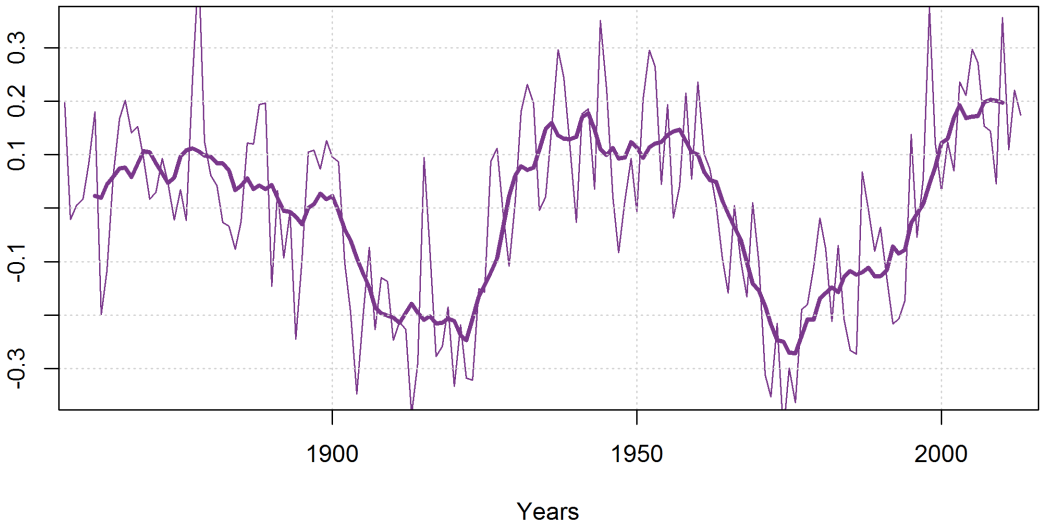
The analysis of the potential predictability of synthectical data a circle of concept, its outcome implied in the data's synthesis. Real meteorological data, however, in infinitely less homogeneous. The real test for this methodology is the application to real time series. This first look into the potential for decadal predictions over Europe includes at two things: The predictability of the mechanisms believed to be the main source of decadal variability in this region and that of a main variable such as temperature.

In Fig. 3.4 the AMO-Index as calculated from the Kaplan SST V2 dataset [*Enfield et al.*, 2001; *NOAA–ESRL*, 2015] is shown in annual resolution and smoothed with a 9 year running mean. The lower panel is the average relative entropy of potential forecasts of various forecast lengths. Again the relative entropy decreases with increasing forecast length as the theory states. The relative entropy of the smoothed AMO-Index (solid line) is higher than the annual AMO-Index (thin line), as smoothing reduces noise. The 95th percentile (dotted vertical line) is reached after a forecast length of about 12 years. 12 years is also within the range what the literature states for the predictability of the AMO [*Knight et al.*, 2005, 2006; *Teng et al.*, 2011]. Therefore, the potential decadal predictability of the source of decadal variability is proven.

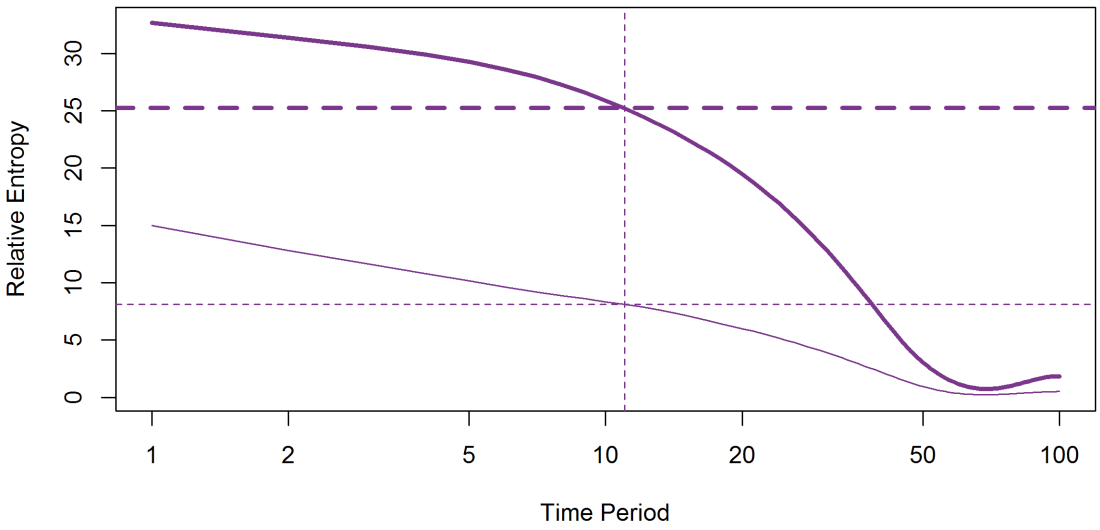
As a next step a temperature time series over Europe will be analysed in the same way to check if there is information transport via teleconnections from the ocean towards land. A grid point with the rough coordinates of Karlsruhe (about 49°North and 8°East) was chosen from the E-OBS V8 [*Haylock et al.*, 2008; *EU-FP6 project ENSEMBLES & ECA&D project* , 2016] observational dataset and is plotted as a detrended daily time series in the upper panel of Fig. 3.5. The lower panel shows the average relative entropy of various forecast lengths. In this case the relative entropy decreases below the 95 percentile (dotted vertical line) after 5 years. The temperature is less predictable than the AMO index, due to information loss during its transport from ocean to land and the temperature being only partially dependent on the teleconnection to the North Atlantic. However the first 5 years of the decadal prediction are worth forecasting in the case of this temperature time series. Please note that, this result is specific to the one location. While mean potential predictabilities for temperature in many PRUDENCE regions of around 5 years have been found (not shown), results can differ locally.

Another point to make here concerns the trend of time series. The implementation of relative entropy does theoretically require stationary time series. In the case of the non-detrended time series (thin solid line) the relative entropy is smaller for all forecast lengths (not shown). The requirement of stationarity does not diminish the applicability of this methodology. The long term trend due to external factors is valuable piece of information. Since climate projections are able to predict the trend, the improvement made by decadal climate predictions should be the accurate representation of amplitude and frequency of multi-year natural variabilities. Following

this distinction between climate projections and climate predictions, the potential for worthwhile decadal predictions is the existence of such low-frequency variabilities, regardless of the model system's abilities to reproduce them. In all following cases, only detrended time series will be considered.

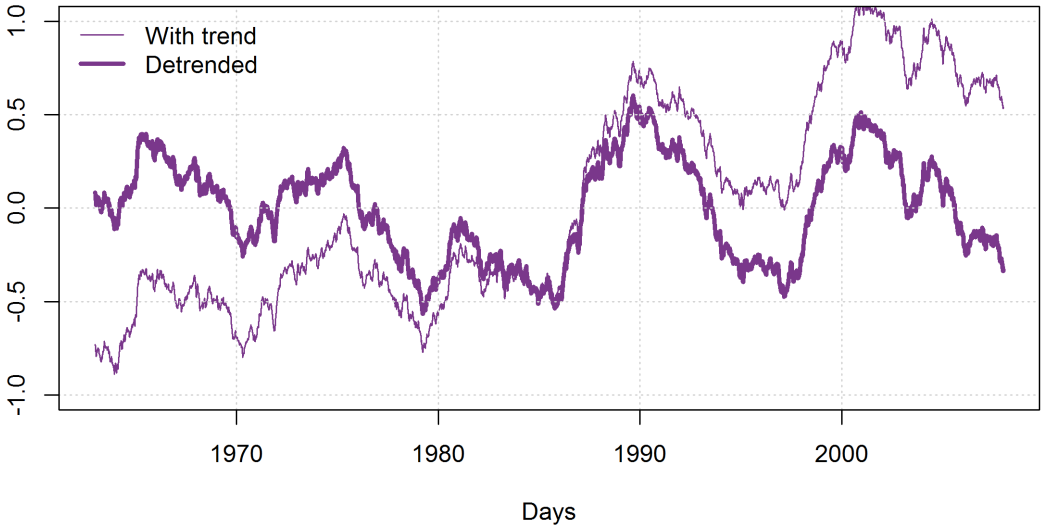


(a) AMO Index

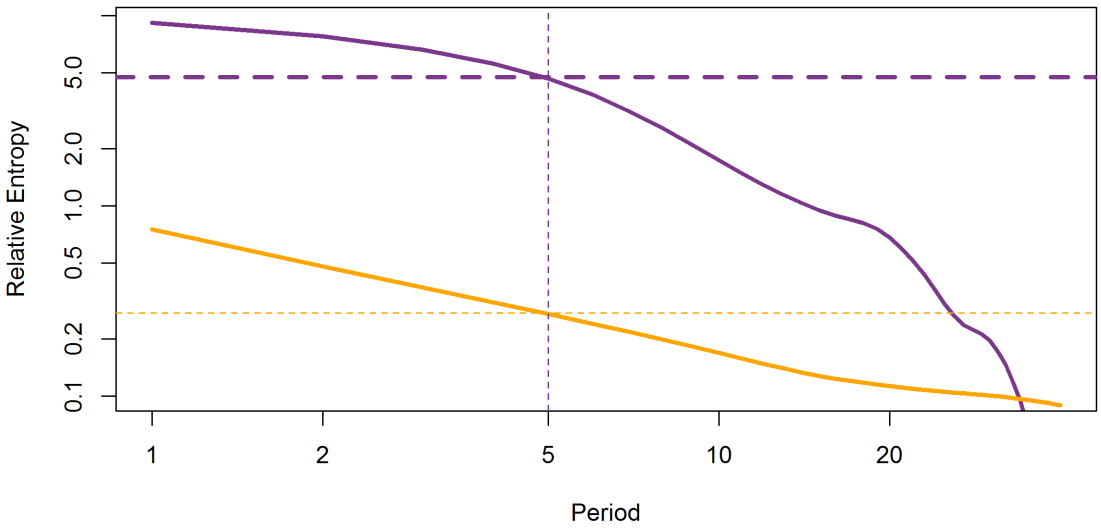


(b) Relative Entropy of the AMO

Figure 3.4.: 158 years of annual Atlantic Multidecadal Oscillation Index (1856–2014) from the Kaplan SST V2 dataset [Enfield et al., 2001; NOAA–ESRL, 2015] (thin solid line) and smoothed with a 9 year running mean (thick solid line) (upper panel) and its average relative entropy of a potential forecast for different forecast lengths (abscissa) (lower panel) is plotted. The vertical line indicates the forecast length at which the relative entropy drops below the 95th percentile.



(a) Daily mean temperature



(b) Relative Entropy of the temperature

Figure 3.5.: 50 years of detrended daily mean temperature at the grid point Karlsruhe (about 49° North and 8° East) from the E-OBS V8 dataset [Haylock et al., 2008; EU-FP6 project ENSEMBLES & ECA&D project , 2016] (thick solid line) and with trend (thin solid line) (upper panel) and its average relative entropy of a potential forecast for different forecast lengths (abscissa) (lower panel) are plotted. In 3.5b the purple line donates indicates the relative entropy as derived from the detrended daily mean temperature filtered by a 5-year running mean (thick purple line in 3.5a, the orange curve is the relative entropy derived from the detrended (unfiltered) daily mean temperature (time series not shown).

3.3. Potential predictability of climate extreme indices in Middle Europe

There is agreement that changes in the frequency or intensity of extreme weather or climate events have a profound impact on environments and societies [*Karl et al.*, 1999; *Eade et al.*, 2012]. So can, for example, the increasing occurrence of extremely warm nights lead to increased mortality and heavy rain to economic loss through flooding. While the assessments of decadal predictions justifiably focus mainly on mean values, the impact of climate change is going to be most severe by the superposition of internal variability and climate change. Additionally, the analysis of extremes requires high resolution observational data sets and model outputs, as well as long time series due to their rare occurrences: Obstacles, that the MiKlip regional decadal prediction system overcomes.

It goes without saying that a skill full prediction of mean values is necessary to predict extremes, since a shift in the mean is likely to be accompanied by changes in the probabilities of extremes. But on a decadal scale both the internal variability and the long-term trend force the progression of a variable. A change in extremes is therefore only partly proportional to the change in the mean. I theorise that the potential predictability of extremes varies with the kind of extremes and is not necessarily the same of its mean value.

After having demonstrated the method of using relative entropy for the quantification of potential predictability in Sec. 3.2, the application of relative entropy will now be expanded. The goal is a selection of theoretically (on decadal time scales) predictable variables by comparing the information they carry in a 10-year forecast. On top of that, further methods as described in Sec. 3.1 are applied. The combined results will lead to a ranking of the predictability of various variables.

The method relative entropy to study the potential predictability uses only observations. The E-OBS observational data set (v. Ch. 2 & App. A.2, *Haylock et al.* [2008]; *EU-FP6 project ENSEMBLES & ECA&D project* [2016]) provides over 60 years of gridded observations of temperature, precipitation and pressure. In this instance both daily temperature and precipitation were used to derive indices using the definitions recommended by the CCl/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI) [*Karl et al.*, 1999; *Peterson et al.*, 2001; *Zhang et al.*, 2005; *CLIMDEX*, 2013] that have been adopted by several projects and organizations including the IPCC [*IPCC*, 2007].

Using relative entropy requires a high data density which is not given by the annual resolved climate index data. Therefore the European land grid points are partitioned into the PRUDENCE regions [*Christensen and Christensen*, 2007]. It is assumed that time series within these regions are climatologically similar enough to form a big data pool for the following statistical analysis. On top of relative entropy several other statistical analyses are deployed:

Table 3.1.: Summary of all extreme climate indices calculated from E-OBS [Haylock et al., 2008; EU-FP6 project ENSEMBLES & ECA&D project , 2016] used in this thesis; as advised by the project CLIMDEX [2013] [Karl et al., 1999; Peterson et al., 2001; Zhang et al., 2005] in addition to the new index TQI (Ticks Questing Index) (v. App. B).

Index	Variable	Definition
Icing Days (IC)	Maximum temperature TX	Annual count of days with TX<0°C
Frost Days (FD)	Minimum temperature TN	Annual count of days with TN<0°C
Tropical Nights (TR)	Minimum temperature TN	Annual count of days with TN>20°C
Summer Days (SU)	Maximum temperature TX	Annual count of days with TX>25°C
Growing Season Length (GSL)	Daily mean temperature TG	Number of days between first span of at least 6 days with TG>5°C and first span after July 1st of 6 days with TG<5°C
Tick Questing Index (TQI) (v. B)	Daily mean temperature TG & daily precipitation PR	Number of days with TG>7°C after at least 7 consecutive days with TG>7°C and daily precipitation pr>1mm

Table 3.2.: Summary of all percentiles of E-OBS variables [Haylock et al., 2008; EU-FP6 project ENSEMBLES & ECA&D project , 2016] used in this thesis

Variable	Summer Percentile (JJA)		Winter Percentile (DJF)	
daily maximum temperature	5 %	95%	5 %	95%
daily minimum temperature	5 %	95%	5 %	95%
daily precipitation	5 %	95%	5 %	95%

1. The correlation with the AMO index [*Enfield et al.*, 2001; *NOAA–ESRL*, 2015] in accordance to the analysis by *Smith et al.* [2012b]. The index is an example for a main driving factor of decadal predictability and its own potential predictability has been proven in the previous section.

2. The time series of the climate extreme indices are filtered using a running mean of 3 to 19 years and the filter length that procures the highest signal to noise ratio (v. App. C) will be used as an indicator for the frequency length of internal variabilities.

3. The second local maximum of the frequency density spectrum of the data is also considered. The second maximum was chosen as the first will inevitably show the inter-anually variability.

4. The level of significance of the χ^2 test for diversity of the distribution of the ten year forecast versus climatology as described in Ch. 3.1 is considered, too.

5. Further more the correlation of the climate extreme indices with historical internal and external global mean effective forcings [*Hansen et al.*, 2005; *NASA/GSFC/HSL*, 2015] and other climate indices (e.g. NAO index, ENSO index etc. (v. App. A.5, A.4) are also considered.

All analysis are undertaken for either each grid point and then averaged over the PRUDENCE regions (correlations) or done for the entirety of the PRUDENCE regions if the methodology requires a high data density (e.g. relative entropy and χ^2-Test)

A variable is hypothetically most suitable for predictions on the decadal scale, if it shows a high explained variance in the case of the AMO and less in the case of the forcings of anthropogenic green house gas emissions. That indicates a strong connection to a multi-decadal variability and weak dependency on the long term climate trend. While both mechanisms are predictable, only the first benefits from an initialization of climate simulations, as opposed to the latter that can be covered by climate projections already in use. In reality, such a simple solution will not occur. Variables over European land especially like temperature are highly dependent on the change in anthropogenic green house gas forcings as observations have shown. Also, both the AMO and the green house gas emission forcing are not easily separated. Especially the increase in global mean temperature in the last few decades that coincides with an increase of the effective forcing due to anthropogenic green house gas emissions also attends a shift towards a positive AMO phase as Fig. 1.2 shows. A positive AMO phase indicates higher than average sea surface temperature in the North Atlantic and strong coupling with the atmosphere over Europe. A change in climate over Europe could therefore very well be a result of multiple complimentary factors whose percentile share cannot be quantified.

A suitable variable would also include a signal to noise ratio that favours large filter windows indicating an internal variability of low frequency. On top of that, the local maximum of the indix' frequency density spectrum would also be located at a low wave number, signifying a slow-acting internal variability. Relative entropy and the significance of distribution diversity are as the name of the former suggest relative measures. Relative entropy has no unit. Both methods are used in comparison. In this analysis they were used to compare e.g. the climate extreme indices altogether and within their respective seasons or variables.

In App. G an example is given for results for the PRUDENCE region Middle Europe. In table 3.3 a summary of some indices and some results for the PRUDENCE region Middle Europe

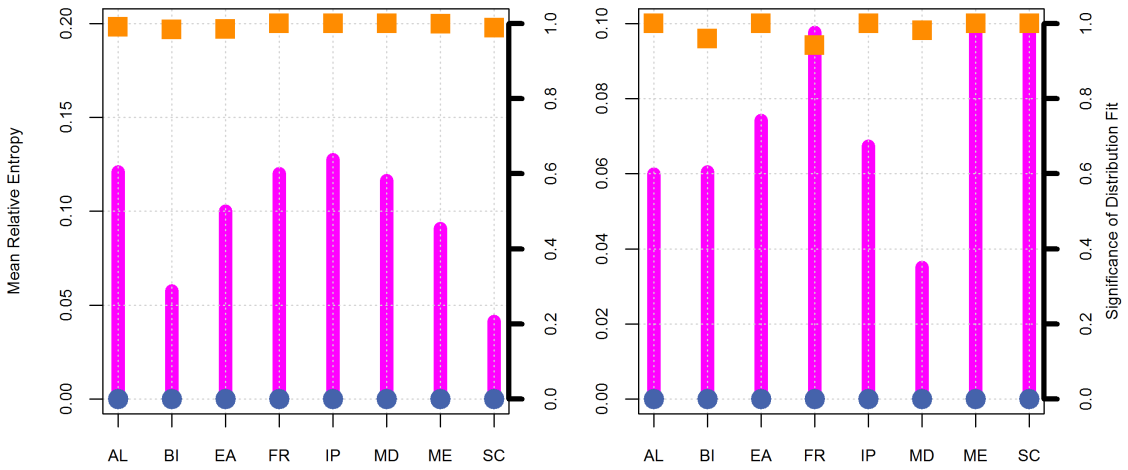
(ME) is shown. In a next step the indices were ranked, where as the best three results of each methodology were given 3, 2, and 1 points respectively (in red). The sum of these points rank the indices. The index Tropical Nights (TR) is not considered in the ranking, as a daily minimum temperature of over 20°C is very rare in the gridded observation data of 25 kilometre resolution. The data density too sparse for the analyses to be performed.

Table 3.3.: Several of the indices recommended by the CCl/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI) [Karl et al., 1999; Peterson et al., 2001; Zhang et al., 2005; CLIMDEX, 2013]are analysed using several statistics: The filter length that produces the maximum of the signal to noise ratio (SNR-Max), the location of the second local maximum of the spectrum (locMax), the relative entropy of the 10-year forecasts vs. the climatology (relE), the significance level the χ^2-Test of distribution diversity of the 10-year forecasts versus the climatology (χ^2-), as well as the correlation with the AMO index [Enfield et al., 2001; NOAA–ESRL, 2015] (rAMO) and an index of global green house gas forcing [Hansen et al., 2005; NASA/GSFC/HSL, 2015] (rGHG). The three best results in each category are awarded points (in red), the sums of these points rank the indices. Note the index TR (Tropical Nights) is not considered in the ranking, as a daily minimum temperature above 20 °C occurs rarely and most analyses are not feasible.

Index	SNR-Max [a]	locMax [a]	relE	χ^2	rAMO	rGHG	Rank
GSL	8 (3)	6	1.86 (1)	0.289 (2)	0.797 (3)	0.765	2
TQI	7 (2)	8 (1)	1.88 (2)	0.27 (1)	0.691 (2)	0.152	3
TR	5	13					-
SU	3	14 (2)	1.92 (3)	0.3 (3)	0.088	0.042 (3)	1
FD	7 (2)	8 (1)	1.88 (2)	0.27 (1)	0.094	-0.156	4
ID	7 (2)	16 (3)	1.76	0.22	0.248	-0.092	5
pr95	5 (1)	7	1.71	0.15	0.233	-0.055 (1)	7
pr90	5 (1)	7	1.76	0.19	0.257 (1)	-0.045 (2)	6

In comparison of the analyses' results for different PRUDENCE regions a regional bias becomes apparent (v. App. G). E.g. for the PRUDENCE region Scandinavia (SC) the analyses cannot be performed for the index Summer Days (SU) as its threshold is rarely reached. Reversely the same becomes an issue when considering the winter indices in southern PRUDENCE regions. A prudent approach would be to consider percentile of temperature instead of threshold variables. In figure 3.6 the relative Entropy of both the 90th percentile of daily minimum and maximum temperature in winter and summer for all 8 PRUDENCE regions are shown. While

the analysis can always be performed when percentiles are used, still, summer temperatures are more predictable in the southern PRUDENCE regions and winter temperatures present a higher relative entropy in the northern PRUDENCE regions.



(a) tx90JJA (b) tn90DJF

Figure 3.6.: Mean Relative Entropy (magenta bars) & Significance of Distribution Fit (orange squares) of 10 year forecast versus climatology of the 90th percentile of daily maximum temperature and daily minimum temperature in summer and winter respectively. The purple circles show the significance of distribution fit as derived from the Wilcox-Mann-Whitney-Test. This test is not suitable for this analysis as it mainly test for the location of the distributions, which are virtually the same when considering the detrended anomalies of climatology and a forecast.

While the ETCCDI indices are very basic and could be universally applied, the threshold indices are defined with a tempered climate in mind: E.g. the winter indices occur rarely in the regions with frost less winters (with few exceptions), as is the growing season length virtually 365 days long in some regions. I will focus on the PRUDENCE region Middle Europe as summarized in table 3.3 from here on in. For most issues there seems to be an agreement over most regions and the following statements can be cautiously generalised.

First of all, the winter variables show lower potential predictability in almost all regions (exception is Scandinavia (SC)) than summer variables. An explanation could be the higher influence of the North Atlanic Oscillation (NAO) on European climate in Winter [*Hurrell*, 1996]. The NAO is a noisy extra tropical phenomenon that may be possible to forecast with some measurable skill. But the NAO has a relatively short frequency of about 2 years which is significantly shorter than the 10-year forecasts considered here. From the decadal perspective a NAO influenced time series could appear noisy and thus less deterministic.

Another factor into the predictability of climate indices is the severity of the extreme. Moderate extremes are favoured in predictability against their more extreme counter parts (e.g. Summer Days versus Tropical Nights, 90th percentile versus the 95th percentile). Also high in potential predictability are the two indices derived from the mean daily temperature: The Growing Season Length (GSL) and the Ticks Questing Index (TQI) are ranked second and third in this analysis. Both indices represent a "moderate" variable but also a memory of this variable. The memory component means essentially a form of a filter (running mean) on the originally variable. So the noise is reduced and the index becomes more deterministic than the variable it is derived from.

Unsurprisingly the precipitation percentiles are less predictable than the respective temperature percentiles. The combination of temperature and precipitation adds predictability significantly (v. Ticks Questing Index).

Conclusion

On important question to answer before undertaking the great work load of running climate models includes what is potentially predictable in the specific context. To assess the potential of predictions up to 10 years that stems from internal climate variability, I introduced the measure of relative entropy. Relative entropy was applied to first synthetical time series to demonstrate the approach and second to actual observations. An analysis of the AMO index, an internal climate oscillation likely to be a major driver of European climate, shows potential to be predictable up to 12 years. The translation of that information into the temperature over Europe decreases the predictability.

In a third step, relative entropy was used to compare several climate indices derived from temperature and precipitation observations. The conclusions drawn from this analysis, mainly that the most potentially predictable variables on a decadal time scale are moderate summer variables, should be generalized only with reservation. Only the case of a forecast of 10 years was considered here, as it is currently custom within the MiKlip project. Studies have suggested that the main skill of decadal predictions can be found in the first pentade. With shorter forecasts the source of predictability may shift towards components of higher frequency and hence other variables. Should the custom change, further investigation into the predictability of variables is in order. Also, there can be a potential for predictability in the second pentade from the higher influence of external factors, e.g. the changes in the green house gas concentration, which this Analysis does not capture due to its use of detrended time series.

Thereby, I demonstrated a universal way of quantifying the potential predictability of variables independent of model performance which will be the focus in the following chapters. First the overall performance of the prediction system used in MiKlip in reference to main variables is discussed before the focus shifts to the prediction of extremes.

4. Verification of Regional Decadal Predictions

Forecasts are issued by many professionals including meteorologist and climatologist to predict the future for a wide variety of purposes. A forecast is like an experiment: One hypothesises a certain outcome given a set of conditions. But an experiment is not complete until its outcome is determined. In the same way, no forecast is complete without finding out if the forecast was successful.

But what makes an a forecast "good"? *Murphy* [1991] distinguished three types of goodness:

- Consistency: The degree to which the forecast corresponds to the forecaster's best judgement about the situation.

- Value: The degree to which the forecast helps a decision maker.

- Quality: The degree to which the forecast corresponds to what actually happened.

The first and second type of "goodness" can not be easily quantified. Although, the usefulness (value) of a forecast was briefly touched upon in the previous chapter. The third type of "goodness", however, the *forecast quality* can be objectively measured. After having discussed the potential for skilful decadal predictions the next logical step is the measurable "goodness" of such a prediction. But forecast quality has many aspects and the quality of a set of forecasts cannot be adequately summarized by a single metric [*Murphy*, 1991, 1993]. [*Murphy*, 1991] described several different "attributes" of forecast quality (v. Tab. 4.1).

Overall skill could be expressed as *bias*, the correspondence between mean forecast and mean observation, *association*, the strength of the linear relationship between the forecasts and observations, and *accuracy*, the level of agreement between the forecast and the observations whereas the difference between the forecast and the observation is the error. All three of these attributes can be combined into one category (*accuracy*) and into metrics, too.

Furthermore, there is *reliability*, average agreement between the forecast values and the observed values conditional on the observation, *resolution*, the ability of the forecast to sort or resolve the set of events into subsets with different frequency distributions, *sharpness*, the tendency of the forecast to predict extreme values, and *discrimination*, the ability of the forecast to discriminate among observations [*Murphy*, 1991; *Murphy and Winkler*, 1992; *Wilks*, 2006]. One could also consider the *uncertainty* of a forecast which is the variability of the observations.

The field of climate predictions is still young and most publications follow *Goddard et al.* [2013] as a guideline for the verification of decadal predictions. *Goddard et al.* [2013] concentrates on the attributes of accuracy and reliability. Hence both will be addressed here, too, to ensure comparability. On top of that, another property of predictions, discrimination, will be considered.

In the following sections a verification framework for regional decadal prediction will be lay out and later applied to the MiKlip ensembles. First the overall skill of mean variables temperature and precipitation are verified as a baseline for further analysis of Climate Extremes Indices (ETCCDI).

Table 4.1.: Definitions of attributes of forecast quality with relevant measures relating to each attribute. (After Wilks [2006] and Murphy and Winkler [1992])

Name	Definition	Graphs and Measures
Accuracy	Overall Skill to which forecasts correspond to observations: Scalar measures are meant to summarize the overall quality of a set of forecasts in one single number	MSE, Correlation
Unconditional Bias	Correspondence between the average forecast and the average observed value of the predictand	
Calibration, Reliability, Conditional Bias	Relationship of the forecasts to the average observation conditional on the forecast	CRPS
Refinement, Sharpness	Unconditional distribution of the forecasts in the calibration-refinement factorization	Graphs
Resolution	Degree to which forecasts sort observed events into groups that are different from each other	
Discrimination	Degree to which forecasts discriminate between occasions	GDSS

4.1. Estimating predictive skill

Skill is the relative accuracy of the forecast over some reference forecast. In comparison to the commonly used correlation coefficient that compares two data sets, this scaled representation of quality takes into account the value of a forecast over an unskilled forecast. Such an unskilled

reference forecast could be random chance, persistence (the most recent set of observations), or climatology, and are easier (cheaper) to obtain than runs of complex models. Every forecast has to be justified by proving its better than the unskilled forecast.

On way of quantifying skill is the use of skill scores. They are based on the Eq. [4.1]

$$\text{Skill Score } SS = 1 - \frac{S_{forecast}}{S_{reference}}, \tag{4.1}$$

whereas S is a measure of error or association, and are universally applicable to many aspects of forecast quality.

Accuracy: Anomaly Correlation

Most commonly used, the *Anomaly Correlation* will be applied to the MiKlip ensembles in addition to skill scores to measure the overall skill of a forecast. A correlation is readily conceivable and simple methods for the test for significance are known and easy to implement. The anomaly correlation (AC) is a measure of association that compares anomaly values of forecast and observation in time. Anomalies in this case are the values derived by subtracting the climatological mean *Wilks* [2006]. There are two forms of the AC. In the following the *uncentred anomaly correlation* will be used. The *uncentered anomaly correlation coefficient* is calculated in accordance of the Pearson correlation coefficient where in contrast to the *centred anomaly correlation* the average over a given map of grid points is not subtracted.

$$AC_U = \frac{\sum_1^t (y_t - \bar{y})(o_t - \bar{o})}{\left[\sum_1^t (y_t - \bar{y})^2 \sum_1^t (o_t - \bar{o})^2\right]^{1/2}} \tag{4.2}$$

Accuracy: Mean Square Error Skill Score

The *Mean Squared Error Skill Score* (MSSS) is one measure used to describe overall skill. The MSSS does not only reflect association between two time series as does the correlation but takes unconditional bias (mean error) into account, too.

The MSSS is defined as follows:

$$MSSS = 1 - \frac{MSE_{fcst}}{MSE_{ref}} \tag{4.3}$$

If the reference is just the climatological mean, the MSSS is composed by 2 terms:

$$MSSS(f, \bar{x}, o) = r_{f,o}^2 - \left[r_{f,o} - \left(\frac{s_o}{s_f}\right)\right]^2 \tag{4.4}$$

Whereas the first therm is the correlation between the forecast and the observation and the second term a conditional prediction bias. The correlation gives a measure of potential skill and the conditional bias represents the magnitude of the observation given the prediction [*Murphy*,

1988]. If the reference forecast is not climatology but e.g. the unitialized runs the MSSS could be rewritten as:

$$MSSS(f,r,o) = MSSS(f,\bar{x},o) - MSSS(r,\bar{x},o) \quad [4.5]$$

$$MSSS(f,r,o) = r_{f,o}^2 - \left[r_{f,o} - \left(\frac{s_o}{s_f}\right)\right]^2 - r_{r,o}^2 + \left[r_{r,o} - \left(\frac{s_o}{s_r}\right)\right]^2 \quad [4.6]$$

The perfect score would be 1. A score of 0 or lower indicates no skill, meaning the reference forecast would be a better choice in predicting the variable.

A problem with the interpretation of the MSSS could be the fact that it does show the skill in accuracy, association and bias. Considering the case of a small correlation - making the r^2-component negligibly small - the main information of the MSSS is the bias. Is then a positive MSSS an indicator for a good forecast given that the bias is small but the predicted variable's development in time can not be associated with the observation?

The answer depends on many factors. E.g. in the field of decadal predictions the accurate representation of the year to year variability is perhaps neither needed nor even intended. A bad correlation between prediction and forecast of an annual variable may even be expected. A small bias in the annual prediction could mean a skilful forecast if a small bias was the intended target. One might even reduce noise with averaging over multiple years, which usually increases association. When then a multi-year forecast is considered, small bias and greater correlation will support the original score and lead to the conclusion, that the forecast under its specific terms was indeed successful.

Reliability: Continuous Ranked Probability Score

The example above shows that forecast skill is not a one-to-one issue. The methods used to verify forecasts depend on the type of forecasts, the verification attributes that can provide answers to questions of interest and the users and their application of the forecast information. Therefore more attributes of forecast quality should be covered by a comprehensive verification framework.

On of these basic property that contributes to the quality of a probabilistic prediction system, such as an ensemble by its very nature is, is the so-called *reliability* [*Wilks*, 2006; *Casati et al.*, 2008]. Reliability is the statistical consistency between the predicted probabilities and the subsequent observations. Various spread-skill relationships have been defined, that essentially measure reliability (although the global spread of the individual predicted spreads can be a measure of resolution). Reliability pertains the relationship of the forecast to the average observation, for specific values of (i.e., conditional on) the forecast. Reliability statistics sort the forecast/observation pairs into groups according to the value of the forecast variable, and

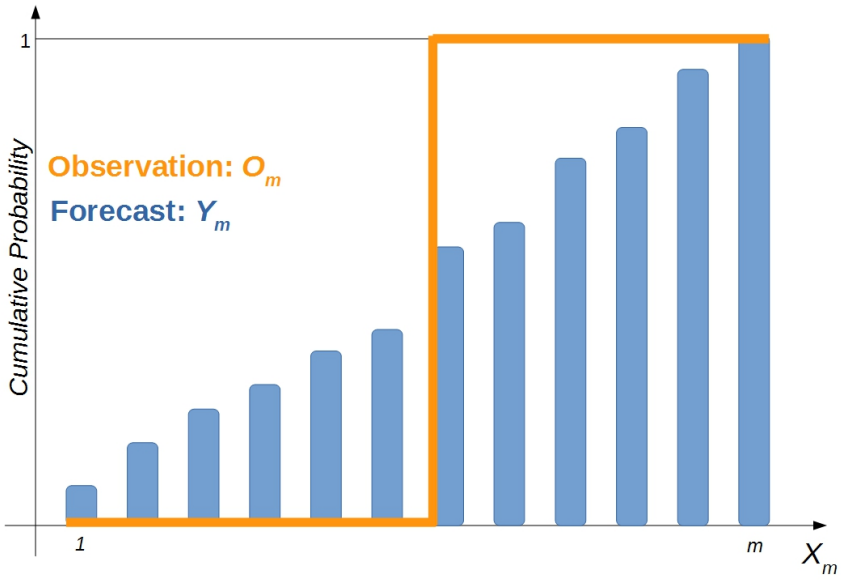


Figure 4.1.: Concept of the Ranked Probability Score after Wilks [2006].

characterize the conditional distributions of the observations given the forecasts [*Wilks*, 2006]. Means to determine the reliability are the talagrand diagram or ranked probability scores among others.

The *Ranked Probability Score* (RPS) is a generalization of the Brier score. It can be defined in a discretized form, if the domain of variation of the variable under consideration is discretized to a finite number of intervals, in which case the score is a finite sum, or alternatively in a continuous form, in which case the score is an integral [*Candille and Talagrand*, 2005].

The cumulative forecasts and observations, denoted Y_m and O_m, are defined as functions of the components of the forecast vector and observation vector, respectively, according to

$$Y_m = \sum_{j=1}^{m} y_j, \quad m = 1, \ldots J \tag{4.7}$$

and

$$O_m = \sum_{j=1}^{m} o_j, \quad m = 1, \ldots J \tag{4.8}$$

whereas both Y_m and O_m are both cumulative functions of m probability components (e.g. ensemble members) that must add to one as illustrated in Fig. 4.1.

The ranked probability score is the sum of squared differences between the components of the cumulative forecast and observation vectors in Equation Eq. [4.7] and Eq. [4.8], given by

$$RPS = \sum_{m=1}^{J} (Y_m - O_m)^2 \quad [4.9]$$

or, in terms of the forecast and observed vector components y_j and o_j,

$$RPS = \sum_{m=1}^{J} \left[\sum_{j=1}^{m} y_j - \sum_{j=1}^{m} o_j \right]^2 \quad [4.10]$$

A perfect forecast would assign all the probability to the single y_j-category corresponding to the event that subsequently occurs, so that the forecast and observation vectors would be the same. The the RPS is zero. Non-perfect Forecasts receive score above zero. The final tern in Eq. [4.9] and 4.10 when $m = J$ is always zero, because the accumulations ensure that $Y_J = O_J = 1$. Therefore worst possible score is $J - 1$.

Regardless of how a forecast probability distribution is expressed, providing a full forecast probability distribution is both a conceptual and a mathematical extension of multi-category probability forecasting, to forecasts for an infinite number of predictand classes of infinitesimal width. A natural approach to evaluating this kind of forecast is to extend the ranked probability score to the continuous case, replacing the summations in Equation Eq. [4.10] with integrals. The result is the Continuous Ranked Probability Score:

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 \, dy \quad [4.11]$$

The CRPS, too, has a negative orientation (smaller values are better), and it rewards concentration of probability around the step function located at the observed value.

In the following only the continuous form of the Ranked Probability will be considered, its skill score (CRPSS) then

$$CRPSS = 1 - \frac{CRPS_{forecast}}{CRPS_{reference}} \quad [4.12]$$

The perfect Score would be 1 in this case, too. A skill of 0 or lower would indicate no skill in comparison to the reference.

Discrimination: Generalized Discrimination Score

Another attribute of quality, *discrimination*, measures whether forecasts differ when their corresponding observation differ, e.g. using the *Generalized Discrimination Score* (GDSS) [*Mason and Weigel*, 2009; *Weigel and Mason*, 2011]. The Generalized Discrimination Score introduced by *Mason and Weigel* [2009] calculates the probability that any two (distinguishable) observations can be correctly discriminated by the corresponding forecasts. So the *GDSS* can be interpreted as an indication of how often the forecasts are correct, regardless on the type of

forecast. For a given set of forecast-observation pairs, the *GDSS* is calculated by first constructing all distinguishable forecast-observation pairs from the data; then asking the question if the forecast can be used to successfully distinguish the observation for each set (⇒Fig. 4.2). The proportion of sets where this is the case yields the *GDSS*. The best Score would be 1 (e.g. all pairs are distinguished). On the other hand, if the forecast does not contain any useful information, than the probability to correctly discriminate between two observations would be equal to random guessing ($GDSS = 0.5$).

For continuous forecasts the discrimination can be calculated using the Kendall's correlation Coefficient τ:

$$\tau = \frac{4P}{n(n-1)} - 1, \qquad [4.13]$$

Where P is the number of concordant pairs of observations and forecasts (i.e. forecasts rightly discriminated) and n the number of observations [*Mason and Weigel*, 2009].

The Discrimination Score then is

$$GDSS = \frac{1}{2}(\tau + 1) \qquad [4.14]$$

Weigel and Mason [2011] completed this study by providing a way to calculate *GDSS* for ensemble forecasts to void the need for post-processing of the ensemble. Two ensemble forecasts $y_s = (y_{s,2}, ...y_{s,m})$ and $y_t = (y_{t,2}, ...y_{t,m})$with m being the number of ensemble members are now ranked either $y_s < y_t$ ($y_s > y_t$) if the probability that a randomly selected member if the ensemble y_s exceeds a randomly selected member of the ensemble y_t is larger (smaller) than 0.5.

To summarize, *Skill Scores* can be used to quantize the quality of forecasts. They are a relative measure that require a reference. In the case of this work the climatology will be used as reference. A generic formulation of skill scores allow for the application towards any forecast property. As stated in the literature there are many aspects of forecast quality. Which of them are addressed is dependent in on the situation. Here I will address 3 of them: Accuracy by using the skill score *Mean Square Error Skill Score* and the correlation, reliability using the *Continuous Ranked Probability Skill Score* and discrimination using the *Generalized Discrimination Score*. Thereby, the bias, the level of association with the observation, both linear (correlation) and with any discriminable pair of observations (Generalized Discrimination Score) and ensemble generation (through reliability) is addressed.

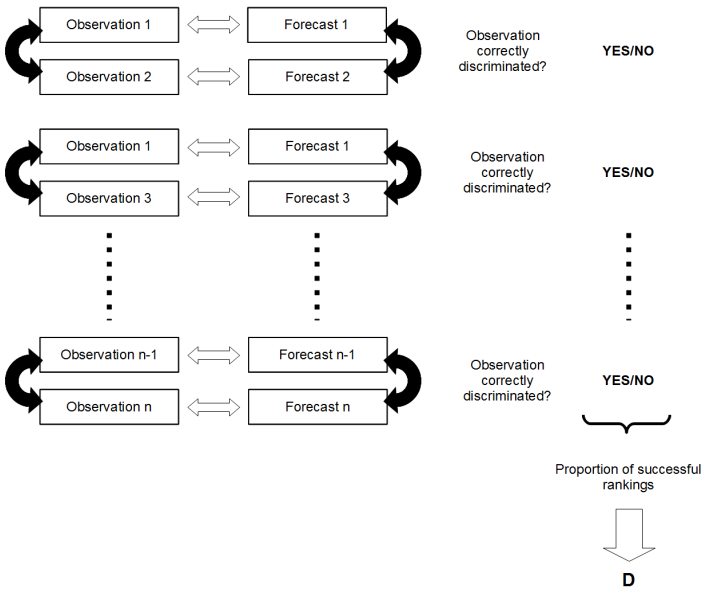


Figure 4.2.: Concept of the Generalized Discrimination Score after Weigel and Mason [2011]

4.2. The relationship between skill scores using an example from MiKlip

Having refined the verification framework for regional decadal predictions, an issue that has to be considered is the general behaviour of the skill measures regarding each other. The formulation of the measures used here can be similar and a certain relationship between them will exist. The question arising now is, whether this relationship is stationary. And: If it is not, can it be used to verify the prediction system?

In Fig. D.8 is shown exemplary, how this problem was addressed: For all land grid points the skill scores, in this particular case the MSSS and the CRPSS were calculated, here for the winter precipitation over Europe of the MPI-ESM b0-EUR022 ensemble. Those were then sorted into bins of 0.2 width. In this example it is clear that there is a linear relationship between both skill scores. A fact that has to be verified by all other models, seasons, skills scores and variables to form a universal statement.

Within Miklip the number of models is limited. So, to have the most ensembles to explore the relationship solely on the differences in models, only decadally initialized (initialized every 10 years, meaning 5 initialization dates for the 50-year time period of MiKlip) ensembles of MPI-ESM and CCLM in a resolution of 0.22°were used, making it a 4 full ensembles of 10 members.

Analysing all 4 ensembles for all seasons and variables (not shown) this linear dependency between skill scores was confirmed. The strength of that relationship is indicated by the scatter

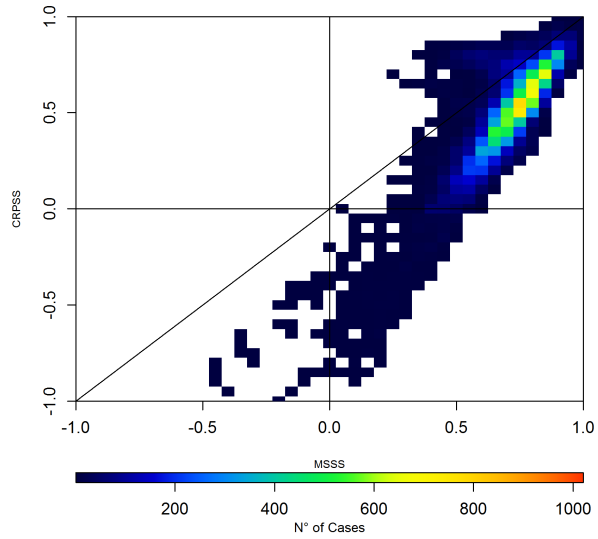


Figure 4.3.: Scatterplot of the relationship of the MSSS and the CRPSS of the winter precipitation over Europe. Shown are all land grid points of the MPI ensemble mean winter prediction over the lead times 1-5 for 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.

of the plot varies between variables. So is the skill of temperature prediction in summer more spread than of precipitation.

The availability of both baselines allows for the analysis of the skill relationships within models, whose ensembles are only distinguished by the initialization method. It becomes apparent that the tilts of the axes of the data point clusters do not differ significantly between the ensembles of one model. Whereas the center of the cluster might move in cases indicating improvement/deterioration of skill between the baselines as well as the intersection with the y-axis, the slope changes little while the clusters axes show conceivably differences between the models. This is the case for almost all seasons, variables and skill scores. Also occurring frequently is a scattering of the cluster towards low skill. Exceptionally high skill in one aspect will most likely also result in high skill other aspects of quality. The worse the skill becomes the more likely become the exceptions of a pairing of a high skill in one aspect and low skill in another.

The exception is the comparison of the Generalized Discrimination Score and the correlation coefficient where there is little difference between variables and seasons. Here in all cases the cluster of data points is streamlined and approaches the identity line. That implies an almost identical pattern of correlation and discrimination, although the correlation is used as a measure of accuracy.

In the cases of accuracy and reliability, when describing the axis of the skill cluster with a degree 1 polynomial:

$f(x) = a_0 + a_1 x$ with $a_0 \neq 0$, whereas x is one skill measure and $f(x)$ the other skill measure;

The intersect (a_0) denotes the absolute skill of the ensemble and the slope (a_1) is a model inherent property that describes the factor with which one skill reacts to a difference in the other skill ($f(\Delta x) = a_1 \Delta x$).

But only having two models in two applications available does not allow for a definitive statement about the relationship. However, the initial assessment is validated by the repetition of the behaviour with almost all seasons, variables and skill scores.

In conclusion, the relationship between skill scores can be used as a first analysis of the prediction system. Though reliability, accuracy, and discrimination are, while equally important prediction qualities, typically considered separately I can show a linear dependency between the skill scores representing the them. I found that to be true for all variables and models of MiKlip (CCLM and MPI-ESM-LR). Only for the comparison of GDSS and correlation the relationship approaches the identity line. For all other skill comparisons the axes of the skill clusters are tilted. The intersect of the axis indicates the overall skill of the prediction system, while the slope of the axes seems change little between the models' applications and is therefore assumed to be a model inherent prediction property.

4.3. The quantification of value added by regionalisation

One major aspect that seperates MiKlip from other decadal prediction research is the big emphasis on the dynamical downscaling of the global decadal climate predictions. As a result, an addition to the verification framework has to be made. Regionalization of climate simulations is costly and requires much computing power. This additional effort has to be justified and the regional models verified.

The value added by regionalization of global decadal predictions becomes an imperative of the verification process [*Kanamitsu and DeHaan*, 2011]. While one might argue, regional climate models add valuable information by definition just because of their higher resolution [*Feser et al.*, 2011], it is not a given that regionalization improves also skill.

The *IPCC* [2013] stated that the modes multi-annual variabilities have strong regional manifestation whose amplitude can be larger than that of human-induced climate change. If models are successful in reproduce these modes, a higher spatial resolution, along with all that entails, can have a positive effect on the predictive skill. But in fields like decadal prediction that require large ensembles of 10 year predictions, the improvement by downscaling has to be carefully weighted against the cost thereof. Additionally, while a potential user might be interested in localized information, the abilities of the regional model to represent the information in comparison to the global model output has to be assessed.

Because of that, an expansion of the verification framework set out in 4.1 will now follow. Two possibilites on how information gain by regionalization are being introduced:

Added Value Version of Skill Scores

On the one hand, the aforementioned skill scores can be used to quantify the added value of regionalization on different forcast qualities. They can be used to determine the spatial pattern of value added.

$$\text{Skill Score } SS = 1 - \frac{S_{regional}}{S_{global}} \quad [4.15]$$

The aded value version of skill scores has the same proporties as the skill scores, values ranges from $-\infty$ to 1. The perfect score is 1 (100% added value), 0 then indicates maintains skill and subzero values reduced skill after regionalisation. An exception would be the Generalized Discrimination Score as its range is from 0 to 1, with 1 the perfect score.

meaning the perfect score is 1 (100 % added value), 0 maintained skill, sub zero values reduced skill. However since the best Discriminatation Score D is 1, formulating the D_{av} the same way of the other skill scores leads to reversed scale. Sub zero = added value.

Added Value Index

On the other hand, the Added Value Index AVI can be used to describe by one single number the percentage of simulated data points of the regional model that are better [*Kanamitsu and DeHaan*, 2011]. It can be used in the temporal as well as in the spatial domain. E.g. the AVI gives the number of gridpoints where the regional model outperforms the global model. The AVI as proposed by *Kanamitsu and DeHaan* [2011] uses the correlation as a measure for model performance. But the skill scores of chapter 4.1 can be used, too.

The issue with this approach is the non-distinction of "better" and "good". While the regional model could be "better" than the global model, i.e. has a high number of gridpoints of higher skill than the global model, the actual skill of the regional model could still be below zero. The literal description of that scenario would be "less bad". The same is true for the use of skill scores to represent the added value.

Of course a possibility can be the modification of the AVI where only the share of gridpoints that have skill to begin with. That obviously lowers the AVI but is a much more important information. Another option would be a fraction of the potential possible AVI:

$$AVI_{fraction\ of\ grid} = \frac{n_{regional} - n_{global}}{N - n_{global}} \quad [4.16]$$

whereas N is the total number of gridpoints, n_{global} the number of gridpoints with skillful predictions of the global model and $n_{regional}$ the regional ones. Thus, $N - n_{global}$ is the potenial

possible improvement the regionalization could achieve. The same approach is also applicable to skill scores instead of gridpoints. IS 1 the best positive skill to be achieve the $AVI_fraction$ is than compounded as follows:

$$AVI_{fraction\ of\ skill} = \frac{S_{regional} - S_{global}}{1 - S_{global}} \quad [4.17]$$

However, this issue and its handling depends on what the prediction wants to accomplish. I.e. can one actually assume a regionalization becoming accurate, when the global prediction showed no skill to begin with? But with the objectives of the forecast experiment clear from the get-go, a suitable verification method can be found. Putting it very broadly, in a field as new as regional decadal climate predictions, the preservation of skill with the downscaling is the first hurdle to take and a simple skill score with the global reference will calculated that. One step further can be taken when weighting the skill score of the regional model by the potentionally possible improvement as given by the global model.

4.4. The Regional Decadal Hindcasts in MiKlip

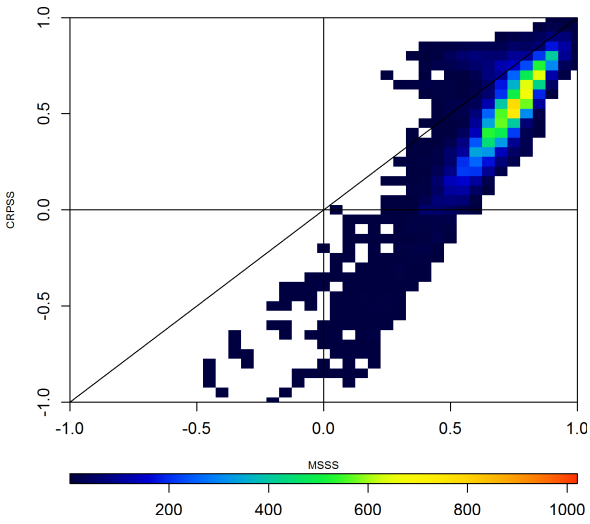
The *IPCC* [2013] stated that most modes of inter annual to inter-decadal variability are now present in climate models. Does that mean, the transport of that information from the areas of the mode (usually oceans) to investigation areas over land will lead to skilful predictions? The previous sections detailed a verification framework to find the answer to that. Now these methods and measures are applied to the MiKlip ensembles as an example of a comprehensive decadal prediction system.

First, the overall skill of the predictions system has to be established, meaning performance of the system in predicting mean variables such as temperature and precipitation. In the following I will lay out the results of thereof of both the decadally initialized EUR22 CCLM ensembles (horizontal resolution of 0.22°) of both baselines and the annually initialized b1-EUR44 CCLM ensemble (horizontal resolution of 0.44°). The EUR22 b0 and b1 ensembles are pasted back to back to create a 50 year time series from which the long time linear trend is removed, while the trend of the annually initialized 01-EUR44 ensemble is not removed. The global ensembles of the MPI-ESM are interpolated to the grid of the regional model. E-OBS [*Haylock et al.*, 2008; *EU-FP6 project ENSEMBLES & ECA&D project* , 2016] is the observational data set used.

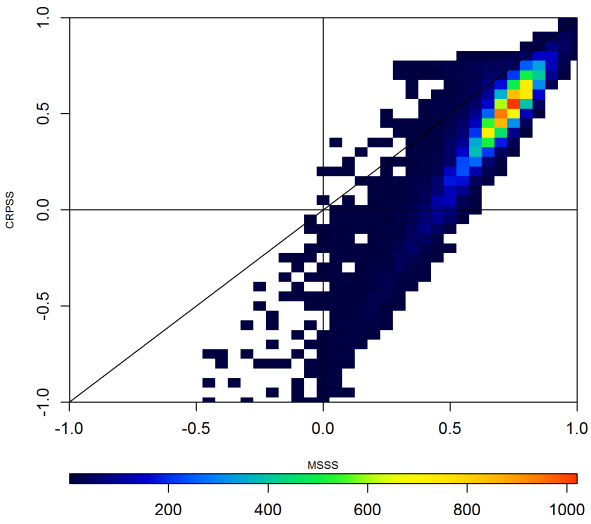
Scatterplots

In section 4.2 a scatter plot of the relationship of of the relationship of MSSS and CRPSS of the precipitation over Europe (Figure D.8) of the 10 member CCLM decadally initialized b0 ensemble was shown (1961–2010). In figure 4.4 the other ensembles (CCLM b1, MPI-ESM b0 & b1) are added. The global ensembles (left panel) are spatially more dispersed, especially

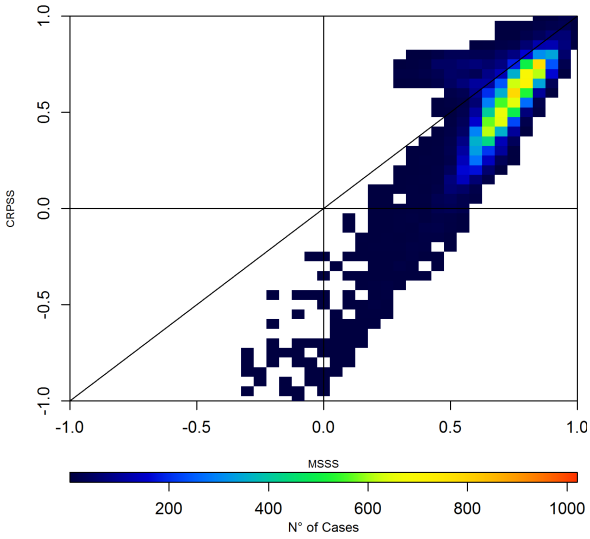
towards lower skill. The tilt of the MPI's axis is less steep. Both facts do not change between the baselines but are not true when other scores, seasons or variables are assessed.



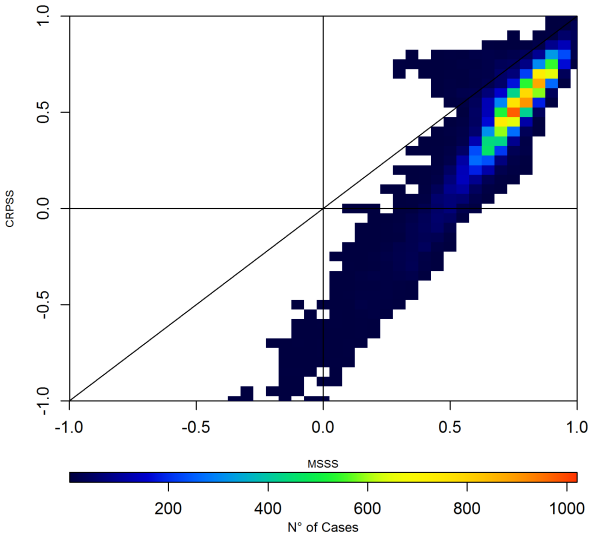
(a) MSSS vs. CRPSS of MPI b0



(b) MSSS vs. CRPSS of CCLM b0

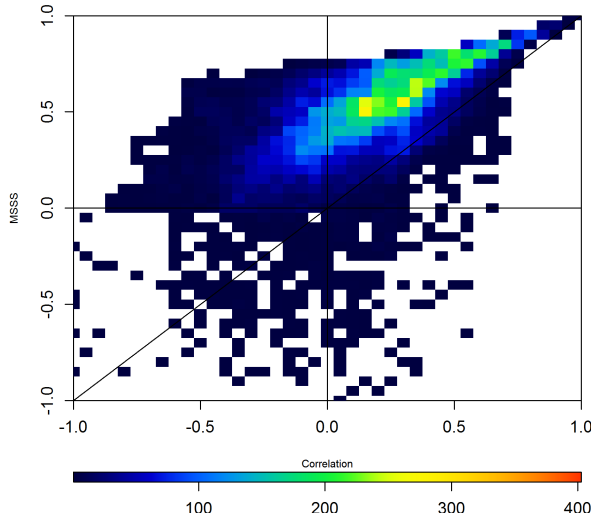


(c) MSSS vs. CRPSS of MPI b1

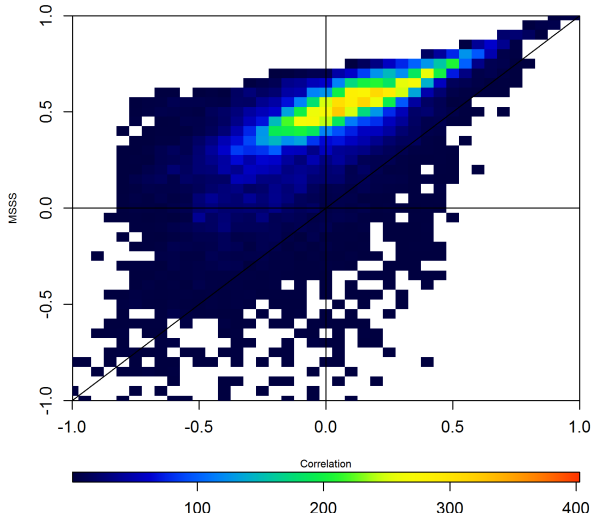


(d) MSSS vs. CRPSS of CCLM b1

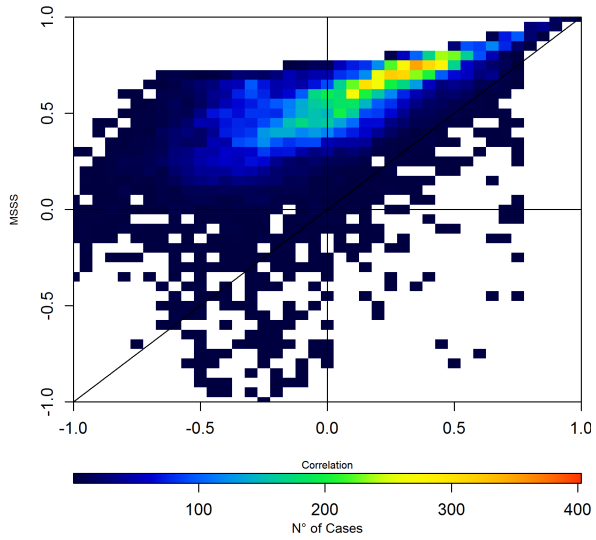
Figure 4.4.: Scatterplot of the relationship of ***MSSS and CRPSS of the winter precipitation*** *over Europe. Shown are all land grid points of the CCLM ensemble mean winter prediction over 5 decades 1961–2010 from the b0 ensemble (initialisation every 10 years) with the long time (50a) trend removed (topleft) and that of the b1 CCLM ensemble (bottom left and the corresponding global ensembles)*



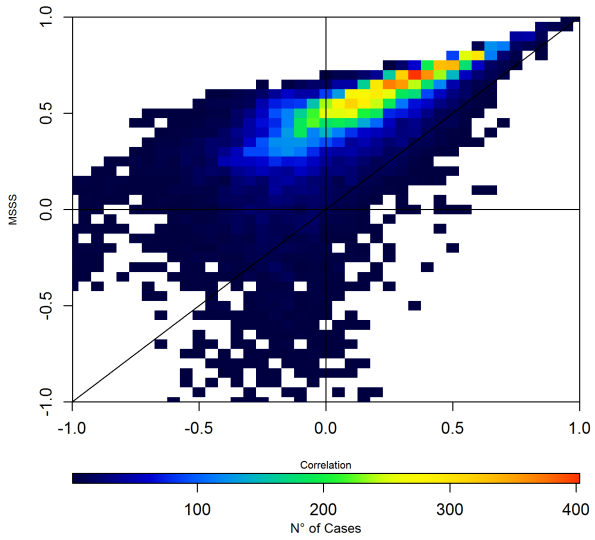
(a) MSSS vs. correlation of MPI b0



(b) MSSS vs. correlation of CCLM b0



(c) MSSS vs. correlation of MPI b1



(d) MSSS vs. correlation of CCLM b1

Figure 4.5.: Scatterplot of the relationship of ***MSSS and the correlation of the winter precipitation*** *over Europe. Shown are all land grid points of the CCLM ensemble mean winter prediction over 5 decades 1961–2010 from the b0 ensemble (initialisation every 10 years) with the long time (50a) trend removed (topleft) and that of the b1 CCLM ensemble (bottom left and the corresponding global ensembles)*

The tilt, the intersect with ordinate of the axes as well as the center of the cluster is of interest when the skill of the ensembles is in question. The number of grid points in the negative portion of the plot (2nd, 3rd and 4th quadrant) is higher for both global models than their regional

counterpart. The intersect of the axis of the point cluster the ordinate is higher for the regional models, indicating an improvement of the CRPSS with downscaling. The center of the point cluster is also shifted towards higher MSSS, again showing an improvement by downscaling. Overall the regional model is performing better in this case.

In comparison, Fig. 4.5 shows the relationships between the correlation coefficient and the MSSS for the same ensembles. The improvement of the correlation by the downscaling is less obvious. The distribution of correlation coefficients does include almost all possibilities from -1 to 1, but the location of the distribution has shifted towards higher correlation between global and regional ensembles by a shift in the point cluster's center and less spatter in the low-skill-quadrants. The discrimination (not shown) does show a similar behaviour.

The difference between the ensemble generations represent themselves in the same way as an improvement between b0 to b1 for both models and all three scores by a shift of the center of the cluster, a higher intersection with the ordinate and less scatter in the lower left quadrant.

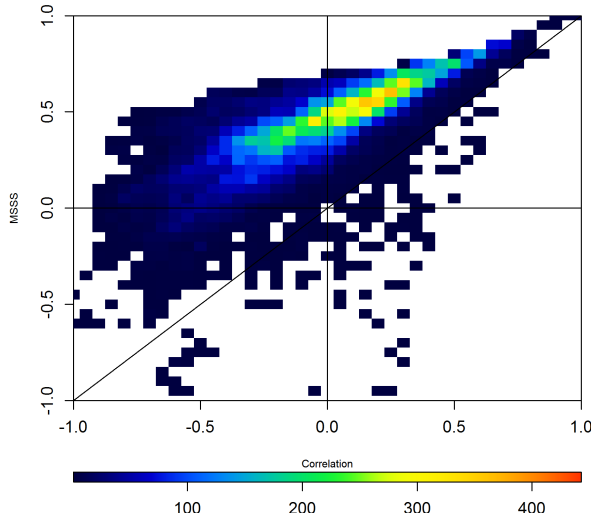
On another note, the same combination of skill scores and ensembles for precipitation in summer reveals a different picture (Fig. 4.6). The skill of the summer precipitation prediction is higher than in winter in general and the regional ensembles show an overall better skill cluster, but an improvement with the development from b0 to b1 can not be rendered. With both the global and the regional ensembles the b0 clusters are moved towards the higher CRPSS (coordination ordinate) and the center moved towards the right (higher MSSS on the coordination abscissa).

In contrast predictions of temperature show a more scattered cluster (E.g. Summer Temperature 4.7). The prediction the likelihood of very accurate and reliable prediction increases as does the opposite and every combination thereof. The difference is largely due to the decrease in reliability even though regionalisation increases reliability in both temperature and precipitation.

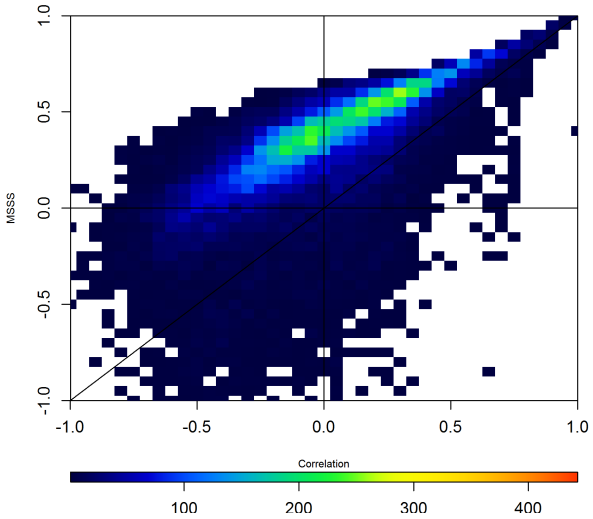
In conclusion is to say, the predictive skill of mean temperature and precipitation can be increased by the means of downscaling (mainly reliability) but no universally valid statement can be made: The skill of decadal predictions are sensitive to regional differences, variable and season.

The spatial distribution of skill

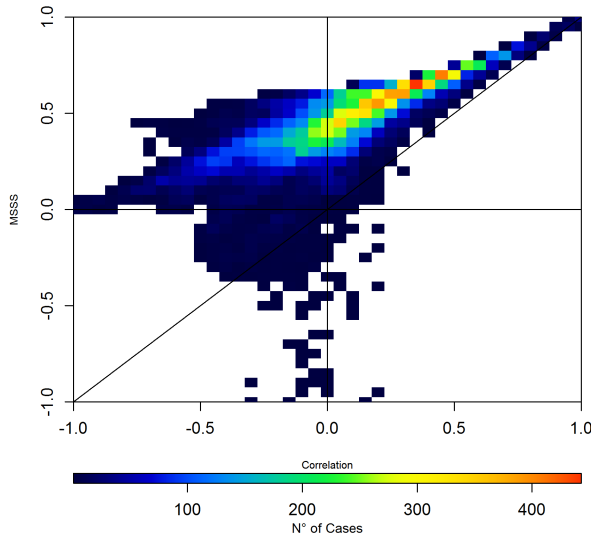
The scatter plots in the preceding section have one major drawback: They do not show the actual spatial distribution of predictive skill or the added value. Hence, the regional distribution of the skill presented previous will now be analysed in their spatial distribution. The goal is to find whether there is a regional component to the spread of skill. In the following the two plots will be shown: As an example the spatial distribution of the discrimination skill for the b1 MPI-ESM ensemble over Europe and the two ensemble generations of the CCLM decadal predictions system (b0-EUR22 & b1-EUR22) of a spatial resolution of 0.22 °, i.e. the same data



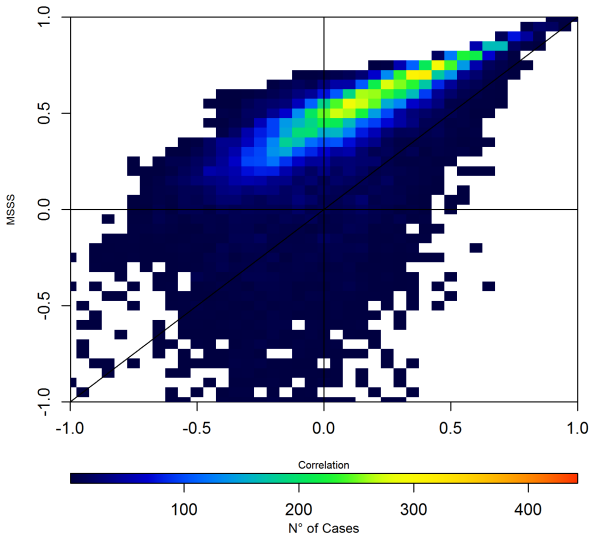
(a) MSSS vs. CRPSS of MPI b0



(b) MSSS vs. CRPSS of CCLM b0



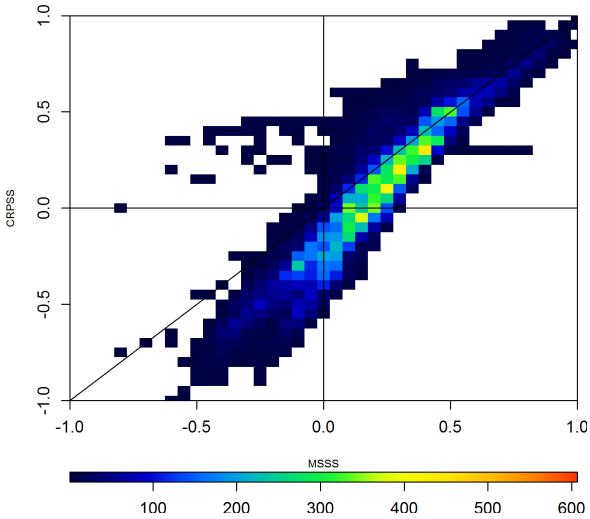
(c) MSSS vs. correlation of MPI b1



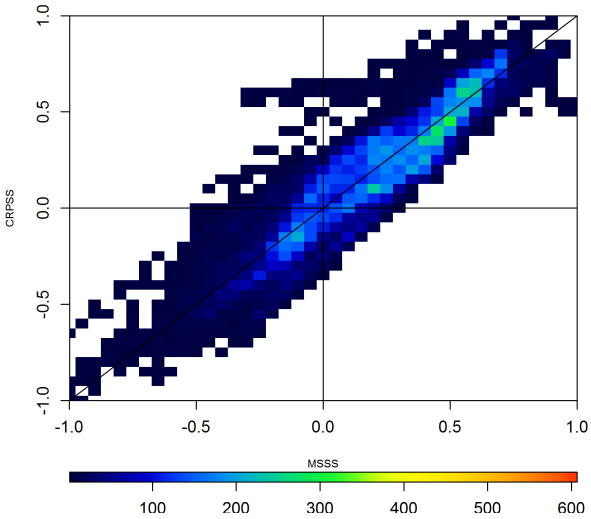
(d) MSSS vs. correlation of CCLM b1

Figure 4.6.: Scatterplot of the relationship of ***MSSS and correlation of the annual precipitation*** *over Europe. Shown are all land grid points of the CCLM ensemble mean annual prediction of precipitation over 5 decades 1961–2010 from the b0 ensemble (initialisation every 10 years) with the long time (50a) trend removed (topleft) and that of the b1 CCLM ensemble (bottom left and the corresponding global ensembles)*

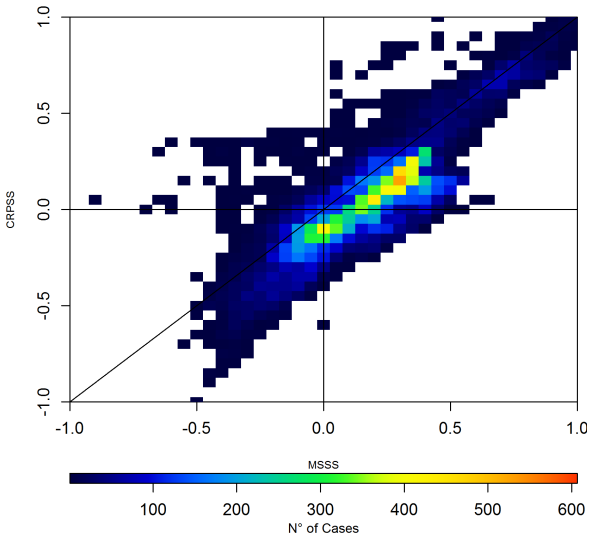
that was used for the scatter plots of the previous section. For the lead years 1–10, meaning the whole decade simulated, summer, winter and the whole year is shown and for comparison the skill for the whole year of the lead years 1–5, too. The Generalised Discrimination Score



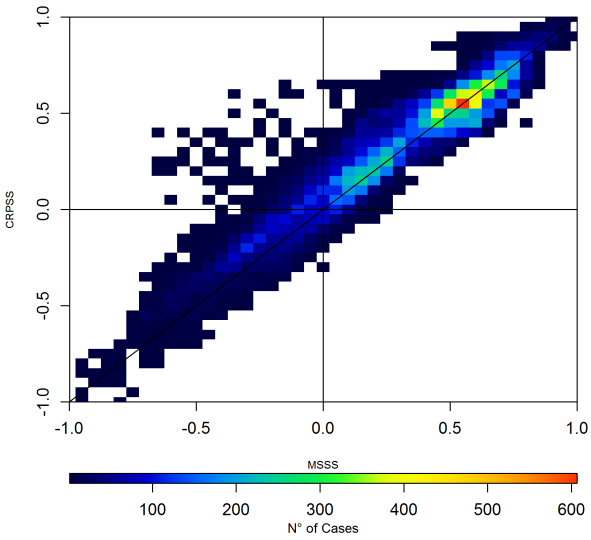
(a) MSSS vs. CRPSS of MPI b0



(b) MSSS vs. CRPSS of CCLM b0



(c) MSSS vs. CRPSS of MPI b1



(d) MSSS vs. CRPSS of CCLM b1

Figure 4.7.: Scatterplot of the relationship of MSSS and CRPSS of the summer temperature over Europe. Shown are all land grid points of the CCLM ensemble mean summer temperature of precipitation over 5 decades 1961–2010 from the b0 ensemble (initialisation every 10 years) with the long time (50a) trend removed (topleft) and that of the b1 CCLM ensemble (bottom left and the corresponding global ensembles)

indicates the ability of a model to discriminate between distinguishable observations. It ranges from 0 to 1, with the a score above 0.5 indicating skill.

Table 4.2.: The skill of the lead years 1–10 of the ***EUR22 decadally initialized ensembles for temperature and precipitation in winter and summer and the whole year*** *be the number grid points with positive skill as a percentage of the total number of grid points. The added value is characterized as the difference of grid points with positive regional skill and grid points with positive global skill. A negative number means, more grid points of the global model show positive skill, i.e. the global model does in general outperform the regional ones. These numbers do not take into account how high the actual skill is, only if it surpasses the limit for a skilful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS*

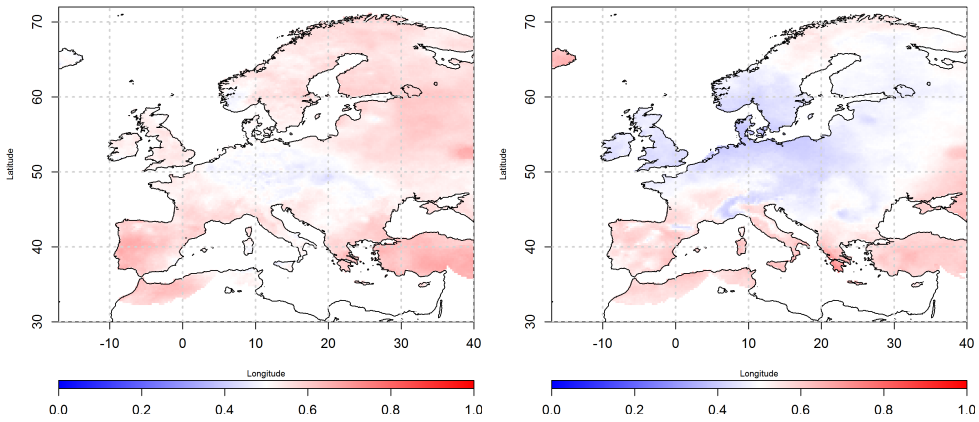
	b0 Temp CCLM	b0 Temp CCLMav	b1 Temp CCLM	b1 Temp CCLMav	b0 Prec CCLM	b0 Prec CCLMav	b1 Prec CCLM	b1 Prec CCLMav
YEAR CRPSS	12	7.9	30.4	5.2	3.5	-0.1	4.8	-4.1
MSSS	15.6	7	50	14.8	6.8	-2	11.7	-7.8
GDSS	51.5	-7.3	91.4	7.4	44.1	1.3	56.4	-8.3
COR	55.2	2.9	93.8	6.5	46	-1.1	55	-8.9
JJA CRPSS	20.7	1.3	16	-3.1	5	1.5	9.7	-5.5
MSSS	26.7	-0.3	25.7	-0.1	10.1	1.5	16.3	-10.5
GDSS	76.8	-0.8	77.9	-12.6	51.1	8.7	62.3	0.2
COR	72.9	1.4	73.6	-16.2	49.3	11.1	64.9	0.5
DJF CRPSS	8.4	5.2	14.5	-1.9	2.7	0.6	7.4	2.3
MSSS	18.6	3.5	18.7	-2.6	6.6	-2.5	13.1	-0.2
GDSS	53	-1.7	58.6	2.2	49.3	-5.1	59.6	-2.4
COR	54.9	2.2	67.9	11.1	52.4	-2.1	57.7	-3.4

Fig. 4.8 illustrates the discrimination skill of the temperature hindcasts. There are many cohering areas of positive skill. The pattern do differ with season. The skill in the Southern Europe is higher in summer than in winter. The same holds true for the East of Europe. Scandinavia on the other hand shows great skill in all seasons and is greatly improved with second ensemble generation (b1 vs. b0: center and right column) and is also improved through downscaling (CCLM vs MPI, left and right column). Some regions show almost consistently low skill: Middle Europe and South Scandinavia. In Middle Europe the skill decreases with the second ensemble generation for both summer and winter, but not in the case of the whole year. The pattern of skill for the lead years 1–10 in comparison to the first pentade (1–5) does not vary significantly even though it is greater for the shorter forecast period in some cases.

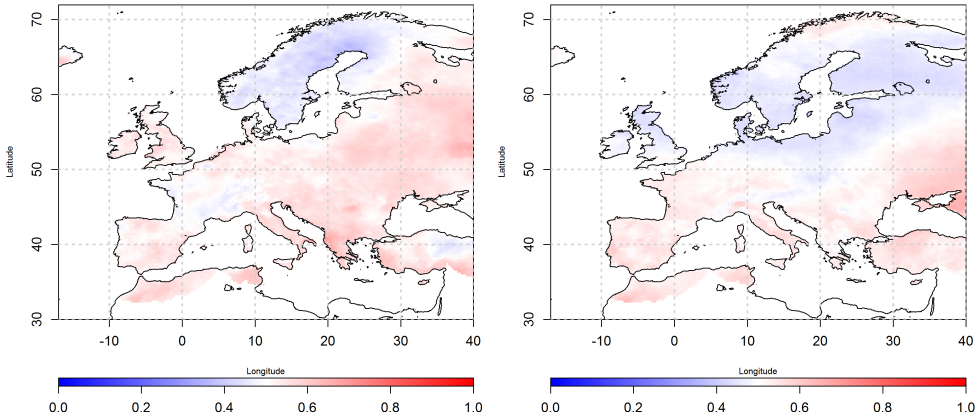
Precipitation hindcasts in Fig. 4.9 reveal a more inhomogeneous pattern. Still, some statements apply here, too. So is the skill for South Europe higher in summer than in winter. In contrast to the temperature, the skill in Southern Scandinavia is higher than in the north, which is further improved with the second ensemble generation. In contrast to the temperature is the skill in Middle Europe mostly positive for the second ensemble generation. The difference between the skill for the full decade and the first pentade is as with temperature in the net value. The prediction of temperature and precipitation is more skilful in the first half of the simulated decade though only in regions of already positive skill. Regions that show no skill in the first 5 years of the prediction are unlikely to be skilful when the last 5 years are added into the analysis. The EUR22 ensembles had the linear trend over the 50 years removed and the influence of external factors diminished, resulting in the source of skill only being the initialization. And the initialization will only be a key driver within the first years of the decade. If information about the net value of the prediction system is asked, the long-time trend and influence of external factors will add to the predictability and subsequently to the skill.

But is the prediction improved with downscaling? The distribution of skill in precipitation becomes more scattered with downscaling, proving that increased spatial resolution does add spatial diversity. Locally that can result in a reduction or increasing of skill. Then only a net count of grid points of positive skill can provide an answer to this question. Tab. E.1 lists the percentages of positive grid points of positive skill scores and correlation for both EUR22 CCLM ensembles as well as the percentage of grid points of the EUR22 CCLM ensembles that outperform the global model. In the case of the GDSS, as shown in the previously discussed figures, only in some cases is there added value.

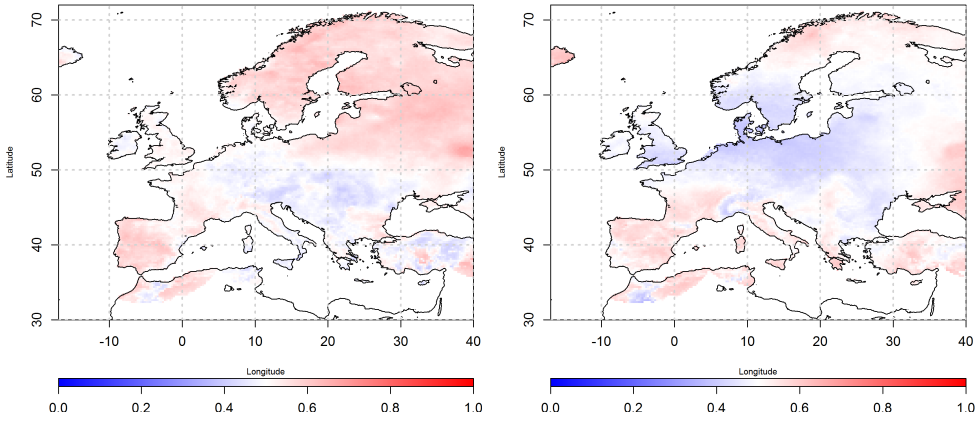
There is added value in discrimination for the summer precipitation in both ensemble generations. That is also true for the correlation skill. In comparison the reliability of summer precipitation is only increased for the first ensemble generation. Whereas the reliability of winter precipitation is improved through downscaling while the GDSS and the correlation are not. To summarise of the 48 cases listed in Tab. E.1 25 show a decrease in the number of grid points



(a) MPI b1 JJA (b) MPI b1 DJF

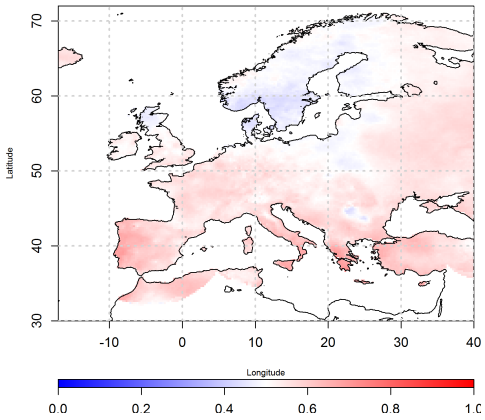


(c) CCLM b0 JJA (d) CCLM b0 DJF

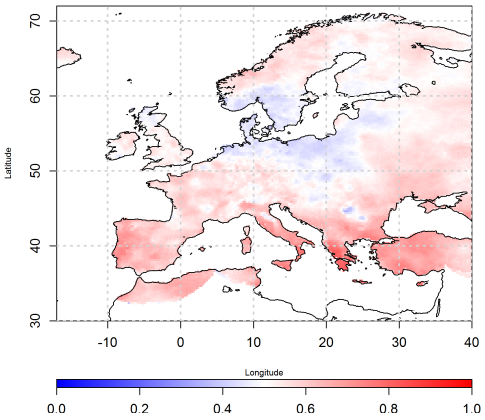


(e) CCLM b1 JJA (f) CCLM b1 DJF

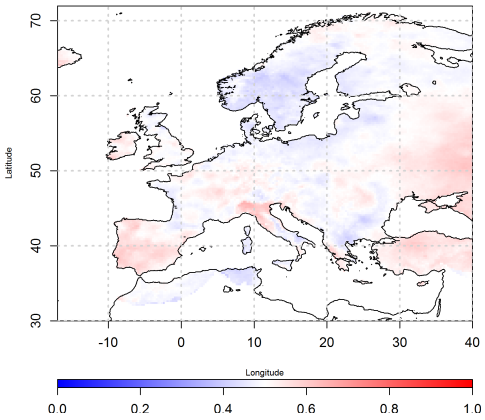
Figure 4.8.: The spatial ***discrimination skill of three EUR22 ensembles of temperature*** *in summer (JJA, left), Winter (DJF, right) and the whole year (continuation) is shown. The b1 ensemble of the decadally initialized MPI-ESM-LR over Europe for 1961–2010 (left) can be compared to the decadally initialized CCLM EUR22-b0 (middle) and the CCLM EUR22-b1 (right).*



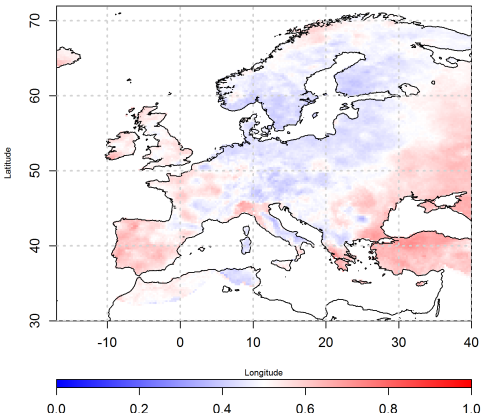
(g) MPI b1 YEAR



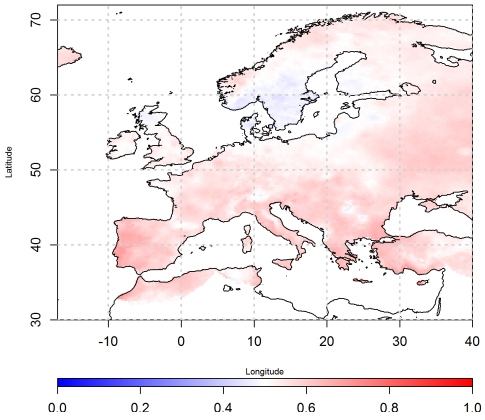
(h) MPI b1 YEAR 1-5



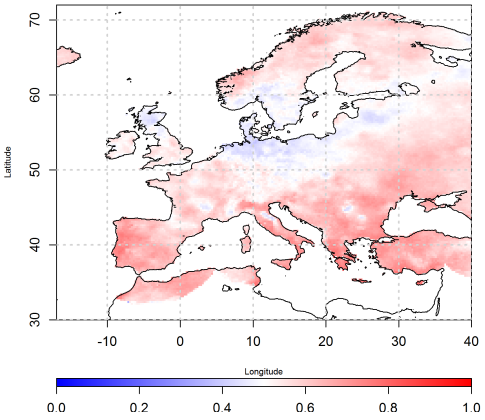
(i) CCLM b0 YEAR



(j) CCLM b0 YEAR 1-5

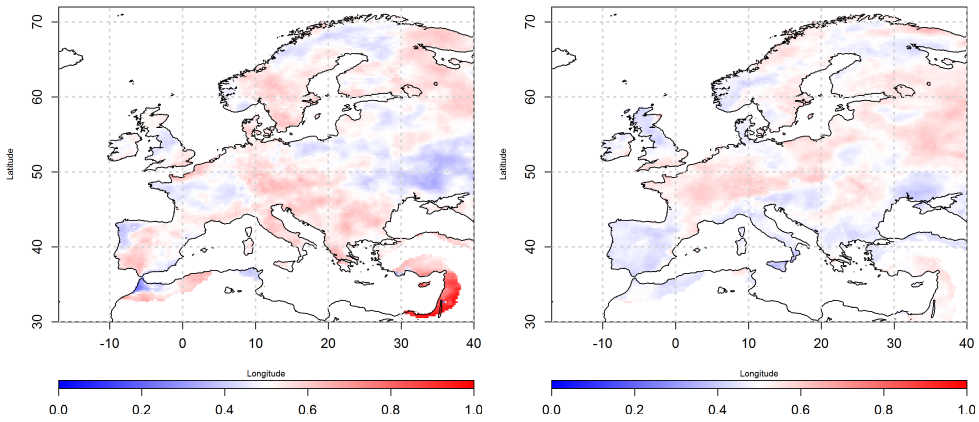


(k) CCLM b1 YEAR

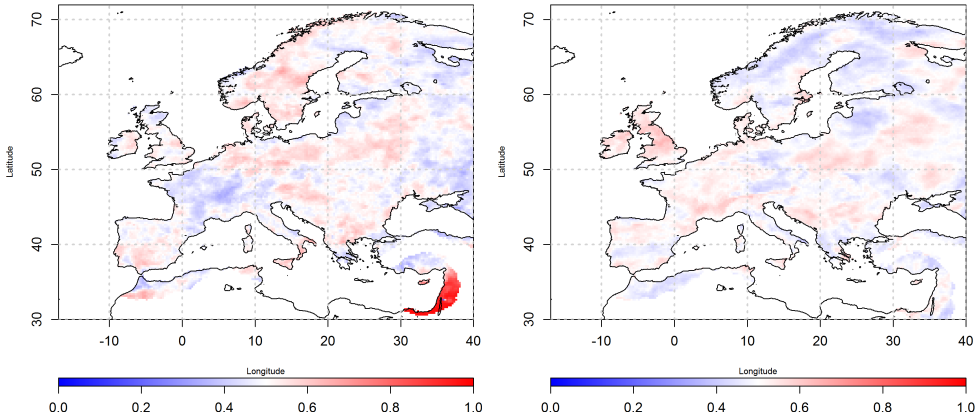


(l) CCLM b1 YEAR 1-5

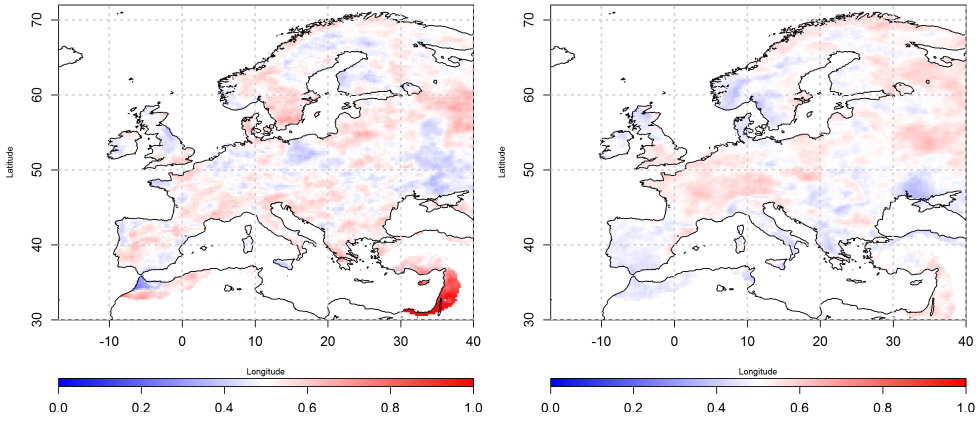
Figure 4.8.: Continued.



(a) MPI b1 JJA (b) MPI b1 DJF

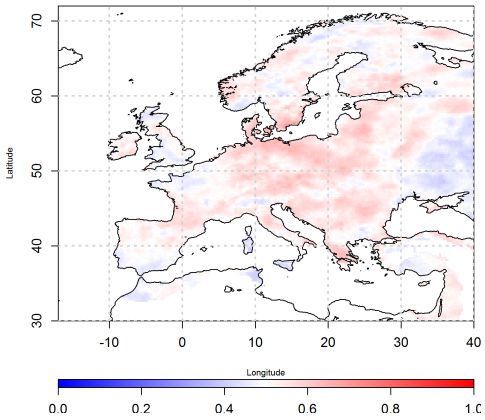


(c) CCLM b0 JJA (d) CCLM b0 DJF

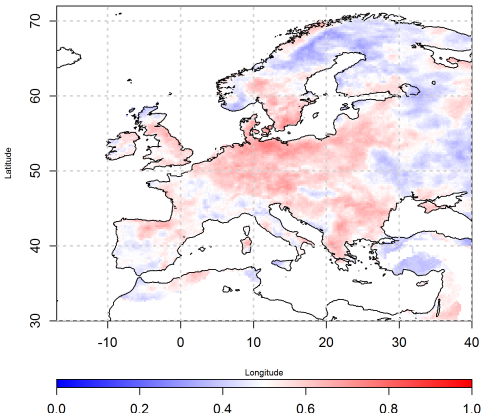


(e) CCLM b1 JJA (f) CCLM b1 DJF

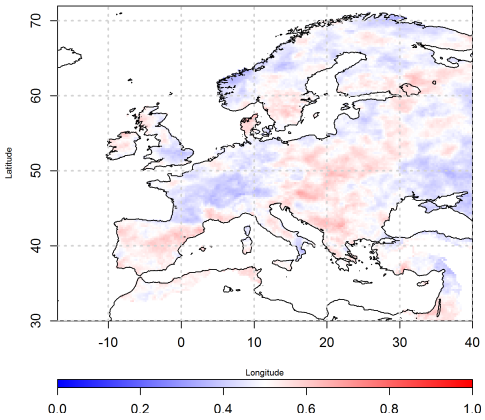
Figure 4.9.: The ***spatial discrimination skill of three EUR22 ensembles of precipitation*** *in summer (JJA, left), Winter (DJF, right) and the whole year (continuation) is shown. The b1 ensemble of the decadally initialized MPI-ESM-LR over Europe for 1961–2010 (upper panel) can be compared to the decadally initialized CCLM EUR22-b0 (middle) and the CCLM EUR22-b1 (lower panel).*



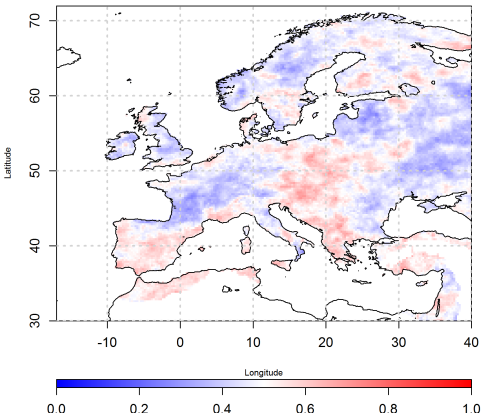
(g) MPI b1 YEAR



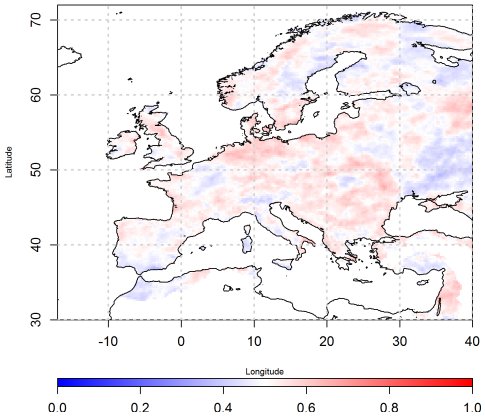
(h) MPI b1 YEAR 1-5



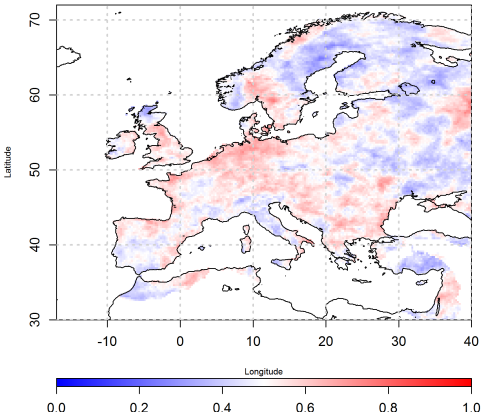
(i) CCLM b0 YEAR



(j) CCLM b0 YEAR 1-5



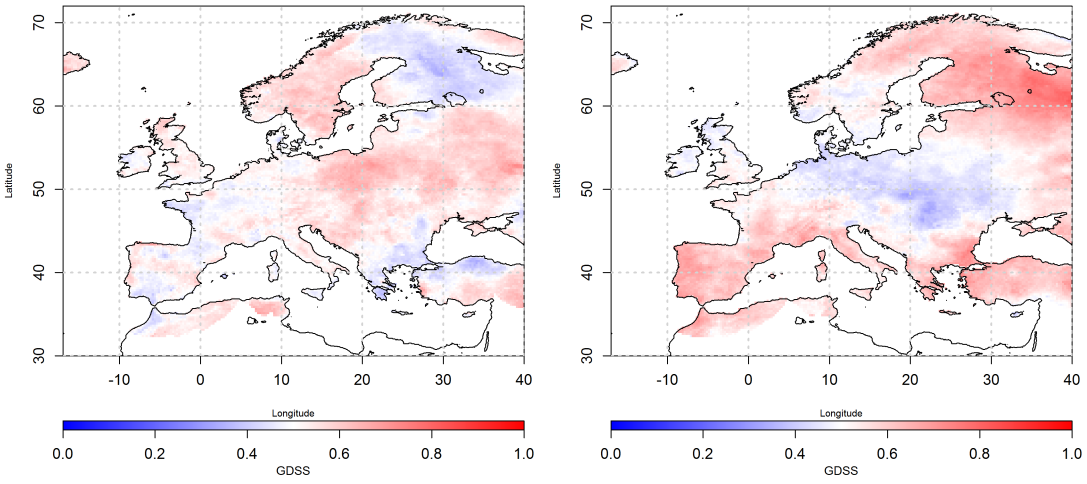
(k) CCLM b1 YEAR



(l) CCLM b1 YEAR 1-5

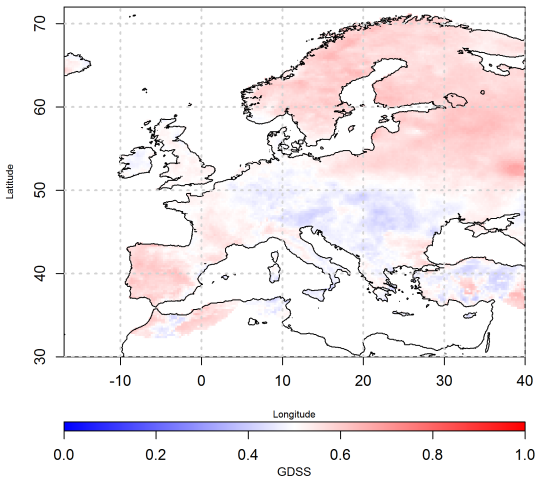
Figure 4.9.: Continued.

of positive skill. In App. E.1 the same table is presented but with the added value characterised as the percentage of grid points of positive global skill that is improved by downscaling. About 30 % of areas with positive GDSS are improved by regionalisation. On the other hand more than half of the correlation coefficients are improved by downscaling except in the case of precipitation of the b1-EUR22 ensemble.



(a) MSSS of CCLM b1 1961–1980

(b) MSSS of CCLM b1 1991-2010



(c) MSSS of CCLM b1 1961-2010

Figure 4.10.: The ***discrimination of hindcasts of summer (JJA) mean temperatures of CCLM over Europe for (a) 1961 – 1980 and (b) 1991 – 2010 of the b1 ensemble*** *is shown. The accuracy is described by the means of the Mean Square Error Skill Score which ranges from 0 to 1, with 1 indicating perfect skill and* $\leqslant 0.5$ *indicating no skill.*

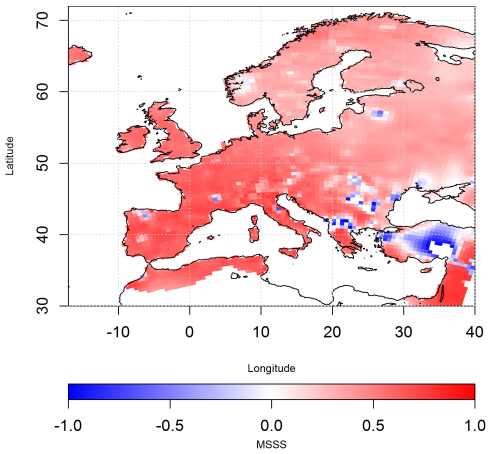
Important to add is that the skill is highly dependent on the decade considered. When in comparison to the previous figures single decades between 1961 – 2010 are analysed, it arises that e.g. the last 2 decade of the investigation period (1991 – 2010) offers more skilful hindcasts than the first 2 (1961 – 1980). Different causes could attribute towards that. For one, the observation data set (E-OBS) is likely to be more faulty in the earlier decades than the latter ones. Or, the prediction system struggles during the first decades, maybe due to the negative phase of the AMO (v. section 3.2) or the different climate trend in the earlier and latter decades. Both reasons could have wide ramifications for future decadal forecasts, especially as not only the net values of skill but also the spatial pattern change as evidenced by Fig. 4.10.

It emerges, that the verification of decadal predictions is not an easy task. The skill differs highly with variable, season and region. As does the added value. While the downscaling does at spatial variability, it will not necessarily widen the area of positive skill but can increase the skill itself. However it has to be noted, that a downscaling is dependent on the skill of the global model. If there is no skill in the global model to begin with any skill produced by the regional model is a product of chance. The lowest level to added value is preservation of skill. A net loss or gain of grid points of positive skill has to be weighted against the increase or decrease in skill value. E.g. the discrimination skill changes with downscaling by a net loss of 12.6% of grid points that are no longer positive, but 38.6% of the grid points of positive skill in the global model show a higher skill score after downscaling. While this analysis does showcase the behaviour of the scores very well, it has its limitations. For example, with the trend removed, a big source for predictive skill is missing. On top of that the decadally initialization does provide not enough starting dates. As is common practise in the field of decadal prediction, I will now use the annually initialized simulations.

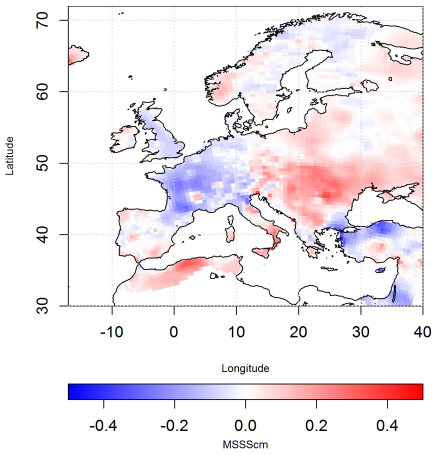
Decadal Hindcasts using annual initialization

By initializing the decadal prediction system every year and composing new time series from means over section of each run the full benefit of the initialization as opposed to the unitialized climate projections can be utilized (v. chapter 2). In contrast to decadally initializations it has the advantage by including more than a few starting dates to increase the potential predictability that is dependent on the starting day. In MiKlip annually initialized the full first generation global ensemble was not downscaled, due to restrictions in computing time. For the second ensemble generation the effort was made, to downscale some annually initialized runs in a courser resolution than discussed before (0.44 °, in the following called b1-EUR44). The following results were calculated using only 5 realizations of the global runs and their regionalisations.

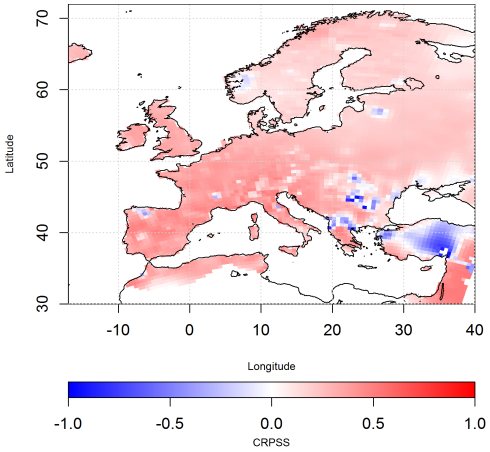
In the literature *Goddard et al.* [2013] the lead times 1, 2–5 and 6–9 had been widely used. The first year is excluded from the multi-year means as it carries the initialization shock and will produce different skill than the rest of the decade. On top of these I also considered lead



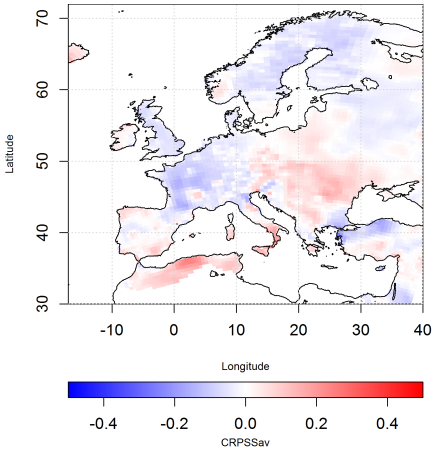
(a) CCLM y2-6 MSSS



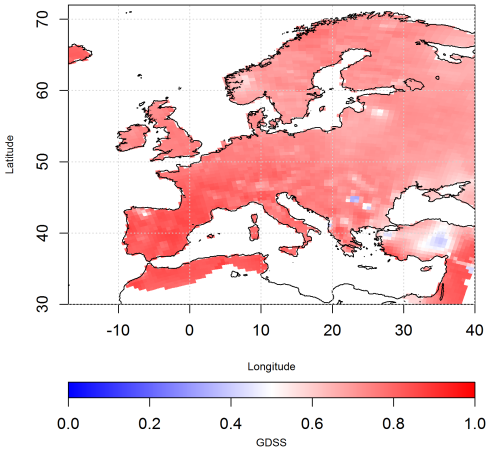
(b) CCLM y2-6 MSSSav



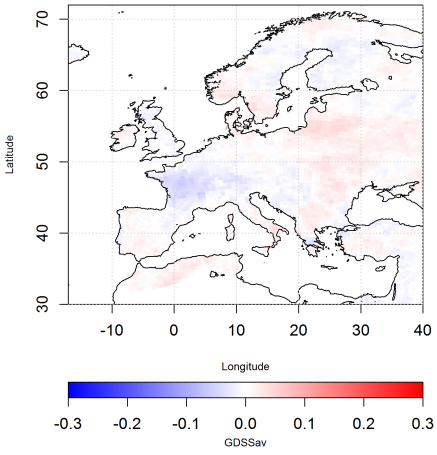
(c) CCLM y2-6 CRPSS



(d) CCLM y2-6 CRPSSav

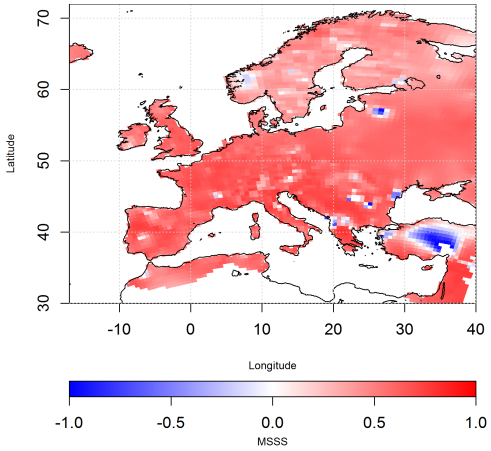


(e) CCLM y2-6 GDSS

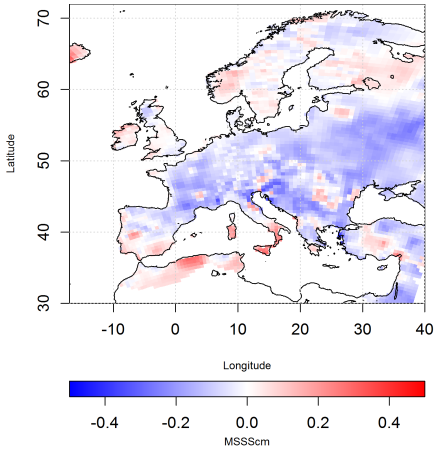


(f) CCLM y2-6 GDSSav

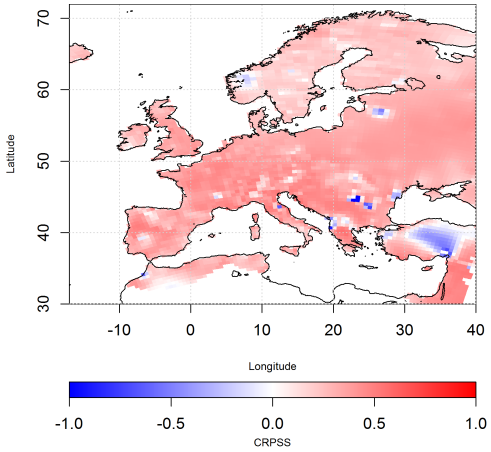
Figure 4.11.: ***Skill of b1-EUR44 hindcasts of annually initialized mean temperature*** *over Europe from 1961 – 2010 containing accuracy (upper panel), reliability (middle) and discrimination (lower panel). Figures on the left show the actual skill scores for the regional model CCLM and on the right the value added to the global model by downscaling is pictured.*



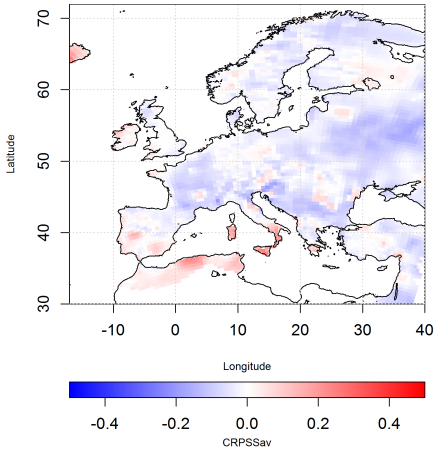
(g) CCLM y6-10 MSSS



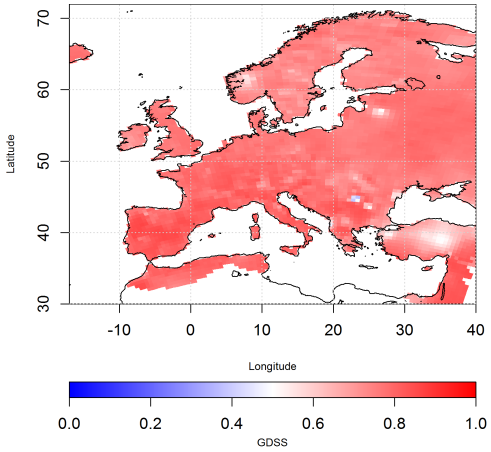
(h) CCLM y6-10 MSSSav



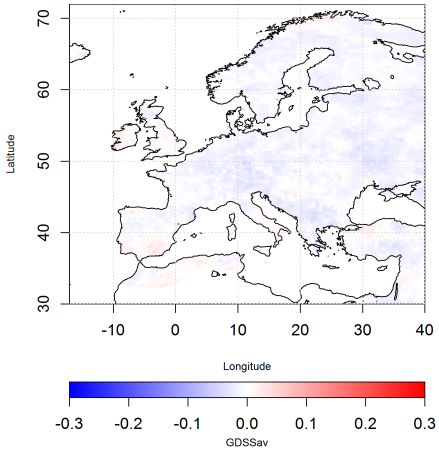
(i) CCLM y6-10 CRPSS



(j) CCLM y6-10 CRPSSav

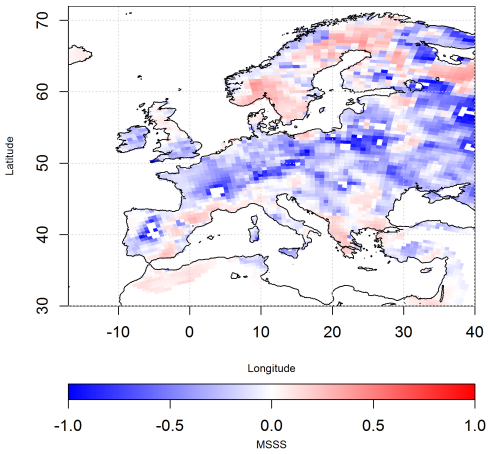


(k) CCLM y6-10 GDSS

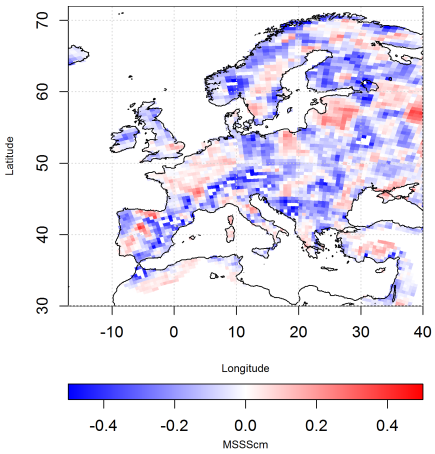


(l) CCLM y6-10 GDSSav

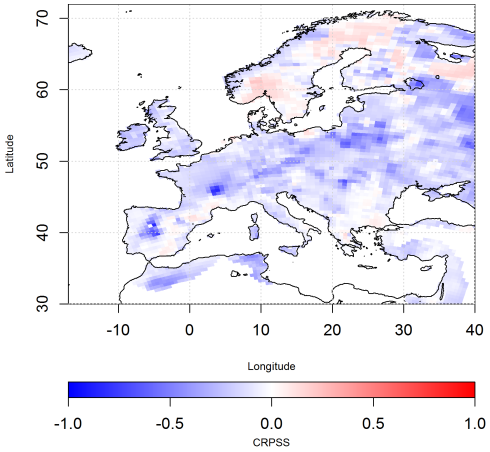
Figure 4.11.: Continued.



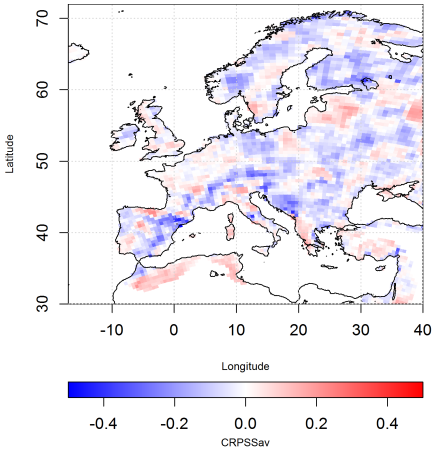
(a) CCLM y2-6 MSSS



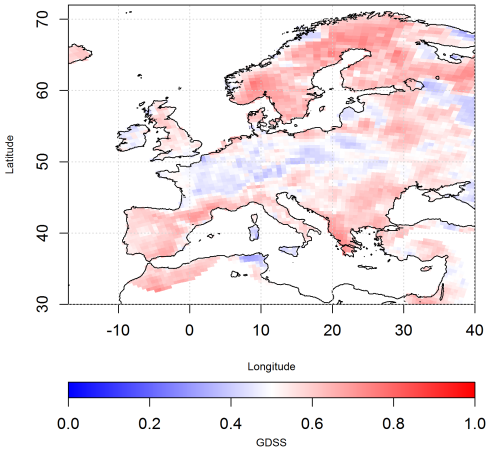
(b) CCLM y2-6 MSSSav



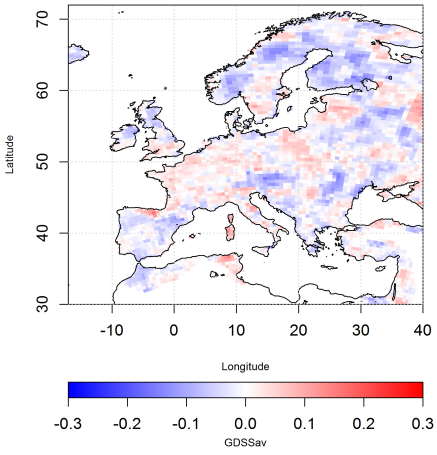
(c) CCLM y2-6 CRPSS



(d) CCLM y2-6 CRPSSav

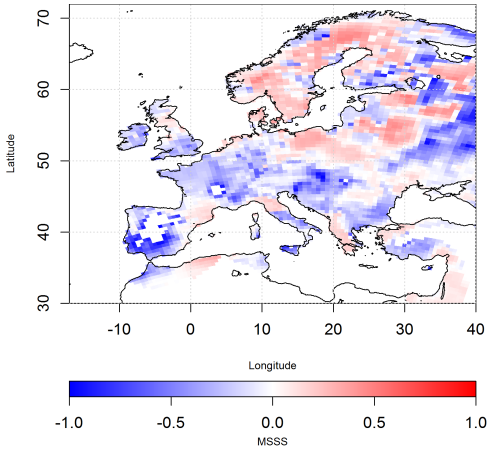


(e) CCLM y2-6 GDSS

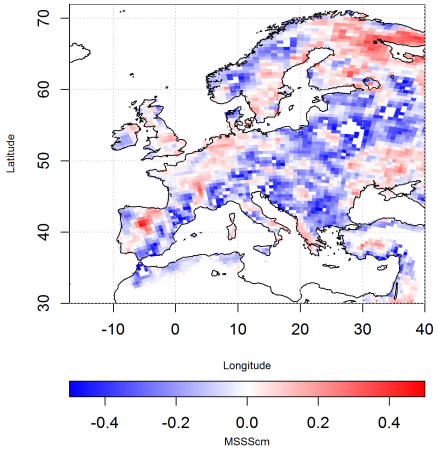


(f) CCLM y2-6 GDSSav

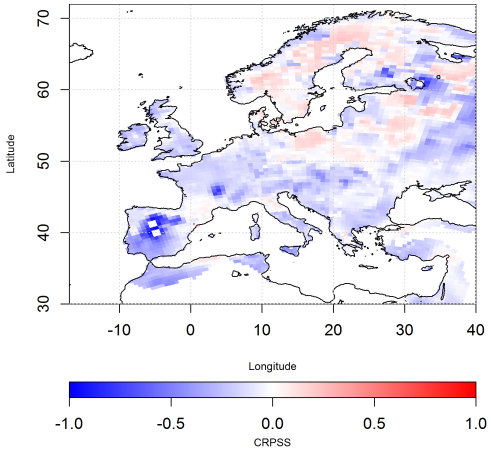
Figure 4.12.: ***Skill of b1-EUR44 hindcasts of annually initialized mean precipitation*** *over Europe from 1961 – 2010 containing accuracy (upper panel), reliability (middle) and discrimination (lower panel). Figures on the left show the actual skill scores for the regional model CCLM and on the right the value added to the global model by downscaling is pictured.*



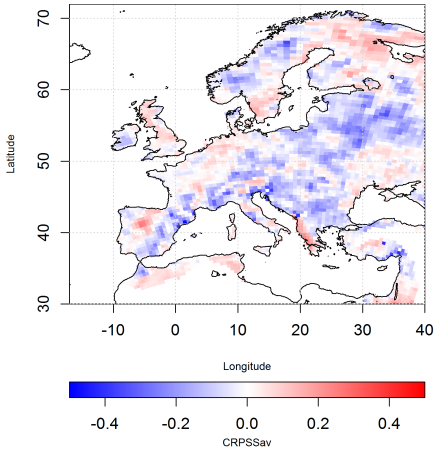
(g) CCLM y6-10 MSSS



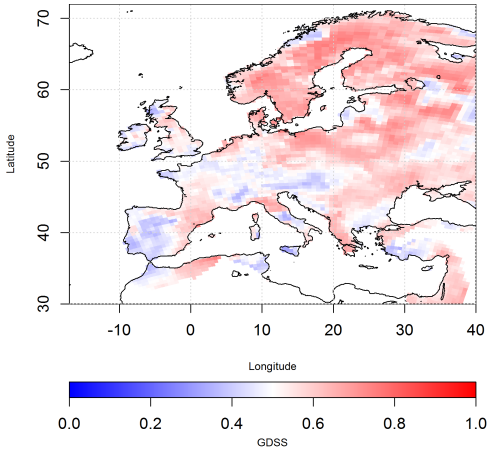
(h) CCLM y6-10 MSSSav



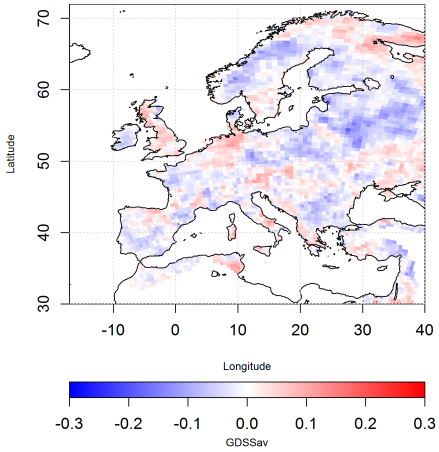
(i) CCLM y6-10 CRPSS



(j) CCLM y6-10 CRPSSav

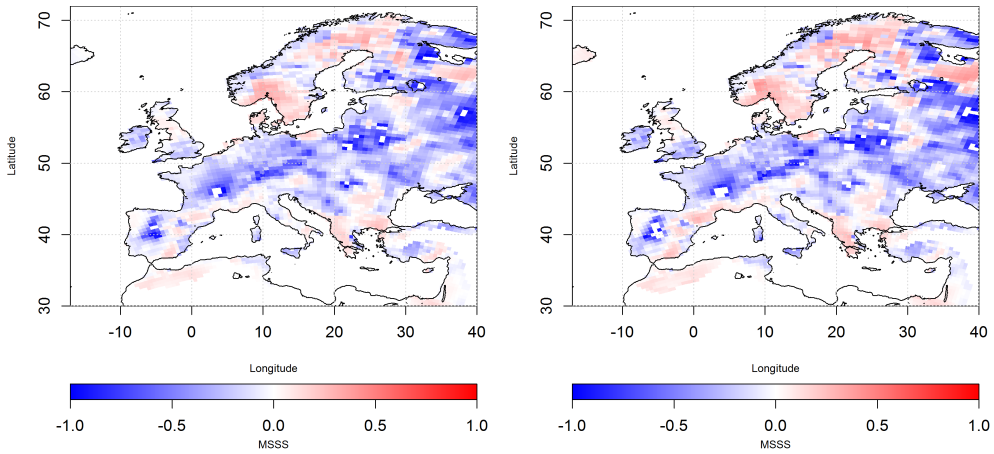


(k) CCLM y6-10 GDSS



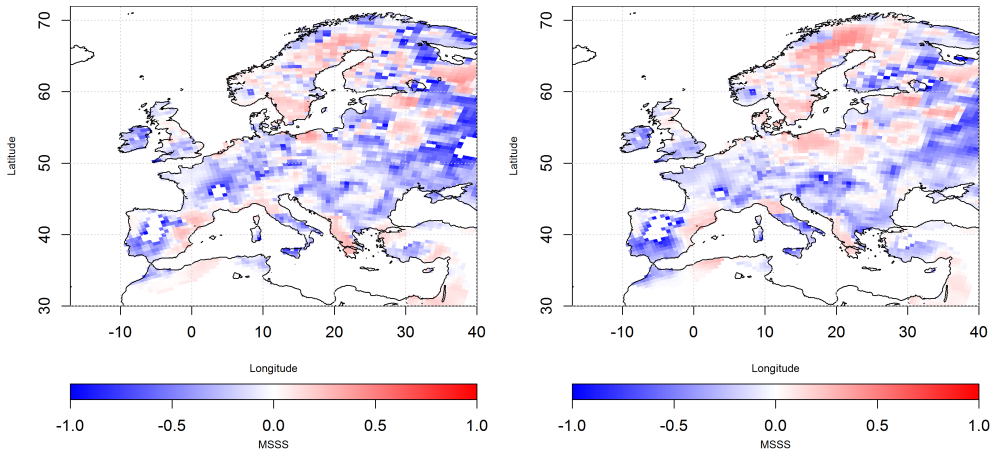
(l) CCLM y6-10 GDSSav

Figure 4.12.: Continued.



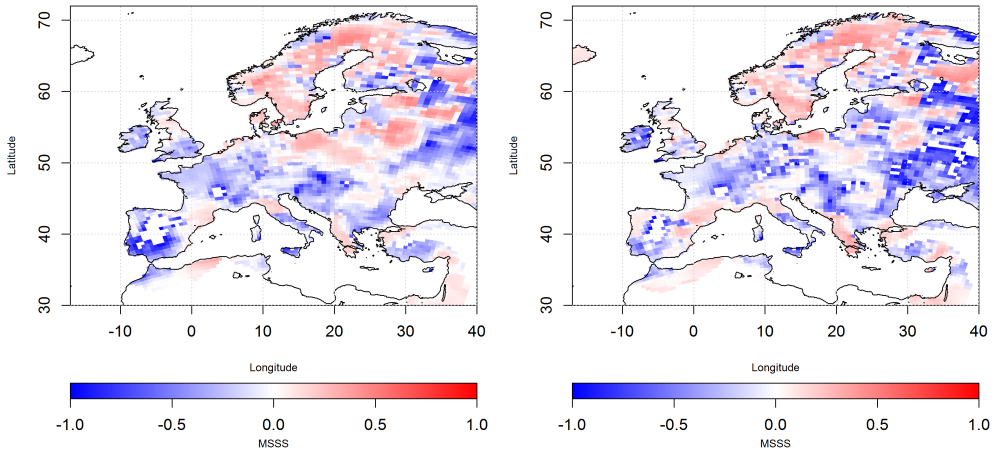
(a) CCLM y2-5 MSSS

(b) CCLM y2-6 MSSS



(c) CCLM y3-8 MSSS

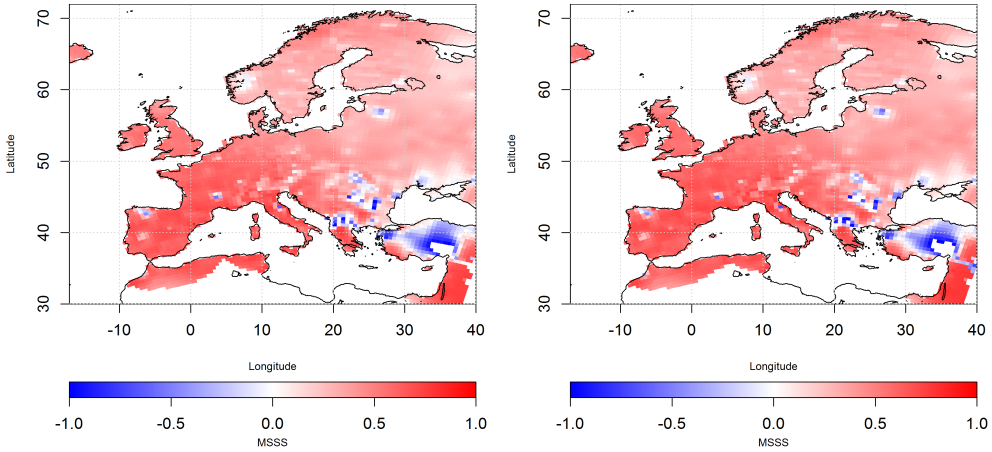
(d) CCLM y6-9 MSSS



(e) CCLM y6-10 MSSS

(f) CCLM y2-9 MSSS

Figure 4.13.: The ***MSSS of b1-EUR44 CCLM hindcasts of annually initialized mean precipitation*** *over Europe from 1961 – 2010 is shown for different lead times averaging.*



(a) CCLM y2-5 MSSS

(b) CCLM y2-6 MSSS

(c) CCLM y3-8 MSSS

(d) CCLM y6-9 MSSS

(e) CCLM y6-10 MSSS

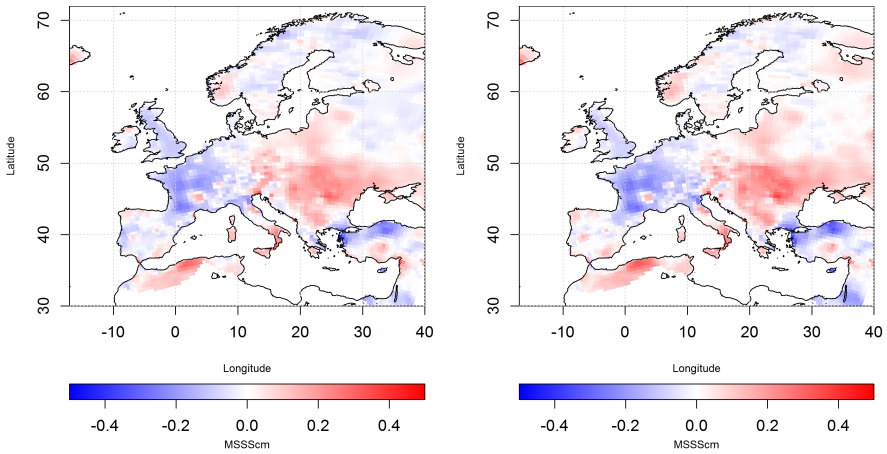
(f) CCLM y2-9 MSSS

Figure 4.14.: Accuracy skill of the ***b1-EUR44 hindcasts of annually initialized mean temperature*** *over Europe from 1961 – 2010 is shown for different lead times averaging.*

times 2–6 and 6–10 to have an average spanning at least 5 years which is probably the extend of the potential predictability for temperature (Ch. 3), as well as 3–8, to have an average over the middle of the decade where neither initialization nor external factors are dominant drivers. Doing that I will be able to ascertain the abilities of the MiKlip regional decadal system within the boundaries of conventional handling of decadal predictions. This time the long term trend was not removed. Parts of the skill will stem from the realistic prediction of external factors such as green house gas concentrations, especially with averages of the latter pentade.

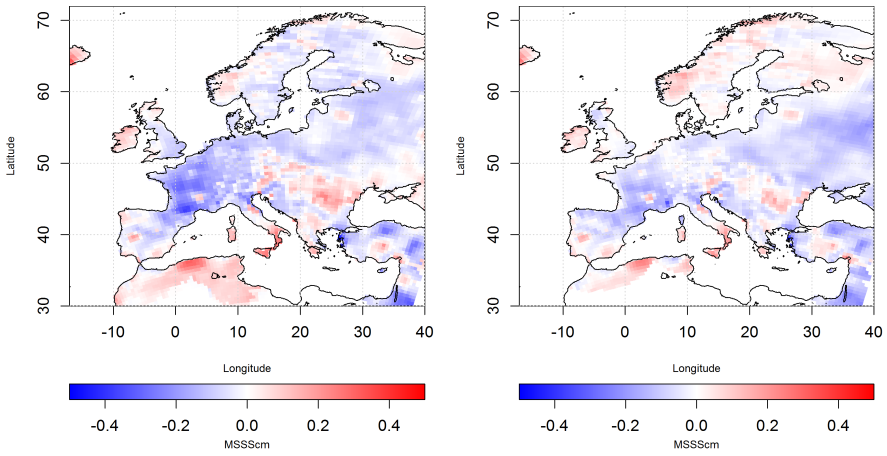
The Fig. 4.11 and 4.12 show the skill scores MSSS, CRPSS and GDSS with the climatology as reference and the added value for two selected lead year averages for the second ensemble generation of CCLM (b1-EUR44). Two variables were considered here: The annual mean temperature and the annual sum of precipitation. The annual initialization improves the skill of temperature predictions significantly in comparison to the decadally initialized ensembles (see Fig. 4.8). There are only small differences between the lead year averages. The skill decreases slightly towards East Europe for lead year averages 2–6. That is not the case for the second half of the decade. Regionalisation does increase the skill in Eastern Europe in the first half of the decade. That improvement through downscaling does vanish later in decade. During the second pentade, as the climate trend becomes the dominant factor driving the decadal climate, local climate variations become less important. Also, in the case of the MSSS score the regional model does overall perform worse than the MPI-ESM, indicating a problem of reproducing the trend. As before, the skill of precipitation presents a less homogeneous pattern. The area of positive discrimination skill is the largest, while the area where the regional model outperforms the global model does in almost all cases exceed 40 percent of the whole investigation area (see also Tab 4.4). For the MSSS and the CRPSS, positive skill scores do not cover more than roughly 40 percent of the area.

To further compare the different lengths of lead year averages the skill score MSSS of the CCLM b1-EUR44 ensemble summarized for 6 different lead year averages in Figs. F.3 and F.4 as well as in Tabs. 4.3 and 4.4 for temperature and precipitation respectively. It does help to illustrate the small differences between the lead time averages. The spatial patterns of the skill do not change. Some locally restraint areas that persistently show no skill in many analyses can be a result of errors in the observations, too. Lead time averages do, however, make a difference in added value of downscaling as illustrated in Fig. F.4. As before, the advantage of the CCLM in Eastern Europe does decline with longer lead year averages as well as lead year averages of the second half of the decade. The best lead year averages is difficult to ascertain, as it changes with score and variable. To summarize it broadly the best averages would probably be not too short (2–5) and not during the middle of the decade (3–8). Lead time averages of covering at least 5 years during the beginning of the decade or the whole decade should produce the best



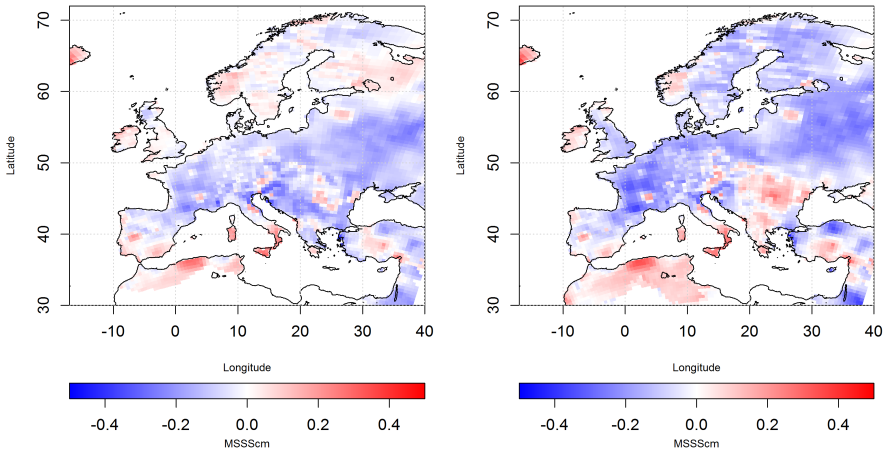
(a) CCLM y2-5 MSSSav

(b) CCLM y2-6 MSSSav



(c) CCLM y3-8 MSSSav

(d) CCLM y6-9 MSSSav



(e) CCLM y6-10 MSSSav

(f) CCLM y2-9 MSSSav

*Figure 4.15.: The **value added** to the accuracy of decadal hindcasts by downscaling the **b1-EUR44 ensemble of annually initialized mean temperature** over Europe from 1961 – 2010 is shown for different lead times averaging.*

result based of the limited b1-EUR44 ensemble. The problem of insufficient evidence for the best lead year average is the topic of the next chapter.

*Table 4.3.: Skill of **b1-EUR44 CCLM ensemble**: The skill of the EUR44 annually initialized ensembles for **temperature** is the percentage of grid points of positive skill and the added value. These numbers do not take into account how high the actual skill is, only if it surpasses the limit for a skilful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS. The result of the use of several lead time averages including the recommendations by Goddard et al. [2013] are analysed.*

	y2-6	y2-5	y2-9	y3-8	y6-9	y6-10
CRPSS	90.2	87	95.1	95	95	95.4
av	47.9	40.5	29.3	32.7	22.6	22.5
MSSS	93.6	92.6	95.7	95.4	96.6	96.3
av	63	53.7	30	38.5	32.7	30.7
GDSS	98.9	98.8	99.3	99.5	99.9	99.8
av	55.3	50.4	32.7	36.6	28	20.4
COR	99.2	99.3	99.5	99.1	99.8	99.8
av	66.4	48.7	21.5	28.2	16.4	16.5

Table 4.4.: Skill of ***b1-EUR44 CCLM ensemble****: The skill of the EUR44 annually initialized ensembles for* ***precipitation*** *is the percentage of grid points of positive skill and the added value. These numbers do not take into account how high the actual skill is, only if it surpasses the limit for a skilful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS. The result of the use of several lead time averages including the recommendations by Goddard et al. [2013] are analysed.*

	y2-6	y2-5	y2-9	y3-8	y6-9	y6-10
CRPSS	10.4	8.2	14.7	11.7	12	16.4
av	37.5	34.8	39.8	38.4	39.2	43.1
MSSS	20.6	17.3	25.7	21.4	21.4	27.3
av	32.8	28.5	33.2	31.7	33	37.7
GDSS	69.3	66.7	68.1	62.4	67.1	68.4
av	45.4	44	40.8	36.9	41.3	41.7
COR	69.3	66.4	68.4	63	68.1	67.5
av	44.3	42.4	39.3	37.1	43.9	43.3

4.5. One model vs. two model ensemble

Early attempts of decadal predictions including the CMIP5 experiments [*IPCC*, 2013; *Hurrell et al.*, 2010; *Taylor et al.*, 2012] focused on a consortium of global modelling groups. That multi-model ensembles can be successful has been shown in many studies. A multi-model can even locally outperform the best performing model within the ensemble [*Weigel et al.*, 2008]. That apparent paradox has been attributed to the fact, that multi-model ensembles do decrease overconfidence, i.e. widen the spread and reduce ensemble mean error.

In light of this can the MiKlip project found lacking, as it only includes one global model. But for some simulations another downscaling with the model REMO does exist. The REMO ensemble is by no means complete and to the extend of the CCLM ensemble but in some instances a comparison can be drawn. Fig. 4.16 the spatial distribution of the skill scores and the correlation of the b0-EUR22 ensembles of CCLM and REMO (10 members respectively) for the decade 2001–2010. The differences are small, as both models do downscale the same global model. However the area of positive skill scores, or significant correlation does increase with the inclusion of the second regional model despite its obvious errors (e.g. in Eastern Europe). This small excursion into multi-model decadal prediction is not complete but is promising. While it is difficult to built multi-model ensembles during a national project like MiKlip, global efforts like in CMIP5 successfully make use of multi-model ensembles and are vindicated.

4.6. Verifying climate indices

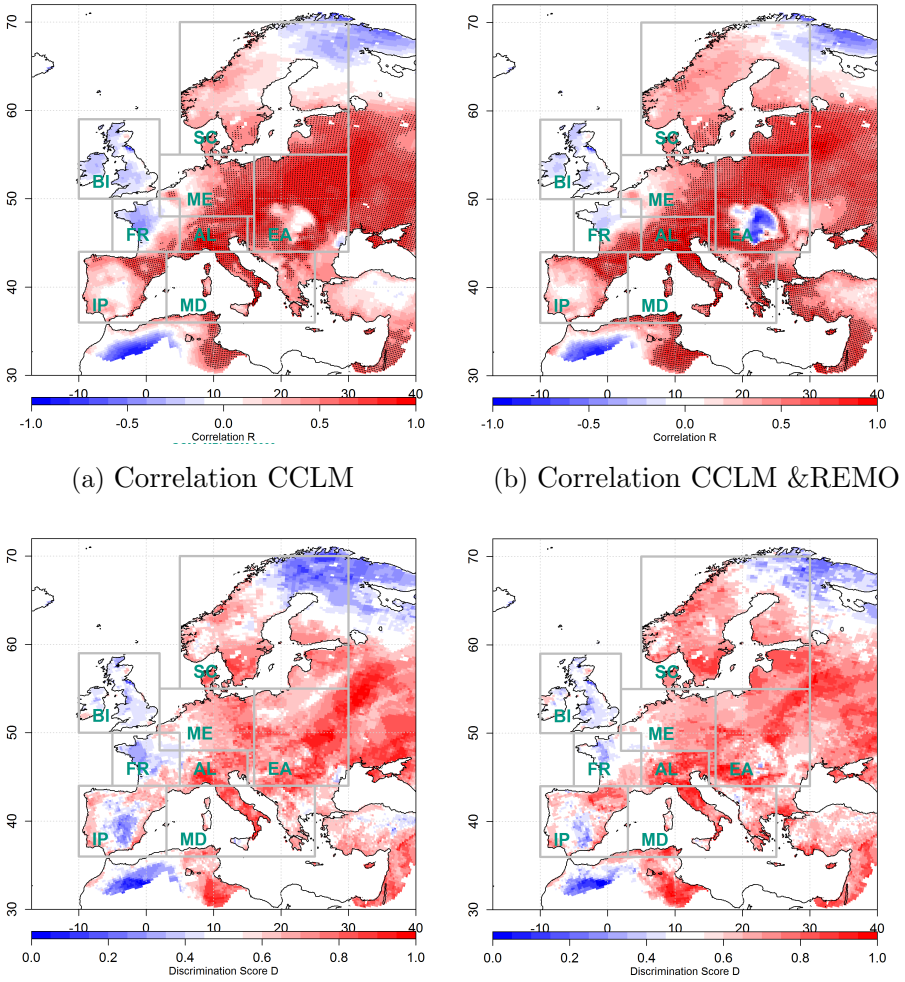
Forecast utility has already been mentioned as a vital part of forecast "goodness". The term "utility" does include both the "the fitness for some purpose or worth to some end" and "something useful or designed for use" [*Merriam Webster Dictionary*, 2016]. Decadal predictions are still a young field of research and the use of such forecasts for actual decision makers in fields like politics and economy is still discussed. But for the sake of a more complete picture about the abilities of the current MiKlip ensemble, one foray into the utility shall be made. In the beginning it has to be mentioned that I have not been in contact with any potential suers of the Miklip products and anything presented here has been selected to my own best judgement.

The performance of the MiKlip ensembles regarding mean annual values has been discussed in length before, but of more interest to potential users would be extremes. Extremes are representative of the tails of the probability distribution and are by definition generally more difficult to realistically represent in climate models [*IPCC*, 2013]. Extreme are often associated with a smaller scale spatial structure that may be better represented by high resolution models, but there are also also large scale and long-duration extremes generally related to persistence of weather patterns. Already the *IPCC* [2007] concluded that, perhaps surprisingly, the global statistics of the extreme events in the current climate, especially temperature, are generally well simulated and models have been more successful in simulating temperature extremes than precipitation extremes. On top of that, MiKlip includes the downscaling of its global prediction system.

To document the skill of the MiKlip system regarding extremes some ETCCDI indices were selected, as well as percentiles of temperature and precipitation and the new index TQI. The analysis does include all indices discussed in Ch. 3 The same skill scores as before were calculated for the global and the regional b1-EUR44 ensembles. Tab. 4.5 summarizes areas covered by positive correlation as a percentage of the whole investigation area.

In accordance to the potential predictability discussed before, the highest scores can be found for summer indices as well as the two indices containing a memory of temperature (GSL and TQI). The added value of downscaling (even rows) shows the same behaviour. The performance improves for the second half of the decade (lead year averages 6-9), when the long term climate trend becomes prevalent and thus extremes more important. Predictions of over the whole decade show no consistent improvement or deterioration of skill compared to the first half of the decade.

It should be noted, that the areas of positive skill in Tab 4.5 are generally smaller than those of the annual mean variables. While the potential predictability of some extremes may exceed the predictability of mean variables, the ability of models to reproduce them might not. On top of that, the analysis of indices that calculated by days exceeding absolute levels like the



(a) Correlation CCLM

(b) Correlation CCLM &REMO

(c) GDSS CCLM

(d) GDSS CCLM &REMO

Figure 4.16.: Verification of the annual means of temperature anomalies over Europe using the correlation and the GDSS of the baseline0 hindcast of the decade 2001–2010. The calculation of the skill scores used the observed climatology as reference and the data was smoothed with a 5 year filter.

ETCCDI indices, require a bias correction because of model bias. The bias correction used here is a relatively simple one that adjusts the location of the prediction's probability distribution. That might not be the best way to correct bias. Only using appropriate percentiles and not correcting the model bias could yield better results.

The added value, on the other hand, is increased compared to the mean variables. The area covered with correlation higher in the regional model than the global one for the leas year average 6–10 of temperature is 16.5 % (Tab. 4.3). In comparison the Growing Season Length (GSL) and the Ticks Questing Index (TQI) are improved over 46.2 % and 52.3 % of the investigation

area by downscaling. Such an improvement is almost always the case, further illustrating the importance and necessity of high resolution predictions.

Conclusion

Before one may generalise statements about out ability to predict the climate up to 10 years, the current prediction system has to be analysed intensely. Both ensemble generations of the global and regional models in MiKlip have now been put through an extensive verification framework. I found, that the skill varies with region, season and variable. In general temperature predictions are more skilful than precipitation. Also summer variables do mostly better than winter variables, while annual means can outperform both. Regional patterns of skill do vary with season, in that the skill in the south of Europe is better in the summer months than in winter. The added value does mostly follow the same pattern as the absolute skill.

In general the regional decadal ensembles of MiKlip show at least a conservation of global skill. Added Value can be found dependent on baseline, region, variable, season and time scale but coverage does not increase beyond 50 percent of the investigation area. Difference between the ensemble generations of MiKlip include an improvement in reliability of precipitation. The skill of the regional model and the added value is also dependent on the global skill. The increase in skill of temperature is small as the original skill of the global model is already high and therefore potential for improvement small also.

When analysing the annually initialized predictions, results form the previous chapter were proven as the skill is improved by choosing longer lead year averages than those proposed by *Goddard et al.* [2013]. The potential predictability of variables analysed before does range from 5-12 years, while the averages of the verification framework only include 4 years. However, that is only a preliminary statement. In the following I will take an extensive look into different averages and their influence on skill.

Table 4.5.: Correlation skill of the ***b1-EUR44 CCLM ensemble*** *in predicting several* ***ETC-CDI recommended climate extreme indices*** *[Karl et al., 1999; Peterson et al., 2001; Zhang et al., 2005; CLIMDEX, 2013] as the number of grid points with positive correlation relative to the total number of grid points (odd rows), while the added value is defined as the percentage of grid point where the correlation with the observation of the CCLM is higher than that of the MPI-ESM regardless of the actual skill. The rows are designated in the following: The first part names the indix, with "p95 jja" being the 95th percentile of the summer months JJA, and the second part naming the variable, the index was calculated upon ("pr" is the precipitation, "tas" the temperature). The indices are defined in Tab. 3.1 on p. 32.*

	y2-6	y2-5	y2-9	y3-8	y6-9	y6-10
p95jja pr	58.5	58	58.2	53	61	62.9
	52.5	50.1	50.1	49.4	48.4	45.4
p05djf pr	45.2	52.8	52.7	43.9	59.9	62.6
	48.2	51	44.8	42	50.4	48.4
p95jja tasmax	66	59.1	49.2	51.1	51.5	47.3
	59.9	49.1	39.5	43.9	40.8	37.3
p05djf tasmax	21.8	30.7	18.6	20.5	36.2	41.3
	48.9	46.5	44.1	42.1	44	47.9
ID tasmax	0	0	0	0	0	0
(Icing Days)	0	0	0	0	0	0
SU tasmax	53.9	63.2	57.1	39.5	57.7	58
(Summer Days)	54.8	47.9	40.5	38.6	39.3	29.8
p95jja tasmin	49.4	51.9	46.4	44.5	62.6	62.1
	39.4	37	37.4	41.1	37.1	37.6
p05djf tasmin	41.7	49.2	26.9	28.8	45.4	46.1
	64.6	66.2	68	59.1	60.5	58.9
FD tasmin	0	0	0	0	0	0
(Frost Days)	0	0	0	0	0	0
TR tasmin	55.4	74.6	61.7	41	87.9	90.8
(Tropical Nights)	53.5	54.1	45.9	46.6	49.7	49.5
GSL tas	55.5	58.3	49.6	50	60.5	63.4
(Growing Season Length)	56.7	55	63	66.1	63.8	62.9
TQI tas	57.8	58	54.1	49	59.9	69.4
(Ticks Questing Index)	52.8	52.7	56.8	53.7	56.5	63.3

5. Consideration of optimal scales for decadal predictions

In 1925, G. U. Yule talked about correlation in his presidential address at a meeting in the Royal Statistical Society:

"t is fairly familiar knowledge that we sometimes obtain between quantities varying with the time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly "significant." [...] When we find that a theoretical formula applied to a particular case gives results which common sense judges to be incorrect, it is generally as well to examine the particular assumptions from which it was deduced, and see which of them are inapplicable to the case in point." [*Yule*, 1926]

Over 90 years ago the problem of tests for significance was already an issue: The usual tests of significance for correlation coefficients require independence of the different observations in both variables, and this requirement is rarely met in the case of time series that are more often than not autocorrelated. Geophysical time series are frequently autocorrelated because of inertia or carry over processes in the physical system. For example, the slowly evolving and moving low pressure systems in the atmosphere might impart persistence to daily rainfall. Generally there is no way around that problem and is therefore widely ignored.

Decadal predictions are typically assessed using multi-year means. The commonly used averaging windows are chosen on the basis of the following assumptions: The first year of a decadal run carries the initialization shock and the relative uncertainty of a variables prediction (esp. temperature) decreases with lead year as the predictions change from an initial state dependence to the forced response to external drivers [*Meehl et al.*, 2009]. Therefore lead year averages are separated into year 1 and the first and second half as well as the whole of the decade sans the first year (1, 2-5,6-9 & 2-9). The overall performance of the MiKlip ensembles has been analysed using these lead time averages as well as some additional one as explained in Sec. 4.4. While the reasoning is sensible, the selection of lead year averages is intuitive following recommendations e.g. given by *Goddard et al.* [2013] and have been adopted into wide spread. But are these averages really the best choice?

Within the regional division of the MiKlip research, we are uniquely equipped to approach this subject matter analytical. Within the MiKlip project several hindcast ensemble have been produced spanning 50 years, including high resolution regional predictions, allowing for an evaluation of all possible averaging windows for their influence on the predictive skill.

In theory all filtering will lead to an increase in skill first and foremost because of a reduction in noise. However, there can be disadvantages of filtering. Once data is filtered too much, not only noise but also the desired signal is removed. Filtering also increases the autocorrelation of the data and consequently decreases the significance of any statistic calculated. Is there an optimum? Before finding an answer, a methodology to describe the significance of autocorrelated data is needed. Then the predictive skill and its significance of filtered hindcasts will be used to find an optimum. Temporal and spatial autocorrelation has to be viewed differently and will be considered in two independent analyses.

5.1. Testing the significance of association between time series

At the beginning of the consideration of optimal scales the significance of skill, e.g. the correlation coefficient, has to be described. To test the significance of statistics by conventional tests one needs the data's degrees of freedom. However, meteorological data – spatial fields as well as time series – is more often than not autocorrelated. In such a case the derivation of the number of degrees of freedom is not trivial. Several methods to test the significance of statistics of autocorrelated data exist:

- The number of degrees of freedom are reduced by a certain amount that takes into account the autocorrelation [*Orcutt and James*, 1948; *Thiébaux and Zwiers*, 1984]. This method is applicable to autocorrelated time series as well as spatial fields [*Bretherton et al.*, 1999].

- The autocorrelation is removed from the data and the statistical analyses and the tests are performed as usual.

- The analyses are performed for the actual data as well as for several randomly constructed data sets that exhibit the same properties as the original data. Two data sets are independent if the correlation is not higher than expected by chance [*Krueger and Lienert*, 1980; *Goddard et al.*, 2013]

Using synthetic time series and fields and testing the above mentioned against each other, several questions are answered:

1. Should the choice for a method be dependent on the length of the time series or the size of the field?

2. Must one first consider the strength of the autocorrelation before choosing the method?

   And ultimately:

3. Is there one method that can be used in either case?

AR(1)-Process

To test the three methods using autocorrelated time series and spatial fields of different length and size with differing autocorrelation a first order autoregressive process AR(1) is used as an algorithm to generate synthetic time series [5.1a] and spatial fields [5.1b] [*Wilks*, 2006].

$$X(t) - \mu_X = \delta(X(t-1) - \mu_X) + \epsilon(t) \quad [5.1a]$$

$$X(x,y) - \mu_X = \delta(X(x-1,y-1) - \mu_X) + \epsilon(x,y) \quad [5.1b]$$

Whereas the predictand is $X(t)$ and the predictor $X(t-1)$ and μ the mean of the data X. δ is the autoregressive parameter that for a AR(1) process is simply the sample lag 1 autocorrelation coefficient.

$$\delta = r_1$$

The right-hand side of Eq. [5.1a] and Eq. [5.1b] consist of a deterministic part in the first term, and a random part in the second term (white noise). ϵ are mutually independent random quantities having a mean $\mu_\epsilon = 0$ and a white noise variance $\sigma^2\epsilon = (1 - \delta^2)\sigma_X^2$.

To determine the most reasonable method for significance testing, a set of synthetic time series was constructed. Therefore Eq. [5.1a] is deployed as follows: Beginning with an initial value $x_0 - \mu = 1$ multiplied with the autoregressive parameter ranging from $\delta = 0.1...0.9$, a randomly generated variable ϵ drawn from a Gaussian distribution with mean zero and variance $\sigma^2 = 0.1...0.9$ the first value if the time series x_1 would be then produced by adding back the mean mu, that is as well a random value from a Gaussian distribution similar to ϵ. The next time series value x_2 is then produced in a similar way, by operating with x_1 and so on until the 120th value is reached.

Non-parametric test using surrogates

A non-parametric test using surrogate time series is used as reference to compare the methods to be introduced in the following. In contrast to every other methods,the non-parametric test does not make use of the t-test that assumes the statistic, here the correlation coefficient, to be distributed following the *Student-t-distribution* [*Bleymüller et al.*, 2004]. The surrogate method is similar to a parametric bootstrapping, used to detect non-linearities in time series. This method specifies a null hypothesis H_0 describing a process and calculates the statistic for the original time and repeats the analysis with the surrogate data set. If the statistic differs significantly from the statistics for the surrogates, the null hypothesis is rejected [*Venema et al.*, 2006b].

The surrogate time series used in this preliminary study were compiled the same way as the original time series. Only the randomly generated white noise part of the AR(1) process

differs between the surrogates. In the case of the synthetical time series the compilation of the surrogates is simple, with more complex data as meteorological time series the methods for conceiving the surrogates become complex and time consuming. The goal of this study is to find an easier and faster way of calculating the significance without forgoing the accuracy of the surrogate test.

The effective number of degrees of freedom

Parametric statistical test require knowledge of the data's independence. If the data is truly independent (lag 1 auto correlation $r_1 = 0$), the degrees of freedom are essentially the length of the time series n. However, is the data autocorrelated ($r_1 > 0$) the *effective sample size* $n' < n$ represents the degrees of freedoms. The effective sample size n' can be calculated as follows [*Wilks*, 2006]:

$$n' = n\frac{1-\delta}{1+\delta} \qquad [5.2]$$

This formula takes into account only the lag 1 autocorrelation coefficient. Similarly, another method of calculating the effective sample size does not only consider the lag 1 autocorrelation coefficient but the whole autocorrelation function. The autocorrelation function $\rho(\tau)$ with τ as the lag theoretically includes all lags $\tau = -\infty...\infty$. Yet, time series in meteorology are finite and the theoretical autocorrelation function not known. *Thiébaux and Zwiers* [1984] showed an applicable way of calculating the effective sample size taking including the complete autocorrelation:

$$n' = n \times \left[\sum_{\tau=-(n-1)}^{+(n-1)} (1 - \frac{|\tau|}{n}\rho(\tau))\right]^{-1} \qquad [5.3]$$

But if testing a covariance of two time series the time series' autocorrelation does not have to be the same. *Orcutt and James* [1948] introduced an effective sample size for testing a correlation.

First one may calculate the variance of the correlation coefficient r

$$var(r) = \frac{1 + r_1 r_1'}{n(1 - r_1 r_1')} - \frac{2r_1 r_1'(1 - r_1^n r_1'^n)}{n^2(1 - r_1 r_1')^2}. \qquad [5.4]$$

Whereas r_1 is the lag 1 autocorrelation coefficient of one time series and r_1' of the other. If this value is equal to or smaller than $var(r) \leqslant 0.25$, one can estimate the effective sample size by the means of

$$n' = \frac{1}{var(r)} + 1 \qquad [5.5]$$

and perform the test. If $var(r)$ is greater than 0.25, it is unlikely that this method is valid. Nevertheless it can be applied when keeping in mind that the test of significance would appear to be stronger than necessary. However, while performing this test several times with different sets of synthetic timeseries, it became apparent that $var(r) > 0.25$ almost never occurs.

Removal of autocorrelation

Surely the most consequential way of dealing with autocorrelation is the removal thereof. Applying the Autoregressive Model [5.1a] backwards one theoretically reduces the data to uncorrelated white noise. The normal *t-test* with $n-1$ degrees of freedom can than be performed. A drawback of this methods is that only a first order autocorrelation is assumed here. But AR(1) processes exhibit higher dependencies despite their name due to the sequential order of their build.

Comparison of methods for significance estimation

To find the best method for significance estimation three sets of 20 time series each are produced following Eq. [5.1a], distinguished by the range of autocorrelation that was assigned: small $r_1 = 0.1...0.3$, moderate $r_1 = 0.4...0.6$ and high $r_1 = 0.7...0.9$. Then the Pearson Correlation Coefficient is calculated between every timeseries of each set, using the whole length of the time series as well as only the first 30, the first 60 and the first 90 time steps of each time series to estimate the influence of the sample size. $20 \times 4 \times 3$ correlation coefficients are derived and tested for significance.

For each length and all three ranges of autocorrelation 20 AR(1) processes were calculated (240 time series). To keep the analysis simple only the 20 time series within one group of autocorrelation were combined 190 times to calculated the correlation coefficient and its significance. The distributions of the 190 correlation coefficients and their respective significances from 4 different *t-tests* were compared to the results of the non-parametric test in consideration of the length of the timeseries and the strength of autocorrelation. I.e. mean squared difference between the results of *t-tests* using

1. method: Effective sample size following *Wilks* [2006] and Eq. 5.2
2. method: Effective sample size following *Thiébaux and Zwiers* [1984] and Eq. 5.3
3. method: Effective sample size following *Orcutt and James* [1948] and Eq. 5.4
4. method: Removal of autocorrelation and performing a standard *t-test*.

and the surrogate were calculated,

$$qq = \overline{(Sig_{surrogates} - Sig_{other})^2} \qquad [5.6]$$

as well as the significance of the diversity of distributions derived from a χ^2-Test. In the case of the degrees of freedom after *Wilks* [2006] and *Thiébaux and Zwiers* [1984] where the degrees of freedom are calculated for each time series seperately and the average out of both was used to perform the *t-test*.

Table 5.1.: Compilation of synthetical time series derived from AR(1) processes.

r_1 \ n	$n = 30$	$n = 60$	$n = 90$	$n = 120$
$r_1 = 0.1...0.3$	20 time series plus 100 surrogates	20 time series plus 100 surrogates	20 time series plus 100 surrogates	20 time series plus 100 surrogates
$r_1 = 0.4...0.6$	20 time series plus 100 surrogates	20 time series plus 100 surrogates	20 time series plus 100 surrogates	20 time series plus 100 surrogates
$r_1 = 0.7...0.9$	20 time series plus 100 surrogates	20 time series plus 100 surrogates	20 time series plus 100 surrogates	20 time series plus 100 surrogates
Complete number of time series: 240				

Dependency on the length of the time series

When considering time series of a length between 30 and 120 time steps, the method of significance estimation does not matter. The number of time time steps $n = 30$ is a usually stated as the minimum sample size below which most common statistical analyses become unreliable. All methods of significance estimation lead to larger spread with longer timeseries, meaning a larger distribution of correlation coefficient and subsequent significances can emerge, the larger the sample size. The difference to the reference method of surrogates also increases with higher sample size. This behaviour is shared by all 4 other methods of significance estimation.

Dependency on the strength of autocorrelation of the time series

I find, higher autocorrelation leads to higher agreement of the three methods using a revised number of degrees of freedom and the reference method (surrogates). Contrarily the method that uses a common *t-test* with data clear of their autocorrelation the difference to the reference method increases with higher autocorrelation. That is due the removal of the first degree autocorrelation when the autocorrelation is actually of a higher degree. The removal does not

yield white noise but only reduces the autocorrelation, therefore making this method the most inaccurate one out of all tested.

Conclusion: Which method to use?

To find the significance of a correlation coefficient a *t-test* is commonly used. But a *t-test* does not take into account the grade of internal organization of the data. Several methods for reducing the number of degrees of freedom have been published in different instances. While the most accurate one is the non-parametric test, that does not rely on the *t-test*, a *t-test* is easy to perform and less time consuming than any surrogate method. Subsequently, the goal is to choose a method of significances estimation as an alternative to the surrogate method. Out of all four methods tested against the surrogate method the two ways of reducing the degrees of freedom as in equations 5.2 [*Wilks*, 2006] and 5.4 [*Orcutt and James*, 1948] show the highest agreements to the reference method. Out of these two I recommend the way described by *Orcutt and James* [1948] as it takes into account both time series' autocorrelation. The reduced number of degrees of freedom by *Wilks* [2006] does only require one autocorrelation. In the case of this analysis the time series correlated are of similar autocorrelation out of simplicity. That, among others, leads to the high success rate of this method. In reality it can not always be expected, that time series have similar autocorrelation and therefore, the method which takes both time series' autocorrelations into account individually should be chosen.

5.2. Optimal temporal scales for analysing regional decadal predictions

With a significance estimation dependent on the autocorrelation of the time series, it is now easy to analyse the scales of predictive skill of the MiKlip ensembles. The goal is to revise the convention within the decadal prediction community towards certain averaging intervals. In a first instance the focus will be on the temporal averaging. *Goddard et al.* [2013] proposed 4-year averages. I will consider averages from 1 year up to 9 years. To find an estimation purely dependent on the averaging window all possible lead years are considered as it is already known that there exist a lead year dependency of skill.

To do that, all 4 MiKlip ensembles were used. For the b0-EUR22 ensembles that are only initialized decadally the decades were considered separately and only the decade 2001-2010 will be discussed. The other 4 earlier decades show the same pattern and exhibit weaker skill in general than the last decade of the hindcasts. Also, for the last decade the full 2-model ensemble exists. The gridded data of annual means were filtered using running means the size of 3, 5, 7 and 9 years. In the case of the b1-EUR44 ensembles all lead year averages of one particular filter window were pooled together and averaged onto the midpoint of the interval. In Tab. 5.2 such a composite is illustrated.

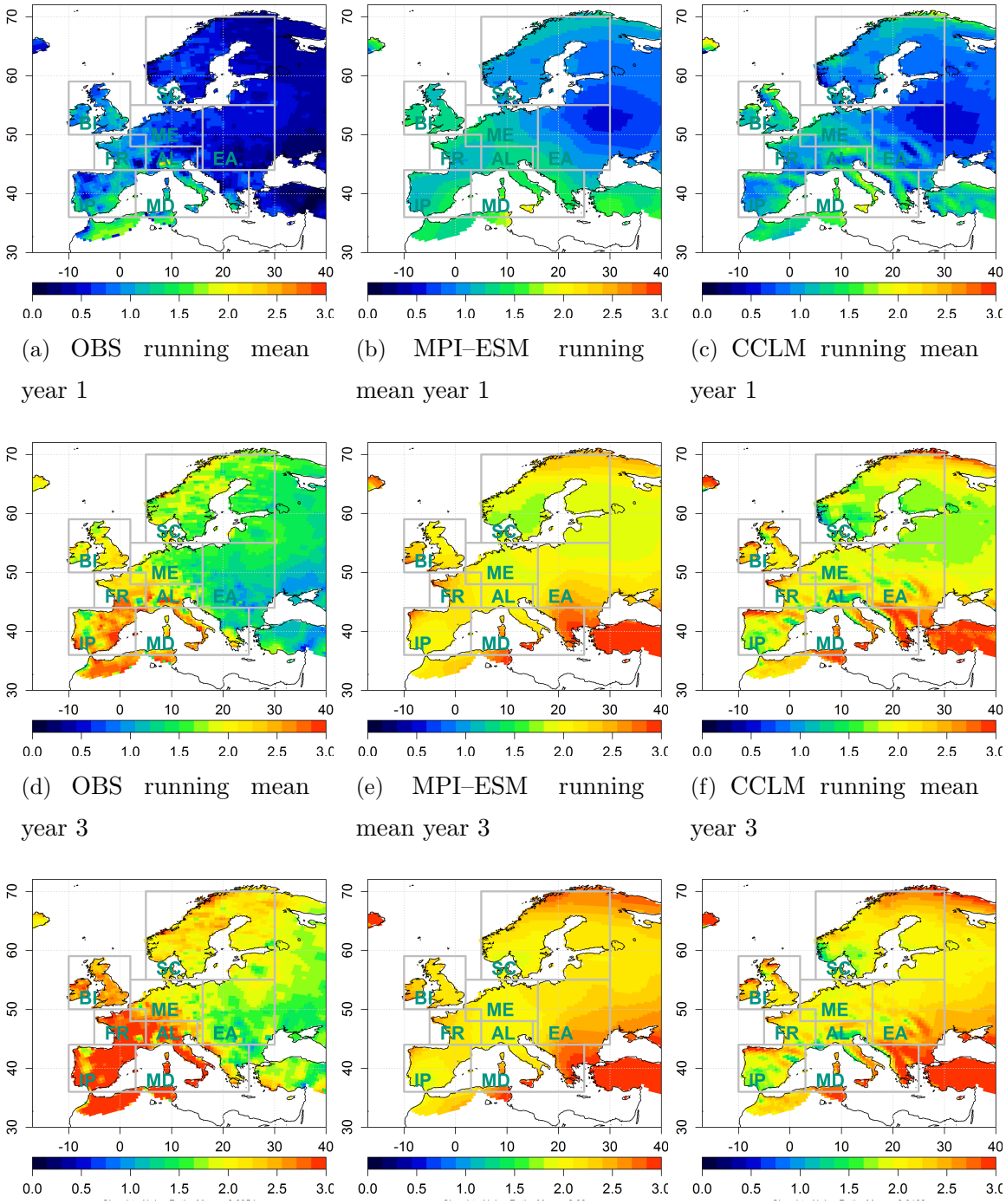
Table 5.2.: A composite for the year 1982 filtered with a three year average using annual initialized decadal hindcasts. When the lead year does not matter, only the length of the average, the following 10 initializations will cover the year 1982. For every year all available runs are pooled together like that. The resulting time series contains values that are the averages over all the 3-lead years episodes from all initializations covering the year. In this particular case the value for the year 1982 is ultimately an average over 24 lead year averages.

Start year of the decadal run	lead time averages							
1982	1-3							
1981	1-3	2-4						
1980	1-3	2-4	3-5					
1979		2-4	3-5	4-6				
1978			3-5	4-6	5-7			
1977				4-6	5-7	6-8		
1976					5-7	6-8	7-9	
1975						6-8	7-9	8-10
1974							7-9	8-10
1973								8-10

Fig. 5.1 shows the influence of filtering. The calculation of the Signal-To-Noise-Ratio is discussed in App. C. The Signal-To-Noise-Ratio of the observations well as the b1-EUR44 MPI-ESM and CLM ensembles is pictured. The filtering increases the Signal-To-Noise-Ratio in almost all regions. However, there are differences between the observations and the hindcasts. The observations of temperature in their unfiltered state with a resolution of one year are noisier than the hindcasts but the improvement by filtering is higher. The Signal-To-Noise-Ratio of global model and regional model is structural similar. The CCLM's lower Signal-To-Noise-Ratio indicates that adding information during regionalisation does not necessarily add value to the signal. Regions where downscaling decreases the Signal-To-Noise-Ratio are places where one might expect a stronger deviation from the global model: The eastern part of the analysis domain and mountains regions, due to the higher resolution and the development of downstream,internal systems. In western Europe, the highest Signal-To-Noise-Ratio can be observed in the observations. That is not the case for both models. An explanation could be, that the teleconnections towards the Atlantic ocean, transporting the information (signal)

towards the continent, are weaker in the hindcasts than in the observations. Also, the regional hindcasts present regions where the Signal-To-Noise-Ratio decreases towards the longer filters e.g. the Balkan Peninsula or the northern coast of the Iberian Peninsula. Already, conclusions about the time scale of the predictable signal itself can be drawn. The Signal-To-Noise-Ratio does increase with small filters and decreases with filters larger than 5 years. Therefore, the signal within the data has likely a frequency of less than 7 years. Generally, signals can be boosted by filtering and the subsequent reducing of noise. But filtering will also reduce the signal, when applied to strongly, as it appears in some regions of the regional hindcasts. The optimum can differ regionally. In the following I will attempt to find a reasonable compromise for an optimal temporal scale for regional (and global) decadal predictions over Europe.

In the case of the b0-EUR22 ensembles only one decade 2001–2010 will be discussed. Here only the filters 2,5 and 7 years are used, as the running mean of 7 years conspicuously shortens the already short time series: From 10 data points to 4. Small sample sizes in general are a problem for the significance of the correlation. *Schönwiese* [2013] states a sample size of $n = 30$ as a minimum for many statistical analyses including correlation. The unfiltered time series is already shorter than the recommended sample size and the correlation's significance will be affected no matter what, while the decrease in significant correlation with filtering is a consequence of the decrease of sample size and the increase in autocorrelation. The results of the other 4 decades of the Milklip analysis time frame, exhibit a similar patterns. Figure 5.2 shows the value added by initialization to the climatology in the form of the fraction of area with positive correlation coefficient in the coloured bars for the running averages of 3, 5 and 7 years, as well as the portion of it that is significant above the 80th percentile. The area of positive correlation for all ensembles is highest for the average window of 5 years. The same is valid for the area of significant correlation. The value added by downscaling is illustrated by the dashed lines (in correspondence to the right y-axis) for the CCLM (green). The added value increases with longer filter windows. However this increase does not coincide with an increase in the fraction of area of significant improvement by regionalisation, meaning that while the fraction of area of outperforming grid points still increases with longer filter, the correlation coefficients are not necessarily significant.



(a) OBS running mean year 1

(b) MPI–ESM running mean year 1

(c) CCLM running mean year 1

(d) OBS running mean year 3

(e) MPI–ESM running mean year 3

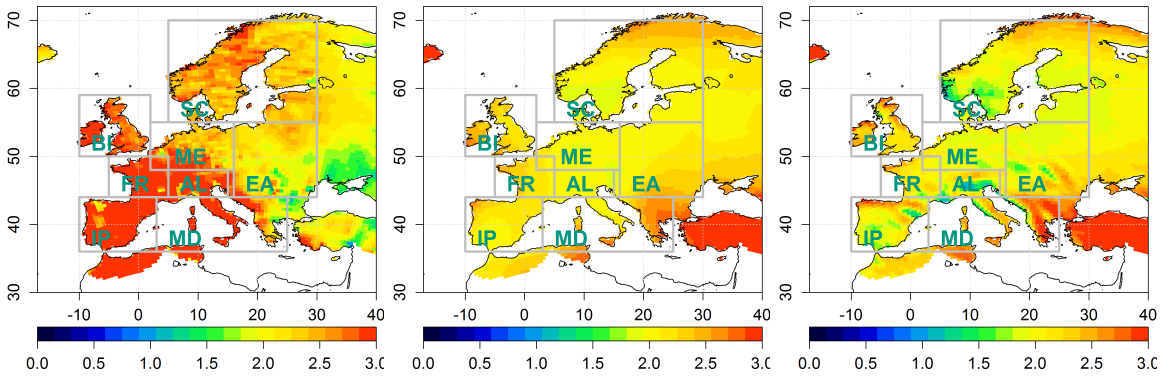
(f) CCLM running mean year 3

(g) OBS running mean year 5

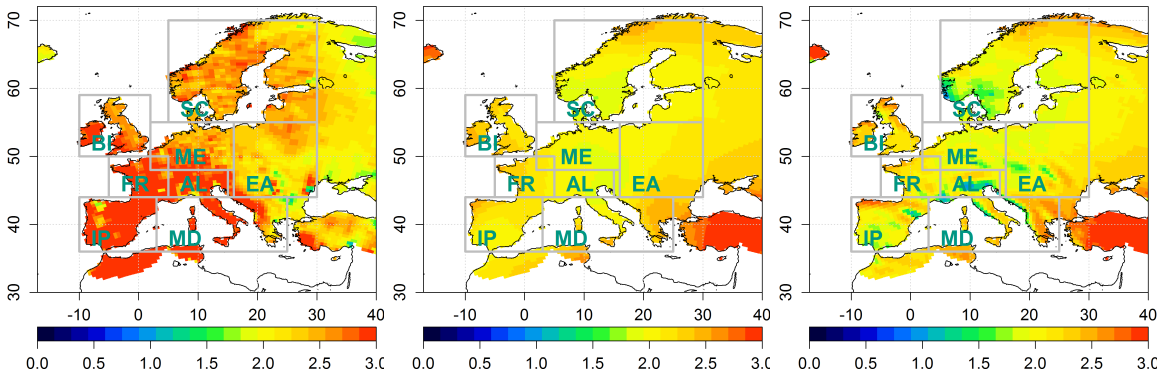
(h) MPI–ESM running mean year 5

(i) CCLM running mean year 5

Figure 5.1.: Signal-To-Noise-Ration of near surface temperature of the b1 EUR44 ensembles of MPI-ESM and CCLM as well as the E-OBS data set following App. C. The averaging windows were of 1, 3, 5, 7 and 9 years of length. Continued on next page.



(j) OBS running mean year 7 (k) MPI–ESM running mean year 7 (l) CCLM running mean year 7



(m) OBS running mean year 9 (n) MPI–ESM running mean year 9 (o) CCLM running mean year 9

Figure 5.1.: Continued

The obvious favouring of the filter 5 years is further vindicated by the results of the skill scores in Tab. 5.3. The table lists the fraction of area of positive skill scores, or in the case of the Discrimination Skill Score above 0.5, and the fraction of the outperforming grid points of the regional model ensemble consisting of 10 Members of the CCLM and 10 members of the REMO regional models for the decade 2001-2010. The conclusion drawn from this small analysis can then be stated as follows: In the case of a small sample size, when the original data already underscores the recommended sample size of the statistic, a medium length filter will increase both skill and significance.

The problem of the small sample size can be addressed, when using annually initialized decadal hindcasts like the b1-ERU44 ensembles. Fig. 5.3 illustrates the influence of filtering on skill scores and correlation as well as the added value (in three different forms) for four variables: temperature and precipitation with and without trend. In the case of accuracy of temperature the improvement between the baselines is apparent. One difference between time series with and without trend apart from the decline in overall skill with detrending is the behaviour with

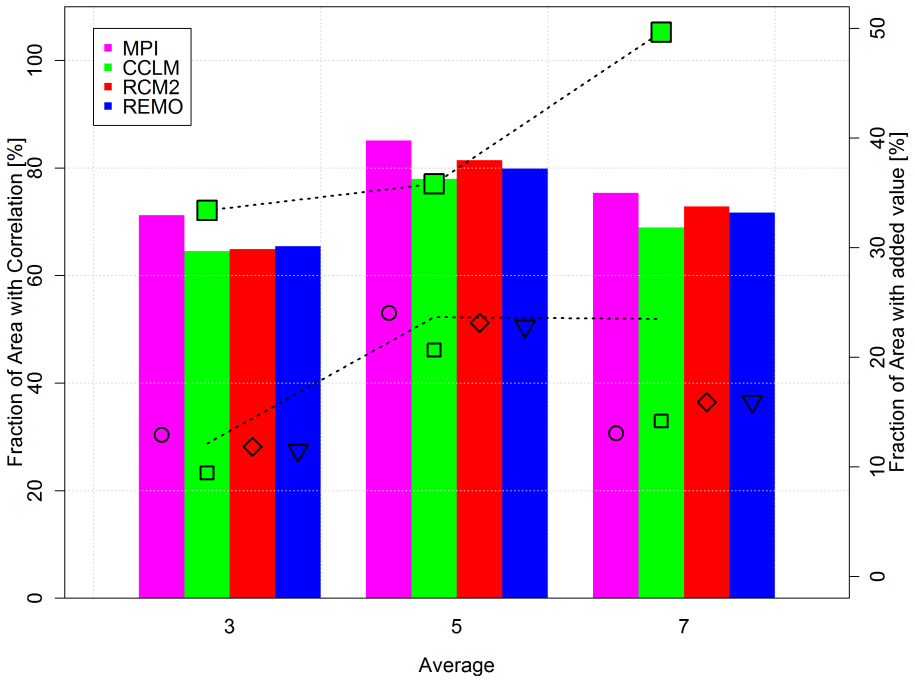


Figure 5.2.: Fraction of Area with a positive correlation coefficient (bars) and a significance of at least 80% (symbols) of the filtered b0 sea surface temperature of CCLM, REMO, both RCMs (RCM2) and MPI as well as the fraction of area where the CCLM correlation is better than MPI-ESM-LR (black, dashed line with green squares) and the percentage thereof that show significant correlation (black, dashed line)]

*Table 5.3.: Fraction of areas with positive skill and added value of **reliability**, **discrimination** and **accuracy** of filtered annual means of near surface temperature of the b0-EUR22 decade 2001 - 2010 (Continuous Ranked Probability Score, Generalized Discrimination Score, and Mean Squared Error Skill Score) for the 20 member regional ensemble of CCLM and REMO.*

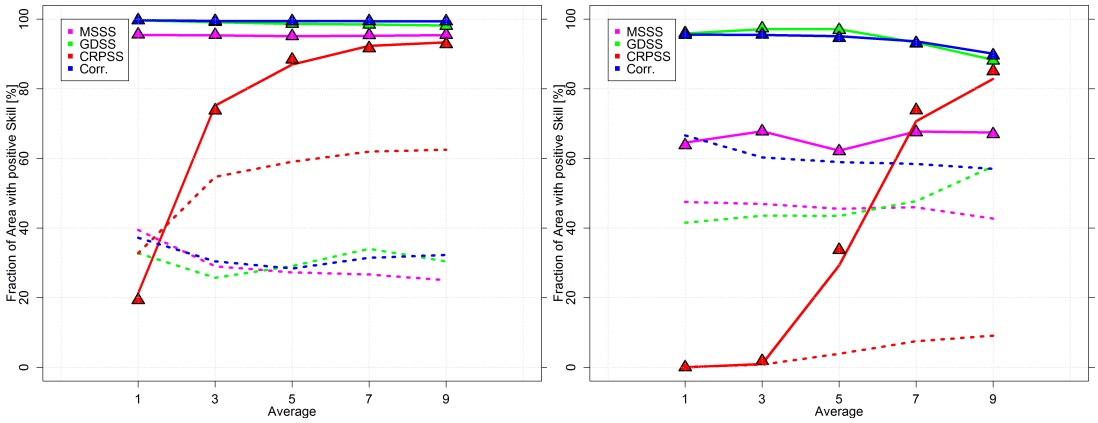
Filter	CRPSS	CRPSSav	D	D_{av}	MSSS	MSSSav
3a	38%	47%	58%	38%	48%	43%
5a	60%	42%	77%	41%	67%	42%
7a	58%	35%	52%	31%	64%	41%

longer windows. While the fraction of area levels out in the case of time series with trend, the skill of the detrended data decreases with longer filters. The extraction of the 50 year trend removes one major source of predictability from the data and the left over decadal signal is too small with long filter windows. The reliability however can be greatly improved by filtering in both cases.

The added value reacts differently to filtering. In the most cases the fraction of area where the regional model outperforms the global one does decrease with longer filter window length. That does not indicate that the regional model performs worse with longer filtering but that the improvement of the global model is bigger.

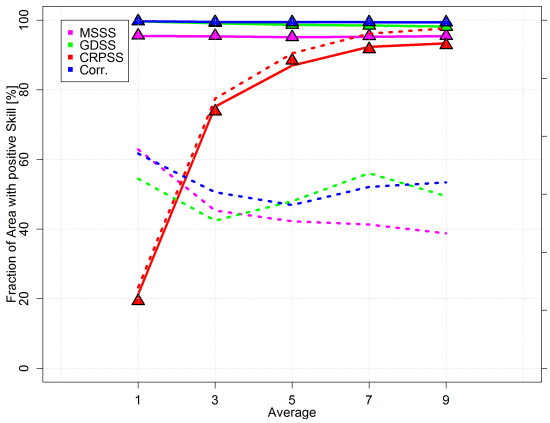
A similar pattern can be found for precipitation. The overall skill of precipitation forecasts is smaller than temperature. There is a big improvement of skill, especially the reliability of the precipitation forecast by regionalisation. The decrease in the fraction of area of positive skill with longer filter windows is more pronounced than in the case of temperature and begins after filter window lengths of 5 years. Detrending the precipitation time series does not change that pattern, but lowers the overall skill of the hindcasts.

Conclusively, it can be stated, that temporal averaging improves skill. In the two cases of decadally and annually initialized hindcast an optimal temporal scale of 5-7 years emerged. Annually initializing increases the scale of signal slightly compared to decadally initialization as it produces longer time series. Detrending temperature does also shorten the time scale where optimal skill can be achieved, as some skill stems from the long term climate trend. In particular, the added value of regionalisation does decrease with longer filter windows except for reliability. Reliability improves with regionalisation and continues to improve with filtering. The decrease in significance is very much dependent on the reduction of time series length as well as the increase in autocorrelation. This stresses the importance of long data sets that can only be achieved by annual initialization.

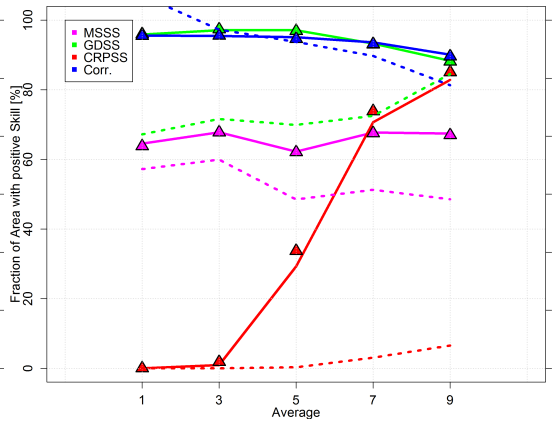


(a) Temperature with trend av1.

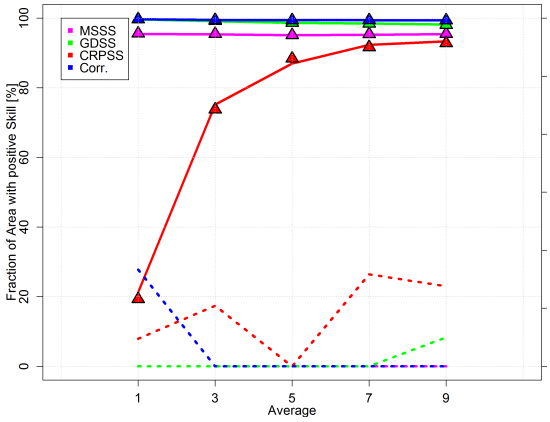
(b) Temperature without trend av1.



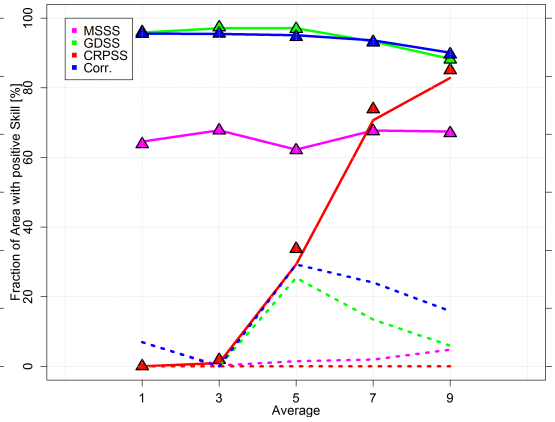
(c) Temperature with trend av2.



(d) Temperature with trend av2.

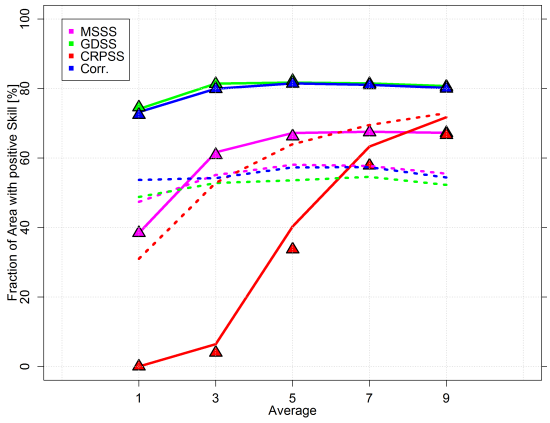


(e) Temperature with trend av3.

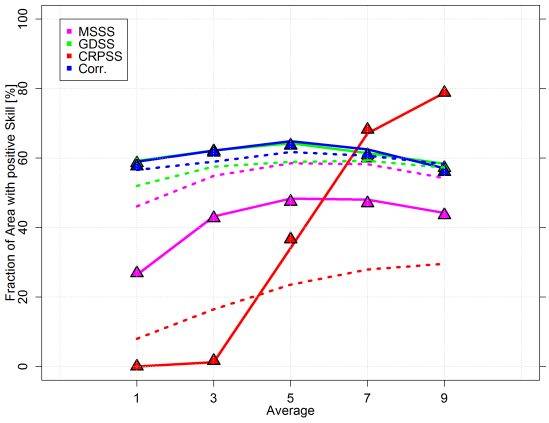


(f) Temperature with trend av3.

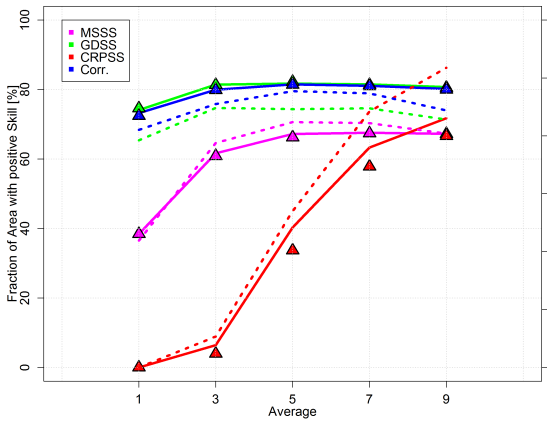
Figure 5.3.: Fraction of areas of positive skill of CCLM (solid Lines), MPI-ESM (Filled triangles) and the value added by Regionalisation (dashed Line, second y-Axis) of mean annual values of b1 surface temperature and b1 precipitation (annual initialization). The three columns represent three different ways of quantifying the added value. The first row shows the added value as the respective formulation of the skill score. The second row (middle) than shows the fraction thereof where the skill of regional model is also positive in addition to being better than the global ensemble. The third row (bottom) then illustrates the added value as it is calculated as the fraction of the potential possible added value v. equation 4.16 in section 4.3)



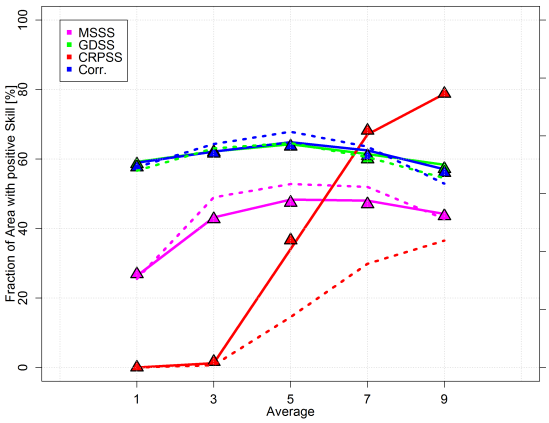
(g) Precipitation with trend av1.



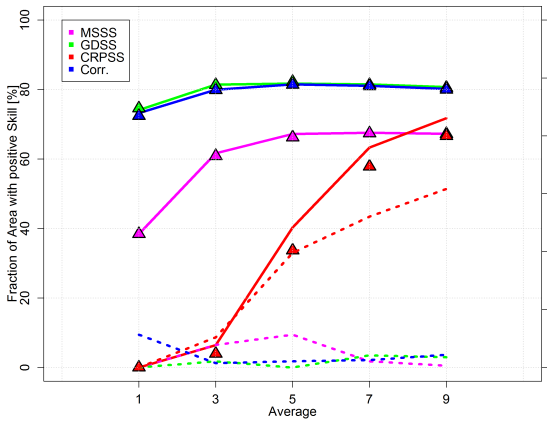
(h) Precipitation without trend av1.



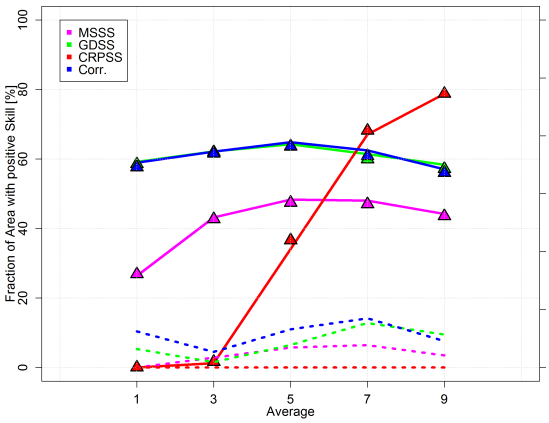
(i) Precipitation with trend av2.



(j) Precipitation without trend av2.



(k) Precipitation with trend av3.



(l) Precipitation without trend av3.

Figure 5.3.: Continued.

5.3. Correlations in space

Following his presidential address about the autocorrelation in time series to the Royal Statistical Society, G.U. Yule was asked by a colleague:

" What about space? Are there not nonsense correlations in space?" [*Johansen*, 2008]

That, of course was a valid point to raise. According to the *First Law of Geography*, "everything is related to everything else, but near things are more related than distant things" [*Tobler*, 1970]. But how can one handle spatial autocorrelation? The simple association to a one dimensional AR-process is not possible when dealing with 2D-fields or even 3 dimensions. Luckily, there are measures to estimate the spatial organization of a field, that are, however, more sophisticated than a simply temporal autocorrelation coefficient. The time component adds another degree of freedom to the analysis. Additionally, the common Pearson's Correlation Coefficient is not optimal to estimate association between fields and a *t-test* can not be performed.

Regardless, high spatial autocorrelation in meteorological data implicates the reduction of significance of skill of the unfiltered as well as the filtered data and should be investigated as well. In the following will I introduce functions for the spatial autocorrelation and describe a way to estimate the significance of the correlation of two 3D fields. Later on, it will be applied to the MiKlip data.

Spatial Organization of Fields

There are several methods to estimate the spatial autocorrelation of fields. The drawback of these methods is that neither produces a correlation coefficient as one is to aspect from the analysis of time series. Therefore the interpretation of the results differs and complicates with the dimensions added [*Bretherton et al.*, 1999].

One method that is the most similar to an autocorrelation of a time series is the Moran's Index I [*Moran*, 1950]. The index is defined as follows:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \qquad [5.7]$$

where N is the number of spatial units in 2-dimensional field of the variable X indexed by i and j, and w_{ij} is an element of a matrix of spatial weights. Simply put, the Moran's Index measures the association of a grid point to all other grid points within the field weighted by the distance. The Moran's Index is oriented like a correlation coefficient with the perfect correlation at 1, complete dispersion at -1 and zero indicating a random spatial pattern.

Another way to illustrate the spatial pattern of a field is the distance between independent grid points mentioned by *Bretherton et al.* [1999]. If the (two dimensional) autocorrelation function r is dependent of the separation distance τ, then is the integral over the r is the correlation radius L and the distance between independent grid points is:

$$2L = 2 \times \int_0^\infty r(\tau) d\tau \qquad [5.8]$$

Simply put, the distance between independent grid points $2L$ is the diameter of the circle around one grid point for which the autocorrelation function approaches zero. Both the distance between independent grid points and the Moran's index are variables that describe a two dimensional field. In the case of the MiKlip data, both coefficients were calculated for every time step and averaged over time as they vary little in time.

Significant correlation between fields

The correlation is possibly the most used analysis to find association between data sets. The calculation of correlation coefficients, e.g. the Pearson's correlation coefficient, for time series is simple and several methods exist to estimate the significance level of the result. With every dimension and subsequent degree of freedom added to the problem the issue becomes more complex. Nonetheless there are methods to estimate the correlation and its significance. In the following, the association of the fields is estimated using the anomaly correlation in all three dimensions. So, the overall number of data points is the product of all grid points and the time steps ($N = x \times y \times t$). The significance of the correlation coefficient is derived from the surrogate method that has been introduced in section 5.1. Surrogates that feature the same properties as the original time series in particular the autocorrelation are generated and the correlation performed again. I use the multivariate adjusted Fourier transform (miaaft) algorithm [*Schreiber and Schmitz*, 2000; *Venema et al.*, 2006b,a] to build a set of surrogate. The miaaft - Algorithm preserves the original distribution, the auto- and cross correlation of the time series and the fields. The resulting distribution of correlation coefficient and the position of the coefficient to be tested within yields the significance level.

5.4. Optimal spatial scales of skill in MiKlip

In recent studies about decadal predictions a spatial average of the hindcasts prior to analysis has been proposed [*Goddard et al.*, 2013]. The average area ranges from $5 \times 5°$to $10 \times 10°$. Utilizing the regional decadal hindcasts of MiKlip these assumptions can be tested extensively. The high resolution of $0.221°$of the decadal initialized runs allows for a gradual spatial smoothing and the skill of the resultant fields. In the following only the results for the baseline0 surface temperature hindcasts from 2001–2010 are shown, as it is the only complete decade with 3 models a 10 ensemble members.

Tab. 5.5a and Fig. 5.5b highlight the sizes of the filter windows used. Every grid point is assigned the value of an average of a centred squared area around itself. The analysis domain

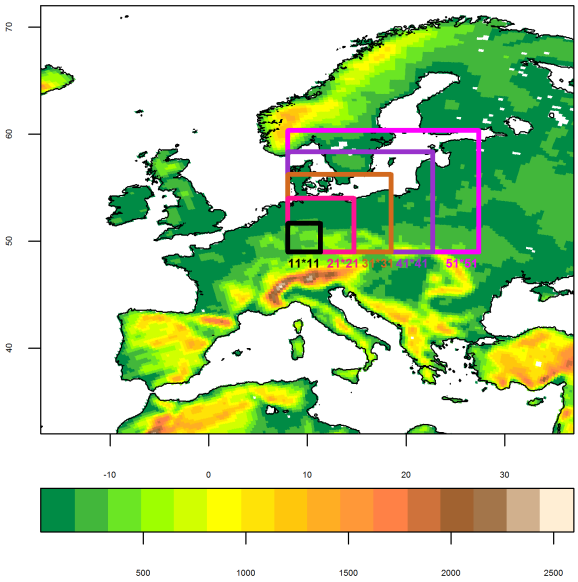
decreases with filtering. The largest filter analysed here is a grid square of 51 × 51 grid points which reduces the domain by 38 percent.

Table 5.4.: Size of the windows used for the spatial filtering in grid boxes, degrees and approximated kilometres. To smooth the fields the following was done: e.g. for the smallest filter, every grid point (without borders) is an average of all grid points 5 boxes in every direction, making the averaging area a field of 11× 11 grid points. In the filters will be labelled using the number of grid boxes around the grid point.

(a)

BOX	Grid point window	Degrees	Kilometres
5	11x11	2.4x2.4	275
10	21x21	4.6x4.6	525
15	31x31	6.8x6.8	775
20	41x41	9x9	1025
25	51x51	11.2x11.2	1275

(b)



The aforementioned statistics, correlation coefficient, significance, the distance of independent grid point and the Morans's Index, are summarized in table 5.5 for monthly means of temperature of the b0-EUR22 ensembles of MPI-ESM, CCLM, REMO and the combined regional ensemble of REMO & CCLM. The unfiltered hindcasts already exhibit an high spatial autocorrelation. On average, the Moran's Index is about 0.9 and the average distance of independent grid point

is about 70 grid points (roughly 15 degrees, 1750 kilometres). The correlation coefficient between the monthly resolved hindcasts and the E-OBS observations is very low and subsequently not significant. Filtering causes the Moran's Index to increase slightly, however, the original spatial organization is very high and the slight increase in skill with filtering also increases its significance. The distance of independent grid point also lengthen slightly with smoothing (e.g. CCLM: from 70 grid points to 75 grid points). These inconsequential effects of spatial filtering on the spatial organization of the data is attributable to the high initial autocorrelation. The filtering windows used here are smaller than the distance between independent grid points. Any influence on the distance by averaging within the distance is going to be small and therefore will the Moran's Index only be effected slightly and the significance level is mostly controlled by the increase in skill.

Fi. 5.4 illustrates correlation coefficient and its significance for the monthly data again and puts it into perspective to the annual means and the temporal filtered annual means of the b0-EUR22 temperature hindcasts form 2001–2010. The skill increases with temporal smoothing as discussed in section 5.2. Spatial autocorrelation changes only marginally when filters are applied. Also, the sample size of the 3-D data is high, contrarily to the time series. The consequence is a neglect-able negative influence on the significance. Hence, the increase of significance due to a better correlation coefficient is more prominent. Generally the significance seems to stagnate from the averaging window 15 (31×31 grid points, 7 °, 775 km) and even drop with the window 25 in some cases.

Though it has been suggested to spatially average, when dealing with (global) decadal predictions, I find, that the influence on skill is relatively small and is largest when the Signal-To-Noise-Ratio is already shifted in favour of signal due to temporal filtering. There is an increase in significance but only in the case of the largest smoothing area for the regional ensemble of CCLM does it reach 90 %.

Nevertheless, in this one example of one decade of temperature the best window sizes are between $6.8 \times 6.8°$or $9 \times 9°$. Which falls within the proposed range of 5 to 10°. On the other hand, any gain in skill has to be weighted against the loss of regional diversity. In the field of regional climate predictions the spatial smoothing is not necessarily desired even though a spatially smoothed regional model can outperform a global model (e.g. CCLM versus MPI-ESM-LR in Fig. 5.4). Ultimately the decision about spatial filters is dependent on the original data and its characteristics as well as the final uses of the prediction.

Conclusion

In general filtering in any dimension reduces noise and increases the signal, leading to an increase in skill. However two points have to be considered: First, at one point the filter becomes to large and the signal, too, is removed from the data, leading in a reduction of skill.

Table 5.5.: Spatial Statistics for the b0-EUR11 temperature hindcasts of the 2001–2010 decade. There are three ensembles of 10 ensemble members of CCLM, REMO and MPI and one two model ensemble of 20 members consisting of CCLM and REMO (2RCM) in monthly resolution. The statistics described in section 5.3 are correlation coefficient (R), significance (SIG), the distance of independent grid point (2L in grid points) and the Morans's Index (I).

(a) Unfiltered Hindcasts

Model	$2L$	Moran's I	R	SIG
CLM	70.361	0.914	0.042	0.77
2RCM	71.446	0.906	0.034	0.79
REMO	64.349	0.898	0.022	0.69
MPI	72.328	0.899	0.02	0.58

(b) Filtered Box 5 (11x11

Model	$2L$	Moran's I	R	SIG
CLM	70.349	0.909	0.047	0.78
2RCM	71.778	0.901	0.037	0.79
REMO	67.769	0.897	0.024	0.66
MPI	70.92	0.894	0.023	0.61

(c) Filtered Box 10 (21x21)

Model	$2L$	Moran's I	R	SIG
CLM	72.922	0.909	0.052	0.82
2RCM	74.304	0.902	0.041	0.79
REMO	71.347	0.898	0.026	0.63
MPI	73.592	0.893	0.028	0.61

(d) Filtered Box 15 (31x31)

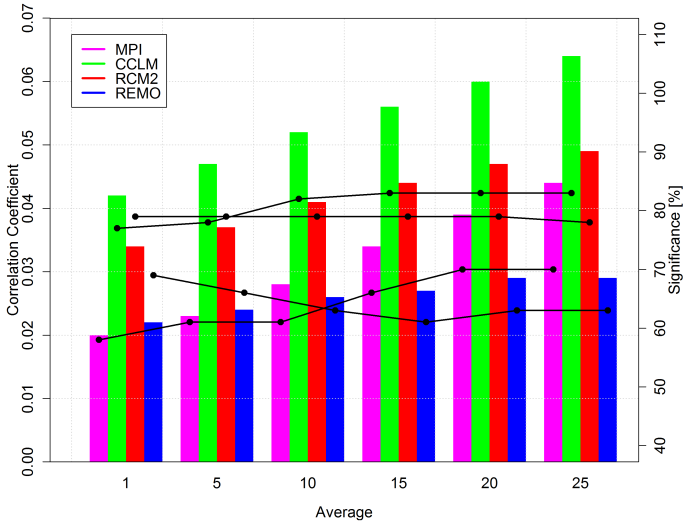
Model	$2L$	Moran's I	R	SIG
CLM	75.66	0.911	0.056	0.83
2RCM	76.438	0.905	0.044	0.79
REMO	74.145	0.902	0.027	0.61
MPI	77.19	0.895	0.034	0.66

(e) Filtered Box 20 (41x41)

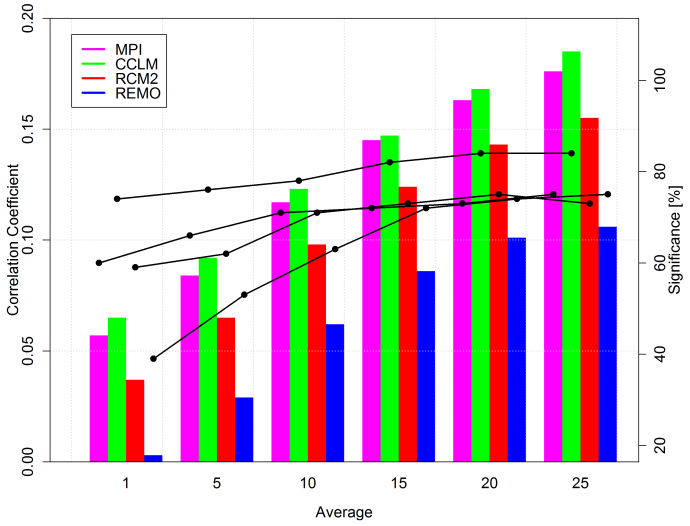
Model	$2L$	Moran's I	R	SIG
CLM	77.04	0.915	0.06	0.83
2RCM	77.607	0.91	0.047	0.79
REMO	75.589	0.907	0.029	0.63
MPI	79.67	0.899	0.039	0.7

(f) Filtered Box 25 (51x51)

Model	$2L$	Moran's I	R	SIG
CLM	75.566	0.92	0.064	0.83
2RCM	74.471	0.915	0.049	0.78
REMO	72.626	0.914	0.029	0.63
MPI	76.283	0.905	0.044	0.7

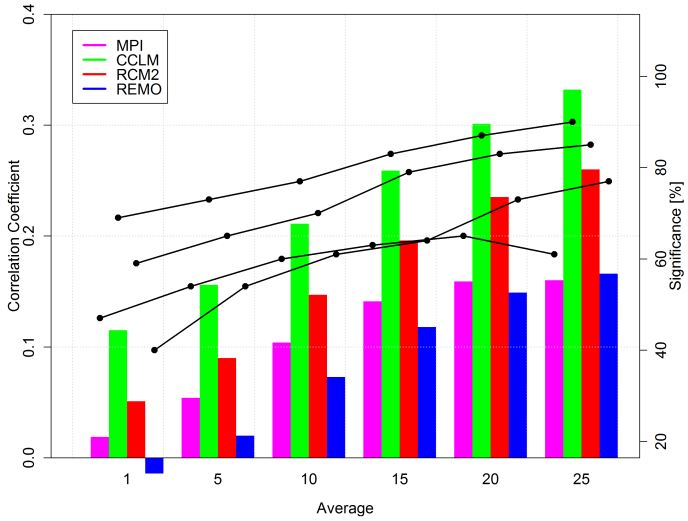


(a) Monthly Means



(b) Annual Means

Figure 5.4.: Correlation coefficient of the b0-EUR22 temperature hindcasts with E-OBS (bars) and its significance respectively (lines) all four ensembles: MPI-ESM (magenta),CCLM (green), REMO (blue) and the combined regional ensemble of CCLM and REMO (2RCM, red)



(c) Annuals Means filtered by a 5 year running mean

Figure 5.4.: Continued.

Second, smoothing increases the internal organization of the data and decreases the significance level of skill. The increase in skill is to be weighted against the drawbacks of filtering, finding an optimum. This optimum is highly dependent on the initial data and its characteristics and the filter.

In the cases of the MiKlip ensembles a temporal running mean (or lead time dependent average) of 5 to 7 years generates the best skill and significance results. The influence on skill and significance by spatial filtering is smaller than in the time domain, since meteorological data is already highly spatially organized in the first place. Averages as suggested in the literature increase the skill only slightly and do not lower the level of significance. The choice of filter should be considered for each case independently on the behaviour of the data and its use. But a case for regional climate predictions can made, in so far, that there is no real disadvantage for high resolution climate predictions towards global, spatially smoothed climate predictions.

With that I close the analysis of the overall skill of the MiKlip ensembles. With the knowledge of the predictability of selected variables, the success with which the MiKlip ensembles are able to predict them as well as the optimal scales for the analysis of such a prediction, the next logical step is the forecast of relevant climate events which will be discussed in the next chapter.

6. Predicting decadal events: An Application

As the summer of 2003 progressed with one meteorological station after another logging breaking temperature records all over Europe, media outlets everywhere labelled it the "summer of the century" [*Gunkel*, 2013]. But an exceptionally hot summer like 2003 is more than a revenue source for ice cream shops, it had devastating consequences for human lives and the economy of Europe. It has been estimated that about 70.000 deaths can be attributed to the high temperatures [*Robine et al.*, 2007] and an excess burden of 10 billion Euros. One of the main problems were that high temperatures had not been considered a major hazard while contingency plans had been made for a variety of natural and man-made catastrophes.

But only three years later, after moderate warm summers, a new heat wave effected Europe in 2006 and again another record breaking summer for the northern hemisphere in 2010 [1].

Studies following the heat wave of 2003 found that the risk of such extreme events happening is increased by the anthropogenic climate change [*Stott et al.*, 2004]. Are such events only a result of an increasing global temperature, or is it the superposition of man-made climate change and internal variability, that has some summers exceed previously observed temperatures to such extent it becomes a hazard? As other major events in the past have been linked not only to climate change but also to the internal variabilities, e.g. the connection of the "Dust Bowl" to the AMO or how droughts in California have been linked to negative phases of the Pacific Decadal Oscillation [*Mantua and Hare*, 2002]. Assuming climate is the result of a combination of factors, decadal predictions could take a major role in preparing for and to mitigate the consequences of extreme events. To successfully predict events like the series of heat waves in the 2000s, a proper definition of an event is needed. Is a definition be found and events in the past identified, the current prediction system can be analysed towards its abilities to predict events.

6.1. A Definition

How can a significant event be defined? In the case of a heat wave that definition is easy: Above average temperature for a number of cohering days. But a heat wave can only lead to a collapse of a health system (as did the summer 2003) if its wide spread. How big has the area affected to be? And is the event the singular head wave of the succession of several abnormal summers within a few years?

[1] Mean Temperature of Germany retrieved from DWD station observations: Archives curated by B. Hussing; Available at `http://www.bernd-hussing.de/Archivdateien/Archiv.htm`

These questions make it apparent, that a definition of an "event" is not a black and white issue. A definition could also be different when another variable is the basis, e.g. precipitation. So, how to define an extreme event as such that it differentiate it from annual extremes as defined by the ETCCDI indices [*Karl et al.*, 1999; *Peterson et al.*, 2001; *Zhang et al.*, 2005; *CLIMDEX*, 2013]?

Climate is largely non-linear, has a large number of degrees of freedom in space and parameter domain, to simplify *Anderson and Ed.* [1996] described a climate time series as a combination of different natural factors:

- Periodic and quasi-periodic phenomena, including the most distinctive, the response to annual cycle in sun-earth geometry, but also decadal-to-century scale climatic signals,
- Aperiodic random variability,
- Low frequency variabilities, including long-term climate trends, and
- distinctive temporal signatures and discrete jumps, e.g. an abrupt "regime shift".

Now, the latter warrants a closer look. Regime shifts could manifest themselves in many way and depending on the definition of a regime. Is a regime characterised as a distinctive range of a value or index the probability density function should be at least bi-modal. A shift would then be the change in modes. The problem is the testing of the statistical significance of such an event, especially in terms of decadal and inter-decadal events, when the record length itself only covers one regime. But there is also the chance of superposition: When a shift in climate coincides with a positive phase of any periodic phenomena and is further upset by a long term trend. That leads to an "unprecedented events" of record lows or highs, a long uninterrupted sequence of the same polarity, a high frequency of record low or high values within an interval or the longest interval between such events [*Anderson and Ed.*, 1996]. The statistical significance of these is also controversial, as even random time series can exhibit record highs and lows and extrema are likely to be clustered together.

Even though a wide range of natural states can occur even without factoring in climate change, the anthropogenic influence on the global climate is likely to change the frequency, intensity, spatial extent, duration and timing of these "unprecedented events" [*IPCC*, 2012]. In fact, such changes have already been observed [*IPCC*, 2012, 2007; *Trenberth et al.*, 2007; *IPCC*, 2013]. E.g. changes in precipitation have likely lead to increases in the number of heavy precipitation events during the second half of the 20th century – even in areas where the total amount of precipitation decreased [*Trenberth et al.*, 2007].

In general, two different kinds of these events can be differentiated: Purely statistical events and events with impacts on societies, economies and ecosystems [*National Research Council*;

Sivakumar et al., 2005; *IPCC*, 2012]. The aforementioned heat waves were both, outliers within the climatological time series and impacting i.a. water availability, the overall health of the population and agricultural yields among others. In the context of useful predictions, events with large impacts are of high interest but have still to be described statistically for the sake of forecast verification.

Sivakumar et al. [2005] categorized the impacts of natural (meteorological) disasters into direct and indirect effects that can be both positive and negative. E.g. the impact of high precipitation on agriculture is varied: The direct effects are negative, with a loss of current crop and damage to facilities. Indirect effects can be both negative such as the low incomes after the fact and subsequently decrease in productivity, and positive, like the fertilization of flood plain soil and the influx of money through relief funds. Overall mostly direct impacts are associated with physical damage and can be short term and temporary. Indirect effects can be long term changes and are hard to quantize.

Statistically speaking, an event occurs, when the climate system is forced to cross a threshold, triggering a transition into a new state at a rate that is faster than its cause [*National Research Council*]. On societal/ecological time scales that means a change abrupt and/or large enough that the climate system has trouble adapting. From the human perspective such events are significant, when

- they persist over years or longer,
- are larger than the typical climate variability and
- affect sub-continental or larger areas [*National Research Council*].

An example into this is given by *Villalba and Veblen* [1998], identifying climatic influences onto tree establishment and mortality in Northern Patagonia. On species of trees was analysed towards its spatial occurrences. Single season events (e.g. extremely dry-warm summers) control local mortality events of the trees, while the regional distribution was linked to decadal climate variations. Of course ecosystem differ, and distinction between the short versus long-lasting climatic variations is crucial to predict forest responses. E.g. the in the mid latitudes prevailing temperate deciduous forests (including Europe) have no limits when it comes to the tolerance of extreme colds, but require winters below 5 °C, a certain moisture and accumulated heat during the growing season [*Prentice et al.*, 1992].

To conclude I introduce 3 possible statistically describable "climate events":

1. The occurrence of a local extrema of multi-year means that covers a certain percentage of the investigation area.
2. A shift in sign of multi-year trends that occurs over a percentage of the investigation area.

3. A multi-year episode of anomalies consistently exceeding the natural variability.

These definitions have been formulated intentionally vague. The numbers of years to calculate the multi-year means and the trends as well as the covered area can yield very different results depending on the application and will have to be adapted to the variable and the area investigated. Even with the limited test over Europe, the occurrences of events were different dependent on variable, years and area selected. Europe is a very heterogeneous area climatological wise partly due to its orography. If the conditions demanded a very large part of Europe to be affected, most events had to be dropped from the statistic. It can be assumed, other regions that are more homogeneous behave differently.

6.2. Hindcasts of selected decadal events

Predictions in hind sight, hindcasts, have the advantage of easy verification. But for any evaluation a statistic has to already exist. There is no reliable statistic for "decadal events" as defined here. But using the E-OBS observations and a range of possible thresholds for the size of the area affected as well as the number of years exhibiting the anomalous behaviour, a refined definition and a statistic can be build. Following the examples given, of the impact of extremely hot summer, the temperature especially in summer will further analysed here.

In general an area threshold of more than 50 % yielded almost no events. 50 % of grid points of the EUR44 ensembles does amount to roughly 4 million square kilometres. Also multi-year values covering at less than 5 years produce rarely events. For the following example the values chosen therefore area thresholds of less than 50 % and 5 year mean. Then, the local maxima of a 5-year running mean of the daily minimum temperature in summer (JJA) that were experienced at at least half the area of Europe are:

1963 1965 1968 1979 1981 1993 2008.

Here one can already see the problem with the statistical expression of "decadal events". Humans perceive heat waves as singular events, but for the largest impact, the anomaly has to persist over several seasons. From this follows that the years, one subjectively remembers might not appear in the statistic. In the case of anomalous hot night temperatures, the years 2003,2006 and 2010 are curiously missing. However, as this is a 5-year running mean, the last so-called event from 2008 does include 2006 and 2010, as the number given by the statistic is the center of the anomalous period. Together do the summers of 2006 and 2010 outmatch the summer of 2003.

In comparison, shifts running 5-year trends of the daily minimum temperature in summer (JJA) whose sign persisted for at least 3 years occurred in

1964 1967 1971 1972 1978 1983 1984 1988 1989 1992 1993 1994 1995
1996 1998 1999,

But only if only 10 % of Europe is enough for it to be called an event. The last shift in trend sign was at the end of the 20th century. While afterwards local maxima still occur and therefore also changes in the sign of the trend, it does so only for a short while before turning again. That could be a sign for anomalies in summer minimum temperature occurring more frequently, the negative trend afterwards only of of short duration.

As the area threshold is a huge factor in determining whether an event takes place or not when only analysing observations, it can be assumed, that spatially highly diverse regional model simulations would yield even fewer events. Because of that, the following analysis will only focus in the PRUDENCE region Middle Europe. The local maxima of the running 5-summer mean temperature that cover at least 50% of the region are then:

`1963 1965 1969 1971 1975 1977 1984 1993 1996 2004 2008.`

To see, whether the CCLM b1-EUR44 ensemble is able to reproduce maxima during these year, all runs of the ensemble whose 2-6 lead year averages include these years were selected. E.g. The event in 1971 is covered by runs starting in 1967 to runs starting in 1975, which amounts to a time series of 5 5-year means. These short time series together from a timeseries of 41 data points and were then analysed the same way like the original observations and all local maxima covering at least 50 % of the PRUDENCE region Middle Europe found.

The verification of non-probabilistic yes/no forecasts can be visualized using a 2x2 Contingency Table [*Wilks*, 2006]. Conventionally, the verification data is displayed in an 2x2 table of absolute frequencies, or counts, of the 2x2 possible combinations of forecast and event pairs. In Tab 6.1 the counts of events from runs covering the years of local maxima of the running 5-summer mean temperature that cover at least 50% of the PRUDENCE region Middle Europe are displayed. The flaw in this analysis is the exclusion of all other runs not covering the aforementioned years. Therefore the count of the false alarms (upper right) could be higher and an accurate account of the rightful predicted non-events (lower right) can not be given.

Table 6.1.: The 2x2 Contingency Table for the prediction of the anomalous summers counts the events from runs covering the years of local maxima of the running 5-summer mean temperature that cover at least 50% of the PRUDENCE region Middle Europe and the corresponding observed maxima.

	Observation yes	Observation no
Forecast yes	9	5
Forecast no	2	25

The most direct measure of accuracy of a non- probabilistic forecasts of discrete events is the Proportion Correct (PC) [*Wilks*, 2006]. In this case the proportion of correctly predicted events and non-events is

$PC = 83\%$.

Of course, when forecasting relatively rare events the cases of correctly predicted non-events are going to be plenty and will skew the skill. The Hit Rate (H) is the number of correctly predicted events in relation of all events. Here, 9 out of 11 events were correctly predicted:

$H = 81\%$

The hit rate can also be regarded as the fraction of those occasions when the forecast event occurred on which it was also forecast, which is equivalent to what has been previously called discrimination. As both the Hit Rate and the Proportion Correct are very high, a look into the false alarms is warranted. The False Alarm Rate (F) and its counterpart, the Failed Alarm Rate (F^{-1}) are the relation of the false alarms to the actual non-events and the relation of the not predicted events to the number of events that occurred respectively.

$F = 20\%$

$F^{-1} = 18\%$

Another score that can be used when the event rarely occurs, is the Critical Success Index (CSI) [*Wilks*, 2006]. The index is the number of correctly predicted events in relation to all predicted events. Then, all successfully predicted non-events are not of consequence.

$CSI = 64\%$

The prediction of extremely hot summers, either in a series or extreme singular events, in Middle Europe with the b1-EUR44 CCLM ensemble can be called a success, albeit with reservations. 9 out of 11 events were correctly predicted, but the variability of the regional ensemble was higher than the observations, as the predictions included 5 additional local maxima. Overall this case study is small with only 11 events and no definitive conclusion about the abilities of the CCLM ensemble can be drawn. Further research into the matter, regarding the definition itself, but also the thresholds regarding the areas as well as the length of the multi-year episodes has to be done. Then further verification of the CCLM ensemble for other variables and extremes for more regions can be carried out.

Conclusion

Extremes that are more than short-lived weather extremes but are happening on a much larger – decadal – scale, are not yet decisively defined. While events with high impact will most likely be of high interest, to properly analyse a prediction system's abilities to reproduce such events, there needs to be a mathematical way to describe them. A very simply way to calculated decadal events is to find local extrema within timeseries filtered multi-year running means. By doing that one does remove high-frequency noise and short-lived extremes but does also take into

account that anomalies have persist be extremely high or have to persist over several seasons to have a great impact. On top of that, an event of decadal scale can not be localized has to be experienced over al larger area. The length of the filter and the size of area will vary depending on the investigation region, the variable, the extreme and the season.

Slightly more complicated definitions could describe large and abrupt changes in the trend or episodes of consistently exceeding the climatological variability. The definition of an event has to be dependent on the desired impact (human/ecological/economical). The tests for events in the E-OBS data using these theoretical definitions do not always yield events though. The choice of definition could therefore also be limited by the simply non-occurance of the desired events.

This application of decadal predictions can only be seen as an example. The definition and the actual prediction and its verification has to be the subject of further studies. The preliminary results for the predictions of extreme summers combined with the high added value to the prediction of extreme indices by the regional prediction system (as discussed in Sec. 4.6) is however promising.

7. Conclusion

With this thesis I set out to explore the potential for regionalised decadal climate predictions especially for Europe and have identified atmospheric variables that exhibit some internal decadal variability and thus a condition for predictability. The study has also sought out to explore statistical methods to evaluate and improve regional decadal predictions. In the literature the problem of decadal climate predictions is generally approached in a similar fashion: Large ensembles of coupled AOGCMs, that are initialized annually and simulate for 10 years, are compared to uninitialized simulations for their performance of multi-year means of (mostly) mean variables, e.g. temperature. In terms of verification, the framework given by *Goddard et al.* [2013] is usually followed, that gives certain guidelines to performance measures and averages. It is assumed that the high skill found over the oceans stemming most likely from low-frequency variations in the oceans is sufficiently transported towards the continents, therefore making climate predictions on a regional level possible.

However, the literature is inconclusive on some vital questions within this discourse, in particular to regional climate predictions:

- In a best case scenario, what can potentially be predicted over land?
- And are the common practises of handling decadal climate simulations (multi-year and spatial averages and a limited number of skill measures) statistically sound?

This study sought to answer these questions while employing 4 different ensembles from the project MiKlip, two of which were from the AOGCM MPI-ESM and two containing runs of the regional climate models CCLM and REMO. The regionalisation of the decadal climate predictions – so far the only attempt to do so – adds another layer of problems to an already complex issue (e.g. how to quantify the added value of increased resolution), but also allows for a detailed analysis of the regional variability of the climate on decadal time scales.

Empirical findings in regards to the current MiKlip regional decadal ensemble are a first step to tackle the problems. The analyses of the four MiKlip ensembles, two global and two regional ensembles, revealed a varied picture in regards to spatial and temporal distribution of skill and added value. In general it can be shown, that all skill measures are highly dependent on each other, their relationship indigenous to the model. The spatial distribution of skill does

vary little between skill scores while their net value do. The added value behaves in a similar fashion, with values changing but not the spatial patterns.

The 4 ensembles were analysed in two different instances: First, the decadally initialized EUR-22 ensembles with 10 members for both generations (b0 & b1) in a resolution of 0.22 °. Here the long term trend over the 50 years of MiKlip's investigation period was removed. Despite the simple experiment design some universal statements have emerged. The most skill can be found in the beginning of the decade (e.g. lead years 1-5 vs. 1-10). As the long-term trend and therefore most of the influence of external factors is removed, the result confirms that the influence of initialization does decrease over the course of a decade.

The spatial pattern, however, do not change with the inclusion of the second pentade. There are differences in skill between the seasons and annual mean variables both in value and spatial distribution. The annual variables are not necessarily a simple superposition of the skill of the individual seasons. One reason could be that the dividing of the year into strictly three-month-interval is too rigid and the dynamics of e.g. the winter climate can extend beyond February. Using annual means and indices that are calculated with data from the whole year could be a better choice.

Temperature predictions show a regional bias when it comes to their seasonal performance. So are the summer hindcasts more skilful in the south than in winter. The PRUDENCE regions Iberian Peninsula (IP) and Mediterranean (MD) show skill in many cases for both temperature and precipitation as does the region in Eastern Europe around the Black Sea. But mostly there is a difference in skill between temperature and precipitation. Precipitation prediction is less skill full in general than temperature predictions and the spatial pattern less homogeneous.

Another issue, that emerges, is the dependency of skill on the initialization date/year. When the decades of the investigation period of MiKlip were investigated separately, dissimilar pattern in skill emerged. This is most likely a result of a combination of factors, e.g. the varying quality of observations and that the predictability fluctuates with the phase of low-frequency internal variabilities the prediction was started on.

The second generation of regional decadal ensembles included 5 members that were annually initialized albeit at a lower resolution (b1-EUR44). The annually started simulations are then averaged over certain years of a decadal run (lead year averages) and the long term trend was not removed. Here skill of both the temperature and the precipitation was mostly higher than in the decadally initialized ensemble. There again is little difference found in spatial distribution among the skill measures. The skill is however slightly dependent in the lead year averages. With the most skill full hindcasts usually found for long averages including the end of the simulated decade. Towards the end of the decade the internal variability, i.e. the initialization, ceases to be the dominant climate driver and the influence of external factors grows. External forces, as green house gas emissions, are a huge source for skill for the prediction of climate for a decade.

A drop in skill can be experienced when either considering too short of an averaging period or averaging over the middle of the decade, where none of the factors are dominant.

In general the regional decadal ensembles of MiKlip show at least a conservation of global skill. The net added value, too, is dependent on ensemble generation, region, variable, season and time scale. The added value is also dependent on the global skill. So can be the increase in skill of temperature small as the original skill of the global model is already high and therefore potential for improvement small.

Both ensemble generations were also investigated for their behaviour with different lengths of filters, regardless of the lead year. Filtering reduces noise and increases the predictability. In the cases of the MiKlip ensembles a temporal running mean of 5 to 7 years generates the best skill and significance results, while the influence on skill and significance by spatial filtering is smaller than in the time domain.

Theoretical implications can be generalised from the investigation of the MiKlip ensembles. The success of decadal predictions stems from the ability of the models to reproduce realistic variations within the slow-acting components of the climate system. I demonstrated an easy to implement test to check the potential stemming from these variabilities and found that the skill from the direct vicinity of these processes is not a guarantee for skilful regional predictions over land. The internal climate oscillation likely to be a major driver of European climate, the AMO, shows potential to be predictable up to 12 years. The translation of that information into temperature over Europe decreases the predictability, in the case of temperature in middle Europe to about 5 years. In terms of the potential of climate extreme indices and percentiles for 10-year forecasts, I found moderate extremes (Summer Days instead of Tropical Nights) most promising, as well as indices that included a memory and/or were a combination of several root variables. The results of this investigation were carried out using only observations and are therefore not dependent on the performance of the MiKlip ensemble.

When it comes to the actual forecast of the variables, the literature stresses the fact, that there are many aspects of forecast quality. The formulation of measurement scores is often quite similar, though the scores do represent different aspects of quality. In the case of MiKlip, I found strong correlations between the skill scores suggested by previous literature on decadal predictions as well as the additional discrimination skill score I introduced. As it was the case for both the regional and the global model, there is no reason that suggests it would be different for other models. In conclusion, I would encourage the usage of a small number of scores that are suited to the situation. A verification that focuses on too many quality aspects could even be a hindrance in communicating the results. E.g. in some situations a correlation coefficient is sufficient if the desired outcome was to simulate the development of a variable in time, in

some other cases it might be of higher interest if the ensemble does adequately represent the variability of the real climate, when a reliability score should be preferred.

The skill can be increased by filtering; it reduces noise and increases the signal. However filters should be applied to the extend that the signal is not too much removed and the autocorrelation becomes not too high for any statistic to be still significant. The increase in skill is to be weighted against the drawbacks of filtering, finding an optimum. This optimum is highly dependent on the initial data and its characteristics and the filter. In the case of temporal filtering, I suggest a slightly longer average window than has been previously in use in the field of decadal climate predictions. That means however, that the averages goes beyond what is likely to be the influence of initialization and will be influenced by external forcing, too. The investigation into the first pentade of a decade is important in exploring the mechanisms and process of internal decadal variability. But still, the external forcing is a big source for predictive skill even on time scales of 10 years and should be involved in the discussion about the net skill.

Meteorological data is highly spatially organized even without any filtering. Averages as suggested in the literature do increase the skill only slightly and do not lower the level of significance. The choice of filter should be considered for each case independently on the behaviour of the data and its use. I do however argue for the feasibility of regional decadal predictions because the insignificant differences between the skill of the regional resolution and the filtered predictions that are of global model resolution.

Recommendations for a regional decadal prediction can be made by summarising and generalising the analyses of the MiKlip ensembles: It is detrimental to ignore the indicators in support of annually initialization. The performance of both the global and the regional model improves with the increased number of starting dates. However, in light of the increased work load that is downscaling, less starting dates can be considered with reservation, when an assessment is made to which extend the skill is dependent on the starting date.

In general a multi-model ensemble will if not increase the net skill, increase the significance of any statistical evaluation. In several instances global decadal prediction ensembles are already consisting of multiple models [*IPCC*, 2013]. The benefits of an amplified regional ensemble (within computational capabilities of course) should be considered, too.

If the desired overall statement of an decadal regional experiment is about an applicable prediction for users, the trend has to be part of the analysis. Also, averages that go beyond the pentades can carry much information and predictive skill. Though, that will be within the discretionary of the user. There will be variables that are more predictable than others: For a 10-year forecast that are moderate extremes preferably with memory. Again, within the context of the wished outcome, other variables might require adjustment of the forecast length.

Future research is still necessary. The field of decadal predictions is relatively young and regional decadal prediction are almost completely unexplored. The downscaling within MiKlip was limited to 2 investigation regions: Europe and Africa. There may be other regions where the climate exhibits strong internal multi-year variability, thus creating high potential predictability. To understand and find these regions, more effort has to be invested into the driving mechanisms in slow-acting components of the climate system as well as atmospheric teleconnections.

In the case of actual forecasts, other issues are mostly unfathomed: The differences between the skill of seasonal and annual means are unsatisfactorily understood. There might be a better way of breaking up a year, forecasts dependent on circulation patterns for example. Additionally, better descriptions on what constitutes climate extremes/climate events, and how they can be mathematically described, is definitely needed. They could be purely defined through statistics, as local extremes, or as an index of their impact on economies, societies or ecosystems, while also addressing how large an area has to be impacted. Such definitions will most likely vary with region investigated. E.g. the ETCCDI recommended indices [*CLIMDEX*, 2013] have turned out to be to narrowly defined, ergo not having much validity in regions outside of mid latitudes. In addition, most models have inherent biases and an exploration of extremes will need further research into the best way of comparing extremes of observations and simulations.

Only tangentially related to the issue of decadal climate predictions is the problem that one faces with observations. Gridded, high resolution, long time period spanning and homogeneous observational data sets are rare and have errors, due to faulty station observations or simply lack thereof. The choice of observations can greatly impact the skill. Ensembles of observations could be a way of reducing some error.

The limitations of this study stem i.a. from the use of only one global model and (mostly) one regional model. The representation of teleconnections, such as the AMO, can differs greatly between models [*IPCC*, 2013]. Also, with the exceptional task of downscaling a decadal prediction system, the investigation period is rather short, especially when filters and averages and only few starting dates are deployed. But statistical significance is dependent on long time series.

Also not discussed in this study are the widely used unitialized runs as the reference forecasts for skill scores. The argumentation is the following: Any skill stemming from external drivers such as greenhouse gas emissions are covered by climate projections. Consequently the value that decadal predictions add is purely on the merits of their initialization on the correct phases of internal variabilities. Then, decadal predictions could only be considered as "good" if they outperform unitialized climate simulations. It is undeniable that decadal predictions should add value to climate projections, however, I argue that climate is more than the superimposition of internal variability and external drivers. Both interact in many ways, that might not be completely understood yet. It is not shown in this work, but in most cases, the skill improves

when the hindcasts are compared to unitialized simulations. To use the climatology as the reference forecast is a stricter test for quality, that is aware of possible non-linear interactions between internal and external factors.

To summarize, my thesis intended on closing gaps in the knowledge about the feasibility and the skill of decadal predictions, especially the downscaling thereof over Europe. I have demonstrated easy tools to first detect potential predictability and then the verify predictions. Within the project the first attempt of downscaling decadal climate predictions was undertook, making the value added by regionalisation a major factor in my study, including the statistical description and the evaluation of a regional decadal climate prediction system.

I have shown that there is indeed potential for decadal predictions over Europe, though it is diminished in comparison to atmospheric variables in direct contact to low-frequency variabilities. I could illustrate the abilities of the MiKlip decadal prediction system both global and regional, that is able to produce skilful prediction of the climate up to 10 years ahead to a varying degree dependent on many factors such as region and season. Taking from the experiences of the MiKlip system a few recommendations for future regional decadal predictions can be made.

This study does also recognise the lack of knowledge about so-called ”climate events”, that are not only predictable but also worth the effort. I made a first excursion into the matter, but much more research is required, including perhaps longer investigation periods to cover more cases and different investigation regions. Though, the general skill demonstrated by the MiKlip prediction system, the potential predictability for many variables, as well as continuing advances in climate modelling hold out the prospect of future prediction of events with great impact on decadal time scales.

8. Bibliography

(), *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change [Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson (eds.)]*.

Abramov, R., A. Majda, and R. Kleeman (2005), Information Theory and Predictability for Low-Frequency Variability, *J. Atmos. Sci.*, *62*, 65–87.

Anderson, D., F. Doblas-Reyes, M. A. Balmaseda, and A. Weisheimer (2009), Decadal variability: processes, predictability and prediction, 47 pp.

Anderson, D. L. T., and J. W. Ed. (1996), Springer Berlin Heidelberg New York.

Balmaseda, M. A., K. Mogensen, and A. T. Weaver (2013), Evaluation of the ECMWF ocean reanalysis system ORAS4, *Quarterly Journal of the Royal Meteorological Society*, *139*(674), 1132–1161, doi:10.1002/qj.2063.

Bellucci, A., R. Haarsma, S. Gualdi, P. J. Athanasiadis, M. Caian, C. Cassou, E. Fernandez, A. Germe, J. Jungclaus, J. Kröger, D. Matei, W. Müller, H. Pohlmann, D. Salas y Melia, E. Sanchez, D. Smith, L. Terray, K. Wyser, and S. Yang (2014), An assessment of a multimodel ensemble of decadal climate predictions, *Clim. Dyn.*, *44*(9), 2787–2806, doi:10.1007/s00382-014-2164-y.

Bleymüller, J., G. Gehlert, and H. Gülicher (2004), *Statistik für Wirtschaftswissenschaftler*, 14th Edition, Verlag Franz Vahlen, München.

BMBF (2007), Die Hightech-Strategie zum Klimaschutz, Bundesministerium für Bildung und Forschung (BMBF), Referat öffentlichkeitsarbeit, 11055 Berlin, accessed 03 December 2015, `http://www.pt-dlr-klimaundumwelt.de/_media/hightech_strategie_fuer_klimaschutz.pdf`.

Boer, G. J., and S. J. Lambert (2008), Multi-model decadal potential predictability of precipitation and temperature, *Geophysical Research Letters*, *35*(5), n/a–n/a, doi:10.1029/2008GL033234, l05706.

Branstator, G., and H. Teng (2010), Two Limits of Initial Value Decadal Predictability in CGCM, *J. Clim.*, *23*, 6292–6311.

Bretherton, C. S., M. Widemann, V. P. Dymnikov, J. M. Wallace, and I. Bladé (1999), The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field, *J. Clim.*, *12*, 1990–2009.

Candille, G., and O. Talagrand (2005), Evaluation of probabilistic prediction systems for a scalar variable, *Quart. J. Roy. Meteorol. Soc.*, *131*, 2131–3150.

Caron, L., C. G. Jones, and F. Dobals-Reyes (2014), Multi-year prediction skill of Atlantic hurricane activity in CMIP5 decadal hindcasts, *Clim. Dyn.*, *42*, 2675–2690.

Casati, B., L. J. Wison, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason (2008), Forecast verification: Current status and future directions, *Meteorol. Appl.*, *15*, 3–18.

Chen, D., M. A. Cane, A. Kaplan, S. E. Zebiak, and D. Huang (2004), Predictability of El Niño over the past 148 years, *Nature*, *428*, 733–736.

Christensen, J. H., and O. B. Christensen (2007), A summary of the PRUDENCE model projections of changes in European climate by the end of this century, *Clim. Change*, *81*(1), 7–30.

CLIMDEX (2013), Climate Extremes Indices, Website, accessed 12 November 2015, `http://www.climdex.org/indices.html`.

Cover, T. M., and J. A. Thomas (2006), *Elements of Information Theory*, 576 pp.

Daniel, M., M. Malý, V. Danielová, BKříž, and P. Nuttall (2015), Abiotic predictors and annual seasonal dynamics of Ixodes ricinus, the major disease vector of Central Europe, *Parasites & Vectors*, *8*(1), 1–12.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart (2011), The ERA-Interim reanalysis: configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *656*, 553–597.

Doblas-Reyes, F. J., I. Andreu-Burillo, Y. Chikamoto, J. Garcia-Serrano, V. Guemas, M. Kimoto, T. Mochizuki, L. R. L. Rodrigues, and G. J. van Oldenborgh (2013), Initialized near-term regional climate change prediction, *Nature Comm.*, *4*.

Doms, G., and U. Schättler (2002), A Description of the non-hydrostatic regional model LM, Part I: Dynamics and Numerics, *Tech.rep.*, Deutscher Wetterdienst, P.O. Box 100465, 63004 Offenbach, Germany, lM_F90 2.18.

Eade, R., E. Hamilton, D. M. Smith, R. J. Graham, and A. A. Scaife (2012), Forecasting the number of extreme daily events out to a decade adead, *J. Geophys. Res.*, *117*, D21,110.

Ebert, E. E., and J. L. McBride (2000), Verification of precipitation in weather systems: determination of systematic errors, *J. Hydrolog.*, *239*, 179–202.

Enfield, D. B., A. M. Mestas-Nunes, and P. F. Trimble (2001), The Atlantic Multidecadal Oscillation and its relationship to rainfall and river flows in the continental U.S., *Geophys. Res. Lett.*, *28*, 2077–2080.

EU-FP6 project ENSEMBLES & ECA&D project (2016), E-OBS gridded dataset, Website, accessed at 2016/1/22 `http://www.ecad.eu/download/ensembles/download.php`.

Feng, X., T. Delsole, and P. Houser (2012), A Method for Estimating Potential Seasonal Predictability: Analysis of Covariance, *J. Clim.*, *25*, 5292–5308.

Feser, F., B. Rockel, H. von Storch, J. Winterfeld, and M. Zahn (2011), Regional climate models add value to global model data, *Bull. Am. Meteorol. Soc.*, *92*, 1181–1192.

Frumkins, H., J. HEss, G. Luber, J. Malilay, and M. McGeehin (2008), Climate Change: The Public Health Response, *Am J Public Health*, *98*(3), 435–445, doi:10.2105/AJPH.2007.119362.

García-Serrano, J., F. J. Doblas-Reyes, R. J. Haarsma, and I. Polo (2013), Decadal prediction of the dominant West African monsoon rainfall modes, *Journal of Geophysical Research: Atmospheres*, *118*, 5260–5279, doi:10.1002/jgrd.50465.

Gilbert, L., J. Aungier, and J. L. Tomkins (2014), Climate of origin affects tick (Ixodes ricinus) host-seeking behavior in response to temperature: implications for resilience to climate change?, *Ecology and Evolution*, *4*(7), 1186–1198, doi:10.1002/ece3.1014.

Giorgi, F., C. Jones, and G. Asrar (2009), Addressing climate information needs at the regional level: The CORDEX framework., *WMO Bulletin*, *58*(3), 175–183.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. J. Mason, B. P. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C. A. T. Ferro, D. B. Stephenson, G. A. Meehl, T. Stockdale, R. Burgman, A. M. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, and T. Delworth (2013), A verification framework for interannuall-to-decadal prediction experiments, *Clim. Dyn.*, *40*, 245–272.

Griffies, S., and K. Bryan (1997), Predictability of North Atlantic multidecadal climate variability, *Science*, *275*, 181–184.

Gunkel, C. (2013), Rekordsommer 2003 – Die vergessene Jahrhundertkatastrophe, Website, Retrieved March 29, 2016,from `http://www.spiegel.de/einestages/jahrhundertsommer-2003-eine-der-groessten-naturkatastrophen-europas-a-951214.html`.

Hansen, J., M. Sato, R. Ruedy, L. Nazarenko, A. Lacis, G. Schmidt, G. Russell, I. Aleinov, M. Bauer, S. Bauer, N. Bell, B. Cairns, V. Canuto, M. Chandler, Y. Cheng, A. D. Genio, G. Faluvegi, E. Fleming, A. Friend, T. Hall, C. Jackman, M. Kelley, N. Kiang, D. Koch, J. Lean, J. Lerner, K. Lo, S. Menon, R. Miller, P. Minnis, T. Novakov, V. Oinas, J. Perlwitz, J. Perlwitz, D. Rind, A. Romanou, D. Shindell, P. Stone, S. Sun, N. Tausnev, D. Thresher, B. Wielicki, T. Wong, M. Yao, , and S. Zhang (2005), Efficacy of climate forcings, *Geophys. Res. Lett.*, *110*, D18,104.

Hansen, J., M. Sato, R. Ruedy, P. Kharecha, A. Lacis, R. L. Miller, L. Nazarenko, K. Lo, G. Schmidt, G. Russell, I. Aleinov, S. Bauer, E. Baum, B. Cairns, V. Canuto, M. Chandler, Y. Cheng, A. Cohen, A. D. Genio, G. Faluvegi, E. Fleming, A. Friend, T. Hall, C. Jackman, J. Jonas, M. Kelley, N. Y. Kiang, D. Koch, G. Labow, J. Lerner, S. Menon, T. Novakov, V. Oinas, J. P. Perlwitz, J. Perlwitz, D. Rind, A. Romanou, R. Schmunk, D. Shindell, P. Stone, S. Sun, D. Streets, N. Tausnev, D. Thresher, N. Unger, M. Yao, , and S. Zhang (2007), Climate simulations for 1880-2003 with GISS modelE, *Clim. Dyn.*, *29*, 661–696.

Hasselmann, K. (1976), Stochastic climate models Part I. Theory, *Tellus*, *28*(6), 473–485, doi: 10.1111/j.2153-3490.1976.tb00696.x.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*(8), 1095+, doi:{10.1175/2009BAMS2607.1}.

Haylock, M., N. Hofstra, A. K. Tank, E. Klok, P. Jones, and M. New (2008), A European daily high-resolution gridded dataset of surface temperature and precipitation, *J. Geophys. Res.*, *113*, D20,119.

Hubálek, Z., J. Halouzka, and Z. Juricová (2003), Host-seeking activity of ixodid ticks in relation to weather variables, *Journal of Vector Ecology*, *28*(2).

Hurrell, J., M. Visbeck, and A. Pirani (2011), CLIVAR Exchanges - Special Issue: WCRP Coupled Model Intercomparison Project - Phase 5 - CMIP5, *Project report*.

Hurrell, J. W. (1996), Influence of variations in extratropical wintertime teleconnections on Northern Hemisphere temperature, *Geophys. Res. Lett.*, *23*, 665–668.

Hurrell, J. W., T. Delworth, G. Danabasoglu, H. Drange, K. Drinkwater, S. Griffies, N. Holbrook, B. Kirtman, N. Keenlyside, M. Latif, J. Marotzke, J. Murphy, G. Meehl, T. Palmer, H. Pohlmann, T. Rosati, R. Seager, D. Smith, R. Sutton, A. Timmermann, K. Trenberth, J. Tribbia, and M. Visbeck (2010), Decadal climate variability, predictability and prediction: Opportunities and challenges, In: *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society*, Vol. 2, Edited by J. Hall, D. E. Harrison, and D. Stammer, p. 2pp, ESA Publ. WPP–306.

Ineson, S., and A. A. Scaife (2009), The role of the stratosphere in the European climate response to El Niño, *Nature Geosci*, *2*, 32–36.

IPCC (1990), Observed climate variations and change, In: *Climate Change: The IPCC Scientific Assessment*, Edited by J. T. Houghton, G. J. Jenkins, and J. J. Ephraums, pp. 195–238.

IPCC (1990), *Report prepared for Intergovernmental Panel on Climate Change by Working Group I [J.T. Houghton, G.J. Jenkins and J.J. Ephraums(eds.)]*, Cambridge University Press, Cambridge, Great Britain, New York, NY, USA and Melbourne, Australia.

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., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC (2012), Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, p. 582 pp., Cambridge University Press, Cambridge, UK, and New York, NY, USA.

IPCC (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]*, p. 1535pp, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jacob, D. (2001), A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin, *Meteorology and Atmospheric Physics*, *77*, 61–73.

Jacob, D., J. Petersen, B. Eggert, A. Alias, O. B. Christensen, L. M. Bouwer, A. Braun, A. Colette, M. Déqué, G. Georgievski, E. Georgopoulou, A. Gobiet, L. Menut, G. Nikulin, A. Haensler, N. Hempelmann, C. Jones, K. Keuler, S. Kovats, N. Kröner, S. Kotlarski,

A. Kriegsmann, E. Martin, E. Meijgaard, C. Moseley, S. Pfeifer, S. Preuschmann, C. Radermacher, K. Radtke, D. Rechid, M. Rounsevell, P. Samuelsson, S. Somot, J.-F. Soussana, C. Teichmann, R. Valentini, R. Vautard, B. Weber, and P. Yiou (2013), EURO-CORDEX: New high-resolution climate change projections for European impact research, *Regional Environmental Change*, *14*(2), 563–578, doi:10.1007/s10113-013-0499-2.

Johansen, S. (2008), Correlation, regression, and cointegration of nonstationary economic time series, *Bulletin of the ISI LXII 2007*, pp. 19–26.

Jolliffe, I. T., and D. B. Stephenson (2011), Introduction, In: *Forecast verification: A practitioner's guide in atmospheric science*, Vol. 2, Edited by I. T. Jolliffe and D. B. Stephenson, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Jungclaus, J. H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, and J. S. von Storch (2013), Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model, *Journal of Advances in Modeling Earth Systems*, *5*, 422–446, doi:10.1002/jame.20023.

Kadow, C., S. Illing, O. Kunst, H. W. Rust, H. Pohlmann, W. A. Müller, and U. Cubasch (2015), Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system, *Meteorologische Zeitschrift*, p. NA, doi:10.1127/metz/2015/0639.

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, *Bull. Amer. Meteor. Soc.*, *77*, 437–371.

Kanamitsu, M., and L. DeHaan (2011), The added value Index: A new metric to quantify the addded value of regional models, *J. Geophys. Res.*, *116*, D11,106.

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.

Keenlyside, N. A., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner (2008), Advancing decadal-scale climate prediction in the North Atlantic sector, *Nature*, *438*, 84–88.

Khodayar, S., A. Sehlinger, and H. Feldmann (2014), Sensitivity of soil moisture initialization for decadal predictions under different regional climatic conditions in Europe, *Int. J. Clim.*, p. (Revised).

Kiewra, D., M. Kryza, and M. Szymanowski (2014), Influence of selected meteorological variables on the questing activity of Ixodes ricinus ticks in Lower Silesia, SW Poland, *JOURNAL OF VECTOR ECOLOGY*, *39*(1), 138–145, doi:{10.1111/j.1948-7134.2014.12080.x}.

Kim, H.-M., P. J. Webster, and J. A. Curry (2012), Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts, *Geophys. Res. Lett.*, *39*, L10,701+.

Kirtman, B., S. Power, J. Adedoyin, G. Boer, R. Bojariu, I. Camilloni, F. Doblas-Reyes, A. Fiore, M. Kimoto, G. Meehl, M. Prather, A. Sarr, C. Schär, R. Sutton, G. van Oldenborgh, G. Vecchi, and H. Wang (2013), Near-term Climate Change: Projections and Predictability, In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Kleeman, R. (2002), Measuring Dynamical Prediction Utility Using Relative Entropy, *J. Atmos. Sci.*, *59*, 2057–2072.

Knight, J. R., R. J. Allan, C. K. Folland, M. Vellinga, and M. E. Mann (2005), A signature of persistent natural thermohaline circulation cycles in observed climate, *Geophys. Res. Lett.*, *32*(20), L20,708+, doi:10.1029/2005gl024233.

Knight, J. R., C. K. Folland, and A. A. Scaife (2006), Climate impacts of the Atlantic Multidecadal Oscillation, *Geophys. Res. Lett.*, *33*, L17,706.

Knight, J. R., M. B. Andrews, D. M. Smith, A. Arribas, A. W. Colman, N. J. Dunstone, R. Eade, L. Hermanson, C. MacLachlan, K. A. Peterson, A. A. Scaife, and A. Williams (2014), Predictions of Climate Several Years Ahead Using an Improved Decadal Prediction System, *J. Clim.*, *20*, 7550–7567, doi:10.1175/jcli-d-14-00069.1.

Kottmeier, C., and H. Feldmann (2012), Project Description for the BMBF MiKlip Programme: Module-C Regionalization (C-50 Regio_Predict), *Project decription*, IMK-TRO, KIT, Karlsruhe, Karlsruhe.

Krueger, H., and G. A. Lienert (1980), Eine exakte nonparametrische Prüfung auf Kovarianz autokorrelierter Zeitreihen, *J.exp.angew.Psych.*, *3*, 460–467.

Latif, M., and N. S. Keenlyside (2011), A perspective on decadal climate variability and predictability, *Deep Sea Research Part II: Topical Studies in Oceanography* , *58*, 1880–1894.

Latif, M., M. Collins, H. Pohlmann, and N. Keenlyside (2006), A Review of Predictability Studies of Atlantic Sector Climate on Decadal Time Scales, *J. Clim.*, *19*, 5971–5987.

Lettenbauer, A. (2013), Verifikation globaler dekadischer Klimaprognosen aus dem CMIP5 Projekt.

Liu, Y., Q. Meng, R. Chen, J. Wang, S. Jiang, and Y. Hu (2004), A New Method to Evaluate the Similarity of Chromatographic Fingerprints: Weighted Pearson Product Moment Correlation Coefficient, *J. Chromatogr. Sc.*, *42*, 545–550.

MacLeod, D. A., C. Caminade, and A. P. Morse (2012), Useful decadal climate prediction at regional scales? A look at the ENSEMBLES stream 2 decadal hindcasts, *Environ. Res. Lett.*, *7*(4), doi:{10.1088/1748-9326/7/4/044012}.

Majda, A., R. Kleeman, and D. Cai (2002), A mathematical framework for quantifying predictablity through relative entropy, *Methods and Applications of Analysis*, *9*, 425–444.

Mantua, N. J., and S. R. Hare (2002), The Pacific Decadal Oscillation, *J. Ocean.*, *58*(1), 35–44, doi:10.1023/A:1015820616384.

Marotzke, J., W. Müller, F. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, K. Prömmel, H. Rust, D. Stammer, U. Ulbrich, C. Kadow, A. Köhl, J. Kröger, T. Kruschke, J. Pinto, H. Pohlmann, M. Reyers, M. Schröder, F. Sienz, C. Timmreck, and M. Ziese ().

Marotzke, J., D. Stammer, U. Cubasch, C. Kottmeier, U. Ulbrich, and P. Becker (2012), MiKlip – A Research Project on Decadal Climate Prediction, *Project decription*.

Mason, S. J., and A. P. Weigel (2009), A Generic Forecast Verification Framework for Admistrative Purposes, *Mon. Weather Rev.*, *137*, 331–349.

Matei, D., J. Baehr, J. H. Jungclaus, H. Haak, W. A. Müller, and J. Marotzke (2012), Multiyear Prediction of Monthly Mean Atlantic Meridional Overturning Circulation at 26.5 N, *Science*, *335(6064)*, doi:10.1126/science.1210299.

Matei, D., H. Pohlmann, J. Jungclaus, W. Müller, H. Haak, and J. Marotzke (2012), Two Tales of Initializing Decadal Climate Prediction Experiments with the ECHAM5/MPI-OM Model, *J. Clim.*, *25*(24), 8502–8523, doi:{10.1175/JCLI-D-11-00633.1}.

Meehl, G., L. Goddard, J. M. Murphy, R. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. Giorgetta, A. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimoto, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer, and T. Stockdale (2009), Decadal prediction: Can it be skillful?, *Bull. of the Amer. Meteor. Soc.*, *90*, 1467–1485, doi:10.1175/2009BAMS2778.1.

Meehl, G. A., L. Goddard, G. Boer, R. Burgman, G. Branstator, C. Cassou, S. Corti, G. Danabasoglu, F. Doblas-Reyes, E. Hawkins, A. Karspeck, M. Kimoto, A. Kumar, D. Matei, J. Mignot, R. Msadek, H. Pohlmann, M. Rienecker, T. Rosati, E. Schneider, D. Smith, R. Sutton, H. Teng, G. J. van Oldenborgh, G. Vecchi, , and S. Yeager (2013), Decadal Climate Prediction: An Update from the Trenches, *Bull. Am. Meteorol. Soc.*

Merriam Webster Dictionary (2016), utility, Website, retrieved March 29, 2016, from `http://www.merriam-webster.com/dictionary/utility`.

Met Office Hadley Centre (2016), HadCRUT4, Website, accessed at 2016/1/22 `http://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/download.html`.

Mieruch, S., H. Feldmann, G. Schädler, C.-J. Lenz, S. Kothe, and C. Kottmeier (2013), *Geosci. Model Dev. Discuss.*, *6*, doi:doi:10.5194/gmdd-6-5711-2013.

Mieruch, S., H. Feldmann, G. Schädler, C.-J. Lenz, S. Kothe, and C. Kottmeier (2014), The regional MiKlip decadal forecast ensemble for Europe: the added value of downscaling, *Geoscientific Model Development*, *7*(6), 2983–2999, doi:10.5194/gmd-7-2983-2014.

Miller, R., G. Schmidt, L. Nazarenko, N. Tausnev, S. Bauer, A. DelGenio, M. Kelley, K. Lo, R. Ruedy, D. Shindell, I. Aleinov, M. Bauer, R. Bleck, V. Canuto, Y. Chen, Y. Cheng, T. Clune, G. Faluvegi, J. Hansen, R. Healy, N. Kiang, D. Koch, A. Lacis, A. LeGrande, J. Lerner, S. Menon, V. Oinas, C. Garcia-Pando, J. Perlwitz, M. Puma, D. Rind, A. Romanou, G. Russell, M. Sato, S. Sun, K. Tsigaridis, N. Unger, A. Voulgarakis, M.-S. Yao, and J. Zhang (2014), CMIP5 historical simulations (1850-2012) with GISS ModelE2, *JOURNAL OF ADVANCES IN MODELING EARTH SYSTEMS*, *6*, 441–477, doi:10.1002/2013MS000266.

Moran, P. A. P. (1950), Notes on Continuous Stochastic Phenomena, *Biometrika*, *37*, 17–23.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, *J. Geophys. Res.*, (117), D08,101, doi:10.1029/2011JD017187.

Müller, W. A., J. Baehr, H. Haak, J. H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J. S. von Storch, and J. Marotzke (2012), Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology, *Geophysical Research Letters*, *39*(22), L22,707+, doi:10.1029/2012gl053326.

Müller, W. A., H. Pohlmann, F. Sienz, and D. Smith (2014), Decadal climate predictions for the period 1901-2010 with a coupled climate model, *Geophys. Res. Lett.*, *41*(6), 2100–2107, doi:{10.1002/2014GL059259}.

Murphy, A. H. (1988), Skill Scores based on the mean squared error and their relationship to the correlation coefficient, *Mon. Weather Rev.*, *16*, 2417–2424.

Murphy, A. H. (1991), Forecast Verification: Its Complexity and Dimensionality, *Mon. Weather Rev.*, *119*, 1590–1601.

Murphy, A. H. (1993), What is a good Forecast? An Essay on the nature of goodness in weather forecasting, *Wea. Forecasting*, *119*, 1590–1601.

Murphy, A. H., and R. L. Winkler (1987), A General Framework for Forecast Verification, *Mon. Weather Rev.*, *115*, 1330–1338.

Murphy, A. H., and R. L. Winkler (1992), Diagnostic verification of probability forecasts, *Int. J. Forecasting*, *7*, 435–455.

Murphy, J., V. Kattsov, N. Keenlyside, M. Kimoto, G. Meehl, V. Mehta, H. Pohlmann, A. Scaife, and D. Smith (2010), Towards Prediction of Decadal Climate Variability and Change, *Procedia Environmental Sciences*, *1*, 287 – 304, doi:http://dx.doi.org/10.1016/j.proenv.2010.09.018, world Climate Conference - 3.

NASA/GSFC/HSL (2015), Global Mean Effective Forcing. USA: Goddard Earth Sciences Data and Information Services Center (GES DISC), Website, accessed at 2015/10/11 `http://data.giss.nasa.gov/modelforce/Fe.1880-2011.txt`.

National Research Council (), *Abrupt Climate Change: Inevitable Surprises*, doi:10.17226/10136.

NOAA–ESRL (2015), AMO (Atlantic Multidecadal Oscillation) Index. USA: NOAA-ESRL Physical Sciences Division, Website, accessed at 2015/10/11 `http://www.esrl.noaa.gov/psd/data/timeseries/AMO/`.

Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger (2012), Decadal prediction skill in a multi-model ensemble, *Climate Dynamics*, *38*(7), 1263–1280.

Orcutt, G. H., and S. F. James (1948), Testing the Significance of correlation between time series, *Biometrika*, *335*.

Panitz, H., G. Fosser, R. Sasse, K. Sedlmeier, S. Mieruch, M. Breil, H. Feldmann, , and G. Schädler (2013), High Resolution Climate Modelling with the CCLM Regional Model, In: *High Performance Computing in Science and Engineering '13*, Edited by W. E. Nagel, D. Kröner, and M. Resch, Springer Berlin Heidelberg New York.

Peterson, T. C., C. Folland, G. Gruza, W. Hogg, A. Mokssit, and N. Plummer (2001), Report of the activities of the Working Group on Climate Change Detection and related rapporteurs 1998-2001, p. 143pp, Comm. for Climatol., World Meteorol. Organ., Geneva, Switzerland.

Pohlmann, H., J. H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke (2009), Initializing Decadal Climate Predictions with the GECCO Aceanis Sythesis: Effects on the North Atlantic, *J. Clim.*, *22*, 3926–3938.

Pohlmann, H., W. A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F. S. E. Vamborg, C. Kadow, S. Illing, and J. Marotzke (2013), Improved forecast skill in the tropics in the new MiKlip decadal climate predictions, *Geophys. Res. Lett.*, *40*(3–4), 5798–5802.

Polkova, I., A. Köhl, and D. Stammer (2014), Impact of initialization procedures on the predictive skill of a coupled ocean–atmosphere model, *Climate Dynamics*, *42*, 3151–3169, doi: 10.1007/s00382-013-1969-4.

Prentice, C., W. Cramer, S. P. Harrison, R. Leemans, R. A. Monserud, and A. M. Solomon (1992), A global biome model based on plant physiology and dominance, soil proporties and climate, *J. Biogeogr.*, *19*, 117–134.

Prichard, D., and J. Theiler (1994), Generating Surrogates Data for Time Series with Several Simultaneously Measured Variables, *SFI Working Paper*, *23*, 1–4.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reyers, M., J. G. Pinto, and J. Moemken (2015), Statistical-dynamical downscaling for wind energy potentials: evaluation and applications to decadal hindcasts and climate change projections, *INTERNATIONAL JOURNAL OF CLIMATOLOGY*, *35*(2), 229–244, doi: {10.1002/joc.3975}.

Rial, J. A., R. A. Pielkesr, M. Beniston, M. Claussen, J. Canadell, P. Cox, H. Held, N. De Noblet-Ducoudré, R. Prinn, J. F. Reynolds, and J. D. Salas (2004), Nonlinearities, feedbacks and critical thresholds within the Earth's climate system, *Clim. Change*, *65*, 11–38.

Robine, J., S. L. Cheung, S. L. Roy, H. V. Oyen, and F. R. Herrmann (2007), 2003 Heat Wave Project: Report on excess mortality in Europe in Summer 2003.

Schlesinger, M. E. (1994), An oscillation in the global climate system of period 65-70 years, *Nature*, *367*, 723–726.

Schönwiese, C.-D. (2013), *Praktische Statistik für Meteorologen und Geowissenschaftler*, Schweizerbart Science Publishers, Stuttgart, Germany.

Schreiber, T., and A. Schmitz (2000), Surrogate time series, *Physica D: Nonlinear Phenomena*, *142*(3–4), 346–382.

SEC (2011), Research joint programming initiative on climate change: 'Connecting Climate Knwoledge for Europe', EUROPEAN COMMISSION, Brussels, accessed 03 December 2015, http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/sec/2011/1254/COM_SEC%282011%291254_EN.pdf.

Sehlinger, A. (2012), Charakterisierung des Einflusses der Initialisierung des Bodens auf mittelfristige Trends im Untergrund und in der Atmosphäre.

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang, D. Bronaugh, and M. G. Donat (2014), Evaluating model-simulated variability in temperature extremes using modified percentiles indices, *Int. J. Clim.*, *34*, 3304–3311.

Sivakumar, M. V. K., R. P. Motha, H. P. Das, and Ed. (2005), Springer Berlin Heidelberg New York.

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy (2007), Improved Surface Temperature Prediction for the coming Decade from a Global Climate Model, *Science*, *317*, 796–799.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife (2010), Skilful multi-year predictions of Atlantic hurricane frequency, *Nature Geoscience*, *3*, 846–849.

Smith, D. M., A. A. Scaife, G. J. Boer, M. Caian, F. J. Doblas-Reyes, V. Guemas, E. Hawkins, W. Hazeleger, L. Hermanson, C. K. Ho, M. Ishii, V. Kharin, M. Kimoto, B. Kirtman, J. Lean, D. Matei, W. J. Merryfield, W. A. Müller, H. Pohlmann, A. Rosati, B. Wouters, and K. Wyser (2012a), Real-time multi-model decadal climate predictions, *Climate Dynamics*, *41*(11), 2875–2888, doi:10.1007/s00382-012-1600-0.

Smith, D. M., A. A. Scaife, and B. P. Kirtman (2012b), What is the current state of scientific knwoledge with regard to seasonal and decadal forecasting?, *Environ. Res. Lett.*, *7*, 015,602.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, and T. R. amd Erich Roeckner (2013), Atmospheric component of the MPI-M Earth System Model: ECHAM6, *J. Adv. Model. Earth Syst.*, *5*, 1590–1601, doi:10.1002/jame.20015.

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, M. Schröder, A. Mazurkiewicz, and A. Hense (2015), Revealing skill of the MiKlip decadal prediction system by three-dimensional probabilistic evaluation, *Meteorologische Zeitschrift*, p. NA, doi:10.1127/metz/2015/0606.

Stott, P. A., D. A. Stone, and M. R. Allen (2004), Human contribution to the European heatwave of 2003, *Nature*, *432*, 610–614.

Süss, J., C. Klaus, F.-W. Gerstengarbe, and P. C. Werner (2008), What Makes Ticks Tick? Climate Change, Ticks, and Tick-Borne Diseases, *Journal of Travel Medicine*, *15*(1), 39–45, doi:10.1111/j.1708-8305.2007.00176.x.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An Overview of CMIP5 and the Experiment Design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498, doi:10.1175/bams-d-11-00094.1.

Teng, H., and G. Branstator (2011), Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM, *Clim. Dyn.*, *36*, 1813–1834.

Teng, H., G. Branstator, and G. A. Meehl (2011), Predictability of the Atlantic Overturning Circulation and Associated Surface Pattern in Two CCSM3 Climate Chnage Ensemble Experiments, *J. Clim.*, *24*, 6054–6076.

Theiler, J., S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer (1992a), Testing for nonlinearity in time series: the method of surrogate data, *Physica D*, *58*, 77–94.

Theiler, J., B. Galdrikian, A. Longtin, S. Eubank, and J. D. Farmer (1992b), Using Surrogate Data to Detect Nonlinearity in Time Series, In: *Nonlinear Modeling and Forecasting*, Edited by M. Casdagli and S. Eubank, pp. 163–188, Addison-Wexley.

Theiler, J., P. S. Linsay, and D. M. Rubin (1993), Detecting Nonlinearity in Data with Long Coherence Times, In: *Predicting the Future and Understanding the Past,SFI Studies in the Sciences of Complexity, Proc. Col. XVII*, Edited by A. S. Weigend and N. A. Gershenfeld, Addison-Wexley.

Thiébaux, H. J., and F. W. Zwiers (1984), The Interpretatin and Estimation of effective Sample Size, *J. Clim. a. Appl. Meteorol.*, *23*, 800–811.

Ting, M., Y. Kushnir, R. Seager, , and C. Li (2009), Forced and Internal Twentieth-Century SST Trends in the North Atlantic, *J. Clim.*, *22*, 1469–1481.

Tippet, M., R. Kleeman, and Y. Tang (2004), Measuring the potential utility of seasonal climate predictions, *31*, L22,201.

Tippett, M. K., R. Kleeman, and Y. Tang (2004), Measuring the potential utility of seasonal climate predictions, *Geophys. Res. Lett.*, *31*(22), L22,201, doi:10.1029/2004GL021575.

Tobler, W. (1970), A computer movie simulating urban growth in the Detroit region, *Economic Geography*, *46*, 234–240.

Tomkins, J. L., J. Aungier, W. Hazel, and L. Gilbert (2014), Towards an Evolutionary Understanding of Questing Behaviour in the Tick Ixodes ricinus, *PLoS ONE*, *9*(10), 1–11, doi:10.1371/journal.pone.0110028.

Trenberth, K. E., P. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. K. Tank, D. Parker, F. Rahimzadeh, J. Renwick, M. Rusticucci, B. Soden, and P. Zhai (2007), Observations: Surface and Atmospheric Climate Change, In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Edited by D. Q. M. M. Z. C. M. M. K. A. M. T. Solomon, S. and H. Miller, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Turner, J. (2004), The El Niño-southern oscillation and Antarctica, *International Journal of Climatology*, *24*, 1–31, doi:10.1002/joc.965.

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 re-analysis, *Quarterly Journal of the Royal Meteorological Society*, *131*(612), 2961–3012, doi: 10.1256/qj.04.176.

Venema, V., F. ament, and C. Simmer (2006a), A Stochastic Iterative Amplitude Adjusted Fourier Transform algorithm with improved accuracy, *Nonlin. Processes Geophys.*, *13*, 321–328.

Venema, V., S. Bachner, H. W. Rust, and C. Simmer (2006b), Statistical characteristics of surrogate data based on geophysical measurements, *Nonlin. Processes Geophys.*, *13*, 449–466.

Vera, C., M. Barange, O. Dube, L. Goddard, D. Griggs, N. Kobysheva, E. Odada, S. Parey, J. Polovina, G. Poveda, B. Seguin, and K. Trenberth (2010), Needs Assessment for Climate Information on Decadal Timescales and Longer, *Procedia Environmental Sciences*, *1*, 275 – 286, doi:http://dx.doi.org/10.1016/j.proenv.2010.09.017, world Climate Conference - 3.

Villalba, R., and T. T. Veblen (1998), Annual- versus decadal-scale climatic influences on tree establishment and mortality in northern Patagonia, In: *The Impacts of Climate Variability on Forests*, *Lecture Notes in Earth Sciences*, Vol. 74, Edited by M. Beniston and J. L. Innes, pp. 145–170, Springer Berlin Heidelberg.

Weigel, A. P., and S. J. Mason (2011), The Generalized Discrimination Score for Ensemble Forecasts, *Mon. Weather Rev.*, *139*, 3069–3074.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008), Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd Edition, Academic Press.

Yang, X., A. Rosati, S. Zahng, T. L. Delworth, R. G. Gudgel, R. Zhang, G. Vecchi, W. Anderson, Y. Chang, T. Delsole, K. Dixon, R. Msadek, W. F. Stern, A. Wittenberg, and F. Zeng (2013), A Predictable AMO-Like Pattern in the GFDL Fully Coupled Ensemble Initialization and Decadal Forecasting System, *J. Clim.*, *26*, 650–661.

Yule, G. U. (1926), Why do we sometimes get nonsense-correlations between time-series?, *J. roy. statist. Soc*, *89*, 1 –64.

Zanchettin, D., O. Bothe, H. F. Graf, S. J. Lorenz, J. Luterbacher, C. Timmreck, and J. H. Jungclaus (2013), Background conditions influence the decadal climate response to strong volcanic eruptions, *Journal of Geophysical Research: Atmospheres*, *118*, 4090–4106, doi:10.1002/jgrd.50229.

Zhang, L., H. Dobslaw, C. Dahle, I. Sasgen, and M. Thomas (2015), Validation of MPI-ESM Decadal Hindcast Experiments with Terrestrial Water Storage Variations as Observed by the GRACE Satellite Mission, *Meteorologische Zeitschrift*, p. NA, doi:10.1127/metz/2015/0596.

Zhang, X., G. Hegerl, F. W. Zwiers, , and J. Kenyon (2005), Avoiding inhomogeneity in percentile-based indices of temperature extremes, *J. Clim.*, *18*, 1641–1651.

Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers (2011), Indices for monitoring changes in extremes based on daily temperature and precipitation data, *Wiley Interdisciplinary Reviews: Climate Change*, doi:10.1002/wcc.147.

9. List of Figures

10. List of Tables

List of abbreviations

Abbreviations that are used in this thesis include:

b0	First ensemble generation (baseline0) of MiKlip
b1	Second ensemble generation (baseline1) of MiKlip
MPI-ESM	The global model MPI-ESM-LR with the horizontal resolution T63
CCLM	The regional model COSMO-CLM
b0-EUR22	Regionalization of the first ensemble generation (baseline0) of MiKlip with a horizontal resolution of 0.22 degrees
b1-EUR22	Regionalization of the second ensemble generation (baseline1) of MiKlip with a horizontal resolution of 0.22 degrees
b1-EUR44	Regionalization of the second ensemble generation (baseline1) of MiKlip with a horizontal resolution of 0.44 degrees

Acknowledgements

Many have contributed in many ways to the completion of this thesis. The support and opportunities given to me at the KIT made this time very enjoyable and educative. Special thanks go to my supervisor Prof. Christoph Kottmeier for his support and mentoring. I am also very grateful to Prof Andreas Fink for co-supervising this thesis. Furthermore I need to thank Hendrik Feldmann, who accompanied me through these years, always had an open ear for problems and time for discussions. Many thanks to my working group colleagues for their support and help. Especially, our head Dr. Gerd Schädler for providing support, discussions and additionally points of view . I'd also like to thank Natalie Laube and Katrin Sedlmeier for not only sharing an office with me, but also chocolate, coffee and tea as well as much laughter. The rest of the working group, namely Hans-Jürgen Panitz, Sebastian Mieruch, Marcus Breil, Hans Schipper, Julia Hackenbruch, Benedict Brecht, Emanual Christner, Pascal Dölcker and Melanie Karremann, were great collegues and made work very enjoyable. I also like to thank my collegues in MiKlip from other institutes, especially Fatemeh Davary-Adalatpanah and Mark Reyers, for being such a delight to work with. I am in debt for all the support I received from Janice, Svenja, Leo, Matt and the rest of my friends, as well as my family. My parents have always supported me in everything I have ever done. For that I am eternally grateful. My endless love belongs to my sister who never wavered in her belief in me.

Appendices

A. DATA

A.1. HAdCRUT4

The temperature time series used in Ch. 1 was derived from the HAdCRUT4 data set.

HadCRUT4 is a gridded dataset of global historical surface temperature anomalies. Data are available for each month since January 1850, on a 5 degree grid. The dataset is a collaborative product of the Met Office Hadley Centre and the Climatic Research Unit at the University of East Anglia. The data is a combination of the CRUTEM4 land-surface air temperature dataset and the HadSST3 sea-surface temperature (SST) dataset. The dataset is composed of an ensemble of 100 dataset realisations, thereby sampling the distribution of uncertainty in the global temperature record.

The data used in this work is the mean annual surface temperature deviations to pre-industrial values (1850–1899) over European land. Europe is defined as the area between 35°to 70 °North and -25°to 30°East, plus Turkey (35°to 40°North and 30°to 45°East) [*Morice et al.*, 2012; *Met Office Hadley Centre*, 2016]

A.2. E-OBS

The observational data set used to verify the MiKlip decadal prediction system is the E-OBS data set. E-OBS is a European land-only daily high-resolution gridded data set for precipitation and minimum, maximum, and mean surface temperature for the period 1950–2006 [*Haylock et al.*, 2008; *EU-FP6 project ENSEMBLES & ECA&D project* , 2016]. The gridded data is available on four spatial resolutions to match the grids of many rotated pole Regional Climate Models (RCMs) currently in use. Each has been designed to provide the best estimate of grid box averages rather than point values to enable direct comparison with RCMs.

E-OBS is derived from daily observations of precipitation, and minimum, maximum and mean surface temperature covering the time period 1950–2006 primarrly by Royal Netherlands Meteorological Institute (KNMI), which also hosts the European Climate Assessment and Data set (ECA&D). The raw data then undergoes a series of quality tests and is gridded using three-step methodology of interpolating the daily data: First interpolating the monthly mean using thin-plate splines to define the underlying spatial trend of the data; kriging the anomalies with regard to the monthly mean and finally applying the interpolated anomaly to the interpolated monthly mean.

A.3. AMO-Index

the AMO-Index shows the multi-year variability over the North atlantic (the Multidecadal Atlantic Ascillation) that has been identifed as a major driving factor for climate in Europe on decadal time scales.

The AMO-Index is the detrended, area weighted average over the North Atlantic, basically 0°to 70°N. The index as taken from *NOAA–ESRL* [2015] is calculated from the Kaplan SST dataset (monthly reanalysis of global SST anomalies (SSTA)) covering monthly data from 1856 to the present [*Enfield et al.*, 2001].

A.4. Climate Extreme Event Indices

The project CLIMDEX [*CLIMDEX*, 2013] aimed to produce a suite of in-situ and gridded land-based global data sets of indices representing the extreme aspects of climate [*Karl et al.*, 1999]. CLIMDEX is developed and maintained by researchers at the Climate Change Research Centre (CCRC), The University of New South Wales (UNSW) funded by the Australian Research Council and the Australian Department of Climate Change and Energy Efficiency through Linkage project LP100200690 and in collaboration with the University of Melbourne, Climate Research Division (Environment Canada) and NOAA's National Climatic Data Center (USA). In line with the project a set of 27 core indices, the ETCCDI indices (Expert Team in Climate Change Detection Indices), were defined and have been adopted into the IPC.

The core indices are derived from daily temperature and precipitation data, either as absolute values or as percentiles. The recommended indices are suposed to cover many aspects of a changing global climate. E.g. Frost Days (FD) sample the winter half-year in all extra-tropcial regions, particularly the beginning and end of the cold season in many continental and high latitude climates [*Peterson et al.*, 2001; *Zhang et al.*, 2005].

A.5. Global Radiative Forcings

Climate forcing, measured in W/m^2, is an imposed change of the planetary energy balance. That could be a change in atmospheric CO_2 or solar radiation. To assess and compare the climate impact of different changing atmospheric constituents effectiveness of climate forcings is needed [IPC]. The global mean response can provide a criterion to help evaluate the degree of imposed climate change that would constitute dangerous anthropogenic interference [*Hansen et al.*, 2005].

The forcing data used here is the forcing set adopted for the GISS model [*Hansen et al.*, 2007; *NASA/GSFC/HSL*, 2015]. The forcings include both direct effects as wel as indirect effect. E.g.

carbon dioxide can directly influence cloud cover and precipitation, without any change in global mean temperature. The total forcing is designated as the effective radiative forcing.

The data set available from [*NASA/GSFC/HSL*, 2015] includes among others annually global means of the effective radiative forcing due to well mixed green house gases, ozone, stratospheric water vapor, aerolsols, albedo and solar radiation. The data used to prescribe or calculate each forcing agent are summarized by *Miller et al.* [2014].

A.6. MPI-ESM

The MPI-ESM is a coupled atmosphereand ocean model, consisting of the ECHAM6 and the ocean model MPI-OM of the Max-Planck-Institute for Meteorology (MPI-M) [*Stevens et al.*, 2013]. The MPI-ESM couples the atmosphere, ocean and land surface through the exchange of energy, momentum, water and important trace gases such as carbon dioxide. It has been included in *IPCC* [2013].

The echam is an atmospheric general circulation model. It was developed from an early version of the global numerical weather prediction model developed at the ECMWF. The MPIOM is the ocean-sea ice component of the MPI-ESM. MPI-OM is a primitive equation model (C-Grid, z-coordinates, free surface) with the hydrostatic and Boussinesq assumptions made. It includes an embedded dynamic/thermodynamic sea ice model and a bottom boundary layer scheme for the flow across steep topography. MPIOM uses a curvilinear orthogonal grid which allows for a variety of configurations[*Jungclaus et al.*, 2013].

A.7. CCLM

The COSMO-CLM is the climate version of the three-dimensional non-hydrostatic limited-area atmospheric prediction model COSMO (Consortium for Small-scale Modeling) of Germany's National Meteorological Service (DWD). It uses hydro-thermodynamic equations to describe compressible flow in a moist atmosphere [*Panitz et al.*, 2013; *Doms and Schättler*, 2002]

The model solves prognostic equations for wind, pressure, air temperature, different phases of atmospheric water, soil temperature and soil water content in rotated geographical Coordinates and a generalized terrain following height coordinate [*Doms and Schättler*, 2002]. A variety of physical processes are taken into account by parametrization schemes.

More information can be found at `www.cosmo-model.org`.

A.8. REMO

REMO is a regional climate model originally developed at the MPI-M to be the atmospheric component of the coupled atmosphere-hydrology model system. It is based on the former numer-

ical weather prediction model of the German Weather Service (EUROPA-MODELL, EM). The MPI-M focused on the development of a regional model suitable for climate modelling [*Jacob*, 2001].

It is based on primitive equations in a terrain-following hybrid coordinate system. To avoid numerical instabilities the vertical advection as well as the vertical turbulent fluxes are treated implicitly. The prognostic variables are surface pressure, horizontal wind components, temperature, specific humidity and cloud water.

B. An example for an application oriented climate index: Ticks Questing Index

In Europe the tick *Ixodes ricinus* is the primary vector of medically important pathogens (e.g. Lyme disease) [*Gilbert et al.*, 2014]. The biting activity and the locomotory of the ticks correlate with the risk of humans and animals contracting tick-bourne diseases. Occurrences of diseases in Europe, such as Lyme disease or encephalitis (TBE), have shifted in the last two decades, mostly attributed to warming. It is known, that the climate is the principal restricting factor to the northern spread of *I. ricinus*. With the overall positive trend in temperature in Europe, the following has been observed:

- Winter activity of ticks in Germany and Sweden,
- accelerated life cycles and increase in population in Latvia,
- occurrence at higher altitudes than previously observed in the Czech republic,
- overall northward spread,
- all along with a higher incidence of tick-bourne diseases [*Süss et al.*, 2008].

In light of the recent changes and the very acute consequences of these diseases, the predicting of the risk of infection is of importance. To do so, it is crucial to know which factors influence the ticks host-seeking (questing) behaviour [*Daniel et al.*, 2015]. There are two major contributors to the distribution and abundance of ticks:

- Biotic factors such as host density and
- abiotic factors like the weather [*Daniel et al.*, 2015]

The environmental factors are most important but the affect on the occurrence of ticks and tick-bourne diseases are not straightforward and vary differ between species, stage, habitat and region. However, there are some common points: All ticks strive for the best conditions to start and finish their life cycle (egg, larvae, nymphs, and adult stages). Ticks need high humidity ($> 85\%$), as they are unable to drink and hydrate through their skin e.g. in the soil or the mat layer. Also, to find a host, they require at least an air temperature of $> 7°C$ [*Süss et al.*, 2008].

The questing behaviour generally increases with temperature until the vapour pressure deficit force the tick to rehydrate [*Tomkins et al.*, 2014].

Several studies have tried to link the questing behaviour of ticks directly to meteorological variables. *Hubálek et al.* [2003] found the activity of ticks highly dependent in the air and soil temperature, as well as the daily cycle thereof, and the relative humidity. A predictive 2-variable model of the relative abundance of host-seeking behaviour based on air temperature and relative humidity explained 32 % of the variance. Sampling ticks at 18 woodland sites in South-west Poland, *Kiewra et al.* [2014] found that number of adult ticks is mostly influenced by solar radiation, air temperature and saturation deficit, all three averaged or summed for the last seven days before collecting the ticks. for the nymphs the meteorological parameters at the day of the sampling were the most important, except wind speed, that does not have any significant influence on tick behaviour both adult and nymph.

Following these result, I introduce the *Ticks Questing Index* as the number of days per year where questing of ticks is most likely. In light of the limited number of variables available as gridded observational data sets, the *Ticks Questing Index* has been defined as:

`The number of days per year with a daily mean temperature` $> 7°C$ `after 6 consecutive days with temperature` $> 7°C$ `and precipitation` $> 0mm$

This index does not convey any information on the actual number of ticks, just the favourability of the environment for potential ticks. This index is only a coarse solution to the problem of the risk assessment of ticks. Regression models as used by *Hubálek et al.* [2003] and the use of other variables like humidity would lead to more refined prediction of tick behaviour. But more research is needed, into the evolution of ticks as well as a the transition from models pertaining towards specific sampling sites to universal model all at-risk regions in Europe.

C. An estimation for the Signal to Noise Ratio

The Signal-To-Noise-Ratio of a time series $X_i = x_1, x_2, x_3, ..., x_n$ has to be guessed. To simplify the calculation the time series anly consist of an autoregressive process (of the order 1) S and white noise ϵ:

$$X_i = S_i + \epsilon_i \tag{C.1}$$

The variance of the white noise is :

$$\sigma_\epsilon^2 = (1 - \delta^2) * \sigma_X^2, \tag{C.2}$$

whereas δ is the autocorrelation coefficient lag 1 of the time series X.

The time series variance is

$$\sigma_X^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1} = \frac{\sum_i ((S_i + \epsilon_i) - \overline{S + \epsilon})^2}{n-1} \tag{C.3}$$

Per definition the white noise has mean zero. Additionally we assume the time series also has mean zero (anomalies), then the variance becomes:

$$\sigma_X^2 = \sum_i \frac{S_i^2 + \epsilon_i^2 + 2 * S_i \epsilon_i}{n-1} \tag{C.4}$$

$$\sigma_X^2 = \sum_i \frac{S_i^2}{n-1} + \sum_i \frac{\epsilon_i^2}{n-1} + \sum_i \frac{2 * S_i \epsilon_i}{n-1} \tag{C.5}$$

If the mean of the white noise as well as the time series is zero, the mean of S will be zero, too. Therefore the sums $\sum_i \frac{S_i^2}{n-1} = \sigma_S^2$ and $\sum_i \frac{\epsilon_i^2}{n-1} = \sigma_\epsilon^2$ are the variances of the autoregressive Process and the white noise respectivily. Now the third term is assumed zero:

$$\sum_i \frac{2 * S_i \epsilon_i}{n-1} = 2 * \frac{\sum_i S_i}{n-1} \frac{\sum_i \epsilon_i}{n-1} \approx 0 \tag{C.6}$$

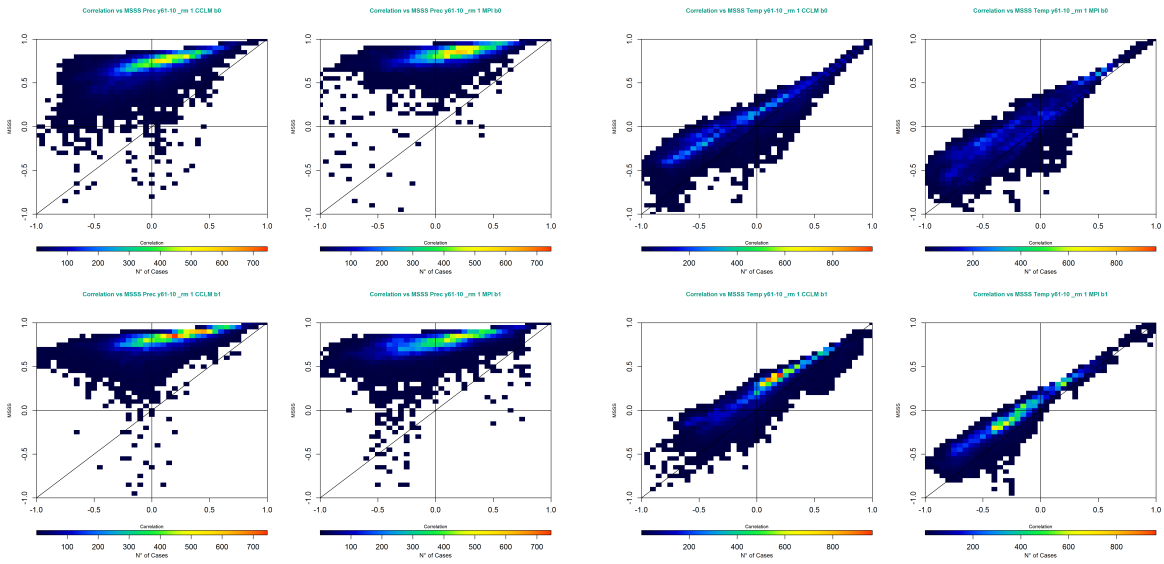
and the variance of the autregressive process becomes:

$$\sigma_S^2 = \sigma_X^2 - \sigma_\epsilon^2 \tag{C.7}$$

Then a simplyfied Signal-to-Noise-Ratio is

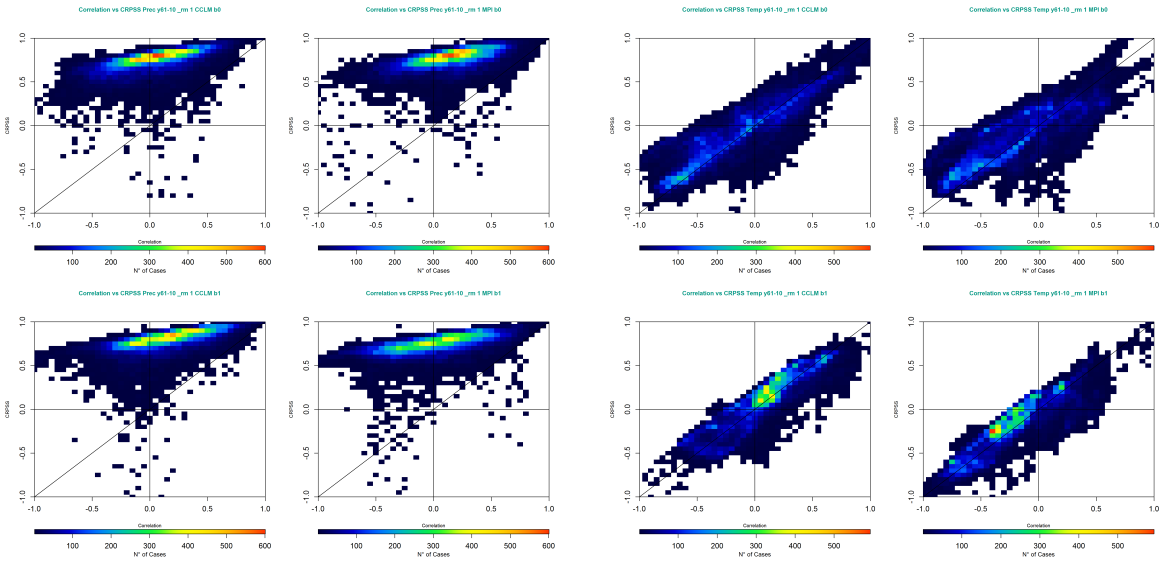
$$SNR = \frac{\sigma_S}{\sigma_\epsilon} = \frac{\sqrt{\sigma_X^2 - [(1-\delta^2) * \sigma_X^2]}}{\sqrt{(1-\delta^2) * \sigma_X^2}} \quad \text{[C.8]}$$

D. Skill Comparison



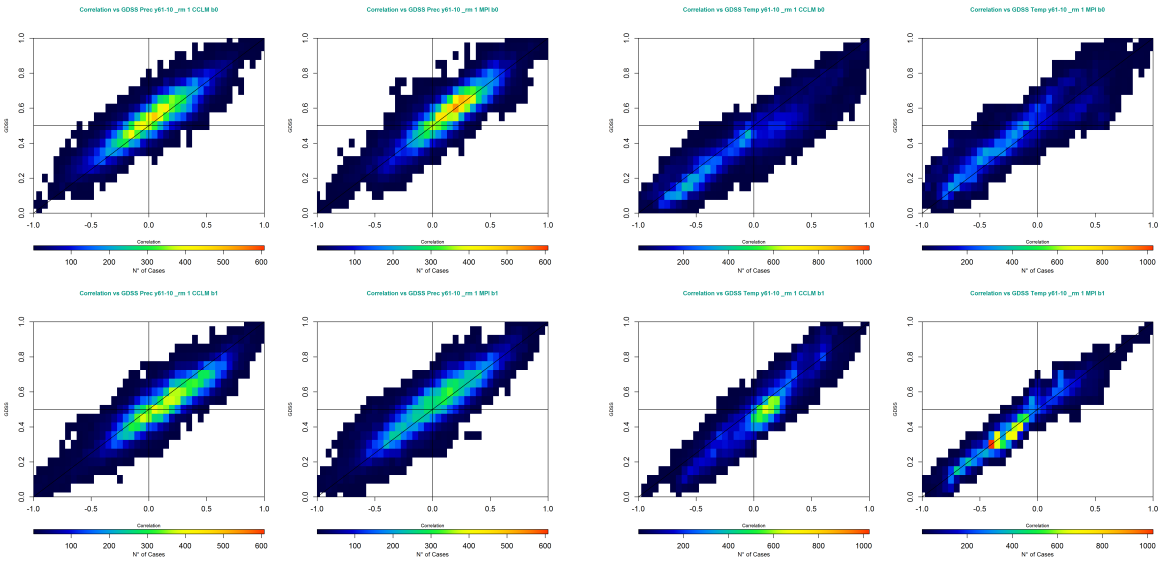
(a) Precipiation (b) Temperature

Figure D.1.: Scatterplot of the relationship of the MSSS and the correlation in winter over Europe. Shown are all land grid points of the CCLM ensemble mean winter prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



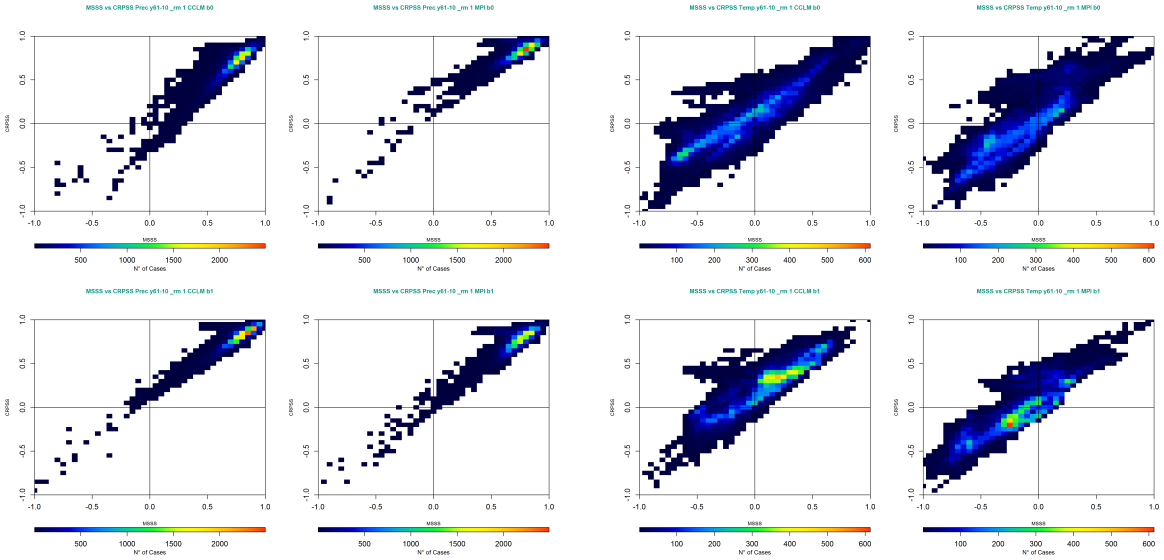
(a) Precipiation (b) Temperature

Figure D.2.: Scatterplot of the relationship of the CRPSS and the correlation in winter over Europe. Shown are all land grid points of the CCLM ensemble mean winter prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



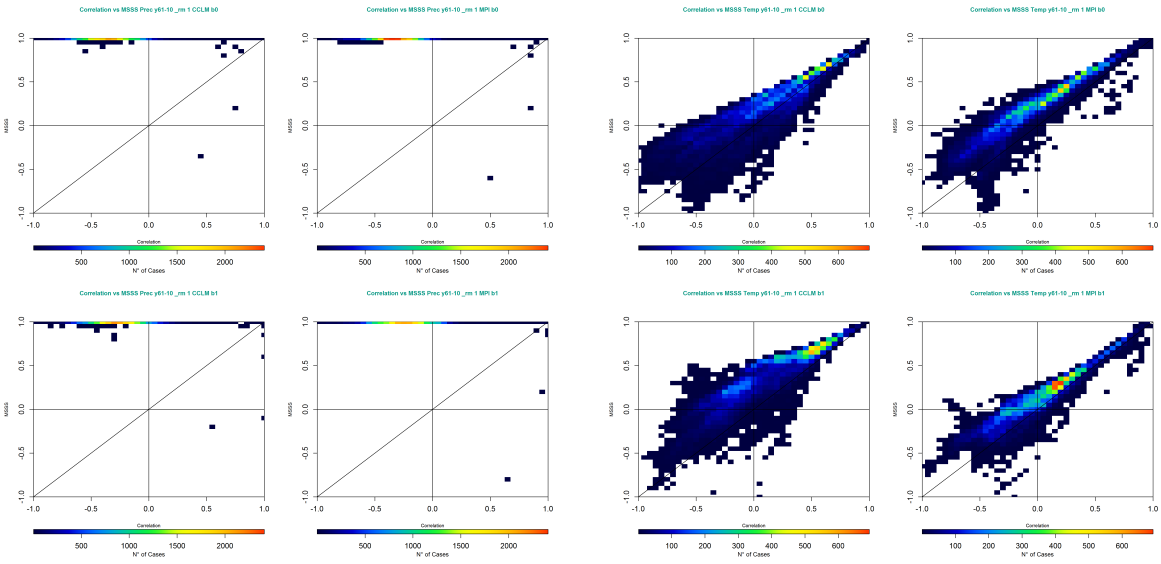
(a) Precipiation (b) Temperature

Figure D.3.: Scatterplot of the relationship of the GDSS and the correlation in winter over Europe. Shown are all land grid points of the CCLM ensemble mean winter prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



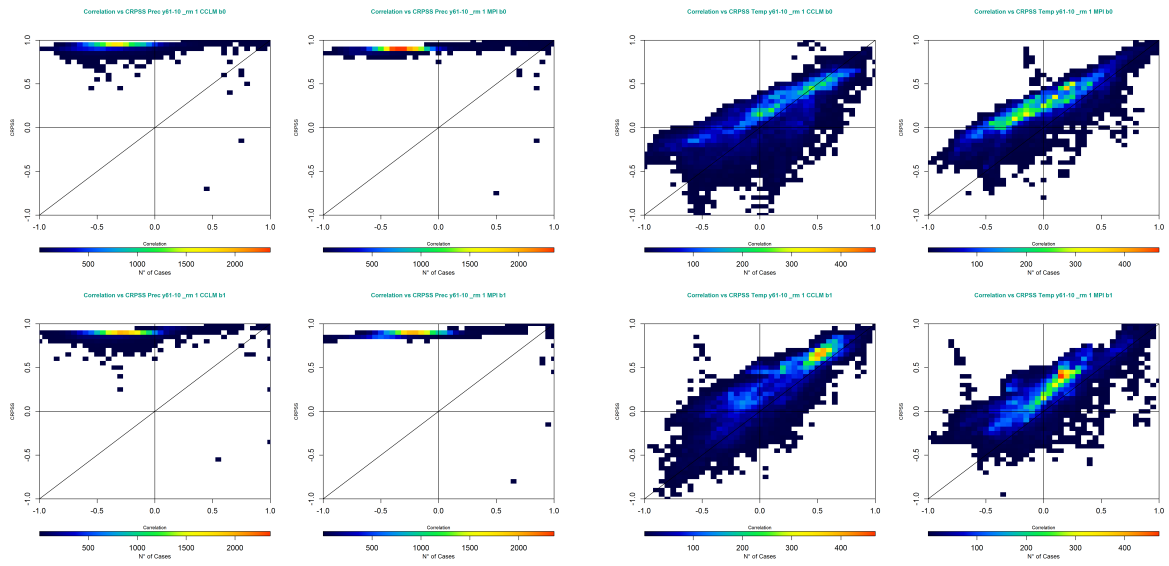
(a) Precipiation (b) Temperature

Figure D.4.: Scatterplot of the relationship of the MSSS and the CRPSS in winter over Europe. Shown are all land grid points of the CCLM ensemble mean winter prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



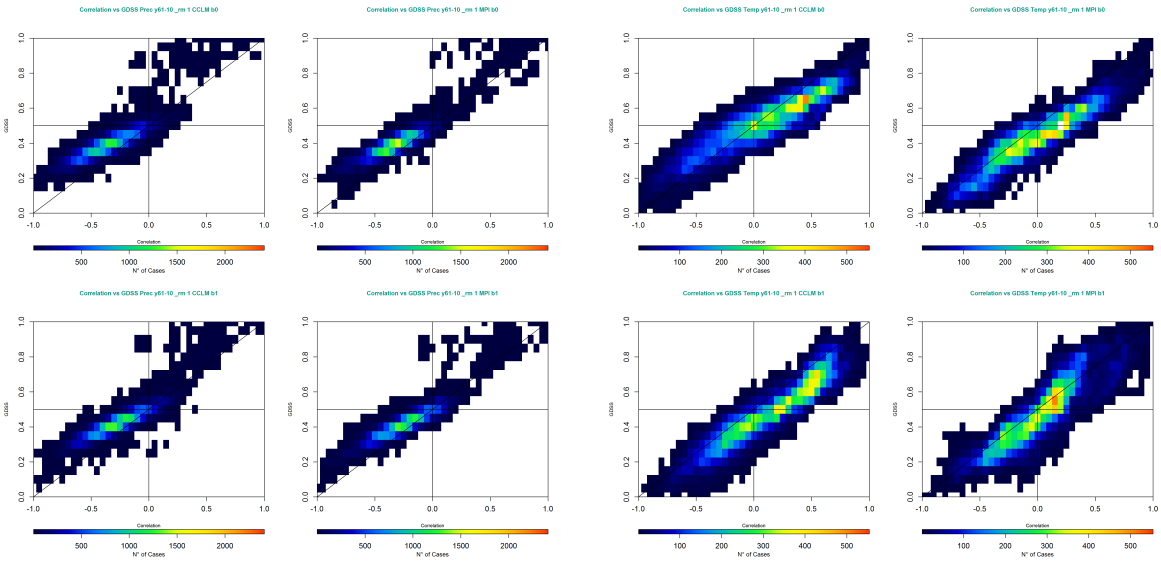
(a) Precipiation (b) Temperature

Figure D.5.: Scatterplot of the relationship of the MSSS and the correlation in summer over Europe. Shown are all land grid points of the CCLM ensemble mean summer prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



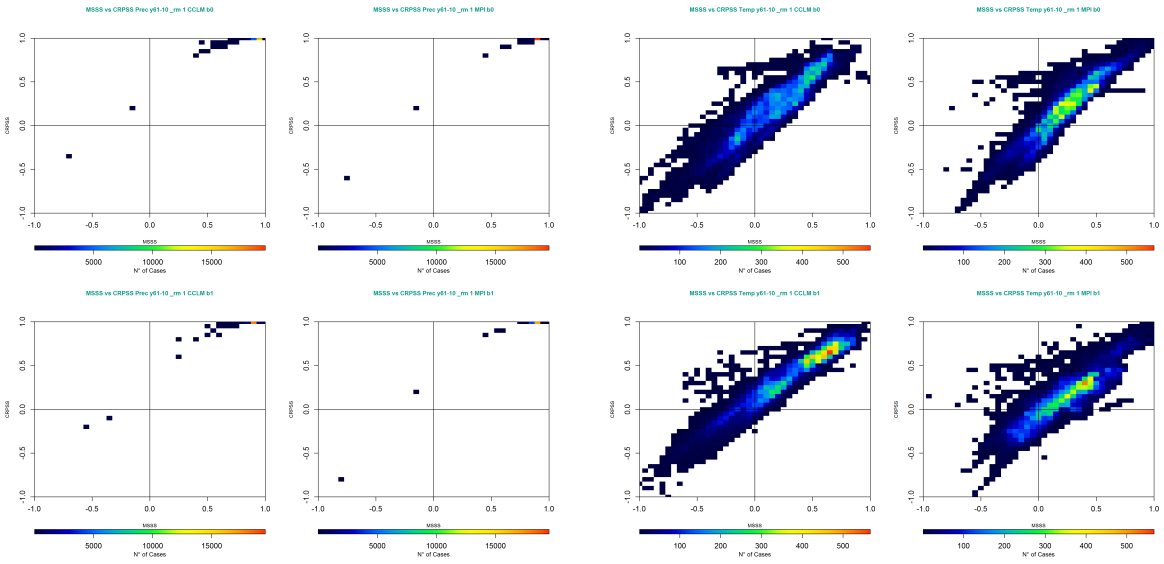
(a) Precipiation (b) Temperature

Figure D.6.: Scatterplot of the relationship of the CRPSS and the correlation in summer over Europe. Shown are all land grid points of the CCLM ensemble mean summer prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



(a) Precipiation (b) Temperature

Figure D.7.: Scatterplot of the relationship of the GDSS and the correlation in summer over Europe. Shown are all land grid points of the CCLM ensemble mean summer prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.



(a) Precipiation (b) Temperature

Figure D.8.: Scatterplot of the relationship of the MSSS and the CRPSS in summer over Europe. Shown are all land grid points of the CCLM ensemble mean summer prediction over 5 decades 1961–2010 from the baseline0 ensemble (initialisation every 10 years) with the long time (50a) trend removed.

E. Verification of the decadally initialized 0.22°ensemble

E.1. Summary of all skill scores for the EUR22 CCLM ensembles

Table E.1.: The skill of the lead years 1–10 of the ***EUR22 decadally initilialized ensembles for temperature and precipitation in winter and summer and the whole year*** *be the number grid points with positive skill as a percentage of the total number of gridpoints. The added value is characterized as the percentage of gridpoints of positive global skill that is improved by downscaling. These numbers do not take into account how high the actual skill is, only if it surspasses the limit for a skillful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS*

	b0 Temp	b0 av Temp	b1 Temp	b1 av Temp	b0 Prec	b0 av Prec	b1 Prec	b1 av Prec
YEAR CRPSS	12	10.5	30.4	21	3.5	3.1	4.8	4
MSSS	15.6	12.5	50	37.5	6.8	5.8	11.7	8.6
GDSS	51.5	32.8	91.4	62.7	44.1	31.4	56.4	30.8
COR	55.2	91.4	93.8	69	46	56.4	55	30
JJA CRPSS	20.7	17.2	16	13.1	5	4.5	9.7	8.4
MSSS	26.7	22.4	25.7	21.2	10.1	8.8	16.3	13
GDSS	76.8	42.9	77.9	38.6	51.1	40	62.3	38.8
COR	72.9	77.9	73.6	42.6	49.3	62.3	64.9	40.3
DJF CRPSS	8.4	7.8	14.5	9.3	2.7	2.5	7.4	6.2
MSSS	18.6	12.5	18.7	11.4	6.6	4.8	13.1	10.4
GDSS	53	30.8	58.6	29.3	49.3	30	59.6	32.7
COR	54.9	58.6	67.9	42.2	52.4	59.6	57.7	31.6

Table E.2.: The skill of the lead years 1–5 of the ***EUR22 decadally initilialized ensembles for temperature and precipitation in winter and summer and the whole year*** *be the difference of number grid points with positive skill as a percentage of the total number of gridpoints. A negative number means, more gridpoints of the global model show positiv skill, i.e. the global model does in general outperform the regional ones. These numbers do not take into account how high the actual skill is, only if it surspasses the limit for a skillful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS*

	b0 Temp	b0 Temp	b1 Temp	b1 Temp	b0 Prec	b0 Prec	b1 Prec	b1 Prec
YEAR CRPSS	22.7	10	33.6	11.4	9.5	-1.6	11.2	-7.6
MSSS	27.7	9.8	51.5	20.4	14.7	-2.1	19.9	-9
GDSS	57.2	-2.9	80.4	1.4	40.4	0.8	48.2	-5.9
COR	61.9	4.2	93.5	9.1	43	0.9	48.1	-5.9
JJA CRPSS	31.7	7.9	23.4	-6.1	10.2	2.2	17.3	-6.1
MSSS	38.3	8.8	34.4	-5.1	15.5	2.6	22.2	-15.1
GDSS	77.8	3.5	73.9	-7.3	50.6	6.1	53	-6.9
COR	68.2	2.1	76.9	-6.4	46.7	8.8	55.4	-6.6
DJF CRPSS	21.9	-1.5	13.8	4.8	7.2	-0.9	17.2	3.6
MSSS	31.7	-2.6	18	2.9	12	-3.7	21.2	-2.7
GDSS	48.4	-3.2	36.2	-2.1	42.1	-0.8	48.9	-4.1
COR	51.1	-1.7	48.5	5.8	42.8	-2.1	49.6	-5.2

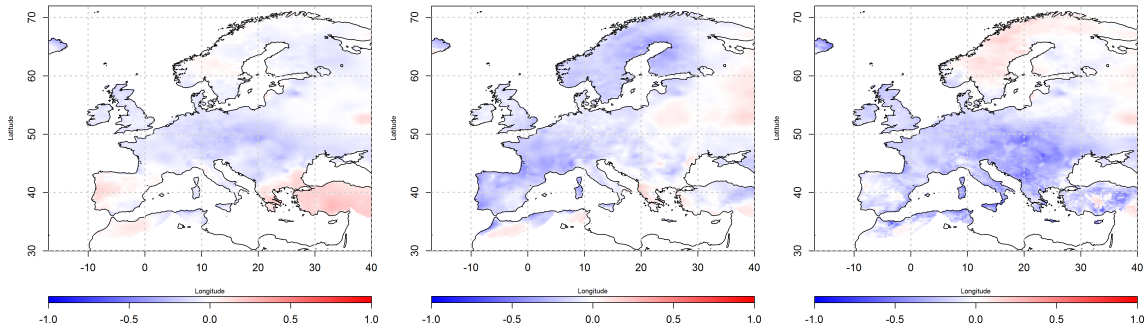
Table E.3.: The skill of the ***EUR22 decadally initilialized ensembles for temperature and precipitation in winter and summer and the whole year in 1961–1980*** *be the difference of number grid points with positive skill as a percentage of the total number of gridpoints. A negative number means, more gridpoints of the global model show positiv skill, i.e. the global model does in general outperform the regional ones. These numbers do not take into account how high the actual skill is, only if it surspasses the limit for a skillful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS*

	b0 Temp	b0 Temp	b1 Temp	b1 Temp	b0 Prec	b0 Prec	b1 Prec	b1 Prec
YEAR CRPSS	22.5	9.6	40	5.9	37.4	-3.6	34.5	-7.6
MSSS	34.2	11.9	51.2	14	36	-6.8	33.4	-10.5
GDSS	59.9	9.8	71.2	11.3	47.9	-2.8	41	-5.5
COR	62.7	5.4	77	6.5	46.8	-6	40.2	-8.6
JJA CRPSS	41.4	10.8	24.7	2.8	32.9	-2.9	38.5	-2.1
MSSS	49.9	22.4	32	7.2	31.3	-3.5	37.6	-2.9
GDSS	79	1.8	70.4	4.4	46.8	-3.2	49.9	2.4
COR	76.2	9	59.3	8.2	50.4	0.1	49.2	3.1
DJF CRPSS	49.4	1	32.5	-5.5	52.9	-1.8	46.5	1
MSSS	54.7	1.2	51.1	-8.6	50.9	-4.1	42.1	-3
GDSS	36.6	-5.3	34.2	-5.6	45.8	-2.5	43.9	-3.6
COR	40.8	0.7	36.8	-3.5	47.3	-1.1	45	-0.9

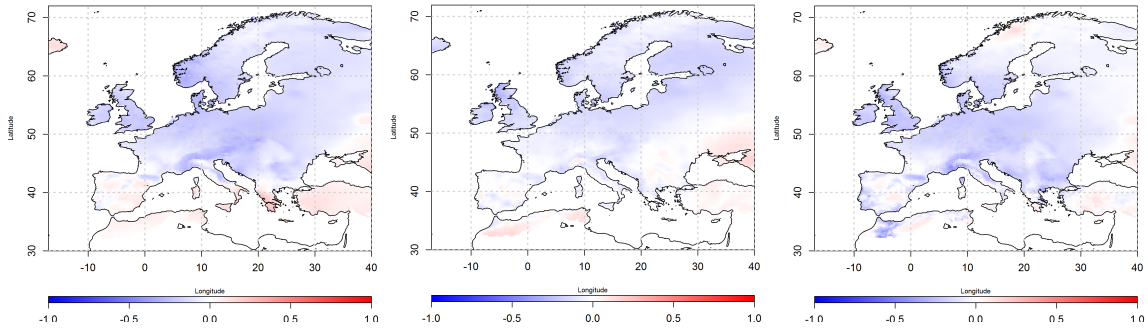
Table E.4.: The skill of the ***EUR22 decadally initilialized ensembles for temperature and precipitation in winter and summer and the whole year in 1991–2010*** *be the difference of number grid points with positive skill as a percentage of the total number of gridpoints. A negative number means, more gridpoints of the global model show positiv skill, i.e. the global model does in general outperform the regional ones. These numbers do not take into account how high the actual skill is, only if it surspasses the limit for a skillful prediction. This limit is zero in the case of the MSSS, the CRPSS and the correlation (COR) and 0.5 in the case of the GDSS*

	b0 Temp	b0 av Temp	b1 Temp	b1 av Temp	b0 Prec	b0 av Prec	b1 Prec	b1 av Prec
YEAR CRPSS	24.7	-0.9	27.3	1.4	15.6	0.2	12	-4.4
MSSS	23	-5.5	29.8	2.4	21.9	-0.5	21.1	-7.1
GDSS	55.1	-9.9	70.8	-7.9	51.2	4	52.6	-5.1
COR	56.8	-7.9	68.4	-8.8	53.5	2	52.9	-4.7
JJA CRPSS	18.5	-6.1	30.6	-5.3	14.5	-0.3	17.8	-9.8
MSSS	19.9	-11.7	42	-11.9	22.2	-1.8	26	-13.8
GDSS	47.3	-8.5	70.4	-3.5	57.5	10.8	54.6	-4.3
COR	45.2	-10.7	75.2	-3.1	55.2	9.7	57.9	-4.1
DJF CRPSS	28.9	-2.5	27.2	4.1	31.1	3.1	29	6.1
MSSS	36.1	0.7	28.1	-1	42.1	0	36.4	5.6
GDSS	70.4	2.3	62.6	-5.2	65.2	-0.8	66.9	-1.4
COR	62.3	7.8	52.3	0	68.7	0.9	67	-0.6

E.2. Spatial distribution of skill scores for the EUR-22 ensembles

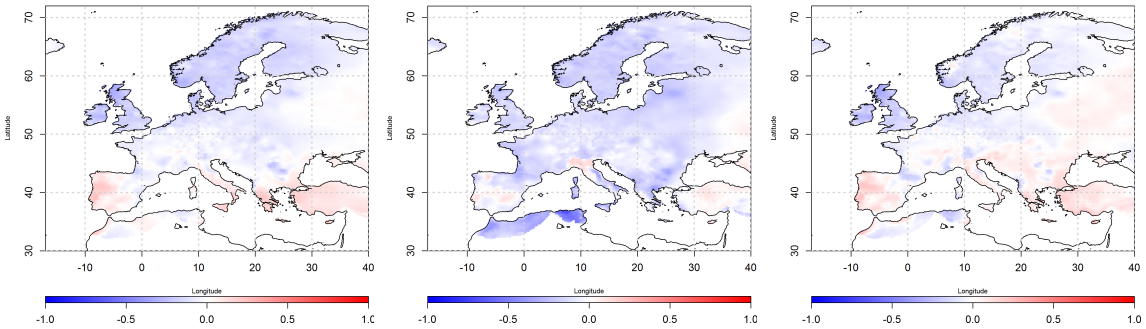


(a) MPI b1 JJA (b) CCLM b0 JJA (c) CCLM b1 JJA

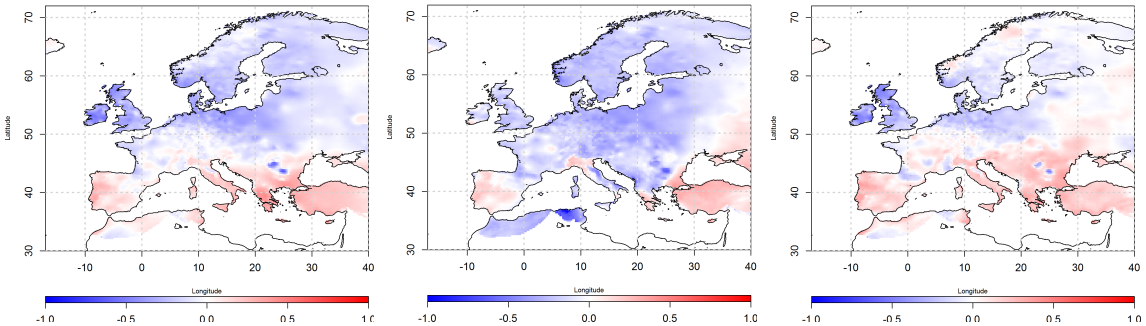


(d) MPI b1 DJF (e) CCLM b0 DJF (f) CCLM b1 DJF

Figure E.1.: The spatial ***accuracy skill of three EUR22 ensembles of temperature*** *in summer (JJA, upper panel), Winter (DJF, second panel) and the whole year (lower panel) is shown. The baseline1 ensemble of the decadally initilialized MPI-ESM-LR over Europe for 1961–2010 (left) can be compared to the decadally initialized CCLM EUR22-b0 (middle) and the CCLM EUR22-b1 (right). The accuracy skill is represented by the the MSSS which ranges from* $-\infty$ *to 1, with 1 indicating perfect skill and* $\leqslant 0$ *indicating no skill.*

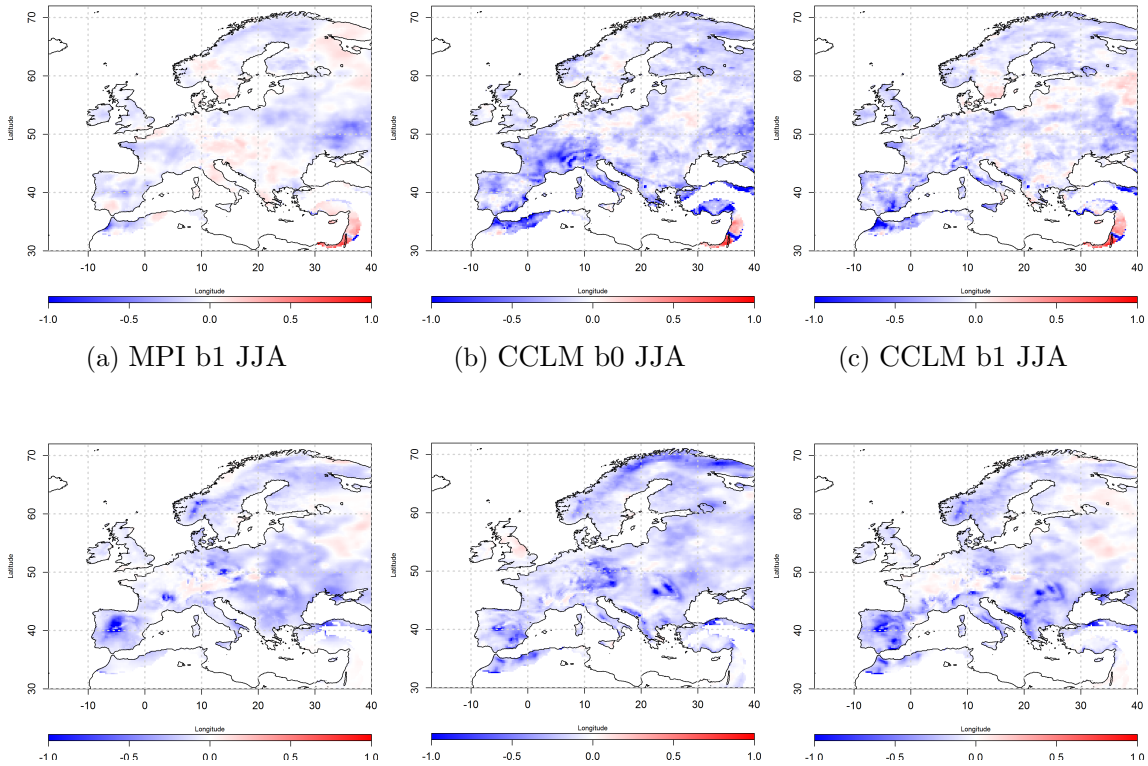


(g) MPI b1 YEAR (h) CCLM b0 YEAR (i) CCLM b1 YEAR



(j) MPI b1 YEAR 1-5 (k) CCLM b0 YEAR 1-5 (l) CCLM b1 YEAR 1-5

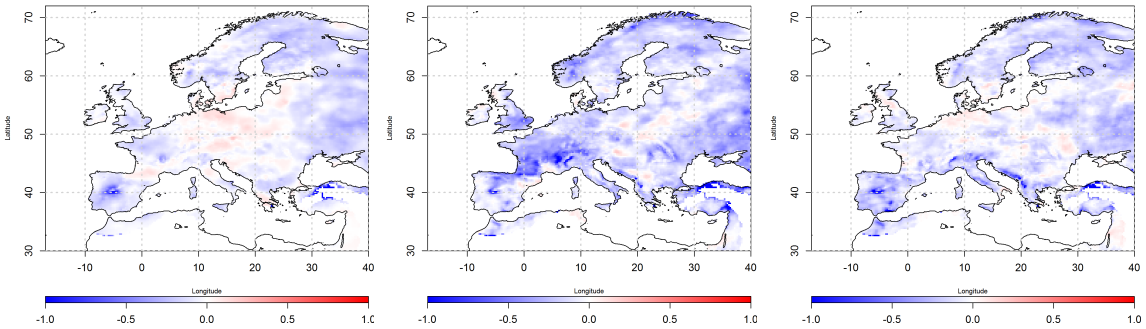
Figure E.1.: Continued.



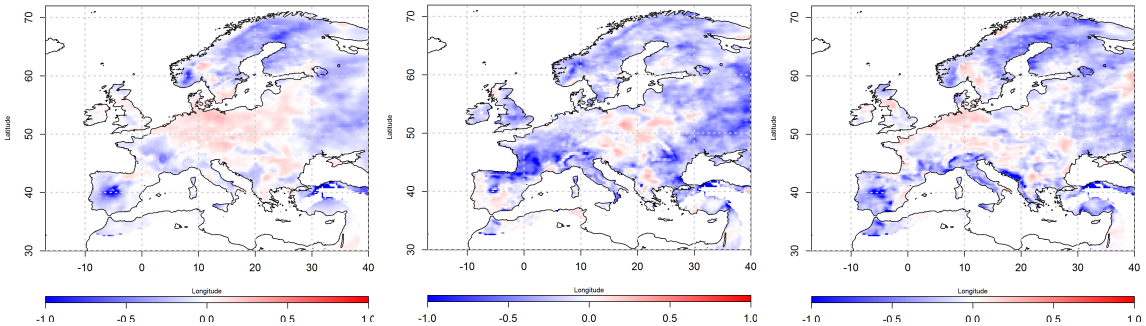
(a) MPI b1 JJA (b) CCLM b0 JJA (c) CCLM b1 JJA

(d) MPI b1 DJF (e) CCLM b0 DJF (f) CCLM b1 DJF

Figure E.2.: The ***spatial accuracy skill of three EUR22 ensembles of precipitation*** *in summer (JJA, upper panel), Winter (DJF, second panel) and the whole year (lower panel) is shown. The baseline1 ensemble of the decadally initilialized MPI-ESM-LR over Europe for 1961–2010 (left) can be compared to the decadally initialized CCLM EUR22-b0 (middle) and the CCLM EUR22-b1 (right). The accuracy skill is represented by the MSSS which ranges from $-\infty$ to 1, with 1 indicating perfect skill and $\leqslant 0$ indicating no skill.*

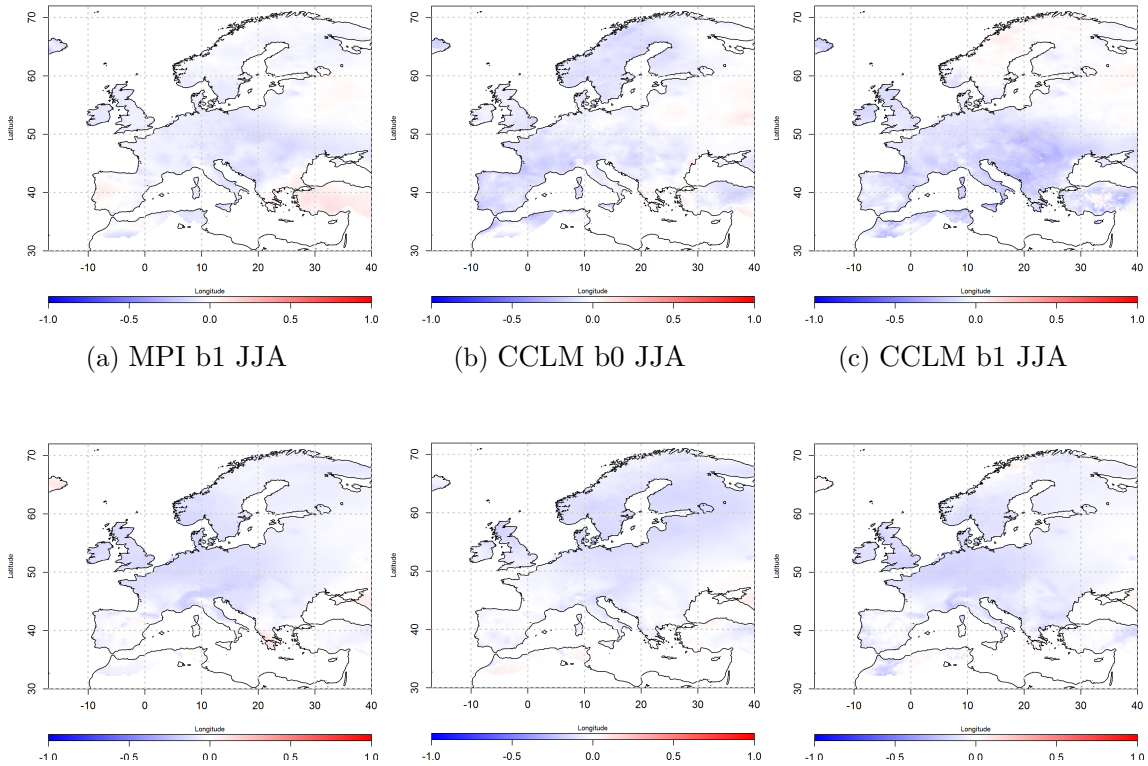


(g) MPI b1 YEAR (h) CCLM b0 YEAR (i) CCLM b1 YEAR



(j) MPI b1 YEAR 1-5 (k) CCLM b0 YEAR 1-5 (l) CCLM b1 YEAR 1-5

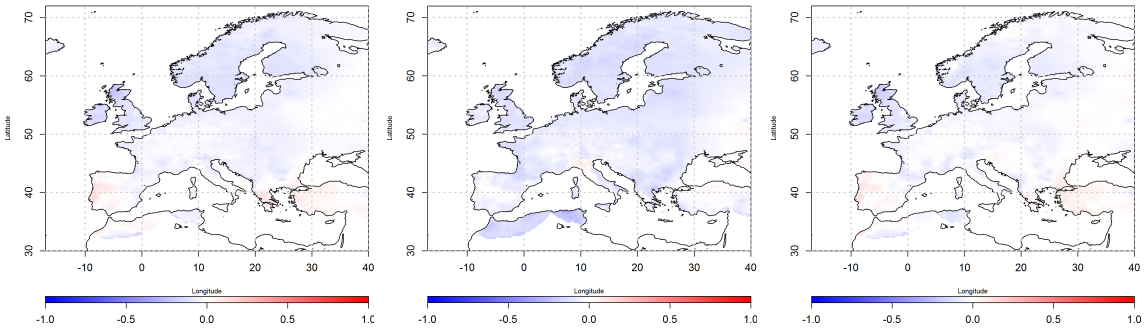
Figure E.2.: Continued.



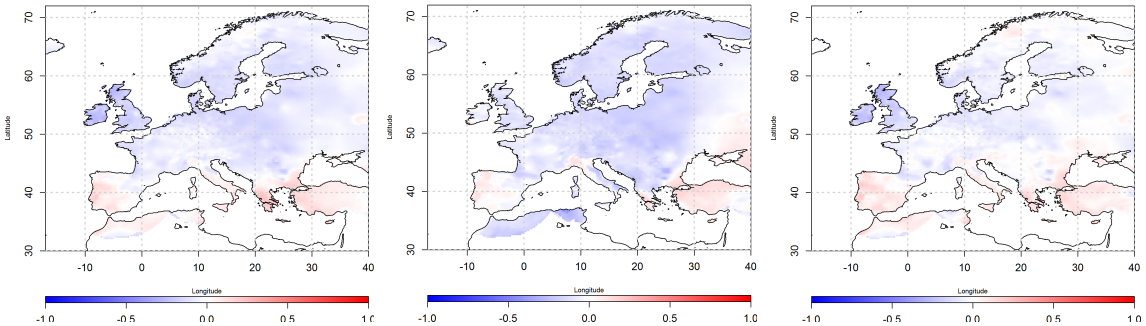
(a) MPI b1 JJA (b) CCLM b0 JJA (c) CCLM b1 JJA

(d) MPI b1 DJF (e) CCLM b0 DJF (f) CCLM b1 DJF

Figure E.3.: The spatial ***reliability skill of three EUR22 ensembles of temperature*** *in summer (JJA, upper panel), Winter (DJF, second panel) and the whole year (lower panel) is shown. The baseline1 ensemble of the decadally initilialized MPI-ESM-LR over Europe for 1961–2010 (left) can be compared to the decadally initialized CCLM EUR22-b0 (middle) and the CCLM EUR22-b1 (right). The reliability skill is represented by the the CRPSS which ranges from* $-\infty$ *to 1, with 1 indicating perfect skill and* $\leqslant 0$ *indicating no skill.*

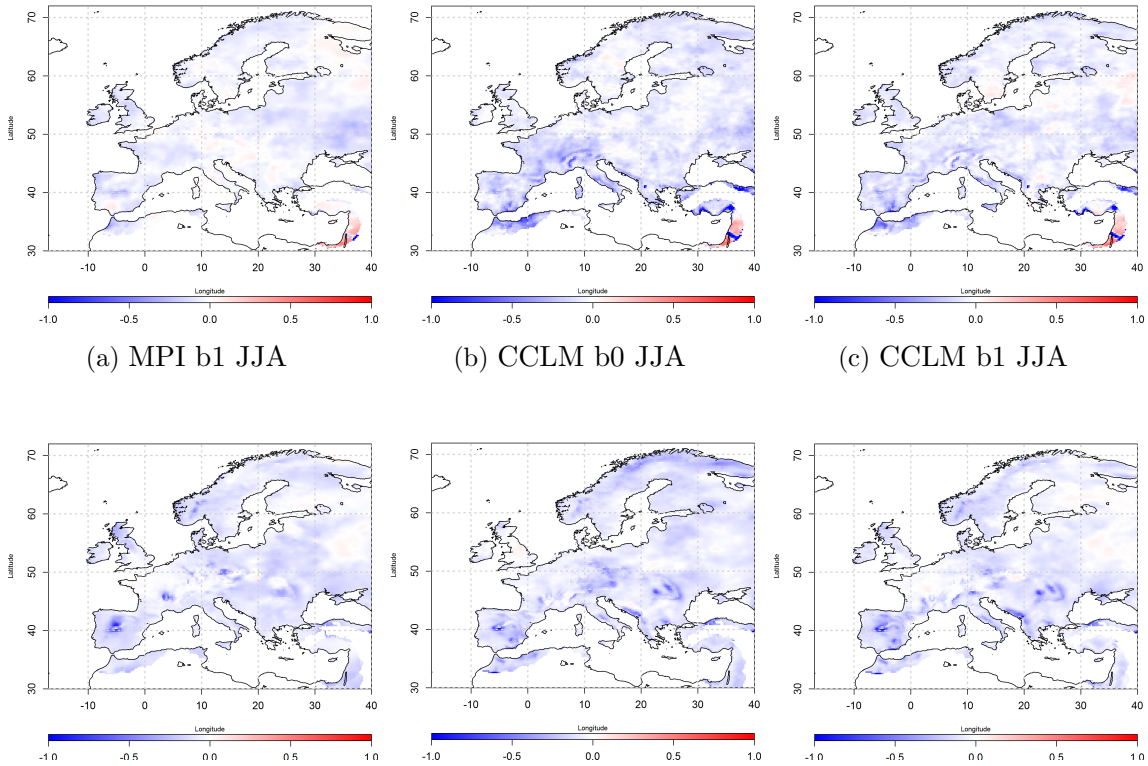


(g) MPI b1 YEAR (h) CCLM b0 YEAR (i) CCLM b1 YEAR



(j) MPI b1 YEAR 1-5 (k) CCLM b0 YEAR 1-5 (l) CCLM b1 YEAR 1-5

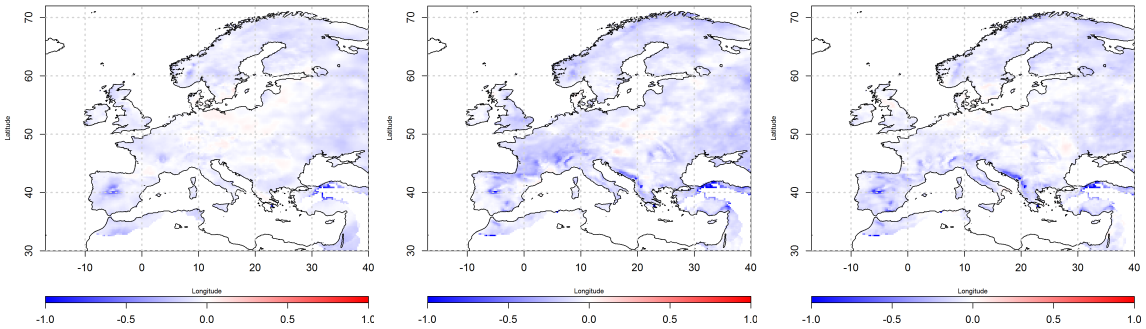
Figure E.3.: Continued.



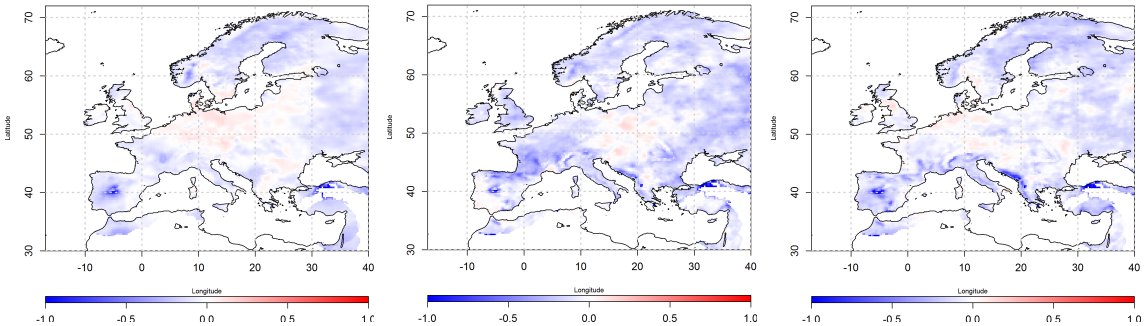
(a) MPI b1 JJA (b) CCLM b0 JJA (c) CCLM b1 JJA

(d) MPI b1 DJF (e) CCLM b0 DJF (f) CCLM b1 DJF

Figure E.4.: The ***spatial reliability skill of three EUR22 ensembles of precipitation*** *in summer (JJA, upper panel), Winter (DJF, second panel) and the whole year (lower panel) is shown. The baseline1 ensemble of the decadally initilialized MPI-ESM-LR over Europe for 1961–2010 (left) can be compared to the decadally initialized CCLM EUR22-b0 (middle) and the CCLM EUR22-b1 (right). The reliability skill is represented by the CRPSS which ranges from $-\infty$ to 1, with 1 indicating perfect skill and $\leqslant 0$ indicating no skill.*



(g) MPI b1 YEAR (h) CCLM b0 YEAR (i) CCLM b1 YEAR

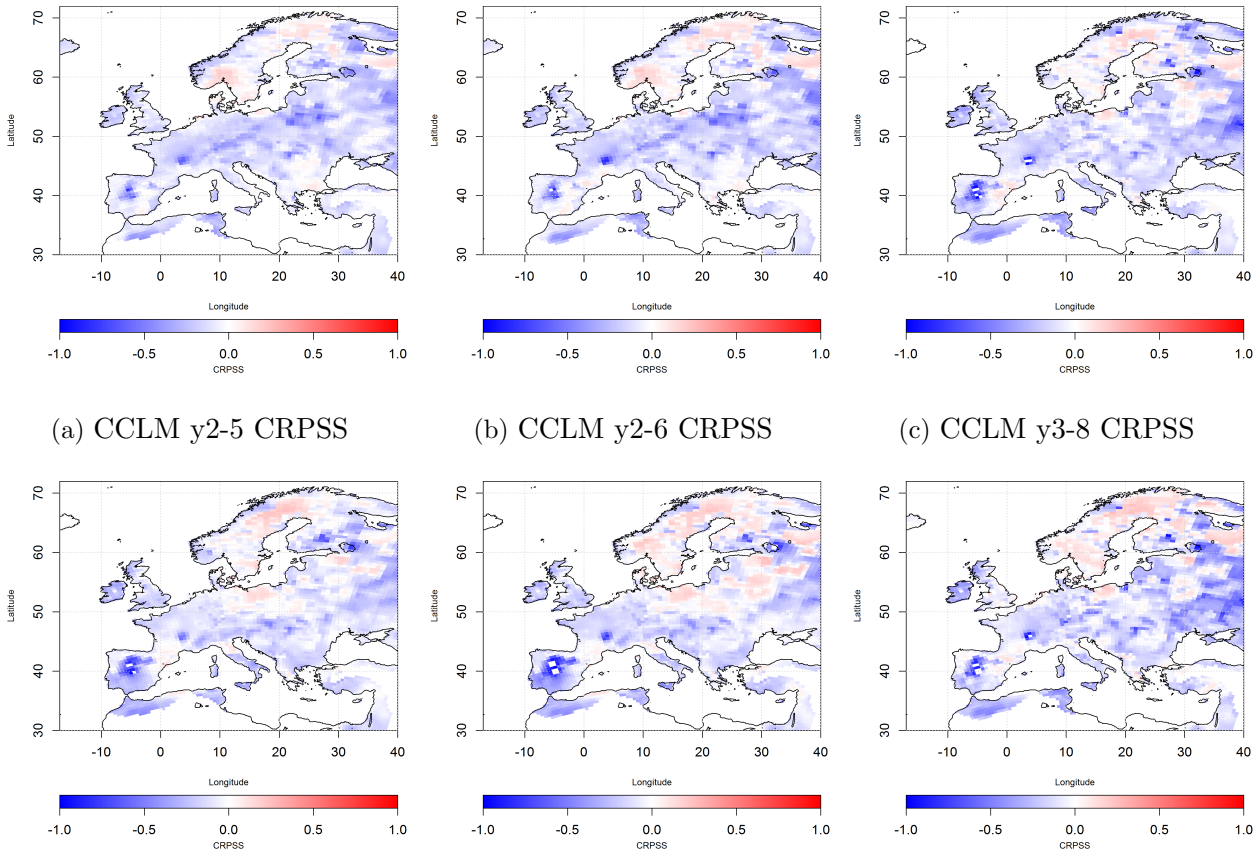


(j) MPI b1 YEAR 1-5 (k) CCLM b0 YEAR 1-5 (l) CCLM b1 YEAR 1-5

Figure E.4.: Continued.

F. Verification of the annually initialized 0.44 °ensemble

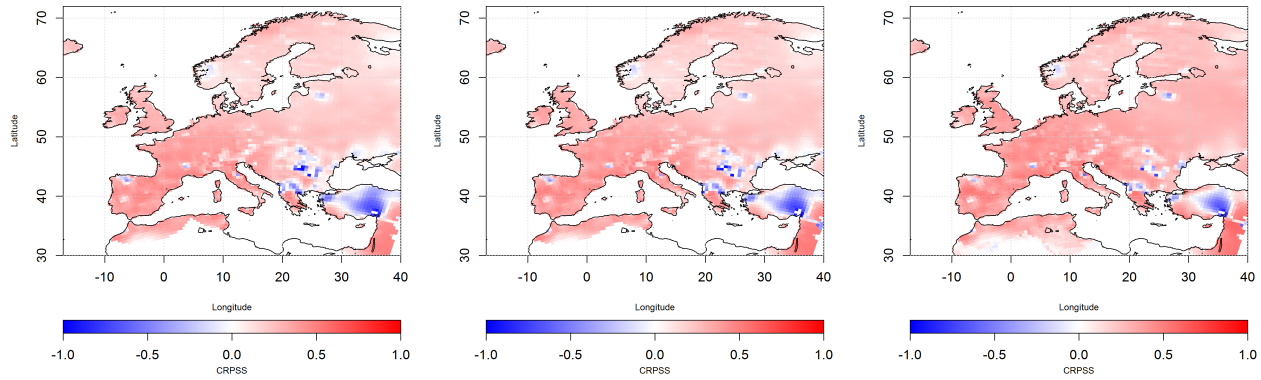
F.1. Spatial distribution of skill scores for all lead years



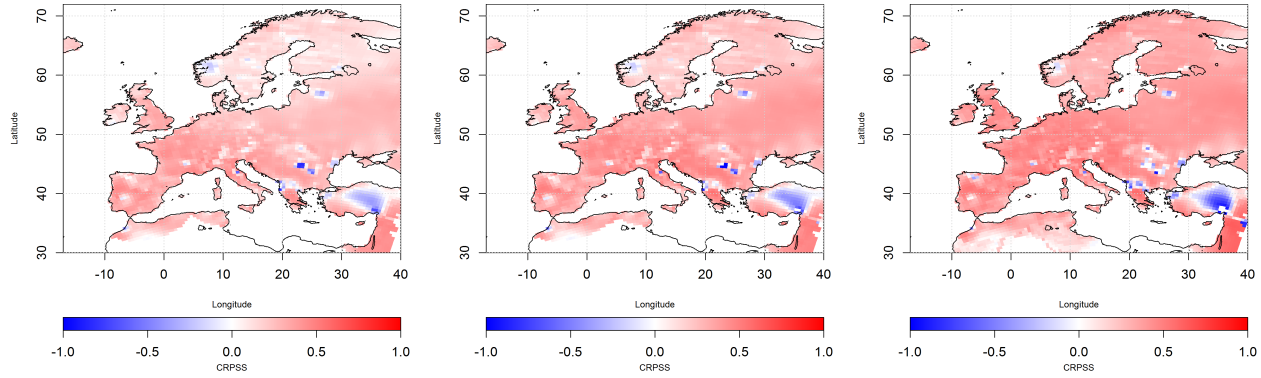
(a) CCLM y2-5 CRPSS (b) CCLM y2-6 CRPSS (c) CCLM y3-8 CRPSS

(d) CCLM y6-9 CRPSS (e) CCLM y6-10 CRPSS (f) CCLM y2-9 CRPSS

Figure F.1.: The ***CRPSS of b1-EUR44 CCLM hindcasts of annually initialized mean precipitation*** *over Europe from 1961 – 2010 is shown for different lead times averaging.*

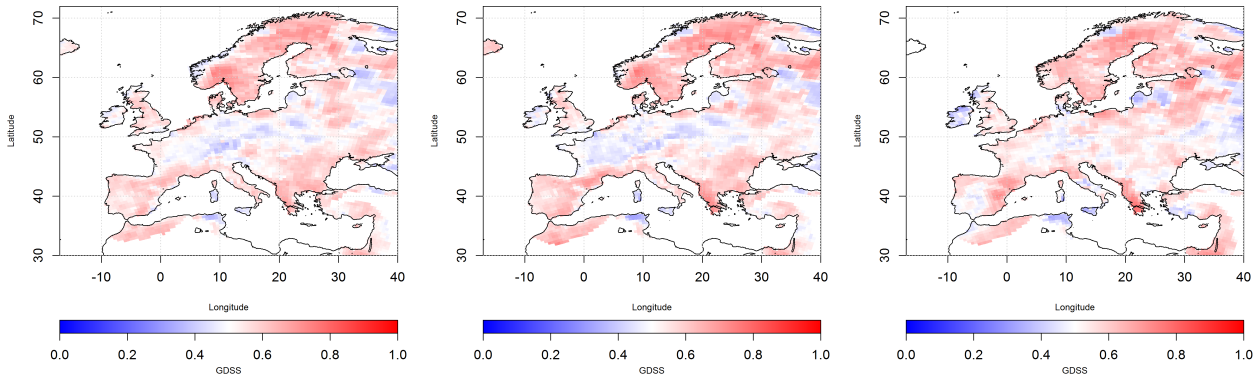


(a) CCLM y2-5 CRPSS (b) CCLM y2-6 CRPSS (c) CCLM y3-8 CRPSS

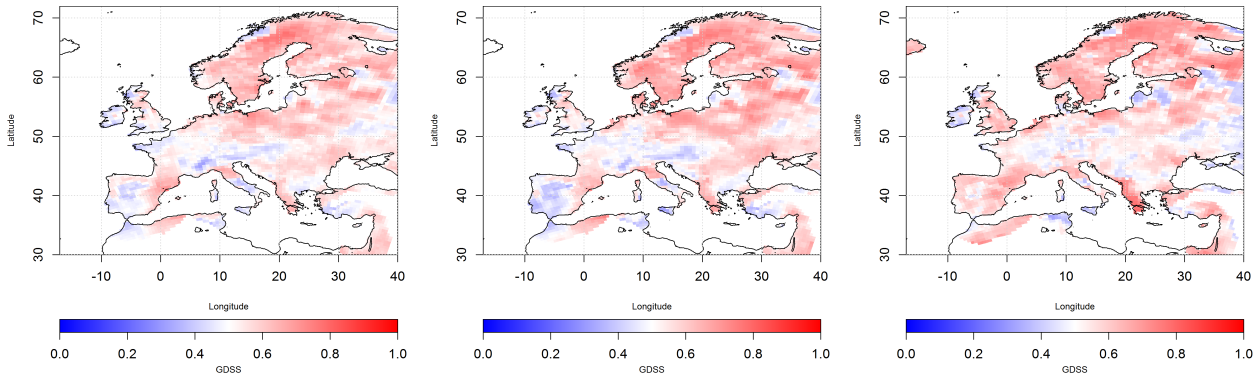


(d) CCLM y6-9 CRPSS (e) CCLM y6-10 CRPSS (f) CCLM y2-9 CRPSS

Figure F.2.: Reliability skill of the ***b1-EUR44 hindcasts of annually initialized mean temperature*** *over Europe from 1961 – 2010 is shown for different lead times averaging.*

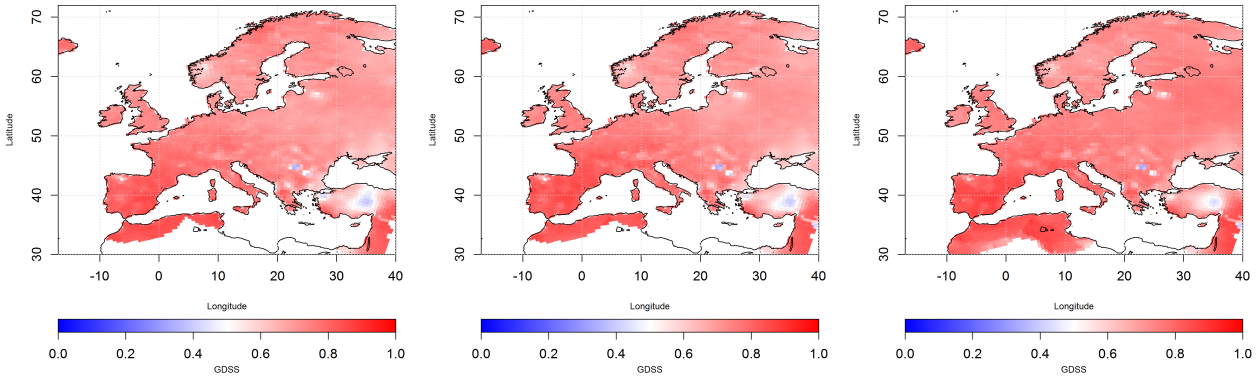


(a) CCLM y2-5 GDSS (b) CCLM y2-6 GDSS (c) CCLM y3-8 GDSS

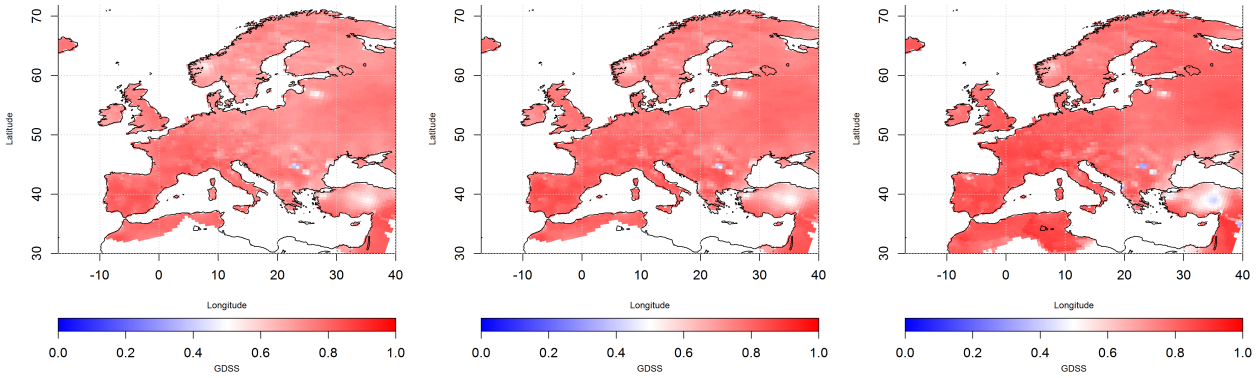


(d) CCLM y6-9 GDSS (e) CCLM y6-10 GDSS (f) CCLM y2-9 GDSS

Figure F.3.: The ***GDSS of b1-EUR44 CCLM hindcasts of annually initialized mean precipitation*** *over Europe from 1961 – 2010 is shown for different lead times averaging.*



(a) CCLM y2-5 GDSS (b) CCLM y2-6 GDSS (c) CCLM y3-8 GDSS



(d) CCLM y6-9 GDSS (e) CCLM y6-10 GDSS (f) CCLM y2-9 GDSS

Figure F.4.: Discrimination skill of the ***b1-EUR44 hindcasts of annually initialized mean temperature*** *over Europe from 1961 – 2010 is shown for different lead times averaging.*

G. An example of the full potential predictability results of the variable Summer Days (SU) for all 8 PRUDENCE regions

Here, a more complete work-up of the time series analysis of one climate extreme event indix, Summer Days (SU), is summarized for all PRUDENCE regions: At first the filter (in years) at optimises the Signal-to-Noise-Ratio is identified. The the second local maxima of the spectral density of both the original time series and the time series filtered with the aforementioned optimum is estimated. Then the mean relative entropy of the all possible 10-year forecast in comparison to climatology is estimated. in the same way, the mean significance level of the χ^2-Test is displayed next. At last are the correlation coefficients with several forcing data, like the AMO-index.

The regional variability of climate extremes becomes apparent. The filter with which the optimum Signal-to-Noise-Ratio can be achieved does vary with region (between 3 and 6 years). The except for the PRUDENCE region Scandinavia (SC) the local maxima of the spectral density is decreased or preserved with filtering, but always indicating a signal with a wave length up to 10 years. For the region Scandinavia (SC) a calculation of the relative entropy was not possible as as Summer Days occurred too rarely in the data.

```
Testing of Information
Data SU tx
--------------------------------
PRUDENCE AL

Local Maxima 1 of S/N Ratio (Filter) 3
Local Maxima 2 of spectral density (original) 12

Local Maxima 2 of spectral density (filtered) 4

Relative Entropy 1.7

Significance X2 Test 0.26
Causalities
```

AMO SOI NAO hur2 ENSO vul ghg solar land

0.343 -0.042 -0.004 0.037 0.111 -0.234 0.121 -0.114 0.127

PRUDENCE BI

Local Maxima 1 of S/N Ratio (Filter) 6

Local Maxima 2 of spectral density (original) 14

Local Maxima 2 of spectral density (filtered) 6

Relative Entropy 1.84

Significance X2 Test 0.27

Causalities

AMO SOI NAO hur2 ENSO vul ghg solar land

-0.326 -0.028 0.463 -0.23 -0.207 0.041 0.133 -0.262 -0.305

PRUDENCE EA

Local Maxima 1 of S/N Ratio (Filter) 4

Local Maxima 2 of spectral density (original) 6

Local Maxima 2 of spectral density (filtered) 6

Relative Entropy 1.83

Significance X2 Test 0.23

Causalities

AMO SOI NAO hur2 ENSO vul ghg solar land

0.367 0.348 -0.584 0.201 -0.132 0.004 -0.027 -0.406 0.407

PRUDENCE FR

Local Maxima 1 of S/N Ratio (Filter) 3

Local Maxima 2 of spectral density (original) 17

Local Maxima 2 of spectral density (filtered) 7

Relative Entropy 1.77
Significance X2 Test 0.25

Causalities
AMO SOI NAO hur2 ENSO vul ghg solar land
0.101 -0.09 0.31 0.163 0.045 -0.215 0.078 -0.094 -0.067

PRUDENCE IP

Local Maxima 1 of S/N Ratio (Filter) 6
Local Maxima 2 of spectral density (original) 10

Local Maxima 2 of spectral density (filtered) 10

Relative Entropy 1.79
Significance X2 Test 0.21

Causalities
AMO SOI NAO hur2 ENSO vul ghg solar land
0.625 -0.213 -0.235 0.377 0.354 -0.057 0.231 -0.03 -0.018

PRUDENCE MD

Local Maxima 1 of S/N Ratio (Filter) 4
Local Maxima 2 of spectral density (original) 8

Local Maxima 2 of spectral density (filtered) 7

Relative Entropy 1.79
Significance X2 Test 0.22

Causalities
AMO SOI NAO hur2 ENSO vul ghg solar land
0.35 0.012 -0.125 0.099 -0.046 -0.04 0.134 -0.044 0.018

PRUDENCE ME

Local Maxima 1 of S/N Ratio (Filter) 3

Local Maxima 2 of spectral density (original) 14

Local Maxima 2 of spectral density (filtered) 5

Relative Entropy 1.92

Significance X2 Test 0.3

Causalities

AMO SOI NAO hur2 ENSO vul ghg solar land

0.088 -0.132 0.107 0.052 0.141 -0.344 0.042 -0.098 0.14

PRUDENCE SC

Local Maxima 1 of S/N Ratio (Filter) 9

Local Maxima 2 of spectral density (original) 6

Local Maxima 2 of spectral density (filtered) 12

Eidesstattliche Erklärung zur selbständigen Fertigung der Dissertation

Ich erkläre hiermit, dass ich die Arbeit mit dem Titel

Regional decadal climate predictions for Europe: Feasibility and Skill

ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Regeln zur Sicherung guter wissenschaftlicher Praxis wurden beachtet (Amtliche Bekanntmachung des KIT Nr. 56, 27.11.2014).

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Karlsruhe, den May 10, 2016

_____

Marianne Uhlig