BMBF Research Programme on Decadal Climate Prediction (MiKlip)

Contribution to MODULE E: Validation and Ensembles

Climate Model Validation by confronting globally Essential Climate Variables from models with observations (ClimVal)

ClimVal Deliverable M7 "Assessment of the representation of selected ECVs in the MiKlip prototype prediction and evaluation system, including sea ice (DS2)"

Submitted by

Daniel Senftleben (Daniel.Senftleben@dlr.de) and
PD Dr. habil. Veronika Eyring (veronika.eyring@dlr.de)
Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)
Institute of Atmospheric Physics

16 December 2014

DLR FONA Decadal Climate Prediction BMBF MiKlip

Verification of Temperature and Sea Ice in the MiKlip Decadal Climate Predictions with the ESMValTool

Daniel Senftleben

Wessling 2014

Ludwig Maximilians University Munich
Faculty of Physics
Meteorological Institute Munich

Verification of Temperature and Sea Ice in the MiKlip Decadal Climate Predictions with the ESMValTool

Daniel Senftleben

Master Thesis

Supervisor: Priv.-Doz. Dr. habil. Veronika Eyring (DLR)

German Aerospace Center
Institute of Atmospheric Physics

August 2014

Ludwig Maximilians Universität München
Facultät für Physik
Meteorologisches Institut München

Verifikation von Temperatur und Meereis dekadischer MiKlip-Klimasimulationen mit dem ESMValTool

Daniel Senftleben

Masterarbeit

Gutachterin: Priv.-Doz. Dr. habil. Veronika Eyring (DLR)

Deutsches Zentrum für Luft- und Raumfahrt
Institut für Physik der Atmosphäre

August 2014

Post-Submission Changes
16 December 2014

Text

- **Chapter 2.2:** the reference to Righi et al. [2014] was added.
- **Chapter 3.2.2:** the paragraph on Eisenman et al. [2014] has been deleted.
- **Chapter 4.1:** a section " 4.1.5 - Statistical Significance Test" has been added.
- **Chapter 5:** statements on skill or differences in skill of statistically not significant magnitude were removed.
- **Chapter 5.1.1:** a paragraph on the agreement in statistically significant values between Pohlmann et al. [2013] and the ESMValTool was added.
- **Chapter 5.2:** a sentence on the exclusive assessments of statistically significant values was added.
- **Chapter 5.2.2:** the paragraph on a missing significance test has been deleted.
- **Chapter 5.3:** the statements in the paragraph on Eisenman et al. [2014] have been weakened.
- **Chapter 5.3.1:** a sentence about the mostly insignificant differences among the decadal prediction systems in the Antarctic has been added.
- **Chapter 5.3.2:** a sentence on the implementation of a significance test has been deleted. Additionally, a significant improvement from pr over long-term simulations was identified in the Ross Sea.
- **Chapter 6:** a paragraph on the implementation of a significance test has been deleted.

Figures

- **Figures 5.1 through 5.11:** crosses have been added denoting skill or differences in skill exceeding the 5-95% confidence level.

- **Figure 5.12:** now, only grid cells containing statistically significant skill are considered for global averages
- **Figures 5.13 through 5.15:** black dots have been added denoting skill or differences in skill exceeding the 5-95% confidence level.

Abstract

Decadal climate predictions, that aim at predicting the time horizon of the next 10-30 years, are a relatively new field of research. An open science topic is whether the initialization of the climate model simulations with observations of the slowly-varying components of the climate system results in more accurate near-term predictions compared to uninitialized long-term simulations. To address this science question, Goddard et al. [2013] introduced a verification system for decadal experiments that enables a quantitative assessment of the model performance from the decadal predictions compared to observations and to uninitialized long-term simulations.

The goal of this thesis is to assess the possible additional predictive skill for near-surface temperature and sea-ice concentrations in the decadal simulations of the Max Planck Institute Earth System Model (MPI-ESM) compared to the uninitialized long-term simulations. To allow this assessment, the verification framework from Goddard et al. (2013) is implemented into the Earth System Model Validation Tool (ESMValTool). The ESMValTool is a software tool developed by multiple institutions that aims at improving routine Earth system model (ESM) evaluation. For this work, in particular the anomaly correlation skill, reliability and accuracy of the simulations are evaluated and tested against each other, the model's uninitialized long-term simulations, and observations.

No further prediction skill in global mean near-surface temperature is found for decadal hindcasts (i.e., retrospective forecasts) in comparison to the long-term simulations, except for the initialization year 1. In the following years, the decadal hindcasts drift to their preferred biased model state resulting in a prediction skill that is similar to that of the long-term simulations. On regional scales however, certain areas such as southwest of the South American continent and the North Atlantic Ocean show a significantly higher predictive skill. This regionality also translates to sea ice.

Further studies are required that expand the proposed metrics and include different variables and additional climate models to provide a concluding answer to the question of whether the initialization of climate models can lead to a higher predictability of near-future climate change.

Contents

Chapter 1

Introduction

The ability to understand and project near-term future climate is of fundamental interest to society. Policy makers require comprehensive and detailed information about the near-term climate state to be able to react to changes in climate. In order to address these demands, work on decadal climate simulations has become a growing international effort over the recent years (e.g., Smith et al. [2007]; Meehl et al. [2009]; Solomon et al. [2011]).

These decadal predictions differ from the long-term climate simulations that are performed since the beginning of the Coupled Model Intercomparison Project (CMIP) in this regard that some slow-varying components of the Earth system, in particular the ocean, are initialized with observations [Meehl et al., 2009; Pohlmann et al., 2009; Goddard et al., 2012]. The same models as used for long-term climate projections are initialized with observational data at the starting time of the decadal simulation. With this initialization, the simulations start from a climate state that better matches the phase of the observed climate.

Decadal predictions have been included for the first time also in the experiment protocol of the 5th phase of CMIP (i.e., CMIP5) [Taylor et al., 2012]. The CMIP5 decadal model simulations are initialized in 1960 and then continuously every year until 2005 to provide a set of hindcast simulations to test the performance compared to observations and to long-term simulations. For each simulation, a minimum number of three ensembles is suggested.

A comparison of the decadal simulations to long-term simulations and to observations has been carried out both qualitatively (i.e., Bräu [2013]) and quantitatively (i.e., Pohlmann et al. [2009]; Müller et al. [2012]; Holland et al. [2013]), with the present work falling into the latter group. More specifically, this thesis is aimed at contributing to the answer of the question, whether the initialization of climate models can lead to more accurate, reliable and skillful predictions of the future climate compared to the uninitialized simulations.

To evaluate the performance, a verification system for decadal climate predictions that has been introduced by Goddard et al. [2013] has been implemented into the Earth System Model Validation Tool (ESMValTool). This verification system calculates multiple skill measures that each aim to quantify a different aspect of forecast quality. The ESMValTool is an open source package that allows for the evaluation and routine benchmarking of Earth system models (ESMs) and is available at `http://www.pa.op.dlr.de/ESMValTool/`. These metrics con-

sist of the anomaly correlation skill, the conditional bias and the mean squared skill score (MSSS), that evaluate the model's hindcast skill, reliability and accuracy, respectively, by testing decadal hindcast or long-term simulations against observations.

The goal of this work is to assess the predictive skill for near-surface air temperature and sea-ice concentrations in the Max Planck Institute Earth System Model (MPI-ESM) decadal prediction system, a coupled climate model embedded in the MiKlip (Mittelfristige Klimaprognosen) project. The MPI-ESM has contributed both long-term [Giorgetta et al., 2013] and decadal [Müller et al., 2012; Pohlmann et al., 2013] simulations to CMIP5. The evaluations for temperature data hereby extend the studies of Pohlmann et al. [2013], who examined the hindcast skill of different MPI-ESM decadal model versions, whereas the sea-ice assessments directly build on the studies of Hübner [2013] and Notz et al. [2013], as well as Bräu [2013], who qualitatively evaluated the representation of sea ice in the MPI-ESM long-term and decadal simulations, respectively.

Before the verification system is used for this assessment, the calculations produced with the verification system implemented into the ESMValTool are compared to previous studies. The MPI-ESM currently provides three different decadal prediction systems. Therefore in a second step, the strategy for the evaluation of both temperature and sea ice is to first find the system with the highest overall predictive skill. Finally, this system is then tested against the uninitialized long-term simulations. The evaluation is done for different lead times, i.e. temporal smoothing of different time ranges within the forecast range of each experiment [Goddard et al., 2013]. This method gives an indication of the dependence of predictive skill on the forecasts' proximity to the point of initialization.

The thesis is structured in the following way. Chapter 2 gives an overview of decadal climate simulations, their difference to long-term projections, initialization methods and drift issues, as well as of the structure and capabilities of the ESMValTool. In Chapter 3, the MPI-ESM and its simulations and the observational datasets used for the evaluation are described. A detailed presentation of the implemented verification system with both its required preprocessing steps and the metric calculations is given in Chapter 4. Results of the temperature and sea-ice assessments are presented in Chapter 5. Chapter 6 closes with a summary and an outlook.

Chapter 2

Scientific Background

2.1 Decadal Climate Predictions

Climate change does not only occur on century time scale, but also impacts on a shorter time range with possible changes of extreme events [IPCC, 2013] or hurricane activity [Goldenberg et al., 2001], for example. Adapting to these near-term changes represents a growing need among society, policy and decision makers, as their planning horizon lies in the magnitude of 10-30 years: the "decadal" time scale [Meehl et al., 2009]. Thus, it is important to closer investigate the changing climate on this time scale. Hereby, natural internal variability plays a major role. Long-term predictions of climate models are aimed at predicting changes of the climate system as a response to changes in natural and anthropogenic forcings [Smith et al., 2014]. Decadal simulations form a new area of research where the models are initialized with observations for the small varying components (such as the ocean) in order for them to start from the correct phase of the climate system's natural variability [Meehl et al., 2014; Smith et al., 2014].

2.1.1 Initialization

Climate predictions of the decadal time range fill the gap between seasonal-to-interannual predictions and long-term projections [Meehl et al., 2009]. They thereby form a combination of two time scales (Figure 2.1). Numerical weather forecasts that predict the atmospheric development of the following days or weeks are initial value problems and thus have to be initialized with observational data. Long-term climate projections on the other hand are uninitialized and free-running and simulate changes in climate under varying boundary conditions for several centuries [IPCC, 2013]. The time horizon of decadal climate predictions lies in between these two, and because of that, they depend on both the initial values and the boundary conditions [Pohlmann et al., 2009].

Variables that are being prescribed during the initialization process, mainly come from slow-evolving components of the climate system, such as the ocean [Pohlmann et al., 2009]. Thus,

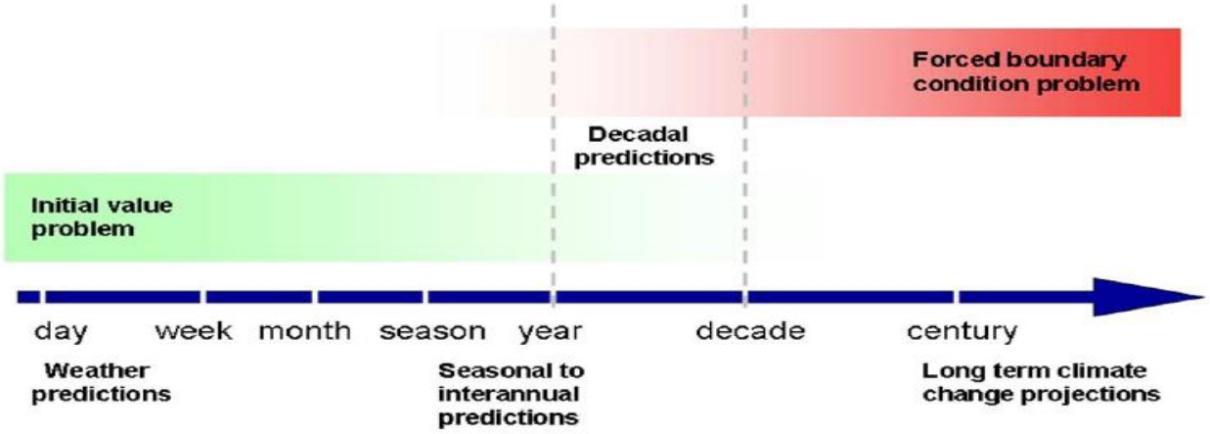

Figure 2.1: *Illustration of the dependence of decadal predictions on both initial values (similar to numerical weather predictions) and boundary conditions (similar to long-term climate projections). Figure from IPCC [2013].*

most models are initialized with observed ocean temperature and salinity fields, but also sea-ice and atmospheric variables may be initialized [Meehl et al., 2013].

For decadal climate predictions, the same models as for long-term simulations are used, and after the initialization, the models are free-running for 10 to 30 years. To assess the predictive skill of decadal experiments, retrospective forecasts, named "hindcasts", are made by integrating from different start dates in the past. The following section describes recommendations for the conduction of these hindcast experiments made by the CMIP5 protocol.

2.1.2 CMIP5 Protocol for Decadal Simulations

Realizations of decadal climate simulations are included in the 5th Phase of the Coupled Model Intercomparison Project (CMIP5) that defines and coordinates a set of climate model experiments aiming at the understanding of past and future climate changes [Taylor et al., 2012]. Adding to recommendations for long-term experiments, CMIP5 for the first time defines decadal prediction integrations (Figure 2.2). Decadal simulations with 10-year simulations initialized every 5 years from 1960 to 2005 and an additional set of hindcasts that are integrated for 30 years each and started in 1960, 1980 and 2005 form the core of the CMIP5 design. Tier 1 simulations are additional experiments and sensitivity simulations, like alternative initialization methods (see also Section 2.1.3), an increased ensemble size and higher initialization frequency.

The Max Planck Institute Earth System Model (MPI-ESM, see Chapter 3.1) has, among other models, performed the decadal CMIP5 simulations. Figure 2.3 depicts a part (1996-2005) of the experimental setup of the MPI-ESM-LR (low resolution) decadal system named baseline-0 (b0-LR) that closely follows the CMIP5 recommendations for the initialization frequency. The vertical axis shows the different decadal experiments with their respective years of initializa-

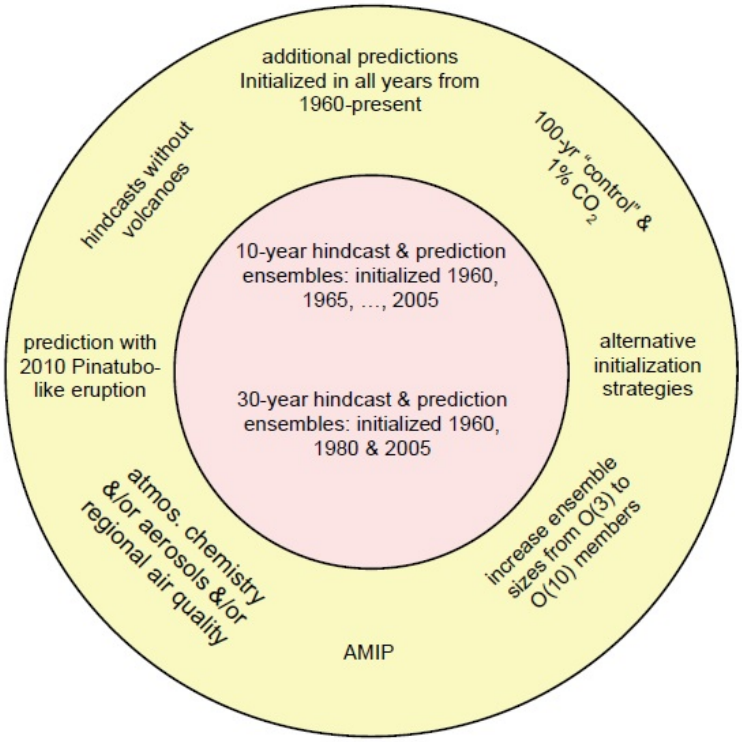

Figure 2.2: *Overview of the core (inner circle) and tier 1 (outer ring) recommendations for decadal prediction experiments by the CMIP5 protocol. Figure from Taylor et al. [2012].*

tion, and the horizontal axis the occurrence of the forecasts in time [Bräu, 2013]. The numbers indicate the number of ensemble members available for each experiment and forecast time. Thus, the b0-LR prediction system is initialized every year with at least 3 ensemble members running for 10 years and an additional 7 members that are initialized every 5 years (indicated by orange shading). For the experiment "decadal1980", the forecast range is longer than ten years, as 3 of the 10 ensemble members initialized in 1960, 1980 and 2005 are prolonged to 30 years (not shown).

2.1.3 Model Drift and Initialization Methods

In addition to differences in the experimental setups of the decadal predictions, initialization techniques differ from model to model. More generally, there are two initialization methods: full-field and anomaly initialization [Meehl et al., 2009; Goddard et al., 2013]. In the full-field method, the model state at the time of the initialization is substituted with the observed state, i.e. the estimated real state [Carrassi et al., 2014]. In contrast, the anomaly technique adds observed anomalies to the model climatology.

Because models are imperfect and run freely after the initialization, the model drifts away from the observed state towards its preferred model state. This problem is illustrated in Fig-

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
File name										
decadal1980	3	3	3	3	3	3	3	3	3	3
decadal1981										
decadal1982										
decadal1983										
decadal1984										
decadal1985										
decadal1986	3									
decadal1987	3	3								
decadal1988	3	3	3							
decadal1989	3	3	3	3						
decadal1990	10	10	10	10	10					
decadal1991	3	3	3	3	3	3				
decadal1992	3	3	3	3	3	3	3			
decadal1993	3	3	3	3	3	3	3	3		
decadal1994	3	3	3	3	3	3	3	3	3	
decadal1995	10	10	10	10	10	10	10	10	10	10
decadal1996		3	3	3	3	3	3	3	3	3
decadal1997			3	3	3	3	3	3	3	3
decadal1998				3	3	3	3	3	3	3
decadal1999					3	3	3	3	3	3
decadal2000						10	10	10	10	10
decadal2001							3	3	3	3
decadal2002								3	3	3
decadal2003									3	3
decadal2004										3
Total running	47	47	47	47	47	47	47	47	47	47

Figure 2.3: *Excerpt of the experimental setup of decadal simulations carried out by the MPI-ESM-LR baseline-0 prediction system, depicting the number of ensemble members running for each experiment and each prediction year between 1996 and 2005. The bottom row depicts the total number of ensemble members available for each year. Table from Bräu [2013].*

ure 2.4: each decadal hindcast simulation starts close to the observed state (black line) due to the initialization. It then gradually "forgets" the observational information, causing it to drift towards the equilibrium state of the uninitialized simulation. In an attempt to correct for this drift, the mean drift (i.e., mean bias) from all experiments is subtracted from each decadal hindcast. This leads to an under-correction of the experiments with starting years 1960-1975, and to an over-correction of the hindcasts starting later than 1985, due to the differing background trend between the model and the observations. Thus, drift corrections need to not only account for the time-independent mean bias, but also for the conditional bias, that includes the estimated model drift over time. This concept is further explained in Chapter 4.2.3.

The problem of the model drift can also be partially overcome by employing the anomaly initialization method, that aims to predict future anomalies by assimilating observed anomalies to the model climatology, keeping the initial model state closer to the observed state (e.g. Smith et al. [2007]). Due to a large diversity of models and experimental setups, it is dif-

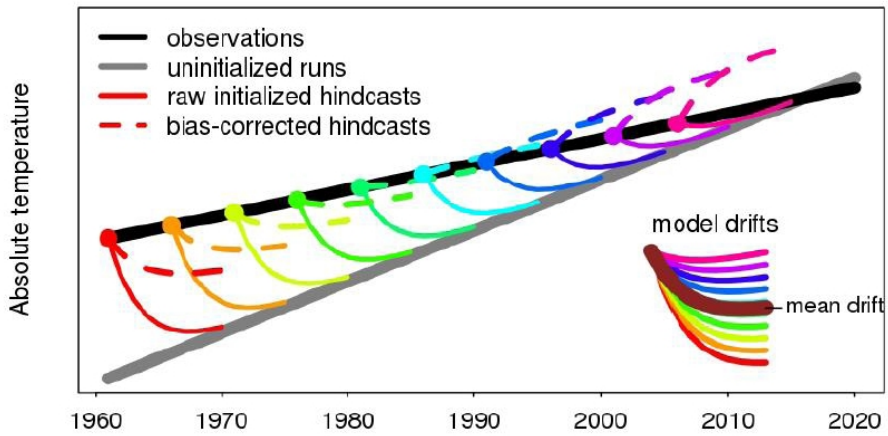

Figure 2.4: *Illustration of bias adjustment for decadal temperature hindcasts (colored solid lines), that drift from the initialized observed state (black solid line) to their imperfect uninitialized model state (solid grey line). This bias is accounted for by subtracting the mean drift (brown line in the inset) from each hindcast simulation, resulting in the bias-corrected hindcasts (colored dashed lines). Figure from Meehl et al. [2013].*

ficult to assess which initialization methods results in the most skillful predictions [Carrassi et al., 2014]. A comparison of the prediction skill between the two initialization techniques is made in this work for decadal model output of the MPI-ESM (see Chapters 5.2.1 and 5.3.1).

2.1.4 Long-term Projections Compared to Decadal Predictions

Long-term climate simulations are started from a control simulation in 1850 and are integrated throughout the historical period into the future [Pohlmann et al., 2009]. In contrast, decadal simulations each start from their individual point in time with prescribed observed variables. This ensures that the decadal simulations start from the correct phase of the natural variability. Figure 2.5 illustrates this effect for global mean temperature anomalies. The ensemble mean (green) of the initialized experiments (purple) starting in 1998 captures the observed (black) local temperature maximum at this point in time much better than the ensemble mean (red) of the uninitialized long-term simulations (yellow). After following the observations quite closely for about three years, after 2001 the decadal ensemble mean begins to drift away from the observed state and closer follows the long-term ensemble mean, not capturing the local minimum of the year 2004, for example.

Hence, for the examination of predictive skill, the proximity of the simulations to its initialization time has to be accounted for. This is done via the selection of different time samples of the forecasts, so-called "lead years". Goddard et al. [2013] recommend a set of four temporal smoothing scales: year 1, years 2-5, years 2-9 and years 6-9. For technical details about the lead-year selection and lead-time average calculation, please see Chapter 4.1.3.

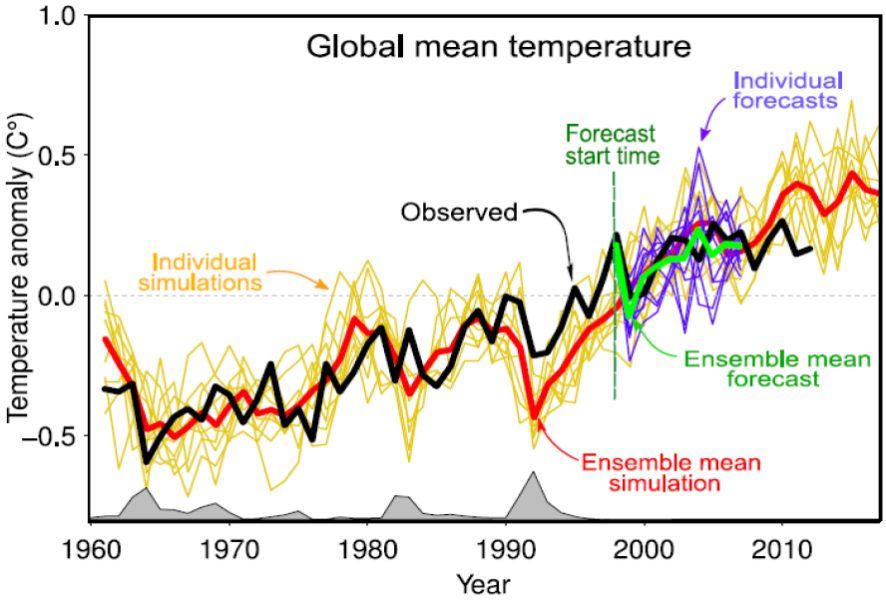

Figure 2.5: *Global mean surface-temperature anomalies calculated with the reference period 1986-2005. Depicted are observations and ensembles of both long-term and decadal simulations together with their respective ensemble means. Figure from [IPCC, 2013].*

Year 1 is closest to the initialization and therefore is expected to have the highest prediction skill. It is excluded from the three other lead-time ranges to reduce the imprint of the initial conditions. The years 2-5 are still within the interannual time scale and possibly dominated by year-to-year variability, whereas the years 2-9 might be able to capture the climate change signal. The years 6-9 are analyzed for an assessment of the lead-time dependence of skill by comparison to years 2-5. In this work, an assessment of lead years 2-6 is only included for global mean predictive skill, together with additional lead-time selections of 1- and 4-year means (see Chapter 5.2, Figure 5.12) which is also recommended by Pohlmann et al. [2013].

2.2 The Earth System Model Validation Tool (ESMValTool)

The Earth System Model Validation Tool (ESMValTool) is a diagnostic and performance metrics tool that facilitates the complex evaluation of ESMs (Righi et al. [2014], `http://www.pa.op.dlr.de/ESMValTool/`). It enables a routine benchmarking and evaluation of single or multiple models. It is designed as a community developed tool to which multiple developers from different institutions contribute. Embedded in a subversion-controlled repository, the implementation of extensions and additional analysis is straightforward.

The ESMValTool is an open source software package and as such only requires open source content from third parties. Its core routines are developed in Python, whereas diagnostic and plot routines are implemented in different languages like NCL, R, or matlab.

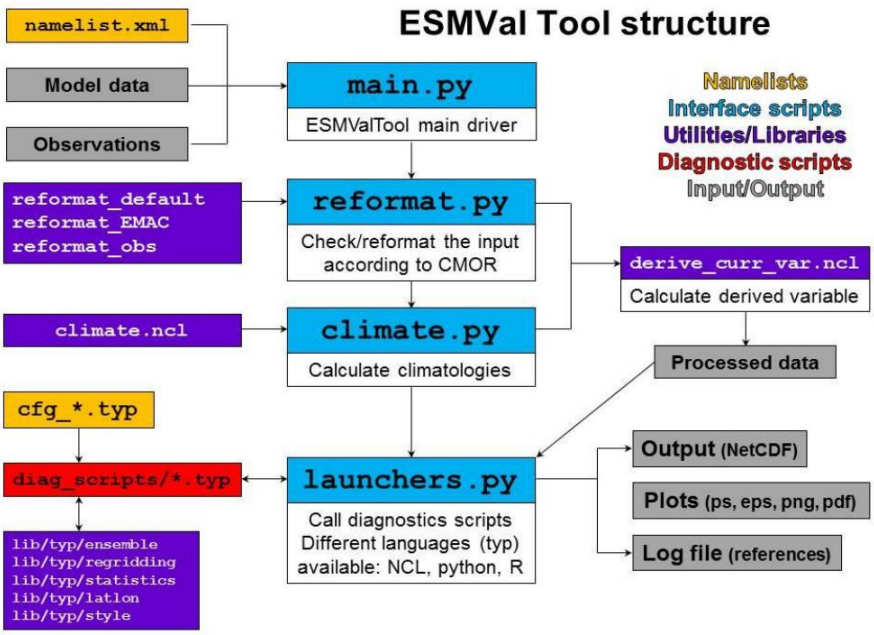

Figure 2.6: *Schematic illustration of the ESMValTool structure with core scripts of the python interface (blue), namelists (yellow), libraries (purple), diagnostic scripts (red) and input/output files (grey).*

Figure 2.6 gives a schematic overview of the ESMValTool structure. The tool is launched by calling the main python routine with a namelist file (top left corner). The namelist, along with general parameters controlling different aspects of the tool, contains lists of the following files: the model and observation data used for this assessment, diagnostic scripts that are to be called and configuration files that contain further specifications required for the diagnostics. The *main.py* then calls the other interface scripts (center column). First, the input data is reformatted to be compliant to the CMOR (Climate Model Output Rewriter) standard. Then, a climatology for each entry in the namelist is calculated. The launcher passes these to the

diagnostic script(s), where the major processing of the input data takes place. The diagnostic script then calls the plot function that produces the results as NetCDF output and/or a graphics file (e.g., ps, eps, png) and writes out a log file containing the references to the applied diagnostic and the tool.

Although the development phase has not yet been concluded, multiple diagnostics of different kinds are already available within the ESMValTool.

The verification framework presented in this thesis has been implemented into the ESMValTool (see Chapter 4) with the goal to make it reusable for other scientist and to run it routinely on CMIP6 decadal simulations. This effort was done in parallel to the development of the MiKlip Central Evaluation System's plugin "MurCSS". The MurCSS-Tool also implemented the Goddard et al. [2013] verification framework. However, in order to perform this thesis, a more flexible use of the verification system was needed. The implementation of the verification system into the ESMValTool has been done as part of this thesis to be able to apply it to other variables and phenomena in a flexible manner. This is especially important for sea-ice analyses that are performed exclusively for the month of minimal sea-ice extent. With the MurCSS-Tool, only annual means can be assessed.

Chapter 3

Models, Model Simulations and Observations for Verification

3.1 MPI-ESM-LR

The MiKlip (Mittelfristige Klimaprognosen) project funded by the Federal Ministry of Education and Research in Germany (BMBF) is aimed at providing climate predictions of the upcoming years to decades. Its Central Prediction System evaluated in this thesis is based on the Max Planck Institute Earth System Model (MPI-ESM). Section 3.1.1 gives a brief overview of the model. In Section 3.1.2 the three decadal prediction systems, and in Section 3.1.3 the long-term model simulations are described. All model systems and simulations that are outlined here are assessed in this thesis.

3.1.1 Brief Model Description

The MPI-ESM is the successor of the ECHAM5/MPIOM coupled climate model. As an Earth system model (ESM), it includes processes of the atmosphere, the land and the ocean that are simulated by different model components. An overview of these modules is given in Figure 3.1 [Giorgetta et al., 2013]. The general circulation model European Center-Hamburg Atmosphere Model version 6 (ECHAM6, [Stevens and Boucher, 2012]) forms the dynamical atmospheric core of the MPI-ESM, while the Max Planck Institute Ocean Model (MPIOM, [Jungclaus et al., 2013]) simulates the ocean. Both modules are coupled through an interface to the Jena Scheme for Biosphere Atmosphere Coupling in Hamburg (JSBACH, [Rieck et al., 2012]) and the Hamburg Ocean Carbon Cycle Model (HAMOCC, [Ilyina et al., 2012]), respectively.

The coupling is performed by the Ocean Atmosphere Sea Ice Soil (OASIS, [Valcke, 2013]). It enables the daily aggregation, interpolation and exchange of fluxes and state variables between the atmosphere (ECHAM6) and land surface (JSBACH), and between the ocean (MPIOM) and the marine biogeochemistry (HAMOCC). The exchanged variables and fluxes include energy, momentum, water and carbon dioxide (CO_2).

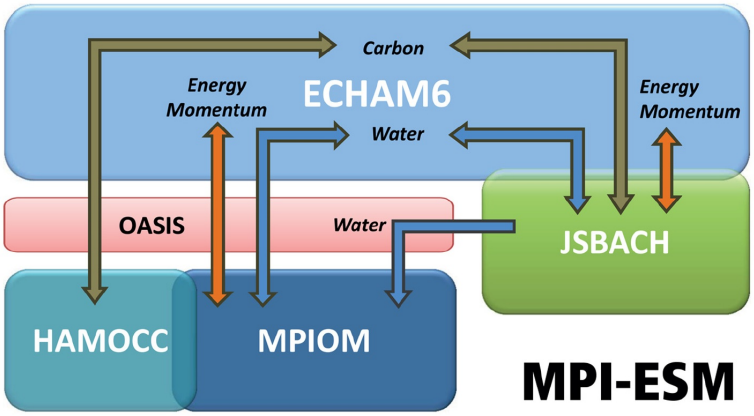

Figure 3.1: *Schematic illustration of the MPI-ESM model components ECHAM6, JSBACH, MPIOM, HAMOCC and OASIS. Arrows indicate the exchange of energy and momentum (orange), CO_2 (brown) and water (blue) between the model components. Figure from [Giorgetta et al., 2013].*

Among the most important improvements of the MPI-ESM towards its predecessor ECHAM5/MPIOM are an advanced treatment of radiative transfer, a better representation of the surface albedo, a better represented middle atmosphere and the inclusion of a coupled carbon cycle [`http://www.mpimet.mpg.de/en/science/models/mpi-esm.html`]. The latter enables the capture of carbon cycle feedbacks and, together with the biogeochemical module, forms the ESM component of the MPI-ESM.

For CMIP5, the MPI-ESM has been implemented with three different model configurations that differ in the components' spatial resolutions (MPI-ESM-LR/-MR) or setups of vegetation and orbit (MPI-ESM-P) [Giorgetta et al., 2013]. In this thesis, the low-resolution version MPI-ESM-LR is being evaluated. It is the most widely used configuration and has the most realizations and start dates for decadal predictions. The simulated atmosphere of the MPI-ESM-LR has a horizontal resolution of 1.9° and 47 vertical pressure levels extending to 0.1 hPa. The ocean is configured with 40 z-levels and a horizontal resolution of 1.5° (near equator) on a bipolar grid with poles at Greenland and the coast of the Weddell Sea.

3.1.2 Overview of the three available MiKlip decadal prediction systems and their simulations

Three different MiKlip decadal prediction systems exist from the MPI-ESM-LR: baseline-0 (b0-LR), baseline-1 (b1-LR) and prototype (pr).

As already discussed in Chapter 2.1.1, decadal climate simulations are initialized with observational data for the slow-varying components of the Earth system. Hereby, two main different

initialization techniques exist: full field and anomaly initialization. For details about initialization methods and general experimental setups of decadal predictions, please see Chapter 2.1. The three prediction systems mainly differ in their employed initialization technique and the ensemble size [Müller et al., 2014].

In the b0-LR prediction system, only the ocean temperature and salinity fields are initialized using the anomaly initialization method on the MPIOM with data from the National Centers for Environmental Prediction (NCEP). The b0-LR decadal predictions experiments are initialized every year from 1961 to 2012, each with three ensemble members integrating for ten years. Every 5 years, an additional set of 7 ensemble members is initialized, three of which are prolonged to 30 years.

The ocean in the b1-LR simulations is also initialized with the anomaly technique, but with ocean reanalysis data from the ocean reanalysis system 4 (ORAS4) from ECMWF [Balmaseda et al., 2013]. In addition, in b1-LR some atmospheric parameters (vorticity, divergence, temperature and air pressure) are initialized from the European Reanalysis data (ERA) [Dee et al., 2011], utilizing the full field technique. The yearly-initialized ensembles each consist of 10 members.

The prototype prediction system uses the same initialization of the atmosphere as b1-LR, but the ocean is initialized applying the full field method with data of temperature and salinity fields taken from ORA-S4 as well, and also from GECCO2 (German contribution to Estimating the Circulation and Climate of the Ocean) [Köhl, 2014]. For each of the yearly-initialized experiments, there are 15 ensemble members for ORA-S4-initialized simulations and an additional 15 members for those initialized with GECCO2. The start dates of both b1-LR and pr cover the time period 1961 to 2013.

For all assessments in this thesis, only the first ten years after initialization of the first three ensemble members (pr: only ORA-S4-initialized ensemble) of every decadal experiment were used.

3.1.3 Overview of the MPI-ESM long-term simulations

A number of long-term experiments have been performed with the MPI-ESM following the CMIP5 experiment protocol [Taylor et al., 2012]. In this thesis, the output of the so-called historical experiment is compared to the results from the decadal hindcast simulations to study whether the initialization improves predictive skill compared to the long-term simulations. The long-term uninitialized historical simulation starts in 1850 and goes until 2005 driven with prescribed natural and anthropogenic forcings [Giorgetta et al., 2013]. The historical simulations include the three ensemble members r1, r2, and r3 that start at the end of the years 1880, 1900 and 1920, respectively, from preindustrial control simulations that have constant forcing.

3.2 Observations

The verification framework applied in this thesis, uses observational data for the evaluation of model performance. For each of the two variables assessed in this work (near surface temperature and sea-ice concentrations), one observational dataset was selected. A brief description of the data is given in this section. Follow-up studies could additionally assess how observational uncertainty influences the results by using alternative datasets.

3.2.1 Temperature Observations

For near-surface air temperature, the observations provided by the HadCRUT3v temperature anomaly dataset are used, covering the time period 1850-2010 [http://www.cru.uea.ac.uk/cru/data/temperature/]. For land regions, monthly mean surface temperatures are derived from over 4800 stations (for recent years) all over the globe with varying density of stations depending on the population of the area [Jones et al., 2014]. For the ocean, sea-surface temperature is derived from merchant and naval vessels and (since the 1980s) from fixed and drifting buoys [Kennedy et al., 2011].

The data consist of monthly means on spatial resolution of 5° times 5° latitude/longitude grid boxes. They are provided as temperature anomalies with respect to the reference period 1961-1990 (period with best coverage). The reason for providing anomalies lies in the different elevations of land stations and the differing methods used for measuring the monthly average temperature. To account for biases due to this problem, every station has its own climatological average of 1961 to 1990 subtracted from each monthly mean. When this period is not completely covered by one station, neighboring stations or other sources of data have been used for compensation [Jones et al., 2014].

Because of different availability in space and time of data for each grid box, the gridded anomalies are calculated over a different numbers of monthly means. This has an impact on the variance of each grid box: averages calculated from fewer observations have a greater variance than those made from many. Thus, a variance adjustment of the data was made (indicated by an appended "v" in the name of the dataset, i.e. HadCRUT3v). The adjustment contains a re-processing of the anomalies with a statistical homogenization of the variances [http://icdc.zmaw.de/crutem_and_hadcrut.html].

3.2.2 Sea-Ice Observations

For the sea-ice evaluation made in this thesis, the Hadley Center's sea ice and sea surface temperature (HadISST) dataset is used [http://www.metoffice.gov.uk/hadobs/hadisst/].

The sea-ice data are derived from different sources including shipping, expeditions, digitized sea-ice charts, operational ice analysis from NCEP and passive microwave retrievals (mainly from the NASA). It is the longest available sea-ice dataset for both Arctic and Antarctic, ranging from 1871 to present. For the time before satellite measurements were available, the sea-ice data is predicted by thresholds of sea-surface temperature records obtained by ships and buoys [Rayner et al., 2003]. Like the surface temperature datasets in the previous section, the sea-ice data are monthly means, but the spatial resolution is larger (1° x 1°).

The combination of many data sources poses both the best strength and greatest weakness of the dataset. Long time ranges can be ensured by including different sources and thereby not being restricted to satellite information, but the continuity of the time series is difficult to assess, leading to breaks in the record that are of no climatic origin. [`https://climatedataguide.ucar.edu/climate-data/sea-ice-concentration-data-hadisst`].

Chapter 4

Verification Framework

4.1 Preprocessing

For the evaluation of decadal climate model output, a number of preprocessing steps are required, which are described in this section. The climate model data evaluated in this Thesis consist of uninitialized longterm simulations and initialized decadal simulations each covering a time period of ten years. Every decadal hindcast experiment contains a number of ensemble members that differ only in their initial conditions. Differences in the initialization method ("full field" or "anomaly" initialization) do not affect the calculations described below [ICPO, 2013].

The comparisons are made by testing the model systems against each other and against observations. For near-surface air temperature (tas), the HadCRUT3v dataset is used for the verification (`http://www.cru.uea.ac.uk/cru/data/temperature/`). This is the same dataset that was applied by Goddard et al. [2013]. It provides temperature anomalies with respect to the reference period 1958-2001. Note that both model and observational data are monthly means.

4.1.1 Ensemble Average

The first step is the calculation of ensemble averages. For each decadal experiment ensemble, the average D over all ensemble members E_i is computed. The resulting array is a function of start time τ, time t, latitude ϕ and longitude ψ:

$$D(\tau, t, \phi, \psi) = \frac{1}{N} \sum_{i}^{N} E_i(\tau, t, \phi, \psi), \tag{4.1}$$

where $i = 1, ..., N$ indicates the ensemble member. The start time τ is the first year of each experiment after its initialization. Thus, it also identifies the decadal experiment, as each of them starts at a specific point in time. For the uninitialized longterm simulations,

the ensemble average is only a function of time, latitude and longitude, as there is only one longterm ensemble.

4.1.2 Cross-Validation and Anomaly Calculation

The next step is the cross-validated anomaly calculation of the decadal ensemble means, which is applied in order to remove the mean bias from the data, following the suggestions given by the World Climate Research Programme Report [ICPO, 2013]. The cross-validation is calculated on a lead-month level by subtracting from every lead month t_l (with $l = 1, ..., 120$) of a given hindcast experiment D_{τ_j} $(j = 1, ..., M)$ the climatological mean of the same lead months from *all other* experiments τ_k $(k = 1, ..., M)$:

$$D_{cv}(\tau_j, t_l, \phi, \psi) = D(\tau_j, t_l, \phi, \psi) - \frac{1}{M-1} \sum_{k \neq j} D(\tau_k, t_l, \phi, \psi). \tag{4.2}$$

Consider, for example, a set of 20 decadal hindcast experiments with yearly initialization, with starting year between 1960 and 1979 (Figure 4.1). The cross-validation is applied to month 1 (first occuring January) of the experiment starting in 1960 by subtracting the average of all other months 1 (196101 from the experiment starting in 1961, 196201 from the experiment starting in 1962 and so on until 197901 from the 1979 experiment). The same is done to month 1 of experiment 1961, from which the average consisting of months 1 from the experiments 1960 and 1962-1979 is subtracted, and so on until 197901 from experiment 1979 that has the average of months 1 of experiments 1960-1978 removed. This procedure is then applied to all other lead months of all experiments, until the final December (lead month 120) of the last experiment (198812 of 1979). In this way, a different average is subtracted from every month in every decadal ensemble mean.

The longterm simulations L and observations O are handled differently in this step. From every time step t_j, the climatological average of the complete time range is subtracted:

$$L_{cv}(t_j, \phi, \psi) = L(t_j, \phi, \psi) - \frac{1}{N} \sum_{j}^{N} L(t_j, \phi, \psi) \tag{4.3}$$

and

$$O_{cv}(t_j, \phi, \psi) = O(t_j, \phi, \psi) - \frac{1}{N} \sum_{j}^{N} O(t_j, \phi, \psi). \tag{4.4}$$

Now, decadal, longterm and observational data can be regarded as anomalies. The anomaly-calculation, performed in a cross-validated manner for decadal hindcast data, is equivalent to the removal of mean bias [Goddard et al., 2013].

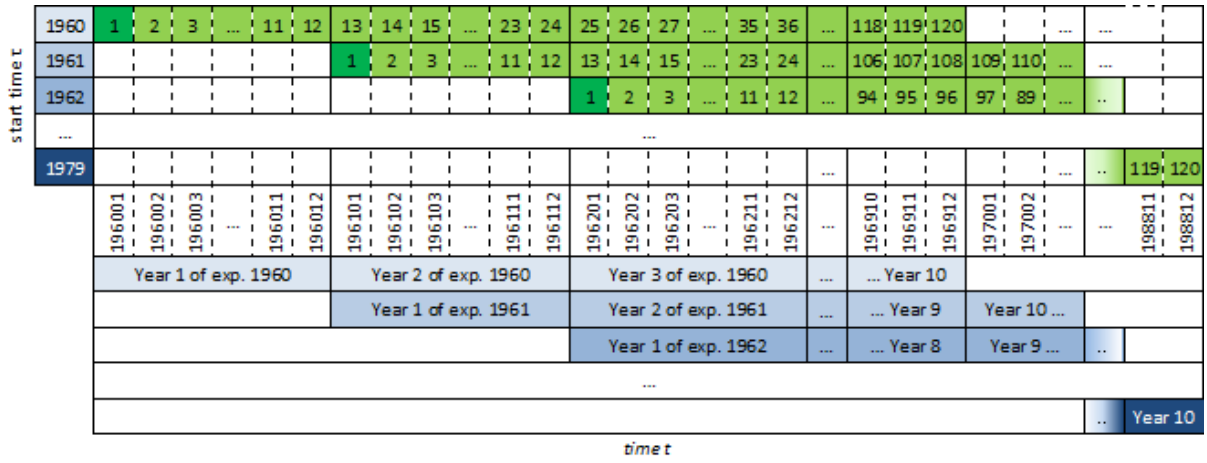

Figure 4.1: *Example table of a set of 20 decadal hindcast experiments with yearly initialization, with starting year between 1960 and 1979. Green colors indicate the lead months of each experiment defined by its respective start year (vertical coordinate), whereby the first occuring January (month 1) of each experiment is highlighted in dark green. The abscissa shows the time coordinate of the data, as year-month.*

4.1.3 Lead-time Selection

As discussed in Chapter 2.1, different time ranges of the decadal experiments are considered for the evaluation of the prediction skill. The selected lead years y_1 to y_2 are averaged out for each decadal (cross-validated) hindcast experiment D_{cv}. This average is then stored in another array $D_{lt}(\tau, \phi, \psi)$ at the position of the start time τ_j of that particular hindcast:

$$D_{lt}(\tau_j, \phi, \psi) = \frac{1}{y_2 - y_1 + 1} \sum_{m=y_1}^{y_2} D_{cv}(\tau_j, t_m, \phi, \psi) \tag{4.5}$$

This procedure is depicted in Figure 4.2. Dark green indicates the lead time selection of years 2-5. For every hindcast experiment defined by its start time (1960,1961,...,1979), the dark green fields are averaged and stored at the start time of the respective experiment. For example, the value stored at the position $\tau = 1960$ is the average of all months of the years 1961 to 1965 (i.e., month 13 to 60) of the decadal experiment started in 1960. The value at $\tau = 1961$ is the average of 1962-1966 of the experiment starting in 1961, and so on until $\tau = 1979$ with the average of 1980-1984 of experiment 1979.

The array L_{cv} (O_{cv}) containing the longterm simulation (observation) anomalies is sampled accordingly, so that its value at a given τ_j is the average of the same respective years in L_{cv} (O_{cv}) as the one at $D_{cv}(\tau_j)$:

$$L_{lt}(\tau_j, \phi, \psi) = \frac{1}{y_2 - y_1 + 1} \sum_{m=y_1}^{y_2} L_{cv}(t_m, \phi, \psi) \tag{4.6}$$

and

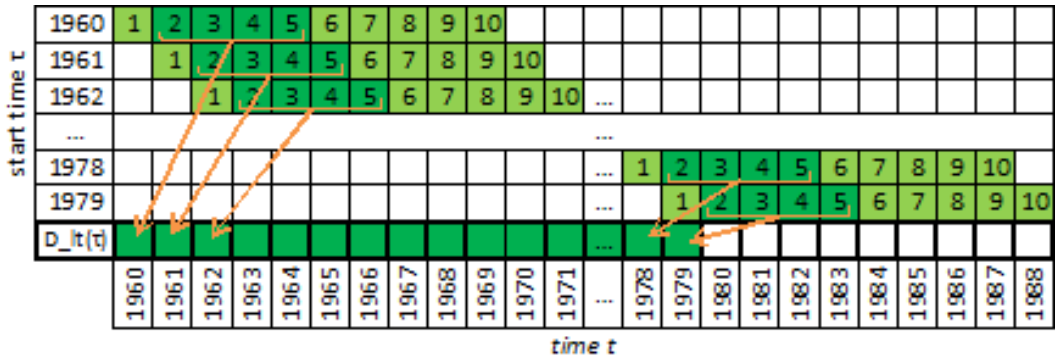

Figure 4.2: *Example table of 20 decadal hindcast experiments with yearly initialization, with starting year between 1960 and 1979. Green colors indicate the location of the lead years of each experiment (vertical coordinate) in time (abscissa). Dark green shading denotes the lead time selection of years 2-5 in every experiment. The bottom row depicts the array D_{lt}, with orange arrows indicating the position τ the lead time averages of each experiment are stored to.*

$$O_{lt}(\tau_j, \phi, \psi) = \frac{1}{y_2 - y_1 + 1} \sum_{m=y_1}^{y_2} O_{cv}(t_m, \phi, \psi). \tag{4.7}$$

The equal treatment of the data ensures a consistent comparison of each decadal hindcast against the corresponding time range of observations and longterm simulations. Note that for long enough lead time ranges and frequent enough initializations (just like in the example shown in Figure 4.2), the lead time selections may overlap, so that some years of longterm and observational data may be included in more than one lead time average.

4.1.4 Regridding

The arrays $D_{lt}(\tau, \phi, \psi)$, $L_{lt}(\tau, \phi, \psi)$ and $O_{lt}(\tau, \phi, \psi)$ are tested against each other for every latitude ϕ and longitude ψ. As the metric calculations are at the grid-point level, which result in a two-dimensional field (latitude, longitude), the model and observational data have to be regridded to the same spatial resolution. The position of the remapping procedure within the preprocessing steps is somewhat arbitrary, but in order not to lose any information from interpolating to a coarser grid before performing the cross-validation and because remapping is quite expensive in terms of computing time (especially for data on irregular grids), it is desirable to perform this operation at the end of the preprocessing when many time steps and ensemble members were averaged out.

Following the example of Goddard et al. [2013], the grid of the observations is used as target grid for the remapping of the model data. The ESMValTool uses bilinear interpolation for remapping a regular rectilinear grid to another (`https://www.ncl.ucar.edu/Document/`

Functions/Built-in/linint2.shtml). Other regridding methods (like area conservation) were tested, leading to the same results (not shown).

4.1.5 Statistical Significance Test

To evaluate the statistical significance of scores and of differences between scores, a non-parametric block-bootstrap algorithm [Wilks, 2011; Goddard et al., 2013; Eade et al., 2014] has been implemented in the ESMValTool. Hereby, the original data is being resampled in the following way: a number of hindcasts is randomly drawn with repitition from the pool of the ensemble averages until the exact same experiment size as before has been reached. This resampled experiment is very likely to be different from the original one, as some of the hindcasts may be included multiple times. To account for temporal auto-correlation, the resampling is done for blocks of 2-5 consecutive hindcasts, the number of which depending on the experiment size and the initialization frequency.

The verification metrics are then calculated for this newly generated experiment, including the preprocessing steps from above. This process is repeated a given number of times (500 for this thesis). In this way, a distribution function of the metrics is obtained for each grid cell and used for the significance test. The p value for the test is represented by the fraction of values that have a different sign than the original value. If p is smaller than or equal to the selected significance level α (5% in this thesis), the score of this grid cell is considered significant for the $(1-\alpha)*100\%$ confidence level. In other words, if for a given grid cell at least 95% of the resampled values have the same sign as the original value, the latter is considered significant.

4.2 Metrics

For the verification system, a number of deterministic metrics were chosen to evaluate the quality of prediction experiments. More specifically, they help to iterate the question whether the initialization of climate models leads to a more accurate prediction of the climate. The metric calculations are performed at the grid-cell level, therefore in the following equations the dependence on the spatial coordinates is omitted for simplicity.

4.2.1 Mean Squared Skill Score

The mean squared skill score (MSSS) measures the accuracy of a test prediction against a reference prediction, such as, for example, uninitialized simulations, or observations. The MSSS is defined via the mean squared error (MSE) between the predictions or hindcasts H_j and the observations O_j Goddard et al. [2013]:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(H_j - O_j)^2, \tag{4.8}$$

where the index $j = 1, ..., n$ represents the time steps of the data. The MSE here includes only the error variance, not the bias error component, since H_j and O_j are given as anomalies for which the mean bias has been removed (see Section 4.1.2).

The MSSS of a test prediction H against the climatological mean of the observations ($\bar{O} = \frac{1}{n}\sum_{j=1}^{n} O_j$) is then defined as the MSE of H over the one of $\bar{O}$, subtracted from 1:

$$MSSS(H, \bar{O}, O) = 1 - \frac{MSE_H}{MSE_{\bar{O}}}, \tag{4.9}$$

making the MSSS a function of the test prediction, the reference, and the observations. The MSSS can be also defined using the uninitialized predictions (P) instead of the observations as reference:

$$MSSS(H, P, O) = 1 - \frac{MSE_H}{MSE_P}. \tag{4.10}$$

A perfect MSSS has the value 1, corresponding to a the mean squared error of the test predictions of value 0. Accordingly, a positive (negative) MSSS indicates an improvement (worsening) in the accuracy of the test predictions over the reference predictions. Note that the MSSS is not symmetric about zero, so a positive MSSS does not imply the same increase in accuracy as the same absolute value of a negative MSSS would imply its decrease.

4.2.2 Anomaly Correlation Coefficient

The MSSS is decomposed with the Murphy-Epstein decomposition to interpret its components: the anomaly correlation coefficient and the conditional bias [Murphy, 1988].

Applying the decomposition to the MSSS for the hindcasts H against the observation climatology $\bar{O}$ results in

$$MSSS(H, \bar{O}, O) = r_{HO}^2 - \left[r_{HO} - \frac{\sigma_H}{\sigma_O}\right]^2, \tag{4.11}$$

where r_{HO} is the correlation coefficient between the hindcasts and the observations, and σ_H and σ_O are the standard deviations of the hindcasts and the observations, respectively:

$$\sigma_H = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(H_j - \bar{H})^2} \qquad \sigma_O = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(O_j - \bar{O})^2} \tag{4.12}$$

The first term of equation 4.11 is the square of the anomaly correlation coefficient that will be further discussed here, whereas the square root of the second term, the conditional bias, is subject of the next section.

The correlation coefficient r_{xy} measures the linear relationship between two datasets x and y. It is a dimensionless quantity ranging between -1 and 1. The correlation coefficient is used in this thesis to evaluate the potential skill of model simulations [Goddard et al., 2013]. Also known as *Pearson product-moment correlation coefficient*, this metric is defined as [Wilks, 2011]:

$$r_{xy} = \frac{1}{n-1}\frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sigma_x \sigma_y}, \tag{4.13}$$

with $\bar{x}$ ($\bar{y}$) being the mean of x_j (y_j):

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j. \tag{4.14}$$

In this thesis, x represents the model data to be tested and y the observations. For the test of decadal hindcast simulations H against a reference model R, the difference between the two correlation coefficients ($r_{HO} - r_{RO}$) is calculated as an estimate for the gain in correlation of H over R.

4.2.3 Conditional Bias

The second term of equation 4.11 is the square of the conditional bias cb representing the reliability of the hindcast/reference prediction and the observations:

$$cb_{HO} = r_{HO} - \frac{s_H}{s_O} \tag{4.15}$$

To evaluate a possible reduction in conditional bias from the hindcasts H compared to the reference predictions R, the absolute values of the conditional biases of both predictions are subtracted from each other ($|cb_{HO}| - |cb_{RO}|$).

For a better understanding of the difference between the conditional and the mean bias, two simplified examples are illustrated in Figure 4.3 [Goddard et al., 2013]. In both panels, a linear temperature trend is considered for the observations (black line). The hindcast data (green

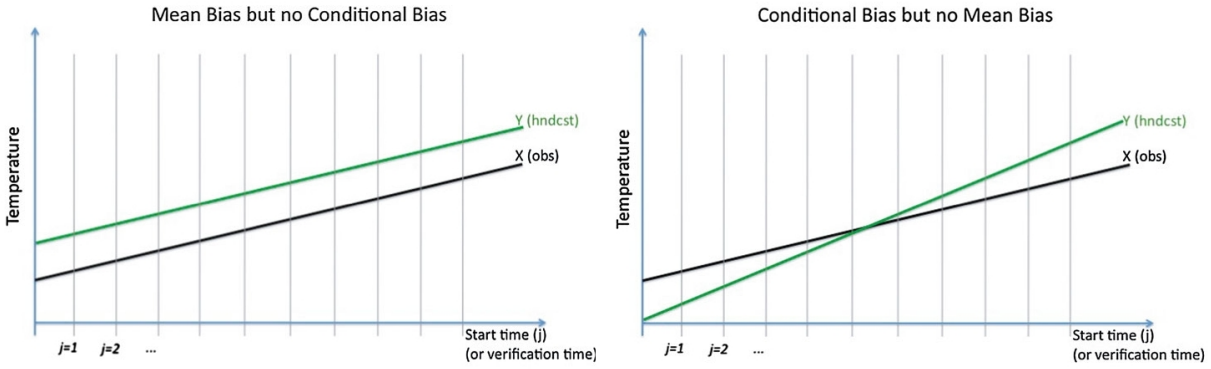

Figure 4.3: *Illustration of the difference between mean and conditional bias of hindcasts (green lines) and observations (black). Vertical grey lines along the abscissa represent the start times of decadal hindcast experiments. Figure from Goddard et al. [2013].*

line) capture the trend perfectly in the left picture, but have a positive mean bias due to the constant offset. In the right example, the model data have a stronger trend and a zero mean bias. The correlation coefficient is 1 in both cases. The conditional bias, on the other hand, is zero in the left case, because here, the variances of the model and observation data are identical ($\frac{s_H}{s_O} = 1$), but negative in the right one, where the variance of the hindcast data is larger than the one of the observations ($\frac{s_H}{s_O} > 1$).

In contrast to the correlation coefficient, the conditional bias depends on the relative magnitude, or expected value, of a time series. Both measures are important for an evaluation of the relative accuracy of climate predictions and will therefore be considered in this thesis.

Chapter 5

Results

5.1 Comparison of the ESMValTool Verification System to Previous Studies

Goddard et al. [2013] developed and applied a verification framework that enables an assessment of the information content in initialized decadal hindcast experiments compared to other initialized or uninitialized simulations and to observations. Pohlmann et al. [2013] applied one of these verification metrics, the anomaly correlation, to near-surface air temperature data from the MPI-ESM model, a model that participated in CMIP5. They assessed the hindcast skill of different initialized MiKlip decadal prediction systems against each other and against the HadCrut3v observational dataset.

In this thesis, the verification system from Goddard et al. [2013] has been implemented into the ESMValTool (see Chapter 2). To test the newly developed code, Section 5.1.1 compares the results of the verification system of the ESMValTool to those published by Pohlmann et al. [2013] using the exact same input data to ensure reproducibility of the implemented metrics. Possible methodological discrepancies that could occur for example through a different ensemble mean calculation, regridding etc. are identified. In Section 5.1.2, results of the additional metrics (mean squared skill score (MSSS) and conditional bias) implemented in the ESMValTool are further compared to those obtained by the MurCSS-Tool, a tool that implemented the Goddard et al. [2013] verification system as well, and that was used by Pohlmann et al. [2013] to calculate the anomaly correlations.

5.1.1 Comparison of the Correlation Metric to Pohlmann et al. 2013

The anomaly correlation is calculated to measure the linear relationship between the hindcasts and the observations. Figure 5.1 shows a comparison of the ensemble mean anomaly correlation for two different MPI-ESM-LR decadal prediction systems (b0-LR and b1-LR) calculated against HadCRUT3v observations as published by Pohlmann et al. [2013] (left) and as calcu-

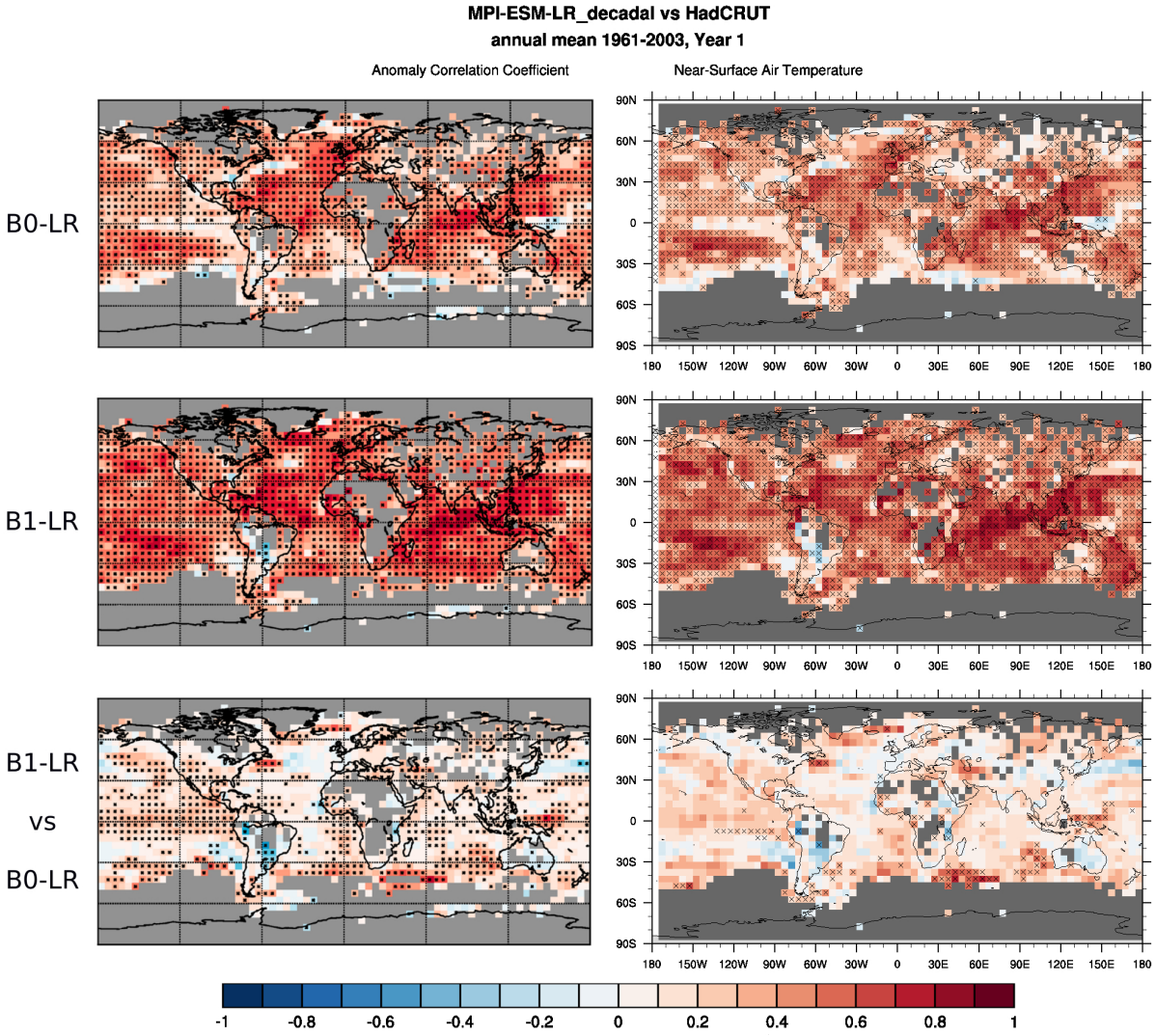

Figure 5.1: *Maps of ensemble mean hindcast skill (anomaly correlation) of near-surface air temperature from two different versions of the decadal prediction system of the MPI-ESM-LR model against observations from HadCRUT3v, as calculated by Pohlmann et al. [2013] (left column) and by the ESMValTool (right column). Initialized simulations from b0-LR (top row) and b1-LR (middle row) against observations and the difference of the anomaly correlation skill between the two model versions (bottom row) is shown. In both systems, lead year 1 of decadal experiments initialized every year from 1960 to 2002 was used. Crosses denote skill or differences in skill exceeding the 5-95% confidence level.*

lated with the ESMValTool (right). From the 43 hindcasts included in this assessment, only the first year after the initialization of each of them is used (i.e., lead year 1). Red colors indicate a positive correlation, i.e. high predictive skill, whereas blue denotes negative predictive skill. As already noted by Pohlmann et al. [2013], lead year 1 has a positive anomaly correlation skill for almost all parts of the globe (top and middle panel of left column). This reflects that the observed warming trend between 1961 and 2012 is well represented in the first prediction year in both prediction systems. The predictive skill from b0-LR to b1-LR is significantly improved

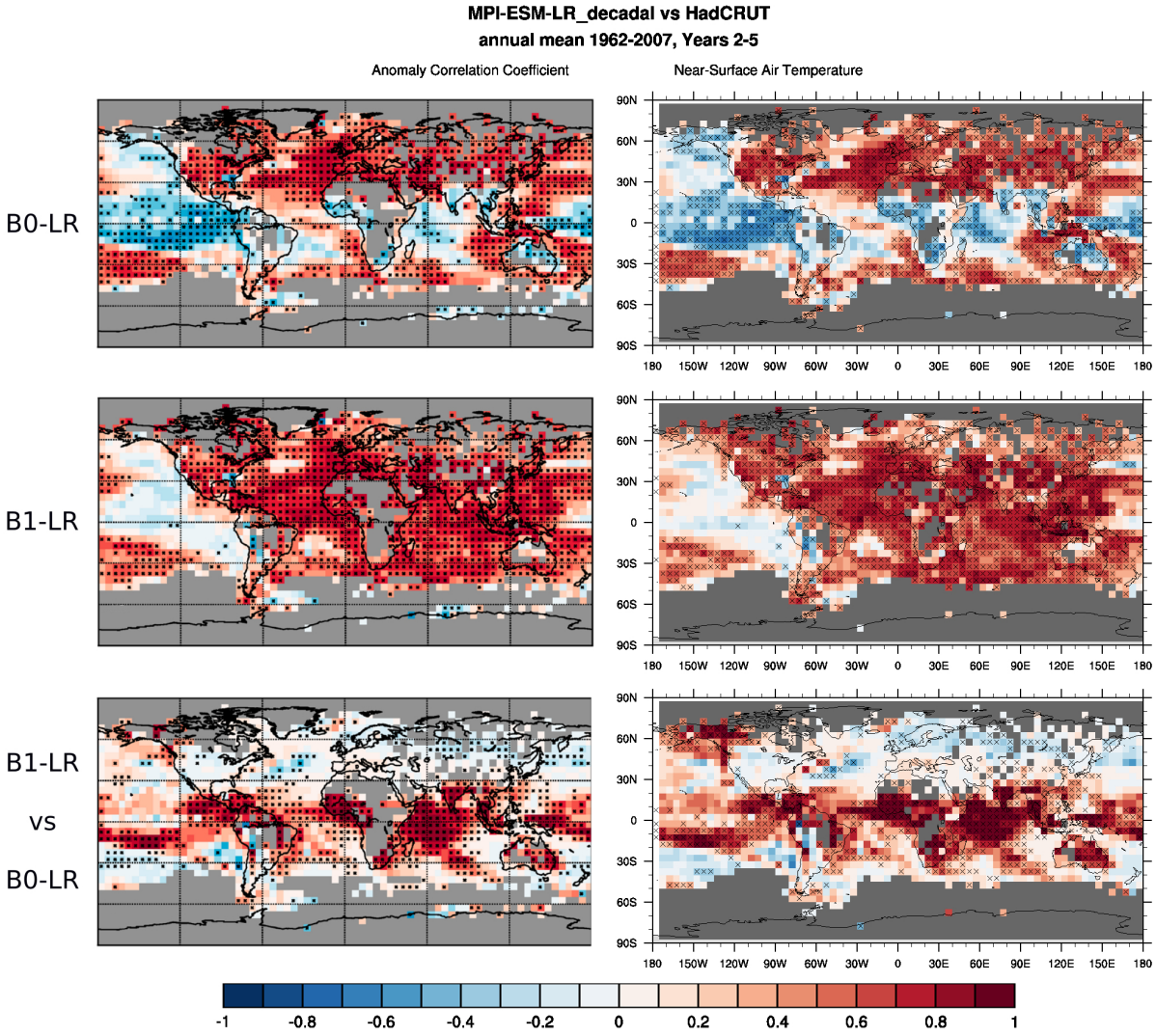

Figure 5.2: *Same as Figure 5.1, but for lead years 2-5.*

over the North Atlantic, the Southern Ocean and the tropical and North Pacific. Only over the South American continent and in the Pacific Ocean west of Japan, b0-LR has a higher correlation skill than b1-LR.

These results are very well reproduced by the ESMValTool (right column). For all three panels, not only patterns match, but also absolute values, with the exception of slightly increased noise. Similarly, grid cells containing values that are statistically significant on the 5-95% confidence level (denoted by crosses) are in very good agreement to the ones found by Pohlmann et al. [2013]. Deviations in statistically significant values, which are especially visible in the difference plot in the bottom row, could be explained by a higher number of bootstrap resamplings calculated here with the ESMValTool.

A similar comparison, but for lead years 2-5, is shown in Figure 5.2. In the b0-LR system (top left panel), negative correlation skill emerges in the tropical and North-East Pacific regions. This is due to a reversed warming trend in the hindcasts of these regions in b0-LR, with warmer

years in the 1990s and 2000s and cooler years in the 1960s and 1970s relative to observations [Pohlmann et al., 2013]. The negative correlation in the tropics is significantly improved in b1-LR (bottom left panel), with positive and significant correlation almost everywhere except the East Pacific region (middle left panel). A sensitivity study found that the skill improvements stem from the different oceanic initializations between the b0-LR and the b1-LR system [Pohlmann et al., 2013] (see also Chapter 3.1).

Again, the ESMValTool is capable of reproducing the aforementioned results (right column). Slight differences occur in b1-LR versus observations (middle row) over the Pacific Ocean, where the ESMValTool underestimates absolute values of both positive and negative correlation.

5.1.2 Comparison of the Verification System to the MurCSS-Tool

The results from Pohlmann et al. [2013], that served for the comparison of the anomaly correlation metric in the previous section, were calculated with the MiKlip Central Evaluation System's plugin "MurCSS" (see `https://www-miklip.dkrz.de/about/murcss/`). The MurCSS-Tool is another system for the analysis of decadal predictions that has implemented the metrics from Goddard et al. [2013] as well. This development was done in parallel to the implementation of the verification system into the ESMValTool. More importantly, the implementation of the verification system into the ESMValTool was done as part of this thesis to be able to apply it to other variables and phenomena in a flexible manner. This is especially important for sea-ice analyses that are performed exclusively for months of minimal sea-ice extent. With the MurCSS-Tool, only annual means can be assessed.

Similar to Section 5.1.1, the results for additional metrics (conditional bias and MSSS) included in both the MurCSS and the ESMValTool are now compared to each other, again using the exact same simulations from the MPI-ESM-LR and the same observations. In contrast to the previous section, lead years 2-9 are chosen, as well as a slightly shorter time period, since the MPI-ESM-LR historical simulations are only available until 2005. In contrast to Section 5.1.1, here the focus is on comparing the predictive skill of initialized decadal hindcasts to the uninitialized simulations.

Figure 5.3 shows the anomaly correlation of near-surface temperature for the MPI-ESM-LR against the HadCRUT3v observations that were also used in Figures 5.1 and 5.2. Compared to the uninitialized simulations, increased skill in the initialized simulations can be seen over large parts of the Northern Atlantic and southwest of South America (Figure 5.3, red areas in lower panels), whereas in other regions there is no additional or even less predictive skill (blue grid cells).

Very good agreement between the results calculated with the MurCSS-Tool and the ESMValTool is found, indicating that the findings of the previous section also hold for different lead

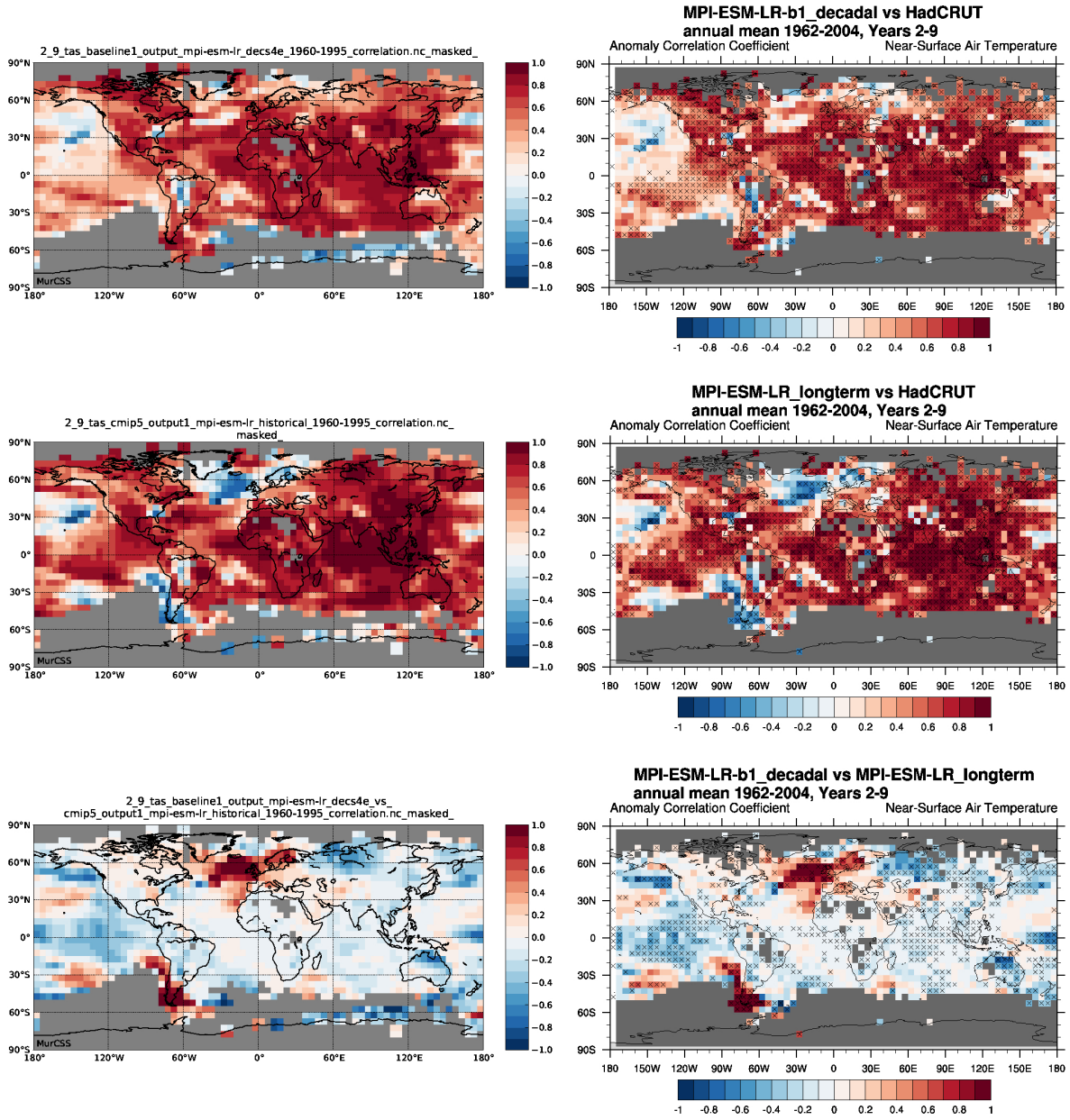

Figure 5.3: *Ensemble mean hindcast skill (anomaly correlation) of near-surface air temperature from b1-LR against observations from HadCRUT3v calculated by the MurCSS-Tool (left) and by the ESMValTool (right). From top to bottom: initialized simulations against observation, uninitialized simulation against observation, and difference between the two anomaly correlations. In both systems, the selected lead time is years 2-9 of decadal experiments initialized every year from 1960 to 1995. Crosses in the ESMValTool's results denote skill or differences in skill exceeding the 5-95% confidence level.*

year selections and the inclusion of long-term simulations. Apart from some additional noise in the anomaly correlation produced with the ESMValTool, not only the geographical pattern is very similar, but also the absolute values over much of the globe.

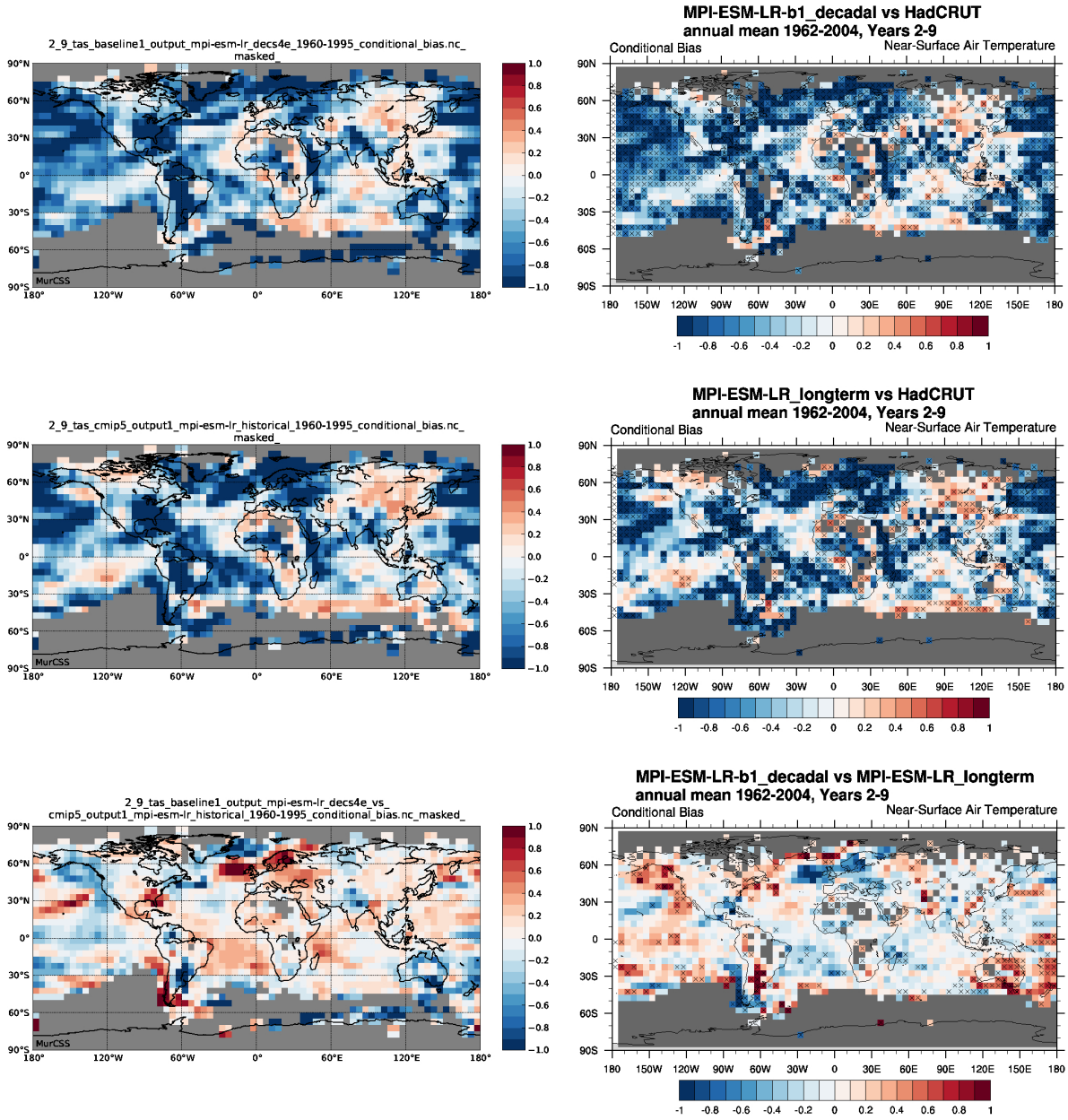

Figure 5.4: *Same as Figure 5.3, but for the conditional bias calculated from the b1-LR (upper row) and the long-term (middle row) simulations against observation from HadCRUT3v calculated by the MurCSS-Tool (left) and the ESMValTool (right). The difference in the conditional bias between the initialized and the uninitialized simulation is calculated from the original values of the individual conditional biases in the MurCSS-Tool (lower left panel) and from the absolute values in the ESMValTool (lower right panel) which explains why the colors are reverse.*

Areas of improvement in correlation skill re-appear in the conditional bias and the MSSS (Figures 5.4 and 5.5). In the North Atlantic as well as in the south-west of South America, the conditional bias is reduced (Figure 5.4, blue areas in lower right panel) and the MSSS increases (Figure 5.4, red areas in lower right panel) from long-term to b1-LR, implying a higher relia-

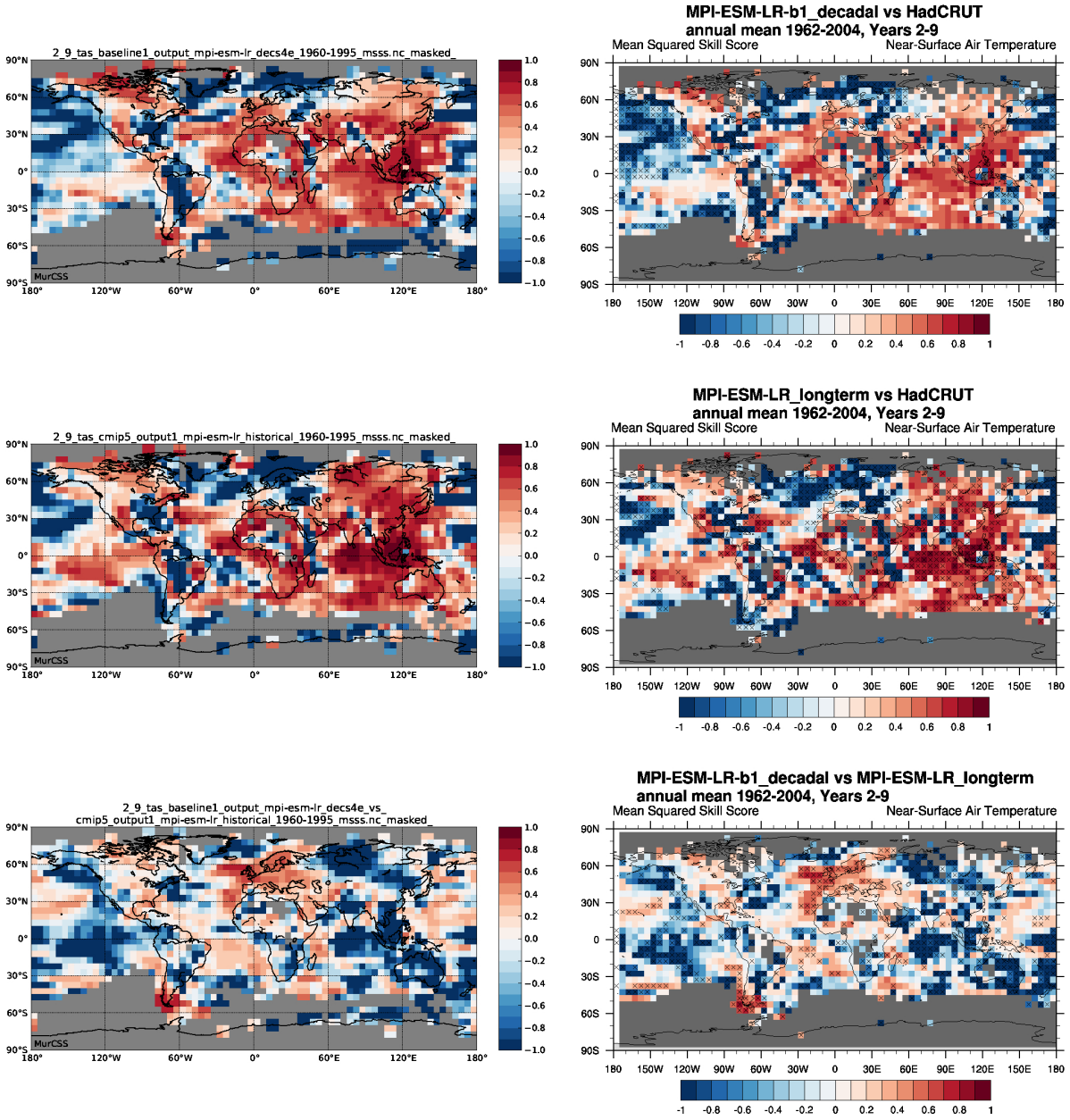

Figure 5.5: *Same as Figure 5.3, but for mean squared skill score (MSSS) calculated from the b1-LR (upper row) and the long-term (middle row) simulations against observation from HadCRUT3v calculated by the MurCSS-Tool (left) and the ESMValTool (right). In the bottom row, not the difference between the decadal and long-term simulations is shown (as for correlation skill and conditional bias), but the MSSS calculated from the decadal hindcasts versus the long-term simulations.*

bility and better accuracy in these regions in the decadal prediction system compared to the long-term simulations.

Similar good agreement as for the anomaly correlation metric is found for the conditional bias and the MSSS between the two verification systems MurCSS and ESMValTool. Note that in

contrast to the definition of Goddard et al. [2013] that uses the absolute difference, the difference in the conditional bias between the initialized and uninitialized simulations in Figure 5.4 is calculated by MurCSS (lower left panel) with respect to the original values instead of the absolute values. This explains the differences in sign compared to the results of the conditional bias as calculated by the ESMValTool (lower right panel) that follows the original Goddard et al. [2013] definition.

5.1.3 Summary of Comparisons

Overall excellent agreement both qualitatively and quantitatively is found between the ESMValTool and the MurCSS-Tool results that were used in Pohlmann et al. [2013] to calculate the anomaly correlations. This confirms that the implementation into the ESMValTool that was done as part of this thesis provides accurate results and can be used for further analyses in subsequent sections of this thesis and in follow-up studies. Slightly increased noise in the ESMValTool could be due to additional smoothing in the results shown in Pohlmann et al. [2013] that is beyond the spatial smoothing that stems from the regridding, like in Goddard et al. [2013] (their Figure 9).

Differences between the two systems are found in the pattern of missing values which does not match perfectly. This is due to a missing specification in Pohlmann et al. [2013] about how missing values were exactly treated. As missing values stem from gaps in the observational dataset, differences in the pattern of missing values between the two systems may be related to a different treatment of missing time steps in the data. When performing temporal averages, Pohlmann et al. [2013] seem to have used a certain threshold representing the number of missing time steps after which the calculated mean returns a missing value as a result. A comparison by eye shows that similar results for missing values are achieved in the ESMValTool if a threshold of 60% is used, i.e. the numerical functions computing averages return a non-missing value only if at least 60% of the observational data in the constructed time series are available.

5.2 Forecast Skill of Ensemble Mean Surface Temperature in MiKlip Decadal Predictions

In this section, the verification system implemented into the ESMValTool is applied to near-surface air temperature data as simulated by different versions of the MiKlip decadal prediction systems. The goal of Section 5.2.1 is to determine which of the three MiKlip decadal predictions systems (b0-LR, b1-LR and pr, see Chapter 3.1.2) performs best in terms of its predictive skill. The results extend the previous comparisons to the MurCSS-Tool in this regard that the newly

available MiKlip prototype system (pr) is included, that was not yet assessed in Pohlmann et al. [2013]. After deciding on the best performing MiKlip system out of the three available systems, in Section 5.2.2 this selected prediction system is then compared to the uninitialized long-term simulations of the MPI-ESM-LR to assess whether the initialization of the model leads to a better performance and more skillful predictions than those obtained with the uninitialized long-term model simulations. For all assessments, only grid cells containing statistically significant values are considered.

5.2.1 Comparison of Different Versions of the MiKlip Decadal Prediction Systems

In order to be consistent with the approach of Section 5.1.1 and Pohlmann et al. [2013], lead years 2-5 are chosen for the assessments in this section. For this lead time, the anomaly correlation skill (Figure 5.6) in b0-LR is positive over most of the mid-latitudes in both hemispheres, but negative correlation skill appears in the tropics and eastern North Pacific with largest magnitudes over the tropical East Pacific (see also Pohlmann et al. [2013]).

An increase in hindcast skill from b0-LR (top left) to b1-LR (middle left) is evident in the difference plot (upper plot in right column), especially in the tropical Atlantic and Indian Ocean, where the correlation becomes positive in b1-LR. Although negative correlation skill is still evident in the Pacific Ocean in b1-LR, the skill has substantially increased compared to b0-LR, especially in the tropical region. In contrast to these improvements, the hindcast skill over North Atlantic Ocean decreases from b0-LR to b1-LR.

With both ocean and atmosphere initialized with the full field method in the pr version (Figure 5.6, bottom left), overall very similar results are found compared to b1-LR, although the correlation skill further increases in the eastern tropical Pacific Ocean and even more strongly in the North Atlantic Ocean (lower right panel).

The conditional bias is used to assess the reliability of the hindcasts (Figure 5.7). Consistent with the regions where the correlation skill was negative, the conditional bias of b0-LR is most strongly negative over the tropical Pacific Ocean and generally negative over most of the globe. From b0-LR to b1-LR, the absolute value of the conditional bias decreases (see also upper right panel in Figure 5.7). The largest improvements are found in the tropical Pacific and Indian Oceans, where the correlation increases as well. b1-LR and pr are overall very similar, with regions where the conditional bias is reduced (East Pacific Ocean and Australia) and where it gets larger (North Atlantic Ocean). Despite a further improvement in these regions in pr (lower right), an increase in conditional bias in the West Pacific and North Atlantic is found.

The MSSS is a combination of the anomaly correlation and the conditional bias and is used to assess the accuracy of the hindcasts (Figure 5.8). If the MSSS is positive (red colors), there is high accuracy, whereas negative values denote low accuracy. Very low accuracy is

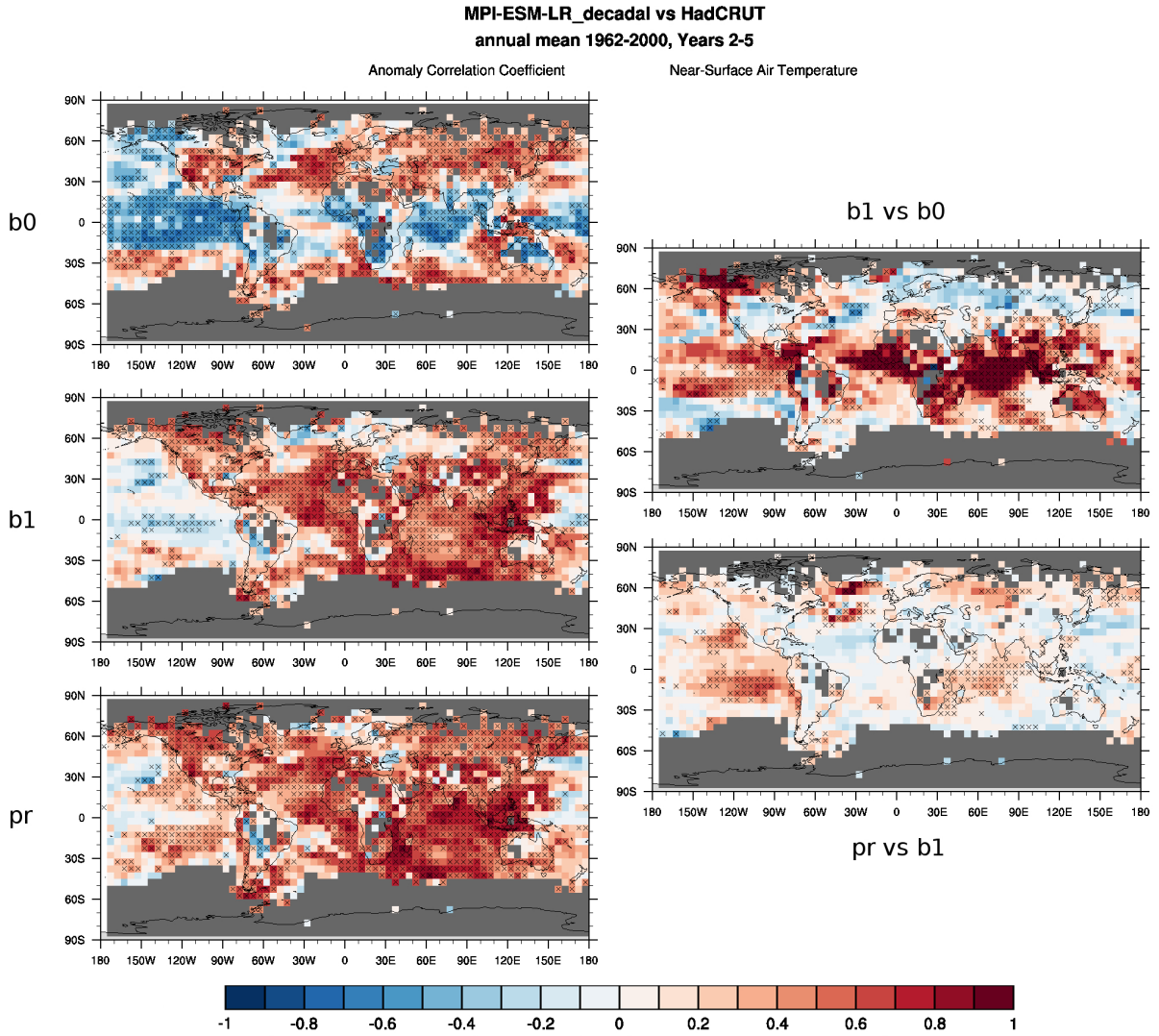

Figure 5.6: *Maps of ensemble mean hindcast skill (anomaly correlation) of near-surface air temperature from the three different decadal prediction systems of the MPI-ESM-LR model against observations from HadCRUT3v as calculated with the ESMValTool. In the left column, initialized simulations from b0-LR (top), b1-LR (middle) and pr (bottom) against observations are shown. The right column depicts the difference of the anomaly correlation skill between baseline-1 and -0 (upper panel) and prototype towards b1-LR (lower panel). The selected lead time is years 2-5 of decadal experiments initialized every year from 1960 to 1995. Crosses denote skill or differences in skill exceeding the 5-95% confidence level.*

found in b0-LR (top left) over the entire Pacific Ocean and the tropics, and also in the North and West Atlantic. In b1-LR (middle left), there is low skill in the same regions, but here of smaller magnitude. An improvement in accuracy occurs especially over the Indian Ocean, where the MSSS becomes positive. For pr (bottom left) even higher accuracy in the Indian Ocean is found and the magnitude of negative MSSS further decreases over the Pacific Ocean.

The right column shows the MSSS of b1-LR tested against b0-LR (top) and pr against b1-

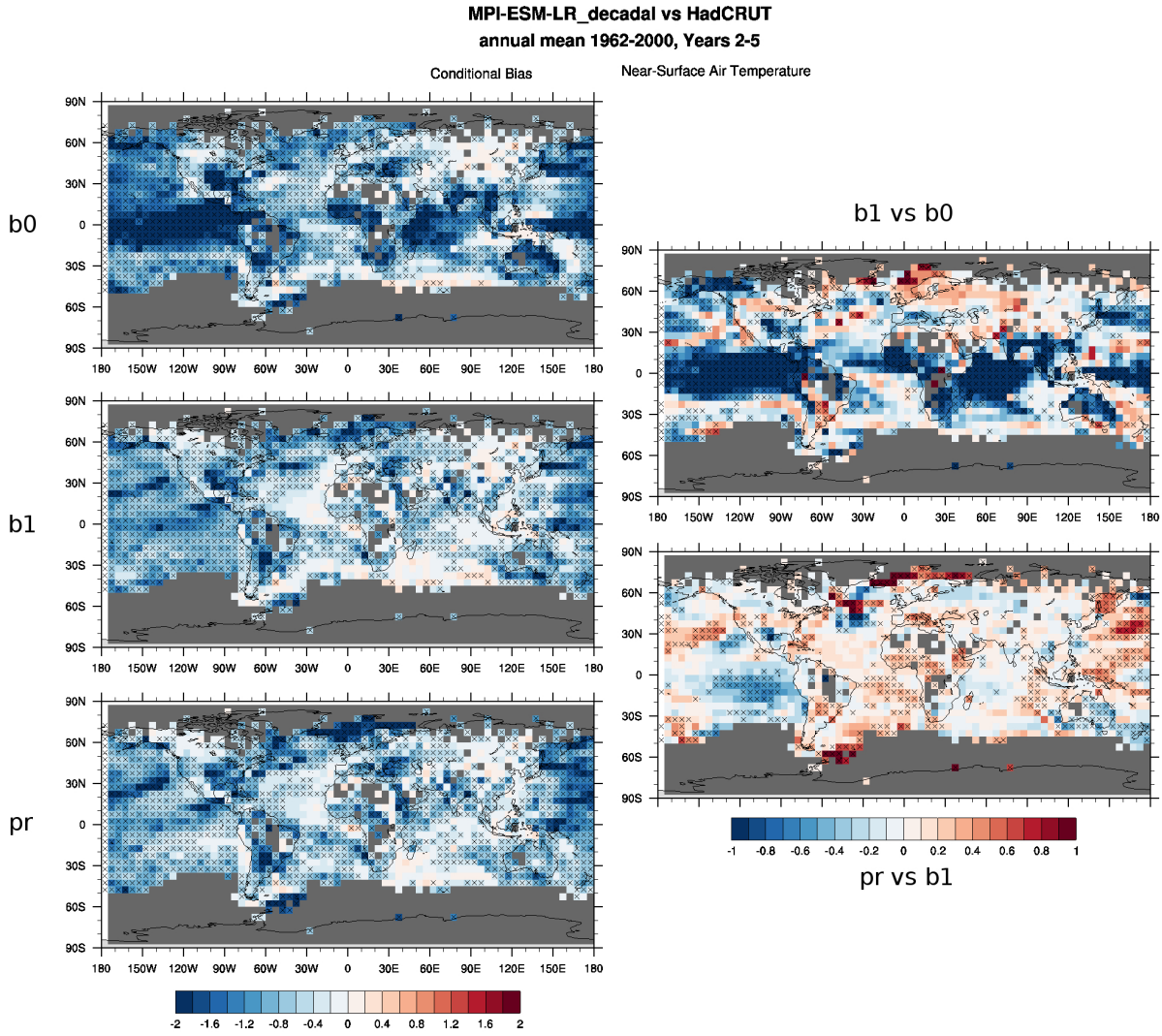

Figure 5.7: *Same as Figure 5.6, but for the conditional bias as calculated with the ESMValTool. The right column depicts the difference of the absolute values of two respective fields in the left column, why here, blue colors denote an improvement, i.e. increase of reliability, and red colors denote the opposite. Also note the different ranges of value in the color scales of the two columns.*

LR (bottom). Positive values (red colors) indicate an improvement in accuracy over the test prediction, blue denotes the opposite. The strongest increase in accuracy is from b0-LR to b1-LR, especially in the tropics, whereas there is less accuracy in some parts of the middle latitudes (both hemispheres, but especially in the northern one). From b1-LR to pr, a further improvement in the East Pacific and Indian Ocean is evident, which corresponds to the findings of the left column.

The results in this section show that for lead years 2-5, b1-LR and pr are performing substantially better than b0-LR, and that the differences between b1-LR and pr are small, with one or the other system performing slightly better or worse than the other one for a specific metric

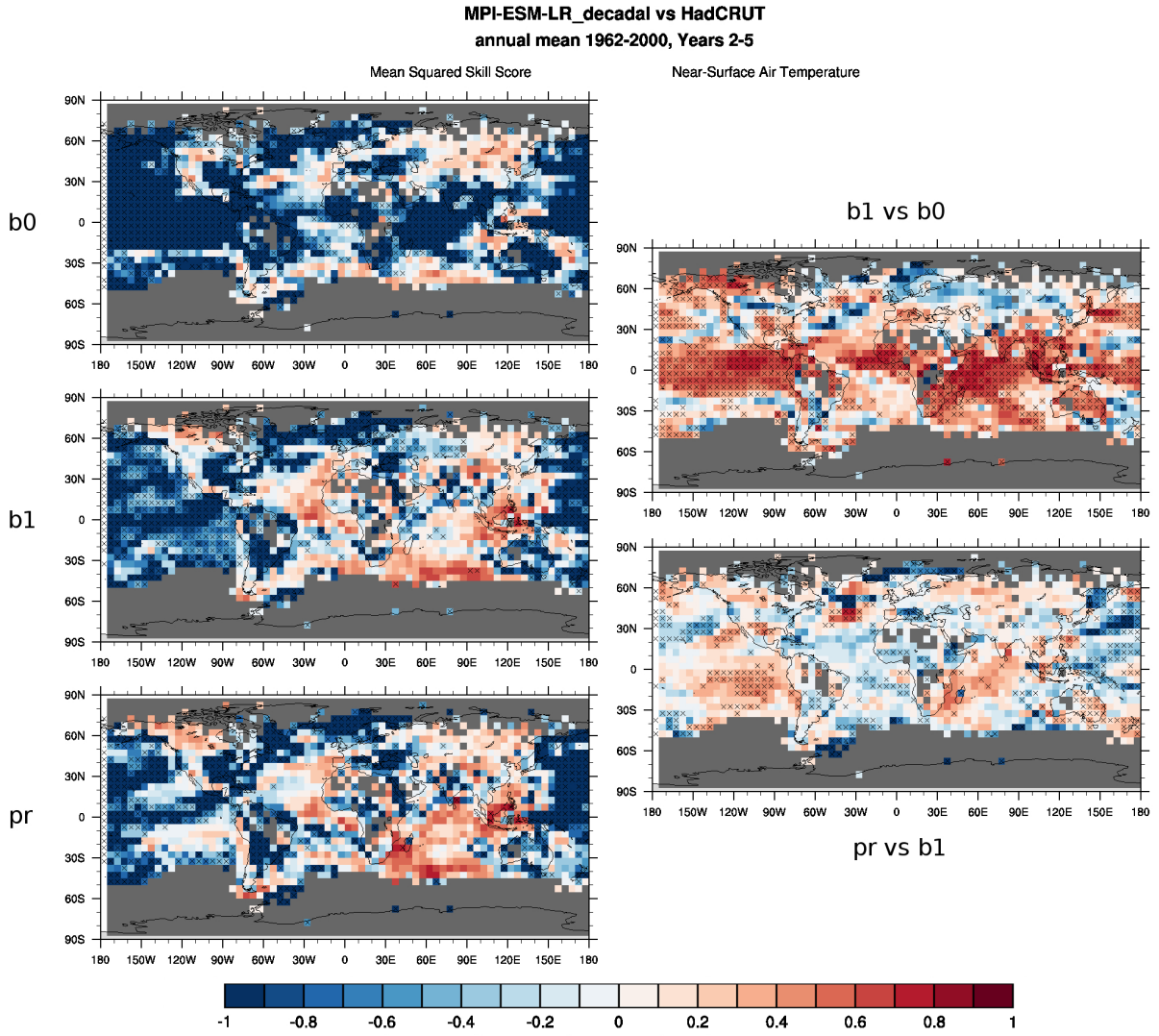

Figure 5.8: *Same as Figure 5.6, but for MSSS as calculated with the ESMValTool. The panels in the right column each depict the MSSS of one decadal model version against another.*

or region. The prototype prediction system is the most recently developed MiKlip system, where only a few studies are available This is why the pr system is chosen for a more detailed comparison to the long-term simulations in Section 5.2.2.

5.2.2 Comparison of the MiKlip Decadal Prototype System to Long-term Simulations

To examine the science question whether the initialization of the MPI-ESM improves the forecast skill, the MiKlip decadal prototype prediction system is compared to the uninitialized long-term simulations in this section.

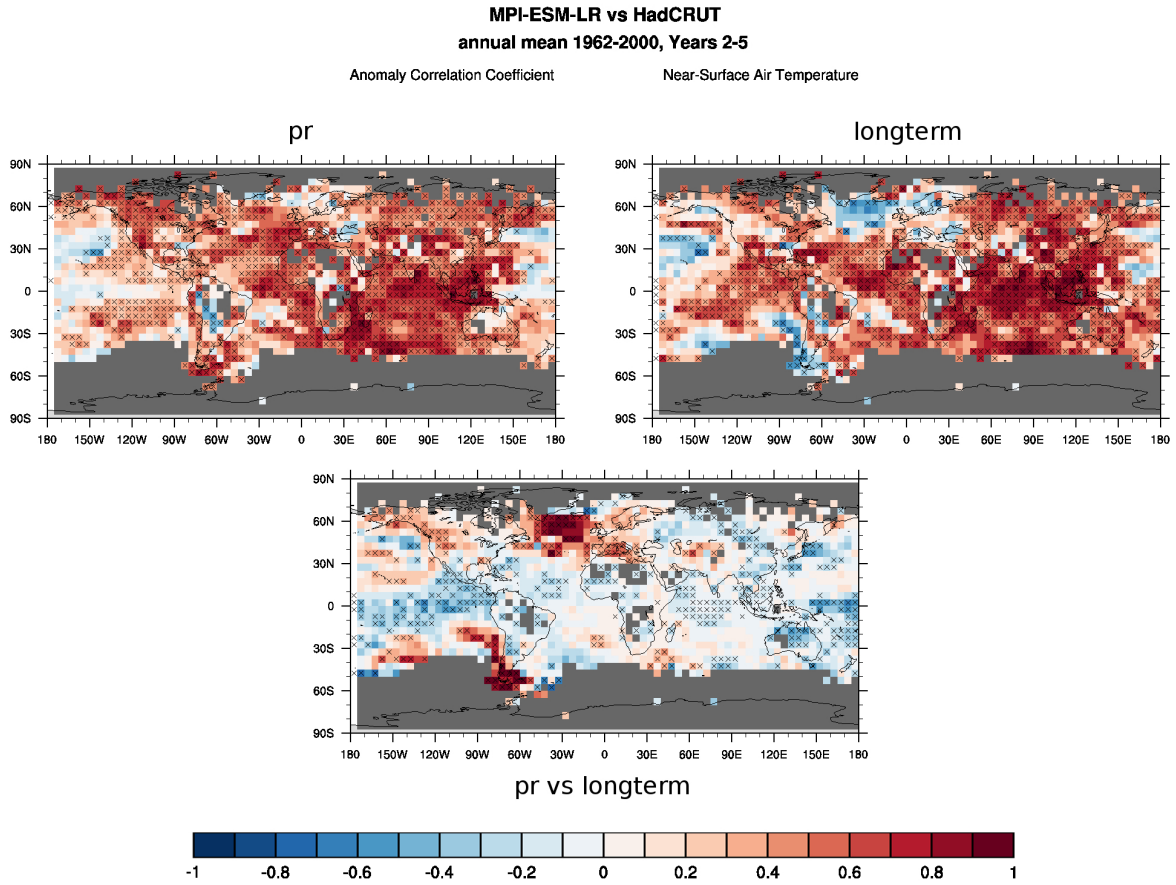

Figure 5.9: *Ensemble mean hindcast skill (anomaly correlation) of near-surface air temperature from MPI-ESM-LR-protype decadal hindcasts (top left) and -LR long-term simulations (top right) against observations from HadCRUT3v as calculated with the ESMValTool. The bottom figure depicts the difference between the two anomaly correlations, with the long-term subtracted from the prototype version. Evaluated are the decadal experiments initialized every year from 1960 to 1995 with a lead-time selection of years 2-5. Crosses denote skill or differences in skill exceeding the 5-95% confidence level.*

The anomaly correlation skill (Figure 5.9) for lead years 2-5 is mostly similar between pr (upper left) and long-term (upper right), which is also visible in the difference plot in the lower row. However, two regions stand out with significant improvements in forecast skill, one over the North Atlantic and one over the south-west of South America. In addition, the forecast skill is slightly higher in pr also over some parts of the North Pacific Ocean, over Northwest America, and the Mediterranean region. The strongest decrease in hindcast skill is found over the tropical Pacific region, but in smaller magnitude than the aforementioned increase.

The conditional bias (Figure 5.10) is again mostly similar between pr and long-term (upper row), with a negative conditional bias dominating much of the globe. The difference in the absolute values of the respective conditional biases (bottom panel) reveals that the conditional bias increases in magnitude by the initialization, especially in the Pacific and in the North and North-West Atlantic Ocean. A reduction in conditional bias, i.e. an increase in reliability, is

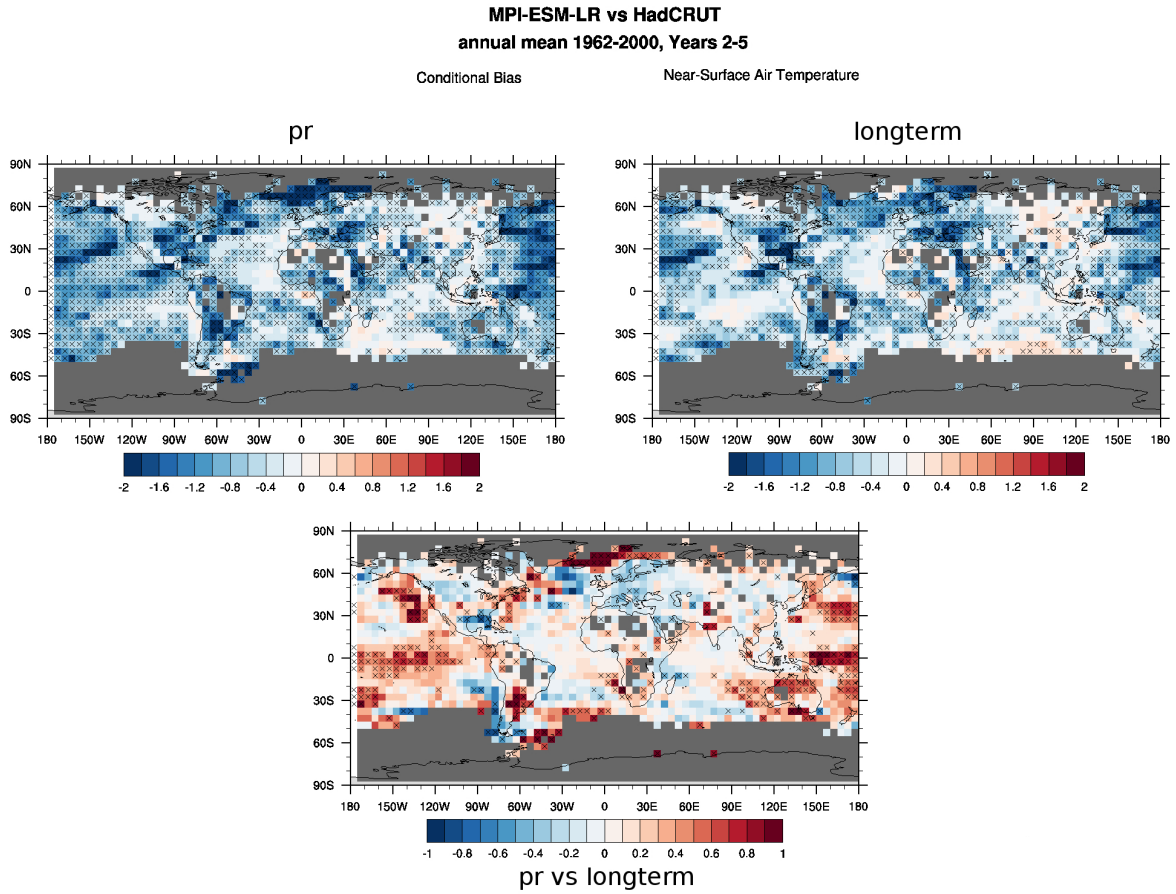

Figure 5.10: *Same as Figure 5.9, but for conditional bias calculated with the ESMValTool from the MPI-ESM-LR-pr and long-term simulations against observation from HadCRUT3v. The difference in the conditional bias between the initialized and the uninitialized simulation (bottom panel) is calculated from the absolute values of the individual conditional biases that are shown in the upper row. This is why in the bottom panel, blue colors denote an improvement, i.e. increase in reliability by initialization. Also note the different color scales.*

found only over the North-East Atlantic Ocean, the Gulf of Mexico and southwest of South America, which corresponds to the higher anomaly skill found in the same regions (compare to Figure 5.9).

The pr and long-term simulations also have similar regions of low and high MSSS values (Figure 5.11), with negative accuracy in the Pacific Ocean and over most parts of the American continents, and positive skill in the Indian Ocean and over East-Asia. The MSSS of pr relative to the uninitialized hindcasts (bottom panel) shows that areas of declined accuracy due to initialization dominate. Only the North-East Atlantic region and, again, the southwest coast of South-America show a notable increase in accuracy.

In conclusion, the initialization of the MPI-ESM can increase the forecast skill measured by the anomaly correlation, the reliability measured by the conditional bias and the accuracy measured by the MSSS of lead years 2-5 of predictions of near-surface temperature in the

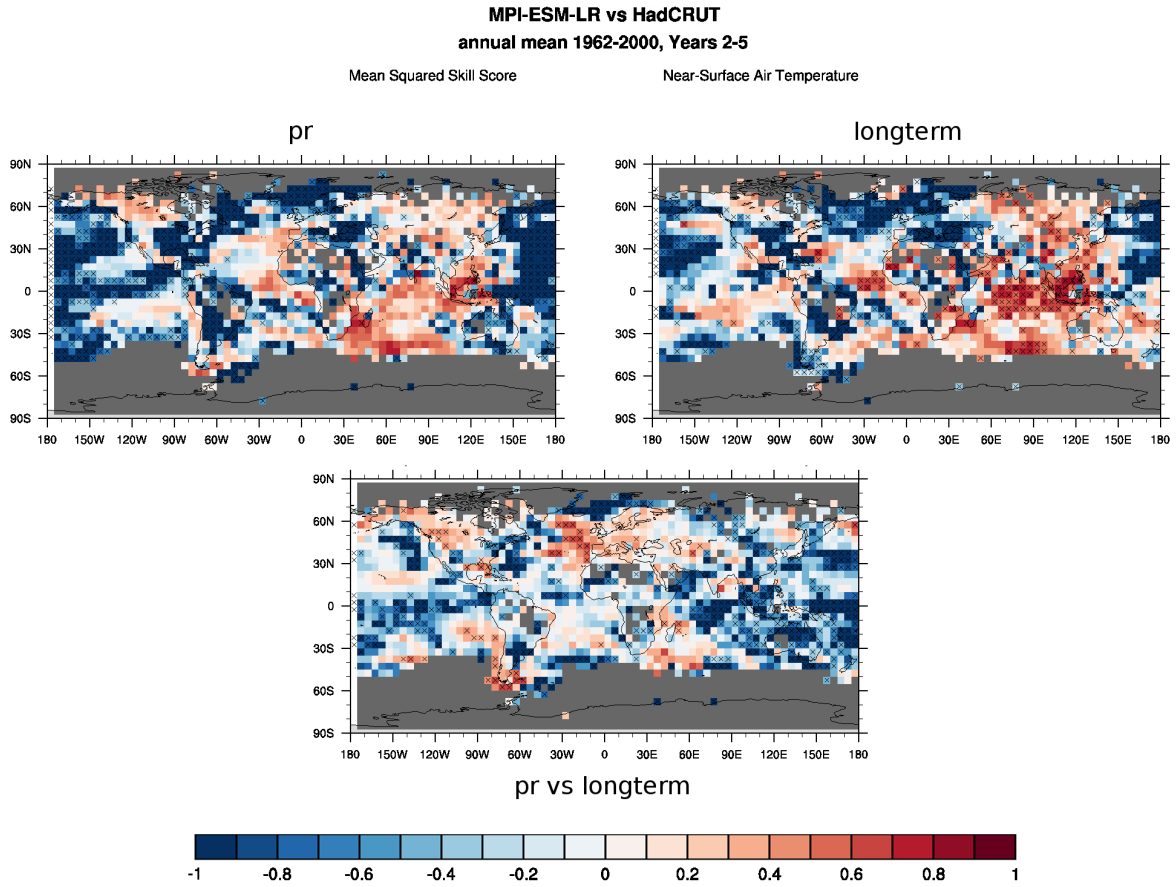

Figure 5.11: *Same as Figure 5.9, but for the MSSS calculated with the ESMValTool from the MPI-ESM-LR-pr and long-term simulations against observation from HadCRUT3v. In the bottom row, not the difference, but the MSSS of decadal hindcasts against long-term simulations is shown.*

North Atlantic Ocean and southwest of South America. For almost all the rest of the globe, no further predictive skill can be derived due to initialization, as for these regions the conditional bias increases and the anomaly correlation and MSSS decrease from the prototype to the long-term model version.

The statements made in this section are all based on the lead years 2-5. Following the example of Pohlmann et al. [2013], Figure 5.12 depicts the global mean of each metric as a function of lead time for annual averages (left) and 4-year averages (right) to also allow an assessment of different lead times. For all metrics, b0-LR performs markedly worse than b1-LR, pr and long-term. Otherwise, the b1-LR and pr hindcast systems are relatively similar to the uninitialized simulations.

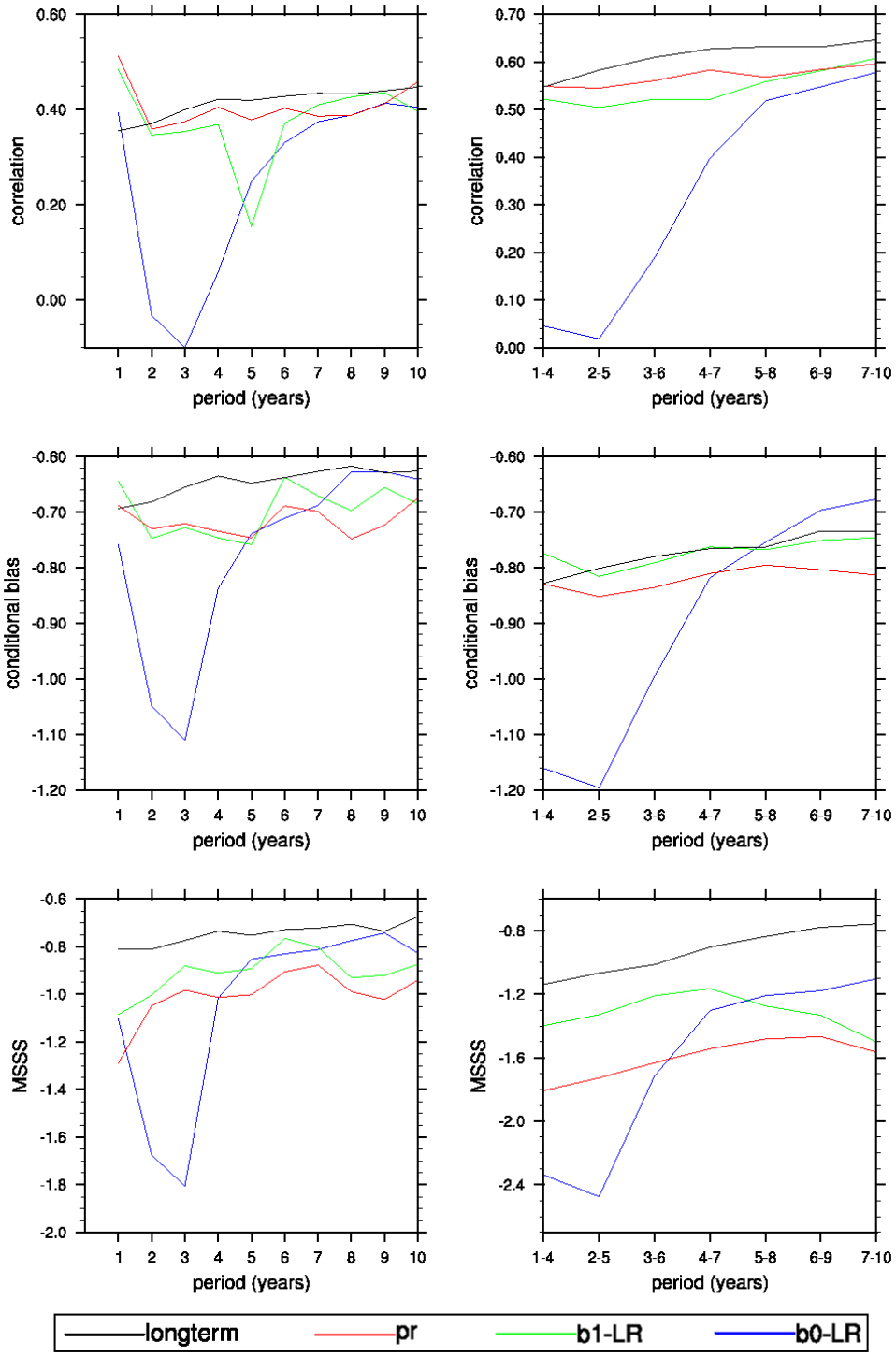

Figure 5.12: *Global mean hindcast skill (anomaly correlation, upper row), conditional bias (middle row) and MSSS (bottom row) as functions of lead time following the example of Figure 3 in Pohlmann et al. [2013]. Tested are the decadal predictions systems b0-LR (blue), b1-LR (green) and prototype (red) and the long-term simulations (black) of the MPI-ESM-LR against observations from HadCRUT3v for annual averages (one-year lead times, left column) and lead times consisting of 4-year averages (right column). The time range is the same as in Figures 5.6 to 5.8.*

5.3 Forecast Skill of Ensemble Mean Sea-Ice Concentration in the MiKlip Decadal Predictions

The observed negative trend in September Arctic sea ice of more than 30% since the late 1970s [Stroeve et al., 2012] was underestimated by CMIP3 climate models [Stroeve et al., 2007; Rampal et al., 2011]. However, efforts of improving simulated sea-ice components and atmospheric circulation have resulted in a better-represented Arctic sea-ice trend in the CMIP5 models [Notz et al., 2013; IPCC, 2013].

In contrast, the CMIP5 models simulate a small sea-ice decline in the Antarctic that is not observed [Turner et al., 2013]. It has to be noted that a change in the satellite data processing for the Bootstrap algorithm dataset has been made in the mid-2000s [Eisenman et al., 2014].

Building on work by Hübner [2013] and Bräu [2013] who qualitatively evaluated the sea-ice representation in long-term and decadal simulations with the MPI-ESM, here the verification system that was tested in Section 5.1 is applied to sea ice. In particular, the anomaly correlation metric is applied to sea-ice concentrations to quantitatively assess the prediction skill in the MiKlip prediction systems compared to the long-term simulations.

Sea-ice concentration, or sea-ice area fraction, is the area covered by sea ice relative to a reference area, i.e. the degree of sea-ice cover of each grid cell [Cavalieri et al., 1996]. From sea-ice concentration maps, sea-ice area and sea-ice extent can be derived, which are important indicators of climate change.

This section follows the structure of Section 5.2, by first identifying the prediction system with the best correlation skill (Section 5.3.1) and by then comparing this system to the long-term simulations (Section 5.3.2). As is common for sea-ice assessments, the focus is on summer-term sea-ice (September in the Arctic and March in the Antarctic), to evaluate the representation of the month with minimal sea-ice concentration. Similar to the results for temperature, only skill or differences in skill of stastically significant magnitude are considered for the assessments.

An important note about the following figures: grid points where the standard deviation over all time steps is zero (for example, grid points that never contain ice in the evaluated calendar month, like September in the Norwegian Sea), had to be set to missing values, as the standard deviation is the denominator in the correlation calculation (see Chapter 4.2.2). A grid point with constant values over the whole time-series (thus leading to a standard deviation of zero for that grid point) is very unlikely to appear in temperature data, but has to be accounted for the sea-ice assessment. This problem is most obvious when looking at minimum sea-ice extents: in March, sea ice exists only on a rather small band around the Antarctic continent.

5.3.1 Comparison of Different Versions of the MiKlip Decadal Prediction Systems

The anomaly correlation of September mean Arctic sea-ice concentrations for lead years 2-5 (Figure 5.13) reveals that pr has the highest overall prediction skill of the three MiKlip decadal

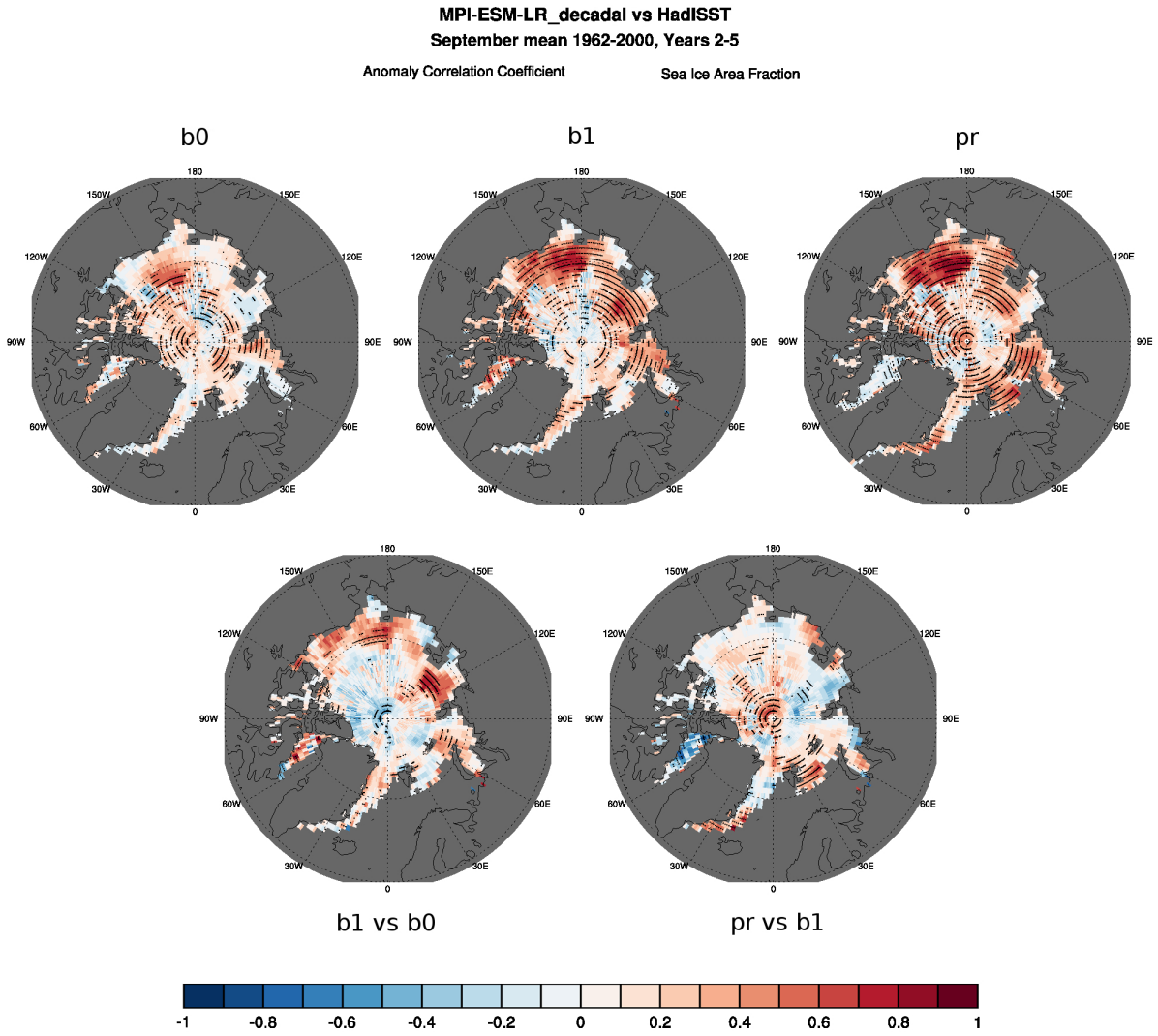

Figure 5.13: *Ensemble mean hindcast skill (anomaly correlation) of September mean Arctic sea-ice concentrations in the three MiKlip decadal prediction systems against observations from HadISST as calculated with the ESMValTool. In the top row, initialized simulations from (from left to right) b0-LR, b1-LR and pr against observations are shown. The bottom row depicts the difference of the anomaly correlation skill between b1-LR and b0-LR (left panel) and pr towards b1 (right panel). The selected lead time is years 2-5 of decadal experiments initialized every year from 1960 to 1995. The 60° N and 75° N circles of latitude (dashed circles) are shown for reference. Black dots denote skill or differences in skill exceeding the 5-95% confidence level.*

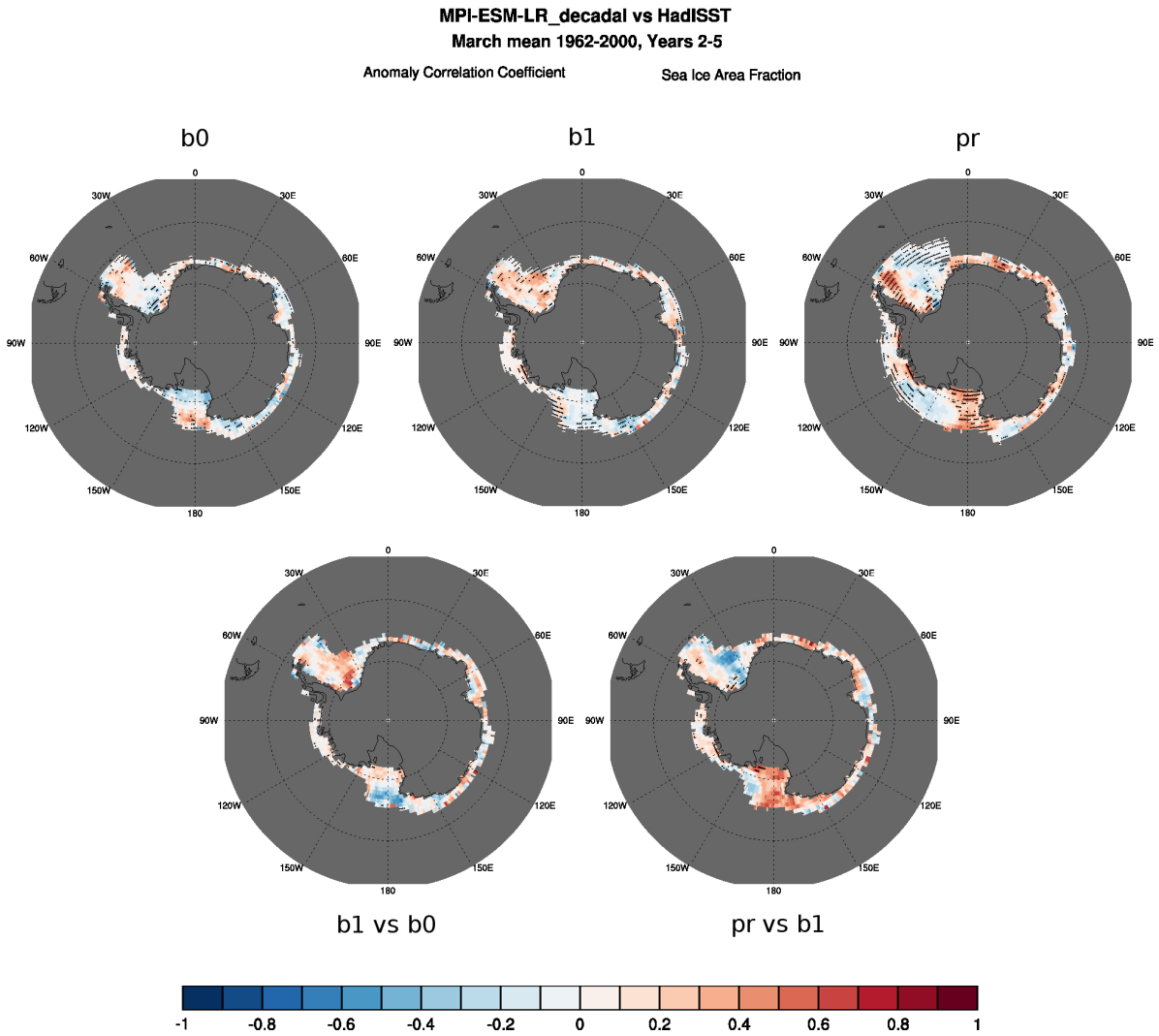

Figure 5.14: *Same as Figure 5.13, but for March mean Antarctic sea-ice concentrations. The 60° S and 75° S circles of latitude (dashed circles) are shown for reference.*

prediction systems (top row). Whereas some grid cells between 90 and 75° N in b0-LR and b1-LR show negative correlation coefficients, pr has a positive hindcast skill almost everywhere over the Arctic. A comparison between b1-LR and b0-LR (bottom left) and pr and b1-LR (bottom right) indicates that pr performs best in terms of hindcast skill around the North Pole region down to 85° N and east of Svalbard, whereas the b1-LR system has higher correlation values north of Canada and Northeast Asia (Beaufort and East Siberian Sea) and in the Baffin Bay and Laptev Sea.

For March mean Antarctic sea-ice concentrations (Figure 5.14, also for lead years 2-5), pr shows overall highest correlation skill again (upper right panel), especially in the Ross Sea. Since the difference plots (bottom row) barely show any significant grid points, final conclusions of possible improvements over predecessor versions cannot be made.

In total, for the lead years 2-5, the pr system appears to have the highest skill among the

three MiKlip decadal prediction systems. Therefore, and for similar reasons as for surface temperature, pr is chosen for the comparison to long-term simulations in the next section.

5.3.2 Comparison of the MiKlip Decadal Prototype System to Long-term Simulations

For temperature, an additional hindcast skill could be achieved by initialization for the lead years 2-5 only for certain regions. To assess, whether this is different for sea-ice concentrations, similar to Section 5.2.2, the decadal prototype system is compared to the long-term simulations in this section.

The prediction skill for September mean Arctic sea-ice concentrations (Figure 5.15, top row) is positive almost everywhere for pr, but the highest correlation coefficients up to 0.9 are found for the long-term simulations over the Barents Sea and around the Queen Elizabeth Islands. The overall high hindcast skill is due to a well-represented Arctic sea-ice decline in both the long-term and decadal simulations of the MPI-ESM [Notz et al., 2013; Bräu, 2013]. The difference between pr and long-term (right panel) reveals that additional hindcast skill in September mean Arctic sea-ice can be achieved through initialization in some regions, in particular at very high latitudes north of 85° N, in the eastern Laptev Sea and between the East Siberian and the Beaufort Sea.

For March mean Antarctic sea-ice concentrations (Figure 5.15, bottom row), pr and long-term show similar positive hindcast skill in the Weddell Sea, but differ substantially in the Ross Sea. Here, the difference plot on the right reveals that for Antarctic sea-ice concentrations of lead years 2-5, additional skill through initialization can be achieved in this region.

In conclusion, the initialization of the MPI-ESM can significantly improve the hindcast skill of sea-ice predictions in a few distinct regions for the lead time years 2-5. As for surface temperature, an additional global mean skill cannot be achieved. It has to be further evaluated, if for different time ranges and lead time selections the regions of improved forecasts remain the same or can even be extended.

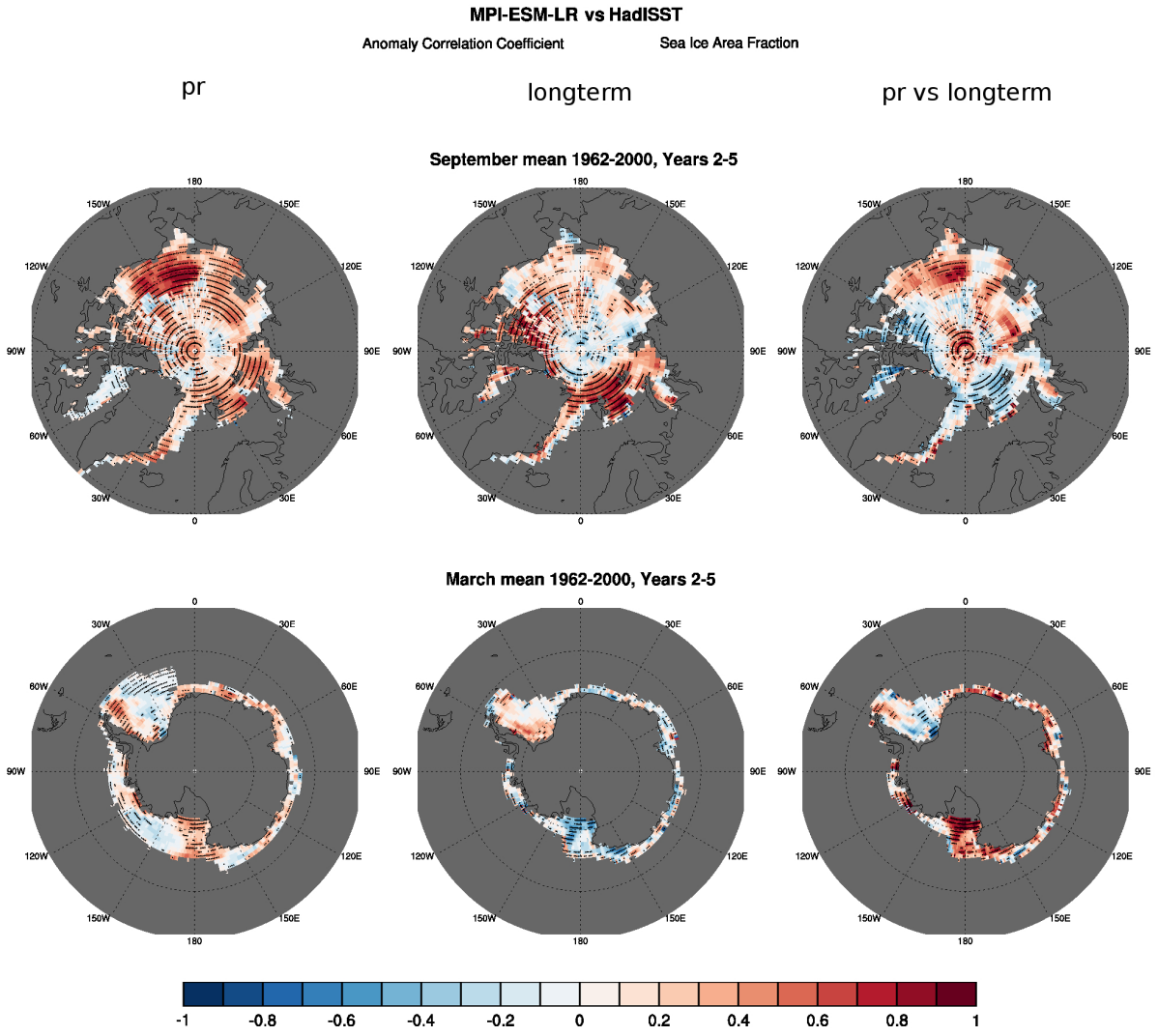

Figure 5.15: *Ensemble mean hindcast skill (anomaly correlation) of September mean Arctic (top row) and March mean Antarctic (bottom row) sea-ice concentrations from MPI-ESM-LR-protype decadal hindcasts (left column) and -LR long-term simulations (middle column) against observations from HadISST as calculated by the ESMValTool. The right column depicts the difference between the two respective anomaly correlations, with the long-term subtracted from the prototype version. Evaluated are the decadal experiments initialized every year from 1960 to 1995 with a lead-time selection of years 2-5.*

5.3.3 Discussion of the Applicability of the Verification System to Sea Ice

Goddard et al. [2013] applied their verification framework to near-surface air temperature and precipitation data. In Figure 5.16, the probability distribution functions (PDF) of global near-surface temperature anomalies from HadCRUT3v (see Chapter 3.2.1), monthly precipitation from GPCCv6 [Schneider et al., 2011] and sea-ice concentrations from HadISST (see Chapter 3.2.2) are shown. The distribution of surface temperature (left panel) is Gaussian, but both precipitation (middle) and sea-ice concentrations (right panel) do not show a Gaussian distribution. Thus the question arises whether the metric calculations applied in this thesis result in utilizable information with other than normally distributed data.

The PDF of sea-ice concentration reveals that almost 80% of the values are zero. This is expectable, as most areas of the Earth's oceans never contain ice. To enable a more detailed look at the distribution of sea-ice concentrations apart from zero, Figure 5.17 depicts the right panel of Figure 5.16 for probabilities of less than 4%. Here, a second maximum at 1 is revealed, making this distribution function qualitatively look like the vertically mirrored precipitation PDF.

Goddard et al. [2013] applied the verification to precipitation in addition to temperature. It was therefore applied here to test the prediction skill of sea-ice concentrations. The resulting qualitative statements probably also hold for variables that are not normally-distributed. Furthermore, the verification system only contains assessments made for each variable individually. Only an inter-comparison of two or more differently distributed variables is likely to cause problems. However, the application limitations of the metrics for non-Gaussian distributions will need to be further examined in follow-up studies.

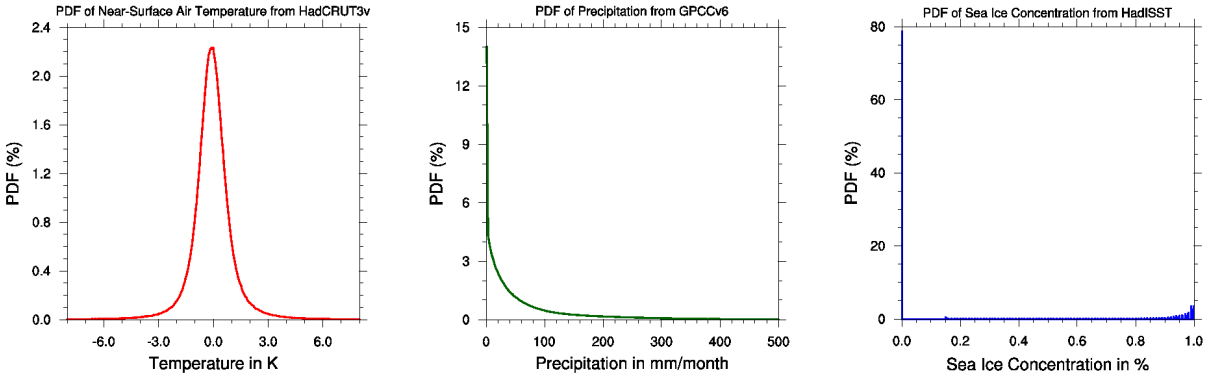

Figure 5.16: *Probability distribution functions (PDF) of global (from left to right) near-surface air temperature anomalies, precipitation over land and sea-ice concentrations for monthly values of all calendar months from different observation datasets. All PDFs were calculated with 1000 bins of equal size.*

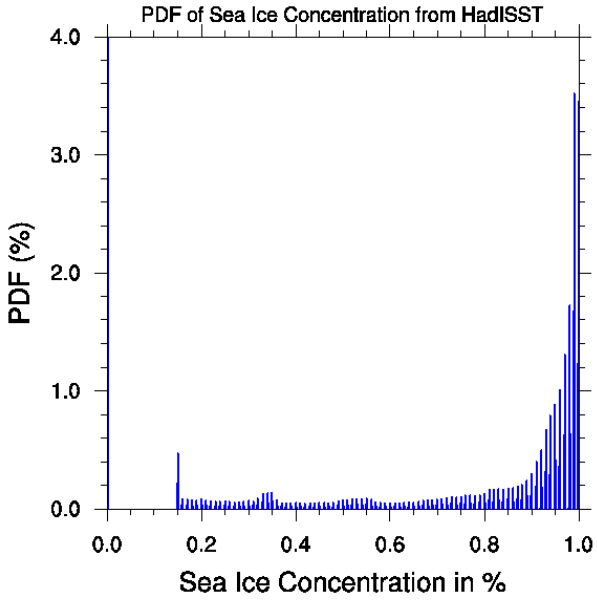

Figure 5.17: *Probability distribution function of sea-ice concentration for monthly means of all calendar months from the HadISST observation dataset. The PDF was calculated with 1000 bins of equal size. The ordinate is only shown up to 4%, as only the bin containing sea-ice concentrations equal to zero has a higher percentage.*

Chapter 6

Summary and Outlook

In this thesis, a quantitative assessment for decadal climate predictions of near-surface air temperature and sea-ice concentration from the low-resolution version of the Max Planck Institute Earth System Model (MPI-ESM-LR) was done. Decadal prediction experiments predicting the near-term future of 10-30 years are carried out with longterm climate models that have to be initialized with observational data due to the dependence on initial conditions. In order to evaluate the science question of whether the initialization of climate models leads to a better predictability of future climate, a verification framework for decadal prediction experiments, introduced by Goddard et al. [2013], has been implemented into the Earth System Model Validation Tool (ESMValTool). The ESMValTool is a software tool developed by multiple institutions which aims at improving routine Earth system model (ESM) evaluation. The verification system allows an assessment of the accuracy, correlation skill and reliability of retrospective decadal forecasts ("hindcasts") compared to observations through the application of the three metrics "mean squared skill score", "anomaly correlation" and "conditional bias", respectively. These metrics were applied to the three MPI-ESM decadal prediction systems, that differ mainly in the employed initialization technique, and the longterm simulations.

Each decadal experiment ensemble is initialized with observational data from a different point in time and then integrated for 10 years without further influence by observations. The hindcasts of the experimental setups evaluated in this thesis are initialized every year from 1960 to 1995. For the evaluation, different time periods ("lead times") within the forecast range of each experiment were assessed in order to explore the dependence of predictability on the distance from the point of initialization.

For both surface air temperature and sea-ice concentration, no further global mean prediction skill can be derived by initializing the MPI-ESM-LR, except for the first year after initialization. Thereafter, the model gradually "forgets" the information content of the observations and drifts towards its preferred biased state. The overall global mean skill of all three metrics is generally comparable to that of the uninitialized long-term projections.

Although for global mean surface air temperature, no increase in predictability by initialization could be found, the opposite occurs on the regional scale for some parts of the globe. For example, for lead years 2-5, decadal predictions of the northwest Atlantic Ocean and southwest of the South American continent show higher correlation coefficients, lower conditional bias

(i.e., higher reliability), and better accuracy than the longterm simulations. The same regionality appears for sea-ice concentration hindcast skill, with the East Laptev Sea and regions within 15° latitude around the North pole showing a significantly improved hindcast skill by initialization for March mean sea-ice concentrations. For September mean Antarctic sea-ice concentration, the Ross Sea could be identified as a region, where the initialization could potentially improve the hindcast skill.

Furthermore, the employed initialization technique can have a strong influence on the prediction skill for both temperature and sea ice. From the three MiKlip decadal prediction systems, the two with additionally initialized atmospheric parameters perform significantly better than the one only initialized with ocean variables.

Additionally, further metrics have to be implemented to more closely evaluate the presented science question. As only deterministic metrics have been employed in this thesis, the verification framework has to be extended by probabilistic metrics as well, like the Continuous Ranked Probability Skill Score (CRPSS). The CRPSS is recommended by Goddard et al. [2013] to address the science question of whether the model's ensemble spread is representative for the prediction uncertainty. Hereby, larger ensembles containing more than only three ensemble members, like presented here, could be helpful to reduce noise and increase the signal-to-noise ratio and thereby the prediction skill [Eade et al., 2014].

Although Jia and DelSole [2012] could identify distinct spatial patterns of predictability of temperature and precipitation that are shared by multiple climate models, additional models and variables have to be included in the assessments to provide a concluding answer to the question, whether the initialization of climate models can lead to a higher predictability of near-future climate change.

To ensure the comparability and reproducibility of the results of such assessments carried out by different science groups, an exact specification of all utilized parameters of the input data (e.g., model and observational data versions, selected ensemble members, or the initialization frequency of the decadal hindcast experiments) and a comprehensive and detailed description of all the preprocessing steps prior to the actual metric calculations are required. Different conventions exist for which techniques to apply in order to adjust the model output, like a mean bias correction or cross-validation. For example, Eade et al. [2014] recently suggested the correction for the Ratio of Predictable Components (RPC) to reduce under-estimation of potential skill and predictability [Eade et al., 2014] as an additional step to the recommendations made by the ICPO's World Climate Research Programme Report [ICPO, 2013]. In contrast, they omit the cross-validation, as they claim it to lead to an under-estimation of the correlation.

Furthermore, Kharin et al. [2012] recommend detrending techniques to avoid inflation of correlation by the capturing of a climate change signal (see also Smith et al. [2014]). As an additional example, in derogation from the suggestions by Goddard et al. [2013], the MurCSS-Tool calculates the difference between the conditional biases of two datasets without consideration of the respective absolute values. This leads to a different interpretation of the results.

All in all, there are many different approaches to evaluate decadal climate predictions. Hence,

there is strong a need for a general and official guideline to ensure comparability and reproducibility of the different studies.

Acknowledgement

This work was performed as part of the BMBF MiKlip Climate Model Validation by confronting globally Essential Climate Variables from models with observations (ClimVal project) and the DLR ESMVal (Earth System Model Validation) project.

First of all, I would like to express my gratitude to Veronika Eyring for enabling this work and supervising me.

Many thanks go to Mattia Righi, who advised me in numerous technical questions and introduced me to the command language NCL.

Big thanks are also due to Dirk Notz and Wolfgang Müller from MPI-M, who supported me in the scientific fields of sea ice research and the evaluation of decadal climate simulations, respectively. I enjoyed a great time in Hamburg.

I want to also thank Sabrina Wenzel, Klaus-Dirk Gottschaldt and Franziska Frank for always having an open ear for me when I had questions of all kinds.

Finally, I thank the Department "Earth System Modeling" of the DLR-IPA for the enjoyable time.

Bibliography

M. A. Balmaseda, K. Mogensen, and A. T. Weaver. Evaluation of the ECMWF ocean reanalysis system ORAS4. *Q.J.R. Meteorol. Soc.*, 139:1132–1161, 2013.

M. Bräu. Sea-ice in decadal and long-term simulations with the Max Planck Institute Earth System Model. Bachelor's thesis, LMU, Munich, 2013.

A. Carrassi, R. J. T. Weber, V. Guemas, F. J. Doblas-Reyes, M. Asif, and D. Volpi. Full-field and anomaly initialization using a low-order climate model: a comparison and proposals for advanced formulations. *Nonlin. Processes Geophys*, 21:521–537, 2014.

D. Cavalieri, C. Parkinson, P. Gloersen, and H. J. Zwally. Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data. Boulder, Colorado USA: NASA DAAC at the National Snow and Ice Data Center, 1996.

D. P. Dee, 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. Hlm, L. Isaksen, P. Kllberg, 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. Thpaut, and F. Vitart. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q.J.R. Meteorol. Soc.*, 137:553–597, 2011.

R. Eade, D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson. Do seasonal-to-decadal climate predictions under-estimate the predictability of the real world? *Geophys. Res. Lett.*, 41, 2014.

I. Eisenman, W. N. Meier, and J. R. Norris. A spurious jump in the satellite record: has Antarctic sea ice expansion been overestimated? *The Cryosphere*, 8:1289–1296, 2014.

M. A. Giorgetta, J. Jungclaus, C. H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, T. Crueger, M. Esch, K Fieg, K. Glushak, V. Gayler, H. Haak, H.-D. Hollweg, T. Ilyina, S. Kinne, L. Kornblueh, D. Matei, T. Mauritsen, U. Mikolajewicz, W. Müller, D. Notz, F. Pithan, T. Raddatz, S. Rast, R. Redler, E. Roeckner, H. Schmidt, R. Schnur, J. Segschneider, K. D. Six, M. Stockhause, C. Timmreck, J. Wegner, H. Widmann, K.-H. Wieners, M. Claussen, J. Marotzke, and B. Stevens. Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Syst.*, 5:572–597, 2013.

L. Goddard, J. W. Hurrell, B. P. Kirtman, J. Murphy, T. Stockdale, and C. Vera. Two Time Scales for The Price Of One (Almost). *Bull. Amer. Meteor. Soc.*, 93:621–629, 2012.

L. Goddard, 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. A verification framework for interannual-to-decadal predictions experiments. *Clim. Dyn.*, 40:245–272, 2013.

S. B. Goldenberg, C. W. Landsea, A. M. Mestas-Nuñez, and W. M. Gray. The Recent Increase in Atlantic Hurricane Activity: Causes and Implications. *Science*, 293:474–479, 2001.

M. M. Holland, E. Blanchard-Wrigglesworth, J. Kay, and S. Vavrus. Initial-value predictability of Antarctic sea ice in the Community Climate System Model 3. *Geophys. Res. Lett.*, 40: 2121–2124, 2013.

M. Hübner. Evaluation of Sea-ice in the Max Planck Institue Earth System Model. Bachelor's thesis, LMU, Munich, 2013.

ICPO - CMIP-WGCM-WGSIP Decadal Climate Prediction Panel, International CLIVAR Project Office Publication Series 150. *Data and Bias Correction for Decadal Climate Predictions*. 5pp, 2013.

T. Ilyina, K. D. Six, J. Segschneider, E. Maier-Reimer, H. Li, and I. Nunez-Riboni. Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth System Model in different CMIP5 experimental realizations. *J. Adv. Model. Earth Syst.*, 5:287–315, 2012.

IPCC - Intergovernmental Panel on Climate Change. *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.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp, 2013.

L. Jia and T. DelSole. Multi-year predictability of temperature and precipitation in multiple climate models. *Geophys. Res. Lett.*, 39, 2012.

P. D. Jones, D. H. Lister, T. J Osborn, C. Harpham, M. Salmon, and C. P. Morice. Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.*, 117, 2014.

J. H. Jungclaus, N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, and J.-S. von Storch. Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI Earth System Model. *J. Adv. Model. Earth Syst.*, 5:422–446, 2013.

J. J. Kennedy, N. A. Rayner, R. O. Smith, M. Saunby, and D. E. Parker. Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850 part 2: biases and homogenisation. *J. Geophys. Res.*, 116, 2011.

V. V. Kharin, G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W.-S. Lee. Statistical adjustment of decadal predictions in a changing climate. *Geophys. Res. Lett.*, 39, 2012.

A. Köhl. Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic. *Q.J.R. Meteorol. Soc.*, 2014.

G. A Meehl, L. Goddard, J. Murphy, R. J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. A. 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. Decadal prediction: Can it be skillful? *Bull. Amer. Meteorol. Soc.*, 90:1467–1485, 2009.

G. A. Meehl, A. Hu, J. M. Arblaster, J. Fasullo, and K. E. Trenberth. Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation. *J. Climate*, 26:7298–7310, 2013.

G. A. Meehl, W. M. Washington, J. M. Arblaster, A. Hu, H. Teng, C. Tebaldi, B. N. Sanderson, J.-F. Lamarque, A. Conley, W. G. Strand, and J. B. White III. Climate System Response to External Forcings and Climate Change Projections in CCSM4. *J. Climate*, 25:3661–3683, 2014.

W. A. Müller, J. Baehr, H. Haak, J. H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J. S. von Storch, and J. Marotzke. Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.*, 39, 2012.

W. A. Müller, H. Pohlmann, K. Kulkarni, K. Modali, F. Vamborg, L. Kornblueh, D. Kleberg, and J. Marotzke. Baseline and prototype: Two exercises to implement global decadal climate predictions. MiKilp Statusseminar, Karlsruhe, 2014.

A. H. Murphy. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Mon. Wea. Rev.*, 116:2417–2424, 1988.

D. Notz, F. A. Haumann, H. Haak, J. H. Jungclaus, and J. Marotzke. Arctic sea-ice evolution as modeled by Max Planck Institute for meteorologys Earth system model. *J. Adv. Model. Earth Syst.*, 5, 2013.

H. Pohlmann, J. H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke. Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Clim.*, 22:3926–3938, 2009.

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

P. Rampal, J. Weiss, C. Dubois, and J.-M. Campin. IPCC climate models do not capture Arctic sea ice drift acceleration: Consequences in terms of projected sea ice thinning and decline. *J. Geophys. Res.*, 116, 2011.

N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, 108, 2003.

M. Rieck, L. Nuijens, and B. Stevens. Marine boundary layer cloud feedbacks in a constant relative humidity atmosphere. *J. Atm. Sci.*, 69:2538–2550, 2012.

M. Righi, V. Eyring, K.-D. Gottschaldt, C. Klinger, F. Frank, P. Jckel, and I. Cionni. Quantitative evaluation of ozone and selected climate parameters in a set of EMAC simulations. *Geosci. Model Dev. Discuss.*, 7:6549–6627, 2014.

U. Schneider, A. Becker, P. Finger, A. Meyer-Christoffer, B. Rudolf, and M. Ziese. GPCC Full Data Reanalysis Version 6.0 at 2.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data. 2011.

D. M. Smith, S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317:796–799, 2007.

D. M. Smith, A. A. Scaife, G. J. Boer, M. Caian, F. J. Doblas-Reyes, V. Guemas, E. Hawkins, W. Hazeleger, L. Hermanson, C. Kit 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. Real-time multi-model decadal climate predictions. *Clim. Dyn.*, 41:2875–2888, 2014.

A. Solomon, L. Goddard, A. Kumar, J. Carton, C. Deser, I. Fukumori, A. M. Greene, G. Hegerl, B. Kirtman, Y. Kushnir, M. Newman, D. Smith, D. Vimont, T. Delworth, G. A. Meehl, and T. Stockdale. Distinguishing the Roles of Natural and Anthropogenically Forced Decadal Climate Variability. *Bull. Amer. Meteor. Soc.*, 92:141–156, 2011.

B. Stevens and O. Boucher. Climate science: The aerosol effect. *Nature*, 490:40–41, 2012.

J. C. Stroeve, M. M. Holland, W. Meier, T. Scambos, and M. Serreze. Arctic sea ice decline: Faster than forecast. *Geophys. Res. Lett.*, 34, 2007.

J. C. Stroeve, V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier. Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophys. Res. Lett.*, 39, 2012.

K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of CMIP5 and the experiment design. *Bull. Amer. Meteorol. Soc.*, 92:485–498, 2012.

J. Turner, T. J. Bracegirdle, T. Phillips, G. J. Marshall, and J. S. Hosking. An Initial Assessment of Antarctic Sea Ice Extent in the CMIP5 Models. *J. Climate*, 26, 2013.

S. Valcke. The OASIS3 coupler: a European climate modelling community software. *Geosci. Model Dev.*, 6:373–388, 2013.

D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, Volume 100 of International geophysics series, ISSN 0074-6142, 2011.

Declaration of Originality

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig verfasst zu haben und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel benutzt zu haben.

München,

_____