Arctic sea ice in Earth system models: decadal hindcast skill and constraints of long-term projections

Daniel Senftleben

Oberpfaffenhofen 2019

Arctic sea ice in Earth system models: decadal hindcast skill and constraints of long-term projections

Dissertation

an der Fakultät für Physik

der Ludwig-Maximilian-Universität

München

vorgelegt von

Daniel Senftleben

aus Dachau

München, 29. Januar 2019

Erstgutachter: Prof. Dr. Veronika Eyring

Zweitgutachter: Prof. Dr. George Craig

Datum der mündlichen Prüfung: 22. März 2019

Contents

Abstract v

Kurzfassung vii

1. Introduction 1
   1.1 The Arctic in a changing climate 2
   1.2 Earth system model simulations and evaluation 5
   1.3 Scientific questions and investigation strategies 8
      1.3.1 Decadal hindcast skill in the Arctic compared to historical simulations 9
      1.3.2 The potential to narrow uncertainties in projections of Arctic sea ice extent 10

2. Decadal hindcast skill in the Arctic 11
   2.1 Previous skill improvements achieved with decadal hindcast experiments 12
   2.2 Model experiments and methods 13
      2.2.1 Model simulations 13
      2.2.2 Observations and reanalysis data 17
      2.2.3 Methods 18
   2.3 Results 20
      2.3.1 Hindcast skill for near-surface air temperature in the North Atlantic 20
      2.3.2 Timeseries of pan-Arctic sea ice area 22
      2.3.3 Hindcast skill in the Arctic 24
   2.4 Summary and discussion 31

3. Constraining uncertainties in projections of Arctic sea ice extent 35
   3.1. State-of-the-art of uncertainty reduction in Arctic sea ice projections 36
   3.2. Methods 37
      3.2.1. Multiple Diagnostic Ensemble Regression (MDER) 37
      3.2.2. Diagnostics 40
      3.2.3. Uncertainty estimation 42
      3.2.4. Models, observations and reanalyses 43

3.3. Application of MDER to Arctic sea ice projections 46
3.3.1. Diagnostic selection and regression model 47
3.3.2. Uncertainty estimation and the potential to narrow model uncertainty 50
3.3.3. Weighting SIE projections and estimating YOD 54
3.4. Summary and discussion 56

4. Summary and outlook 59
4.1. Summary of the results 59
4.2. Conclusions and outlook 61

A. Acronyms 63

B. Supplemental information on Chapter 3 65

C. Integrated author's references 73

References 74

Acknowledgments 83

Abstract

The summer Arctic sea ice cover has retreated by about 50% in the last 30 years. This rate of change is unprecedented in Earth's history and is at least partly due to an anthropogenic impact. It has critical implications for the climate worldwide. Earth system models are parameterized computer models that aim at simulating and projecting global climate and its change. Experiments with these models under common forcings and protocols are coordinated as part of the Coupled Model Intercomparison Project (CMIP). However, uncertainties in CMIP simulations of the Arctic climate are large in both, short- and long-term projections. This thesis investigates how observations can be used to improve Arctic climate simulations by analyzing and evaluating Arctic climate parameters in different types of climate simulations from Earth system models participating in CMIP Phase 5 (CMIP5). For this purpose, the Earth System Model Evaluation Tool (ESMValTool) developed by scientists worldwide to facilitate the challenging evaluation of Earth system models, is enhanced and applied.

The thesis consists of two parts that are based on studies conducted by the thesis author. Part one investigates whether the initialization of Earth system model simulations with observations leads to more skillful predictions of Arctic climate. It focuses on the relatively new field of decadal climate predictions that have a forecast horizon of only 10 years. In contrast to long-term climate simulations, decadal climate predictions are initialized with observations so that the simulations start from the observed phase of natural variability. To analyze possible improvements in prediction skill by initialization, a complex verification system for retrospective decadal predictions ("hindcasts") was developed, implemented into the ESMValTool, and applied to Arctic hindcasts of an Earth system model. It is shown that initialization improves surface temperature skill in the North Atlantic and along the east coast of Greenland, with root-mean-square errors against observations reduced by up to 50%. Accordingly, the skill for integrated sea ice area was increased in the Greenland Sea. However, these improvements were only found for winter and only for the first few forecast years after initialization. In other Arctic regions and other seasons, no significant improvements were found. On the contrary, especially in the marginal ice zone and the Kara and Barents Seas, there is significant degradation in skill by initialization. This can be explained by the so-called initialization shock, a mechanism that causes the model to drift to and overshoot its biased intrinsic climatology after the initialization. The thesis also identifies ways to improve decadal climate predictions in the Arctic, for example by including sea ice parameters in the initialization process and by additionally initializing the model in summer instead of only at the end of the year.

Part two of the thesis focuses on the long-term development of the Arctic and is aimed at reducing uncertainties in multi-model Earth system model projections of the 21st century Arctic sea ice extent. The multiple diagnostic ensemble regression (MDER) method is used to constrain the CMIP5 projections with observations. The main advantage of the MDER method is the iterative regression algorithm that selects the diagnostics that best predict the future target variable from a set of given diagnostics and calculates a regression model to constrain the projections with observations. Applying this method to Arctic sea ice projections from 29 CMIP5 models, model uncertainty in multi-model mean results for the September Arctic sea ice extent could be reduced by up to 50%. Furthermore, model weighting reduces the projected multi-model mean sea ice extent by 20% (1 million km²) and predicts an earlier near-disappearance of Arctic sea ice by more than a decade (from 2076 to 2062) for a high greenhouse gas concentration scenario compared to the unweighted multi-model mean. This suggests a faster retreat of Arctic sea ice than previously estimated. However, the results of the thesis also show that model uncertainty – together with scenario uncertainty and irreducible internal variability – remains too large for exact predictions of the first year of near-disappearance of summer Arctic sea ice.

The thesis demonstrates the potential of improving Earth system model simulations of Arctic climate with observations. Initializing climate models with observations is a relatively new field that needs to be further investigated. Finding emergent constraints for Arctic climate parameters may greatly enhance the potential reduction in uncertainties of multi-model projections when using observation-based analysis methods like MDER. While projection uncertainties are still large, most studies, including this thesis, suggest a more pessimistic future for Arctic sea ice. Hence, the conclusions are that mitigation strategies to reduce Arctic warming need to be intensified and that it becomes increasingly crucial to further improve the understanding of the Arctic climate system.

Kurzfassung

Die Fläche des arktischen Sommermeereises hat in den letzten 30 Jahren um ca. 50% abgenommen. Diese Veränderungsrate ist in der Erdgeschichte beispiellos und ist zumindest teilweise auf einen anthropogenen Einfluss zurückzuführen. Die Veränderungen in der Arktis haben enorme Auswirkungen auf das weltweite Klima. Erdsystemmodelle sind parametrisierte Computermodelle, die das Ziel haben, das globale Klima und dessen Veränderungen zu simulieren. Das Coupled Model Intercomparison Project (CMIP) koordiniert die Forschungsexperimente mit diesen Modellen unter gemeinsamen Leitlinien. Die Unsicherheiten in den CMIP-Simulationen des arktischen Klimas groß sind – sowohl auf Kurzzeit- als auch auf Langzeitskala. In dieser Dissertation wird untersucht, wie Beobachtungen zur Verbesserung der arktischen Klimasimulationen genutzt werden können. Hierfür werden arktische Klimaparameter aus verschiedenen Typen von Klimasimulationen der Erdsystemmodelle aus CMIP Phase 5 (CMIP5) mit Beobachtungen analysiert und evaluiert. Zu diesem Zweck wird das Earth System Model Evaluation Tool (ESMValTool) erweitert und benutzt, das von Wissenschaftlern weltweit dazu entwickelt wird, die komplexe Evaluation von Erdsystemmodellen zu erleichtern.

Die Dissertation besteht aus zwei Teilen, die auf Studien basieren, die vom Autor der Dissertation durchgeführt wurden. In Teil eins wird untersucht, ob die Initialisierung von Erdsystemmodellsimulationen mit Beobachtungen zu erhöhter Vorhersagegenauigkeit („Skill“) des arktischen Klimas führt. Der Fokus liegt dabei auf dem relativ neuen Feld der dekadischen Klimavorhersagen, die einen Vorhersagehorizont von nur zehn Jahren haben. Im Gegensatz zu Langzeitsimulationen werden dekadische Klimavorhersagen mit Beobachtungsdaten initialisiert, so dass sie von der beobachteten Phase natürlicher Klimaschwankung aus starten. Um mögliche Verbesserungen im Skill durch Initialisierung zu untersuchen, wurde ein vielschichtiges Verifikationssystem für retrospektive dekadische Klimavorhersagen („Hindcasts“) entwickelt, in das ESMValTool implementiert und auf dekadische Hindcasts eines Erdsystemmodells angewendet. Es zeigt sich, dass die Initialisierung den Vorhersageskill von Oberflächentemperaturen im Nordatlantik und entlang der Ostküste von Grönland verbessert, wobei die mittlere quadratische Abweichung zu Beobachtungen um 50% reduziert wird. Dementsprechend hat sich der Skill für die integrierte Meereisfläche im Grönlandmeer erhöht. Allerdings zeigen sich diese Verbesserungen nur im Winter und nur für wenige Jahre nach Initialisierung. Tatsächlich zeigt sich eine signifikante Skillabnahme besonders am Meereisrand und in der Kara- und Barentssee. Dies kann mit dem sogenannten Initialisierungsschock erklärt werden - einem Mechanismus, der das Modell dazu bringt, nach der Initialisierung zu seinem verzerrten intrinsischen Klimazustand zurück zu driften und diesen zu überschießen. Die Arbeit zeigt

Möglichkeiten auf, die dekadischen Klimavorhersagen in der Arktis zu verbessern, wie beispielsweise durch eine Berücksichtigung von Meereisparametern beim Initialisierungsprozess und eine zusätzliche Initialisierung des Modells im Sommer anstatt nur am Ende des Jahres.

Teil zwei der Dissertation beschäftigt sich mit der Langzeitentwicklung der Arktis und hat das Ziel, Unsicherheiten in Klimaprojektionen der Entwicklung der arktischen Meereisausdehnung im 21. Jahrhundert zu reduzieren. Hierfür wird von der Multiple Diagnostic Ensemble Regression (MDER) Methode Gebrauch gemacht, die mithilfe von Beobachtungen die Bandbreite der CMIP5-Klimaprognosen einschränkt. Der zentrale Vorteil von MDER gegenüber anderen Analysemethoden ist der iterative Regressionsalgorithmus, der diejenigen Diagnostiken aus einer Menge an vorgegebenen Diagnostiken auswählt, die die Zukunftsvariable am besten vorhersagen können. Mit dem hieraus errechneten Regressionsmodell lassen sich dann die Prognosen mittels Beobachtungen einschränken. Durch die Anwendung dieser Methode auf Prognosen des arktischen Meereises von 29 CMIP5-Modellen konnte die Modellunsicherheit in Multimodellsimulationen der Ausdehnung des arktischen Septembermeereises um bis zu 50% verkleinert werden. Darüber hinaus zeigt das beschränkte Multimodellmittel im Vergleich zum ungewichteten Mittel eine um ca. eine Million km² (20%) kleinere Eisfläche und ein früheres Verschwinden des arktischen Sommermeereises um mehr als eine Dekade (von 2076 auf 2062) in einem hohen Treibhausgasszenario. Dies deutet auf einen schnelleren Rückgang des arktischen Meereises hin, als bisher angenommen. Allerdings zeigen die Ergebnisse der Dissertation auch, dass die Modellunsicherheit – zusammen mit nicht-reduzierbarer interner Klimavariabilität und Unsicherheiten durch verschiedene Annahmen zukünftiger Emissionsszenarien – zu groß für genaue Vorhersagen des Zeitpunkts bleibt, an dem das arktische Sommermeereis zum ersten Mal verschwindet.

Die Dissertation legt das große Potential dar, Erdsystemmodellsimulationen des arktischen Klimas mittels Beobachtungen zu verbessern. Klimamodelle mit Beobachtungen zu initialisieren, ist ein relativ neues Forschungsgebiet, das noch weiterer Erforschung bedarf. Das Auffinden sogenannter „emergent constraints“ für arktische Klimaparameter kann die Reduktion von Unsicherheiten mittels beobachtungsbasierten Analysemethoden wie MDER stark verbessern. Obwohl Prognoseunsicherheiten noch immer groß sind, deuten die meisten Studien, so wie diese Arbeit, auf eine pessimistischere Zukunft des arktischen Meereises hin. Die Arbeit zeigt somit, dass Mitigationsstrategien gegen die arktische Klimaerwärmung weiter ausgebaut werden müssen und es zunehmend wichtiger wird, das Klimasystem der Arktis besser zu verstehen.

1. Introduction

"We need to save the Arctic not because of the polar bears, and not because it is the most beautiful place in the world, but because our very survival depends upon it."

Lewis Gordon Pugh, 2007

These were the words of British environmentalist and endurance swimmer Lewis Gordon Pugh after his 2007 swim across the Geographic North Pole in – 1.7 °C cold Arctic water (Lamb 2008). He was the first human being to swim at the North Pole and his goal was to raise awareness of the fast-shrinking ice cover of the Arctic Ocean, his favorite place on Earth.

This thesis aims to contribute to the world-wide research effort on understanding and predicting the recent and future development of the Arctic climate, with a focus on Arctic sea ice. It is a model evaluation and analysis study examining climate simulations from state-of-the-art Earth system models together with observations to assess the models' ability to reproduce the observed Arctic climate and to constrain their future projections. More specifically, the thesis investigates how observations can be used to improve simulations of the Arctic climate and is focused on two scientific topics. Firstly, possible benefits from a relatively new type of climate simulations that are initialized with observations are assessed for selected Arctic climate parameters. And secondly, the potential of observation-based constraints to narrow uncertainties the multi-model projections of the 21st century development of Arctic sea ice is investigated.

This introductory chapter describes the importance of Arctic sea ice for Earth's climate, explains the basics of global climate simulations, presents the two main scientific questions of this thesis, and outlines how the analysis is performed.

1.1 The Arctic in a changing climate

The Arctic is most commonly defined as the area north of the Arctic Circle (approx. 66° 34' N latitude[1]). It consists of the Central Arctic Ocean and its adjacent seas, surrounded by land belonging to eight different countries (Norway, Sweden, Finland, Russia, the United States, Canada, Greenland, and Iceland). Its ecosystem is unique on Earth and the roughly four million human inhabitants[2] have adapted to its extremely cold conditions.

The Arctic Ocean has been at least partly covered by sea ice without interruption for the past 2.6 million years (Knies et al. 2014). Sea ice is a key component of Earth's climate system (Notz 2015). It forms from freezing sea water and directly influences the large-scale atmospheric and oceanic circulation. Sea ice is distinctively colder than the surrounding open ocean water, altering the heat flux from the ocean into the atmosphere (Singarayer et al. 2006). Without sea ice, the heat flow into the atmosphere is mainly driven by turbulent heat fluxes from the ocean and can amount up to hundreds of W m^{-2} in winter. In contrast, the heat flux with sea ice present consists almost entirely of heat conduction, leading to heat fluxes into the lower atmosphere that are about an order of magnitude smaller in similar conditions, depending on the ice thickness. In summer, radiative processes have a stronger impact on atmospheric temperatures: since the reflectivity of sea ice is about five times higher than that of open sea water - and even more than that when it is covered with snow - sea ice effectively cools the Arctic, especially in cloud-free regions (Perovich 2002). Together, both effects result in a strong temperature gradient between Arctic and midlatitudes in both, summer and winter. Sea ice also impacts on the oceanic circulation by altering the input of kinetic energy from wind stress and of potential energy by affecting ocean temperature and salinity. Formation of sea ice creates high-saline brine that drains into the ocean, increasing the salinity and thereby the density of the upper ocean layer in winter (Weeks 2010). This creates a downwelling motion and forms cold, high-saline bottom water, increasing the outflow into adjacent seas and driving the global ocean circulation (Aagaard and Carmack 1989). In contrast, melting ice in summer creates a low-saline upper ocean layer, stabilizing the ocean stratification.

The Arctic has not always been ice covered. We know from ocean sediment cores that throughout Earth's history, there have been warm periods without any sea ice, as well as ice ages with extensive ice coverage (Polyak et al. 2010). These changes are caused by natural variations in solar radiation input, the composition of the atmosphere and changes in ocean circulation and surface albedo. However, these changes have taken place over the course of millennia. In contrast, Arctic sea ice has been retreating rapidly during the last 30 years: satellite observations show that the summer Arctic sea ice extent was reduced by half since

[1] https://nsidc.org/cryosphere/arctic-meteorology/arctic.html
[2] https://arctic.ru/population/

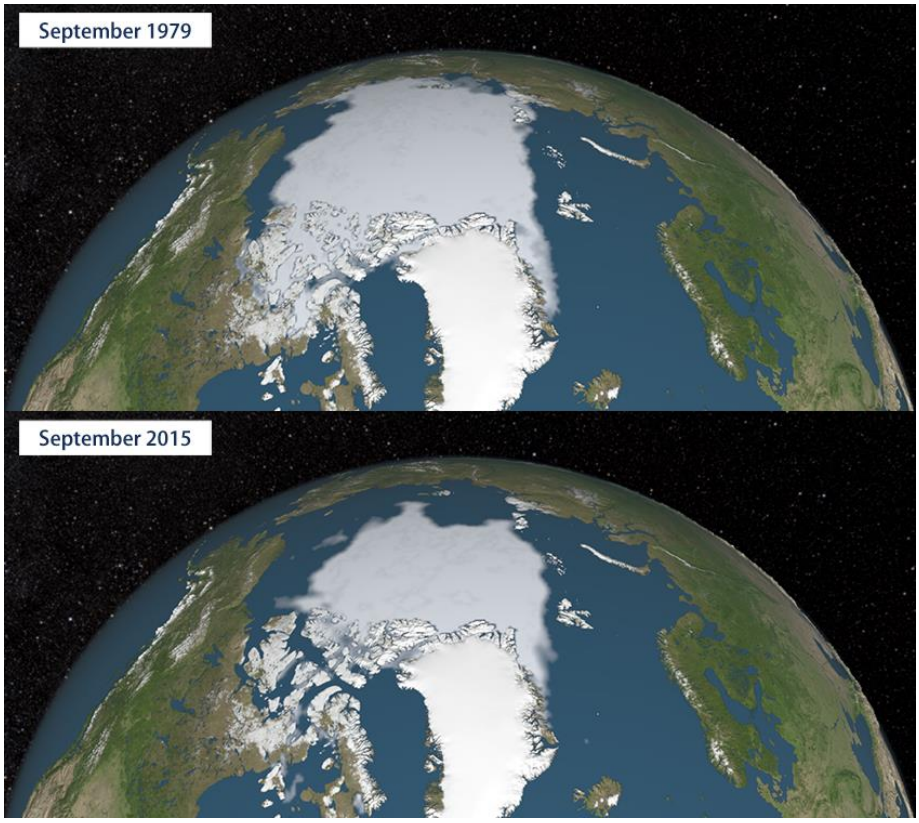

Figure 1.1. The retreat of summer Arctic sea ice between September 1979 and September 2015. Figure from NASA[3].

the 1980s (e.g., Stroeve et al. 2007) from almost 8 million km² in September 1980 to less than 4 million km² in 2012. The strong reduction in September Arctic sea ice extent is visualized by NASA in Figure 1.1, based on satellite data. The eight years with the lowest Arctic sea ice extents ever observed were all recorded after 2006 (NASA/Goddard-Space-Flight-Center 2018). This rate of change is unprecedented in what is known from Earth's history, and is unlikely to have been caused by natural variability alone (Ding et al. 2017). Instead, it is one of the most striking indicators of global warming, which is most likely due to increasing greenhouse gas concentrations in the atmosphere since the industrial revolution (IPCC 2013). In fact, climate change has manifested in the Arctic more than anywhere else on Earth. Since 1900, Arctic air temperatures have risen by about twice the amount of the global mean (Bellucci et al. 2015). The accelerated warming in the Arctic is called polar amplification, and is mainly due to the ice-albedo feedback (Holland and Bitz 2003): retreating sea ice exposes more open sea water, reducing the surface albedo in the Arctic. Thus, more incoming solar radiation is absorbed increasing the heat input into the upper ocean layer, which in turn melts more sea ice, and the feedback loop starts over again. These

[3] NASA (National Aeronautics and Space Administration). 2016. NASA's Goddard Space Flight Center Scientific Visualization Studio. https://svs.gsfc.nasa.gov

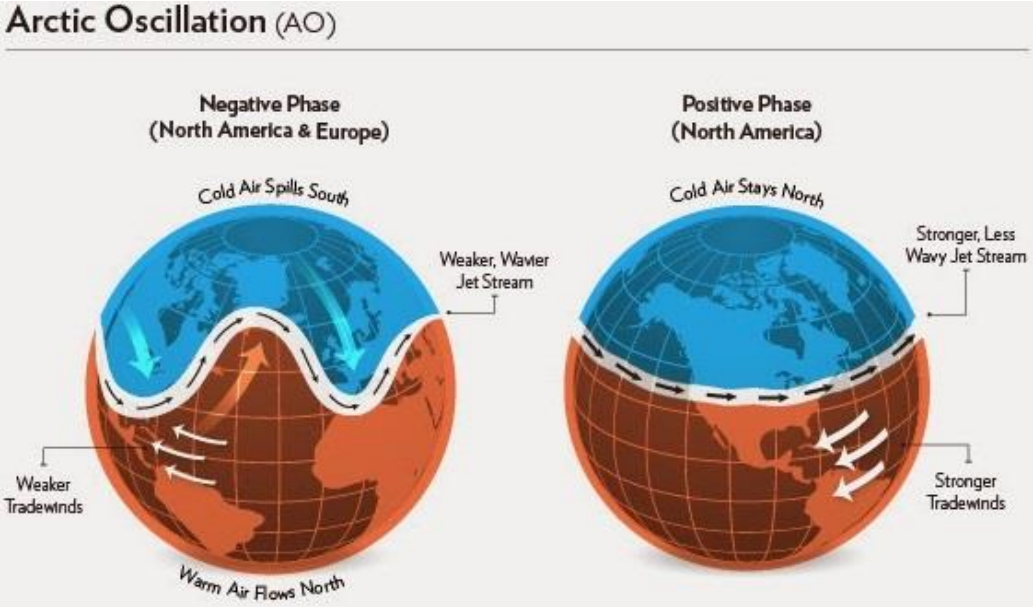

Figure 1.2. The different configurations of the Arctic jet stream in the different phases of the Arctic Oscillation[4]. A shift to a more frequent negative phase with increased exchange of heat and moisture between the Arctic and the midlatitudes is expected (Notz 2015).

effects also cause the ice to become thinner: about 70% of the Arctic sea ice now is seasonal ice (Kwok 2018), which is decisively more sensitive to climate forcing than multi-year pack ice. Both, the decrease in sea ice cover and the thinning of the ice, lead to a reduction of the thermodynamic insulation of the air masses over the Arctic Ocean (see above), which further weakens the temperature gradient between Arctic and midlatitudes. This leads to another feedback loop that amplifies the Arctic warming by a dynamical mechanism: the smaller temperature gradient weakens the Arctic jet stream and destabilizes it (Francis and Vavrus 2015). Thus, the jet's streamlined shape changes to a more frequent high-amplitude configuration. The wavier shape further amplifies the exchange of heat between Arctic and midlatitudes (see also Figure 1.2).

In summary, the thermodynamic, radiative, and dynamical processes discussed above all lead to an accelerated warming in the Arctic, which not only affects Arctic inhabitants and stakeholders, but also has critical implications on a global scale (Vihma 2014). For example, the melting of Arctic in-land ice sheets and glaciers (e.g., Greenland) leads to global sea-level rise (e.g., Church et al. 2004; Holgate and Woodworth 2004; Jevrejeva et al. 2006). Furthermore, the thawing of sub-sea permafrost in the shallow Arctic shelves is expected to release high amounts of methane (e.g., Archer and Buffett 2005; Shakhova et al. 2007), a greenhouse gas that has a 3.7-times higher global warming potential than CO_2 (Lashof and Ahuja 1990). Additionally, North American and European weather patterns are influenced by the weakening of the polar jet stream, leading to an increase in extreme events, especially in

[4] http://weatherwilly.blogspot.com/2014/10/arctic-oscillation-set-to-tank.html

the persistence of winter cold spells in North America (Francis et al. 2017). Finally, changes in salinity and temperature in the Arctic Ocean may alter the global ocean circulation, since they affect the connections of the Arctic Ocean to the Atlantic and the Pacific Ocean (e.g., Shimada et al. 2006; Våge et al. 2008).

All of the above shows that it is critical to understand the future evolution of the Arctic. Numerical climate models aim to simulate the past and future development of the most relevant climate variables of the Earth system. The following section describes the fundamentals of climate modelling and the evaluation of climate simulations with observations.

1.2 Earth system model simulations and evaluation

Coupled climate models that simulate the main components of Earth's climate system (atmosphere, ocean, land surface and ice, as well as the carbon cycle and possibly additional Earth system components) are called Earth system models. Earth system model simulations (commonly called "climate simulations") are solutions to differential equations of thermodynamics and fluid motion of Earth's climate variables in space and time[5]. In the code of the computer models, the underlying mathematical, physical, biological, and chemical principles are parameterized since constraints in computing power make explicit solutions of all equations impossible. The model output is comprised of time-space resolved prognostic variables and typically provided as gridded datasets.

One component of Earth system models is a sea ice module. There are sea ice modules of varying degrees of complexity, from simple thermodynamic models to more sophisticated ones that also include dynamic processes. The quality of a sea ice model is strongly dependent on (a) the quality of the atmosphere and ocean model components that provide to the sea ice model the relevant parameters, such as air and sea surface temperatures, salinity, wind, and radiation quantities, and (b) on the parameterization schemes used, for example for lead opening, the albedo scheme, and the interfaces between the model components.

The parameterized thermodynamic and dynamic processes of the sea ice model determine the sea ice quantities for each grid cell. The thermodynamic part of a sea ice model calculates the ice formation (melt and growth) and is based on the principle of conservation of energy. A schematic overview of the relevant thermodynamic processes in a sea ice model is given in Figure 1.3 (Tsamados et al. 2015). For example, the sea ice thickness in each grid cell is determined by a balance between heat conduction ($F_{condbot}$, $F_{condtop}$) and incoming

[5] https://soccom.princeton.edu/content/what-earth-system-model-esm

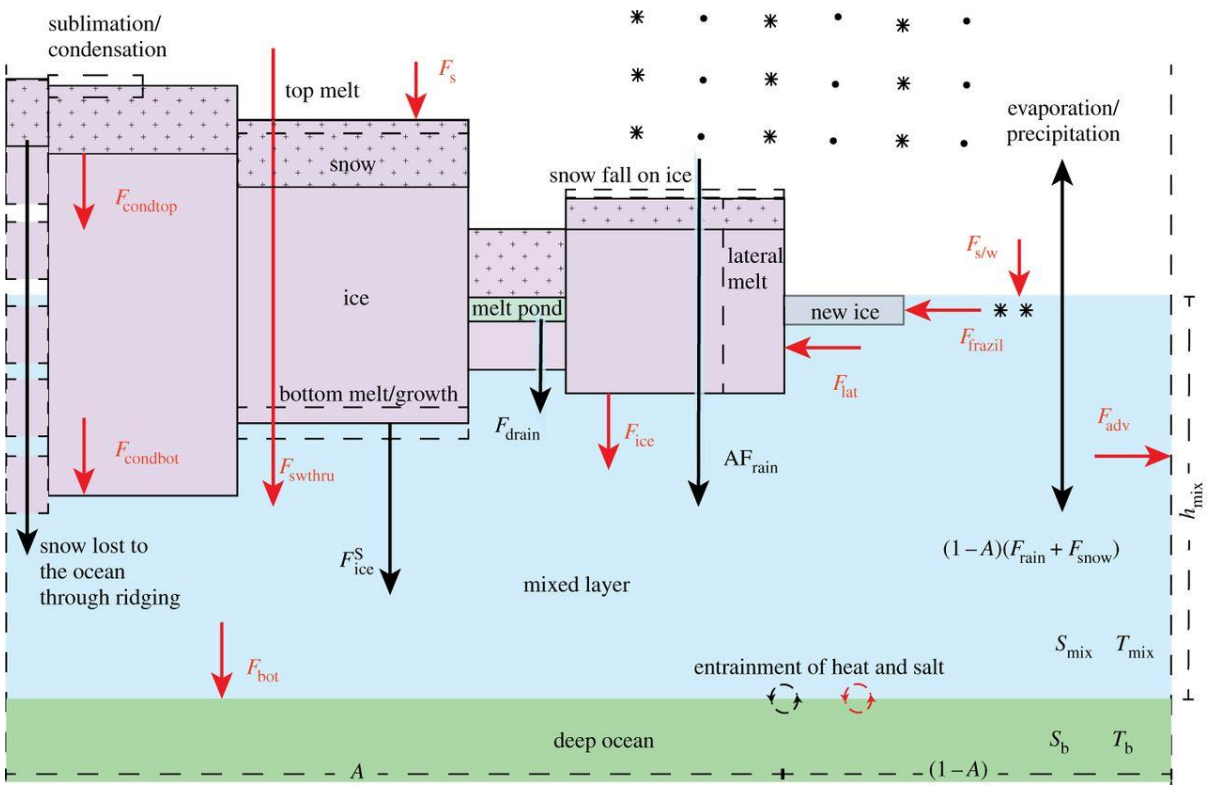

Figure 1.3. Schematic of the thermodynamic sea ice processes parameterized in the Los Alamos community sea ice model (CICE). Figure from Tsamados et al. (2015).

fluxes (F_{ice}, F_s) both from the top and the bottom of the ice. The dynamic part of a sea ice model determines the ice motion (including the ice import to the Arctic basin and ice export to the Atlantic/Pacific Ocean) and is based on the principle of conservation of momentum.

Different parametrizations used in the different components of Earth system models lead to sometimes substantial differences in their simulated climate. The Coupled Model Intercomparison Project (CMIP) coordinates and standardizes the world-wide research effort on climate modelling. It was established in 1995 by the World Climate Research Program (WCRP) to improve climate modelling and to assess climate change. Providing, for example, the data basis for the Intergovernmental Panel on Climate Change's (IPCC) assessment reports (e.g, IPCC 2013), CMIP has become an essential component of global climate science. One common standard of CMIP is that each model calculates a number of different realizations (so-called ensemble members) that only slightly differ in their initial conditions to simulate natural variability under common forcings and protocols. This thesis uses data output from about 30 Earth system models that participated in CMIP Phase 5 (CMIP5, Taylor et al. 2012). Globally, simulations of the 20th-century climate until 2005 (so-called historical simulations) from the CMIP5 model ensemble generally show good agreement with observations, for example in the representation of temperature and precipitation trends (Kumar et al. 2013). Figure 1.4 shows timeseries of September Arctic sea ice extent (SIE)

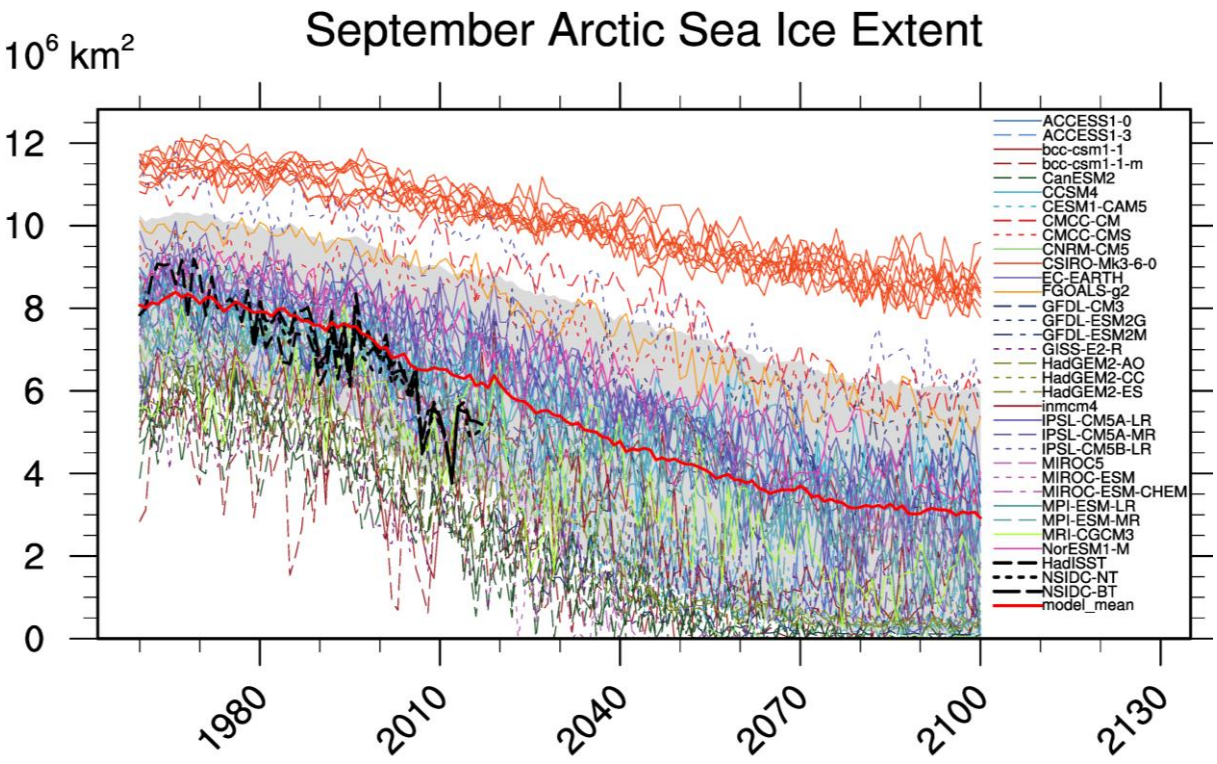

Figure 1.4. Timeseries (1960-2100) of September Arctic sea ice extent (million km²) as simulated by the majority of CMIP5 models (colored lines) and as observed by different instruments (black lines). From 2006 on, the projections assume the RCP 4.5 scenario. Individual ensemble members of each model are shown with the same color. The thick red line is the multi-model mean and the grey shading indicates the models' standard deviation. Figure updated from Lauer et al. (2017), their Figure 5.

from 31 Earth system models that participated in CMIP5. SIE is calculated as the integrated area over all oceanic grid cells in the Arctic with at least 15% sea ice coverage and the monthly mean September SIE is commonly chosen to represent yearly minimal sea ice conditions in the Arctic. The timeseries reveal that the CMIP5 inter-model spread is large for SIE simulations, with a difference of about 9 million km² between the model with the largest simulated SIE and the model with the smallest SIE, a number twice as large as the observed 2018 September SIE (NASA/Goddard-Space-Flight-Center 2018). Similarly, the standard deviation of the CMIP5 ensemble averages at ± 2 million km² around the multi-model mean. The main reason for the large spread in Arctic sea ice simulations are differences in model structure and parameterization schemes in the sea ice modules and in the atmospheric and oceanic model components. When compared to the observations, the multi-model mean realistically represents the observed mean SIE between 1960 and 2000, but shows an unrealistically small trend after that time period. Accordingly, Stroeve et al. (2012b) find that the observed negative trend in Arctic SIE is underestimated by most CMIP5 models, with 16% of them simulating no trend at all.

Nevertheless, a physically plausible connection between the Arctic sea-ice retreat and the increase in atmospheric carbon dioxide concentrations can be confirmed by model simulations (e.g., Notz and Marotzke 2012). Using the Earth system models to project the future evolution of the Arctic sea ice is challenging since the projections are subject to large uncertainties. To represent the influence of the unforeseeable future greenhouse gas emissions, the CMIP5 protocol divides the 21st century climate projection experiments into four sets, each of which assumes a different future greenhouse gas concentration scenario as external forcing for the projections. These scenarios are called representative concentration pathways (RCP) and are assigned a number equal to the amount of radiative forcing in W m^{-2} by the end of the year 2100 (van Vuuren et al. 2011). The RCPs cover the range of scenarios from assuming strong emission mitigation (RCP 2.6) to "business as usual" (RCP 8.5). The climate projections in Figure 1.4 after 2006 are calculated assuming the moderate greenhouse gas concentration scenario RCP 4.5.

Quantitatively assessing the realism of the Earth system model simulations is a pressing need among the climate science community and for the public. The open-source software package Earth System Model Evaluation Tool (ESMValTool, Eyring et al. 2016b) was developed to facilitate the routine evaluation of Earth system models by providing diagnostics and performance metrics for selected variables and processes. It is a community-developed software tool that uses observational data to benchmark the models either against other models or predecessor versions of the same model. The ESMValTool also contains statistical methods to assess and improve multi-model climate projections for example by reducing model uncertainty in the multi-model mean results. ESMValTool v1.0 was released in 2016, with contributions from over 30 scientists and developers from over 20 different institutions world-wide. The ESMValTool was used and enhanced for this thesis. Where not stated otherwise, all figures in this thesis were created with ESMValTool v1.0. The implementations made by the author are documented in Eyring et al. (2016b) and Lauer et al. (2017). ESMValTool v2.0 will be released in 2019 (Eyring et al. 2019a) and is particularly aimed at analyzing model output from CMIP6 (Eyring et al. 2016a). ESMValTool v2.0 will also include the diagnostics of this thesis.

The different past and future simulations of Arctic climate from the different Earth system models together with observations of different kinds form the data basis of this thesis to tackle its two main scientific questions, which are described in the following section.

1.3 Scientific questions and investigation strategies

Uncertainties in Earth system model simulations of Arctic climate are large in both the historical simulations and the future projections, and in both short-term predictions and long-term projections. The overarching goal of the thesis is to investigate how observations

can improve the simulations of Arctic climate in three ways: for model evaluation, model initialization, and constraining model projections. The thesis consists of two main parts. Part one addresses the past, late 20th-century Arctic climate and evaluates the models' skill, i.e., the agreement of model simulations with observations, of a particular type of climate simulations that are initialized with observations and have a prediction horizon of only 10 years. Part two focuses on the Arctic's future development and uses observations and the information on model skill from historical climate simulations to improve multi-model mean long-term projections of the 21st century. Each of the two parts is based on a publication in a peer-reviewed journal; the lead author of both of these is the author of this thesis (see Appendix C). For all analyses, the ESMValTool has been enhanced and then applied.

1.3.1 Decadal hindcast skill in the Arctic compared to historical simulations

Decadal climate predictions are a relatively new type of climate model experiment that have been included into the CMIP5 protocol. Whereas long-term climate simulations aim to calculate all aspects of the world's climate over centuries from a parameterized physical and chemical framework and given boundary conditions (like external forcings), the forecast horizon of decadal climate predictions is years to a decade. On these timescales, the natural variability of the climate system is a crucial factor. Decadal climate predictions consider this factor by initializing the model with observations, i.e. starting the model simulations from an observed climate state, similar to weather forecasts (Meehl et al. 2009). Thus, the main difference between long-term climate simulations and decadal predictions is that the latter are also dependent on the initial conditions. Part one of this thesis assesses the capability of retrospective decadal predictions (so-called hindcasts) from a selected model to reproduce the observed Arctic climate of the late 20th century. This hindcast skill is then compared to the skill of long-term climate simulations performed with the same Earth system model for the same time period. Part one aims to answer the following scientific question:

Q1: *Can model initialization with observations improve Arctic climate predictions?*

The evaluation of the complex decadal hindcast experiment setup requires a special kind of evaluation system (CLIVAR 2011; Goddard et al. 2013), which was implemented into the ESMValTool as part of this thesis. The evaluation system is based on two different metrics to assess hindcast skill (correlation and mean squared error) and was designed to ensure a fair comparison with the prediction skill of long-term historical climate simulations. This part of the thesis focuses specifically on simulations of one particular climate model, the Max Planck Institute Earth System Model (MPI-ESM, Giorgetta et al. 2013; Stevens et al. 2013), which conducted a large number of hindcast experiments.

1.3.2 The potential to narrow uncertainties in projections of Arctic sea ice extent

Part two investigates the future development of Arctic sea ice extent. CMIP5 climate simulations suggest that a transition to a seasonally ice-free Arctic is likely to occur within this century, but current projections exhibit large uncertainties (see also Figure 1.4). The aim of the thesis' second part is to reduce these uncertainties and to give a more robust estimate of the first year in which the Arctic will become essentially sea-ice free in summer. The total uncertainty in Arctic sea ice projections is separated into its three components internal variability, model uncertainty and scenario uncertainty (Hawkins and Sutton 2009). The potential to narrow the types of uncertainty is investigated with the multiple-diagnostic ensemble regression method (MDER, Karpechko et al. 2013; Wenzel et al. 2016), which uses information on different Arctic climate parameters from the observable time period to constrain the future projections of Arctic sea ice extent. Scientific question number two is:

Q2: *Can observation-based constraints improve multi-model projections of 21st-century Arctic sea ice extent?*

Again, the ESMValTool had to be enhanced by the implementation of a method to separate the sources of uncertainty, the MDER method, and their applications to sea ice

Both scientific questions are answered in this thesis, which is structured in the following way: Chapter 2 addresses *Q1* and is based on the author's peer-reviewed publication in *Meteorologische Zeitschrift* (Senftleben et al. 2018). Chapter 3 deals with *Q2* and is based on another publication by the thesis author, that is currently under review by *Journal of Climate* (Senftleben et al. 2019). For both chapters, the respective publications were adapted and considerably extended to fit the style and structure of the thesis. Chapter 4 provides a summary of the thesis' main findings and gives an outlook into the possible future of Arctic sea ice and related research.

2. Decadal hindcast skill in the Arctic

This chapter is based on Senftleben et al. (2018), see Appendix C.

State-of-the-art Earth system models, such as those participating in CMIP5, include detailed representations of sea ice and their performance shows substantial improvements with respect to previous generations of models (Stroeve et al. 2012b; Flato et al. 2013; Notz et al. 2013). Nevertheless, the reliability of climate model results on time scales particularly relevant to society and policymakers (10-20 years) still needs to be further assessed and improved. On these time scales, the natural internal variability of the climate system plays an important role. In the field of decadal climate predictions, which typically focuses on a time range of up to 10 years, the goal is to better capture the effects of internal variability of the climate system by initializing the models with observations of slowly-varying components of the Earth system such as ocean variables (Meehl et al. 2009; Smith et al. 2010).

This chapter aims to answer scientific question *Q1* and examines the skill of retrospective decadal climate predictions (so-called “hindcasts”) of Arctic sea ice. It is investigated whether initialization used in decadal hindcasts leads to more skillful predictions in comparison to uninitialized historical simulations. Here, the sea ice hindcast skill is quantitatively assessed from decadal and historical climate simulations performed with the Max Planck Institute Earth System Model (MPI-ESM, Giorgetta et al. 2013; Stevens et al. 2013) as part of the German Federal Ministry of Education (BMBF) project MiKlip (“Mittelfristige Klimaprognosen”, mid-term climate predictions, Marotzke et al. 2016). Therefore, a verification system for decadal hindcasts was developed, implemented into the ESMValTool, and applied to Arctic climate variables. Additionally, the skill of the model in reproducing observations of two additional variables that are closely related to sea ice is investigated: sea surface temperature (SST) as a proxy for the oceanic influence, and near-surface air temperature (TAS) for the atmospheric component.

The chapter is structured as follows: Section 2.1 gives an overview on existing studies in the field of decadal climate predictions and Section 2.2 describes the observations, model experiments and methods used for the assessment of hindcast skill. In Section 2.3, the results of the hindcast skill assessments of Arctic TAS, SST and sea ice predictions are presented. The chapter concludes with a summary and discussion in Section 2.4.

2.1 Previous skill improvements achieved with decadal hindcast experiments

Numerous studies found that the initialization of climate models can improve the predictability of near-surface air temperature in certain regions (e.g., Müller et al. 2004; Pohlmann et al. 2009; Doblas-Reyes et al. 2013; Goddard et al. 2013; Jia and Delsole 2013; Meehl and Goddard 2013). Such an improved prediction skill stems from a more accurate simulation of the actual natural variability of the Earth system, since the atmospheric and oceanic initialization allows the model experiments to start from the correct phase of relevant modes of natural variability (Müller et al. 2012). Most improvements by initialization are found in the North Atlantic region. For example, Robson et al. (2018) found significant improvements by decadal hindcast experiments conducted with the HiGEM model (Shaffrey et al. 2009), especially in the representation of ocean heat content in the North Atlantic Subpolar Gyre when compared to uninitialized historical experiments with the same model. Here, the decadal predictions demonstrate the role of the ocean as a driver of climatic changes in the North Atlantic region and lead to an improved predictability of North Atlantic SSTs and sea surface height variability. Furthermore, decadal hindcasts from the MPI-ESM model, which is used in this chapter, showed an improvement in the representation of the Atlantic Meridional Overturning Circulation (AMOC) compared to historical simulations (Müller et al. 2017). In CMIP's upcoming phase (CMIP6, Eyring et al. 2016a), the decadal climate prediction project (DCPP, Boer et al. 2016) is dedicated to the coordination of world-wide decadal climate prediction experiments.

In the Arctic, previous studies found the prediction skill of sea ice improving on seasonal timescales by initializing the model with sea ice observations. For example, Day et al. (2014a) found that the inclusion of sea ice thickness in the model initialization process can significantly increase the predictability of Arctic sea ice concentration and extent and reduces forecast errors in Arctic near-surface air temperatures. The sea ice prediction skill hereby depends on the season. Germe et al. (2014) showed that winter Arctic sea ice in initialized simulations is potentially predictable for up to several years, but for summer the potential predictability does not exceed 2 years. A potential predictability of Arctic sea ice volume and extent of up to 3 years was also reported by Tietsche et al. (2014) in a set of four global climate models. Yeager et al. (2015) analyzed the prediction skill of decadal trends in Arctic sea ice extent and showed that initialization improves Arctic winter skill in predicting the trend in sea ice extent, especially in the Atlantic sector of the Arctic. Hindcast experiments with the Norwegian Climate Prediction Model that only initialized SSTs (Counillon et al. 2017; Wang et al. 2018) showed skill in predictions of Arctic SIE for only one year after initialization, but also improved the predictability of the North Atlantic subpolar gyre. Moreover, Warner et al. (2017) assessed the relationship between Arctic sea ice and the North Atlantic Oscillation (NAO) with the UK Met Office decadal prediction system. They found seasonal improvements in the sea ice forecasts when the model is initialized in

November, which correlated with the following winter NAO index. In summary, initialized decadal hindcasts improve Arctic climate predictions especially for short lead times of seasons to few years, and the processes behind these improvements are yet poorly understood.

A recent study (Meehl et al. 2018) found a plausible physical connection between decadal variations in the tropics and Arctic sea ice loss, which also illustrates the importance of the decadal time scale for Arctic sea ice. They investigated why the observed loss rate of Arctic SIE after 2000 was by a factor of 2-3 larger (depending on the season) than the trend before 2000 (1979-1999). Sensitivity experiments with an atmospheric climate model conducted by Meehl et al. (2018) showed a tropical decadal variability pattern as a contributor to the accelerated Arctic sea ice loss: the Interdecadal Pacific Oscillation (IPO). The IPO is a natural oscillation pattern of Pacific SSTs with phases that can last for 20 years each. The IPO was in a positive phase between 1977 and 1999, and entered a negative phase in 2000, which lead to warm SST anomalies in the tropical Atlantic Ocean with increased convective heating. This influenced atmospheric teleconnection patterns between the tropics and the Arctic, with an increase in Arctic surface winds and subsequent sea ice drift in the Arctic. It is shown that increased ice drift leads to a thinning of the ice and a shrinking sea ice cover. As a consequence, the ice-albedo feedback and polar amplification mechanism further accelerated the decrease in SIE. Since 2014, the IPO is in a positive phase again. Whether this may slow down the sea ice retreat in the coming decade, is a scientific question that could be investigated with decadal climate predictions.

The evaluation study presented in this chapter contributes to the research effort on decadal predictions in the Arctic and aims to improve the understanding of the physical processes explaining a possible improvement in Arctic hindcast skill through model initialization.

2.2 Model experiments and methods

The decadal hindcast experiments and the historical simulations that are analyzed in this chapter (see Table 2.1) were calculated by the MPI-ESM, which is described in the following section together with the description of the decadal hindcast experiments. The MPI-ESM simulations are benchmarked against observational datasets presented in Section 2.2.2, and the decadal verification system is detailed in Section 2.2.3.

2.2.1 Model simulations

MPI-ESM is a coupled Earth system model with an atmospheric, an oceanic, a land biogeochemistry and a marine biogeochemistry component and has been used in this configuration for contributions to CMIP5, see Figure 2.1. Simulations from MPI-ESM are

Table 2.1. Overview of MPI-ESM simulations analyzed in this study (EM: ensemble members).

Model run	Time period	#EM	Comment
MPI-ESM-LR prot decadal hindcasts	10 years after initialization (1961 – 2023)	15 + 15	Decadal hindcasts initialized with GECCO2 / ORAS4 reanalyses data with 15 ensemble members each; sea ice not initialized
MPI-ESM-LR historical simulations	1850 – 2005	10	Uninitialized historical simulations with MPI-ESM-LR (CMIP5)
MPI-ESM-LR RCP 4.5 simulations	2006 – 2100	10	RCP 4.5 simulations with MPI-ESM-LR (CMIP5), used to extend MPI-ESM-LR historical runs until 2013
MPI-ESM-HR preop decadal hindcasts	10 years after initialization (1961 – 2023)	5	Decadal hindcasts in which SIC was additionally initialized
MPI-ESM-HR historical simulations	1850 – 2005	5	Uninitialized historical simulations with high-resolution configuration MPI-ESM-HR
MPI-ESM-HR RCP 4.5 simulations	2006 – 2100	5	RCP 4.5 simulations with high-resolution configuration MPI-ESM-HR, used to extend MPI-ESM-HR historical runs until 2013

analyzed from its low-resolution configuration (MPI-ESM-LR) with the atmospheric component ECHAM6 (Stevens et al. 2013) resolved horizontally at 1.9° x 1.9° and vertically with 47 levels up to 0.01 hPa (T63L47). The ocean component is the Max Planck Institute Ocean Model (MPIOM) (Jungclaus et al. 2013) configured with a bipolar orthogonal curvilinear C-grid (Marsland et al. 2003) with one pole over South Greenland and another one over Antarctica, at a nominal resolution of 1.5°, and 40 vertical levels. Part of the MPIOM is a dynamic and thermodynamic sea ice model based on Hibler (1979). The dynamics in the sea ice model are calculated with a parameterized viscous-plastic ice rheology scheme. The thermodynamics on the ice surface are determined by the balance of thermal fluxes from the atmosphere into the ice, outgoing longwave radiation, and heat conduction through the ice. Thereby, ECHAM6 provides the thermal fluxes between surface and atmosphere, and calculates the SSTs considering a melt pond scheme (Notz et al. 2013). Melting occurs in the model when the surface temperature of the ice is greater than 0 °C. Note that in grid cells that contain sea ice, the SST is set to -1.9 °C. At the bottom of the ice, a balance between the ice's conductive heat flux and the oceanic heat flux (provided by

MPIOM) determines the heat exchange between the ice and the ocean water, and ultimately, the ice thickness.

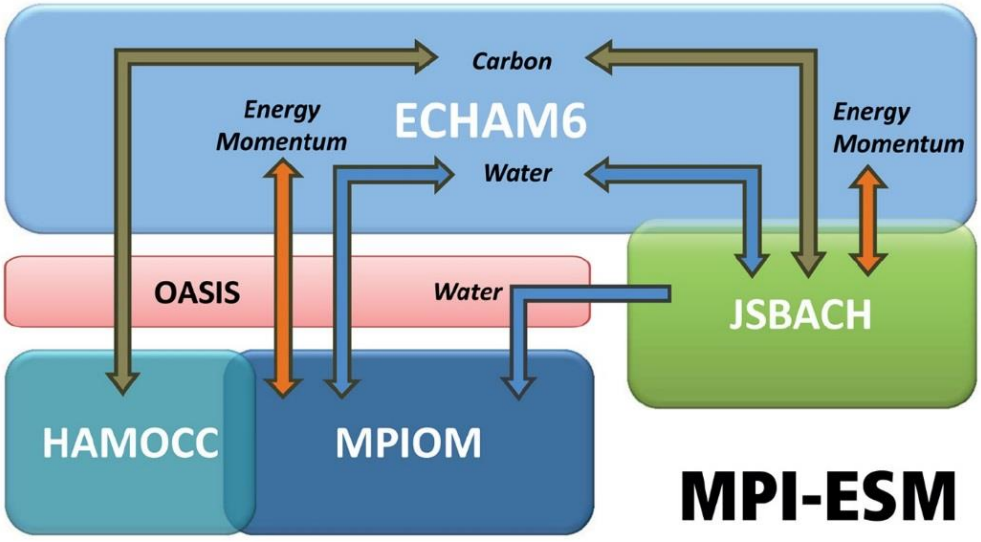

Figure 2.1. Overview of the MPI-ESM model components. Figure from Giorgetta et al. (2013).

The MPI-ESM-LR forms the basis of the MiKlip decadal prototype prediction system (hereafter MPI-ESM-LR prot, Marotzke et al. 2016; Kröger et al. 2017). The MPI-ESM-LR *prot* simulations consist of 30 decadal ensemble members divided in two sets that differ in the reanalyses used for initializing the ocean model component: 15 ensemble members are initialized with observationally-based data from the Ocean Reanalysis System 4 (ORAS4, Balmaseda et al. 2013) and 15 with reanalysis data from the German contribution to Estimating the Circulation and Climate of the Ocean (GECCO2, Köhl 2015). The ensemble members are initialized with slightly different initial conditions each by applying a 1-day lagged initialization. MPI-ESM-LR *prot* uses the so-called full-field initialization technique. First, the reanalyses are nudged into the coupled model to perform a so-called assimilation run. This removes model biases by keeping the model close to the observed climate state during the assimilation run. Then, the decadal hindcasts are started from the assimilation runs and freely integrate ten years into the future. For the atmosphere, vorticity, divergence, temperature and sea level pressure from the ERA40 (Uppala et al. 2005) and ERA-Interim reanalyses (Dee et al. 2011), and for the ocean, 3-dimensional temperature and salinity fields are nudged in the assimilation run. No sea ice variables are nudged. The model is initialized with data from the assimilation run on the simulated January 1st each year between 1960 and 2013. This results in a set of 54 hindcast experiments, each simulating a different, but overlapping, ten-year period. Figure 2.2 schematically illustrates this approach used for the MPI-ESM-LR *prot* experimental setup. As stated above, each hindcast experiment consists of 30 ensemble members, meaning that in total 1620 individual simulations are performed with MPI-ESM-LR *prot*.

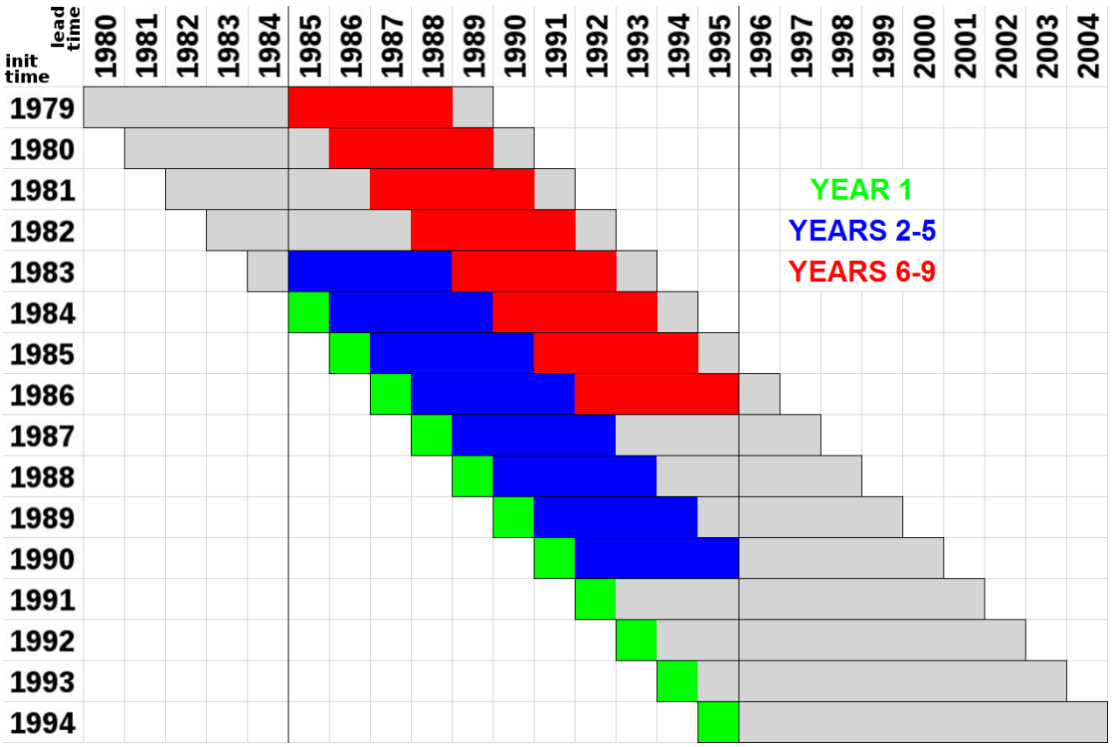

Figure 2.2. Illustration of a snippet of the MPI-ESM-LR *prot* hindcast experiment structure. The boxes represent the hindcast experiments running for ten years each, with the y-axis being the years of initialization of each experiment and the x-axis the lead years (i.e. simulated years). Lead-year samples (colored boxes) are selected from the hindcast experiments so that climatologies constructed from these lead-year samples all cover the same time period (1985-1995 in this illustration). Each hindcast experiment from the MPI-ESM-LR *prot* system consists of 30 ensemble members (not depicted).

In addition, an ensemble of ten uninitialized, long-term historical climate simulations have been performed with MPI-ESM-LR following the CMIP5 experiment protocol (Taylor et al. 2012). In this chapter, the output of the historical simulations is compared to the initialized decadal hindcasts to assess whether the initialization improves the hindcast skill. The historical simulations cover the time period 1850-2005 and are driven with prescribed natural and anthropogenic forcings (Giorgetta et al. 2013). For this chapter, each of the ten ensemble members of the historical simulations has been extended until 2013 with the corresponding ensemble member of the projection under the RCP 4.5 concentration scenario (Thomson et al. 2011).

Furthermore, historical simulations and decadal hindcasts were performed within MiKlip Phase 2 also with the preoperational system based on the high-resolution model configuration (MPI-ESM-HR *preop*) and are also assessed in this chapter. In MPI-ESM-HR *preop*, sea ice concentrations were also initialized in addition to the atmospheric and oceanic variables mentioned above, while sea ice thickness was not initialized. Compared to MPI-ESM-LR, the spatial and vertical resolution of the atmosphere is doubled in MPI-ESM-HR

(T127L95), and the nominal resolution of the ocean grid is improved from 1.9° to 0.4° (TP04/L40). All decadal hindcast experiments from MPI-ESM-HR *preop* as well as the historical simulations from MPI-ESM-HR consist of 5 ensemble members each and cover the same time period as the ones of the MPI-ESM-LR simulations. The MPI-ESM-HR historical simulations were also extended until 2013 with their corresponding RCP 4.5 simulations. All model simulations assessed in this chapter are summarized in Table 2.1.

2.2.2 Observations and reanalysis data

As reference datasets used for assessing the model prediction skill, different observations or reanalyses are used and described in this section for each variable analyzed. Observations are obtained solely through measurements with all kinds of instruments, and reanalyses are created by assimilating and reprocessing observations with a modern numerical weather forecast / ocean model. While reanalyses therefore are inevitably influenced by the model's parameterizations, they have the advantages of covering a longer time period than many observations and providing global data coverage. This is essential for regions that are difficult to observe, as for example close to the North Pole, an area which is rarely retrieved by satellites due to their frequently used inclined sun-synchronous orbits and hard to reach for ship or buoy measurements.

For the evaluation of TAS, the European Centre for Medium-Range Weather Forecast Re-Analysis Interim (ERA-Interim) data (Dee et al. 2011) are used from 1979 to 2013. In addition, HadCRUT4 (Hadley Centre / Climate Research Unit Temperature records 4, Morice et al. 2012) data are used, which cover the time period 1850-2012 and are derived for land regions all over the globe from up to 4800 stations (Jones et al. 2012), and for the ocean from merchant and naval vessels, as well as fixed and drifting buoys (Kennedy et al. 2011).

The SST observations are taken from the Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) (Rayner et al. 2003), which is a global reanalysis product using the Met Office Marine Data Bank (MDB) and, from 1982, also the Global Telecommunications System (GTS).

Observations of sea ice concentration (SIC), i.e. the area fraction of sea ice cover within each grid cell, are not very reliable before 1978 when satellite products became available. The SIC analysis is therefore based on satellite retrievals from the National Snow and Ice Data Center (NSIDC, Walsh et al. 2015). These observations cover the period from 1978 to present. The satellite SIC data are processed by two different retrieval algorithms: the Bootstrap (NSIDC-BT, Comiso 2000) and the NASA-Team (NSIDC-NT, Cavalieri et al. 1996) algorithm. The main difference lies in the treatment of melt ponds that are nearly indistinguishable from open water in the satellite data. Whereas in NSIDC-BT sea ice concentrations are synthetically increased in summer to account for undetected melt ponds, NSIDC-NT does not contain such

a correction. Since NSIDC-BT could potentially introduce a positive bias in melt-pond-free areas, both products are used as a reference in the sea ice evaluation (see also Notz (2014)).

2.2.3 Methods

For the analysis of the model output, a system of evaluation diagnostics for decadal climate predictions was developed and implemented into the ESMValTool (Eyring et al. 2016b). The evaluation system is partly based on Goddard et al. (2013) and the recommendations given by CLIVAR (2011), and is described in this section.

The evaluation system is applied here to assess the prediction skill of the model for TAS, SST, SIC and sea ice area (SIA). The latter is not readily available and has been derived as the area integral of SIC over a certain region. Due to the strong seasonality of Arctic sea ice, September and March means are analyzed separately. The former typically represents the Arctic minimum ice conditions (summer), whereas the latter serves as a proxy for the winter situation with maximum sea ice area.

The prediction skill of the four variables is assessed for each grid cell with the anomaly correlation coefficient (ACC, also known as Pearson product-moment correlation coefficient) and the root-mean-square error (RMSE), using monthly mean values of the model and a reference dataset:

$$ACC(x,y) = \frac{1}{n-1}\frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sigma_x \sigma_y} \tag{2.1}$$

$$RMSE(x,y) = \frac{\sum_{j=1}^{n}(x_j - y_j)^2}{n} \tag{2.2}$$

Hereby, x is the model simulation and y the reference dataset, index j references the point in time (year), $\overline{x}$ ($\overline{y}$) is the time average of x (y), and σ_x (σ_y) the standard deviation in time of x (y). Wherever possible, both metrics are applied to all variables, since the assessment of skill can be dependent on the choice of the metric (Hawkins et al. 2015).

In order to assess a possible reduction in RMSE (i.e., improvement) by initialization, the RMSE skill score is introduced. It is defined as a function of the ratio between the RMSE of decadal hindcasts and the RMSE of historical simulations:

$$RMSE_{skill} = 1 - \frac{RMSE_{decadal}}{RMSE_{historical}} \tag{2.3}$$

Thus, positive (negative) values of this metric give the fraction of improvement (deterioration) in skill of decadal hindcasts over historical simulations. Grid cells for which $RMSE_{historical}$ equals 0 have been excluded from the analysis. It is important to note that differences between hindcasts and historical simulations only stem from the initialization since both external forcings and model components are identical (Marotzke et al. 2016).

The data processing also includes the computation of an ensemble average for each hindcast experiment (30 members for *prot*; 5 for *preop*) and the historical simulations (10 members for *prot*; 5 for *preop*). In the following, ACC and RMSE are always calculated for ensemble means. To apply the metrics to each grid cell, modelled sea ice and SST are first interpolated from their native irregular ocean grids to a regular 1°x1° grid using the distance-weighted regridding method from the Climate Data Operators package (CDO[6]). Higher resolutions of the target grid were tested but were found to have no impact on the results. As recommended by CLIVAR (2011), a cross-validated bias correction is applied to the decadal ensemble means in order to remove the mean bias from the data (e.g., Gangstø et al. 2013). Hereby, for every ensemble mean hindcast, from each lead month the average over the same lead months of the other hindcasts is subtracted. Simple anomalies are calculated from the ensemble mean historical simulations and observations by subtracting their respective means over the analyzed time periods.

In order to estimate the dependence of hindcast skill on the forecast time, i.e. the time distance from the initialization point, different time samples were selected from each hindcast, the so-called lead years (see Figure 2.2). Following previous studies (e.g., Kim et al. 2012; Müller et al. 2012; Goddard et al. 2013), a set of three lead-year samples is analyzed: year 1, years 2-5 and years 6-9 with one climatology constructed for each lead-year sample averaged over all decadal hindcasts. Hereby, the lead-year-1 climatology consists of the first year of each hindcast experiment, whereas the climatology for lead years 2-5 (6-9) consists of the average over the years 2 to 5 (6 to 9) from each hindcast experiment. For the historical simulations and the observationally-based reference datasets, the lead-year climatologies are constructed by taking the same years as the ones used for calculating the hindcast climatologies from their respective timeseries. In most of the aforementioned studies, the lead-year climatologies were constructed by sampling over slightly different time periods depending on the selected lead time. This could cause a bias in the assessed prediction skill in particular for variables such as sea ice extent, which have a large year-to-year variability: a lead-year climatology could, for instance, include an exceptional year with a large positive or negative anomaly that is well predicted, but which a different climatology does not include. This would artificially increase the ACC skill of the former climatology. It is therefore important to sample the same time period for all lead-year climatologies. Thus, the lead-year climatologies were calculated in a way that they all cover the same time period

[6] https://code.mpimet.mpg.de/projects/cdo

(i.e., 1979-2013). A drawback of this is that the climatologies are shorter and only include a subset of the decadal hindcasts for certain lead years.

The statistical significance of the ACC differences is assessed by applying a two-sided t-test with a Fisher r-to-z transformation. For the differences in RMSE skill scores, the original data is resampled with a non-parametric block-bootstrap algorithm (Wilks 2011; Goddard et al. 2013; Eade et al. 2014): n hindcasts are randomly drawn with replacement from the pool of the ensemble-averaged hindcasts, with n being the original experiment size. This resampled set of hindcast experiments is very likely different from the original one as some of the hindcast experiments may be included multiple times. To account for temporal auto-correlation, the resampling is done for blocks of five consecutive hindcast experiments. The RMSE skill scores are then calculated for this newly generated sample including all processing steps. The test is against the null hypothesis that $RMSE_{skill} = 0$, i.e. $RMSE_{decadal} = RMSE_{historical}$ at the 95% confidence level. Statistically, 5% of the grid cells are expected to contain false positives, which is why only clusters of multiple grid cells are discussed that include significant skill scores of the same sign. Nevertheless, the result of a statistical test has to be supported by physical reasoning and shall not be the only criterion to distinguish a signal from noise.

2.3 Results

In the following, the TAS hindcast skill for the North Atlantic is analyzed and the role of different reference datasets (Section 2.3.1) discussed. Then, timeseries of September SIA from decadal hindcasts, historical simulations and observations are compared to each other for different lead-year climatologies in the Arctic, in order to identify possible biases and drifts (Section 2.3.2). Finally, possibly increased hindcast skill compared to the uninitialized experiments is assessed for temperature and sea ice in the Arctic with different metrics in different regions (Section 2.3.3).

2.3.1 Hindcast skill for near-surface air temperature in the North Atlantic

The ACC of TAS from decadal hindcasts and historical simulations for the lead years 2-5 is shown in Figure 2.3. Similarly to Marotzke et al. (2016), the ACC is calculated for the MPI-ESM-LR *prot* system against HadCRUT4 data and a statistically significant skill improvement due to initialization in the North Atlantic Ocean can be confirmed (Figure 2.3b). This leads to the hypothesis of possibly improved skill in the Arctic (Section 2.3.3). In addition, the results for the MPI-ESM-HR *preop* system are shown as well. Since HadCRUT4 data have gaps in the Arctic, both model systems are additionally compared to ERA-Interim data (right column of Figure 2.3).

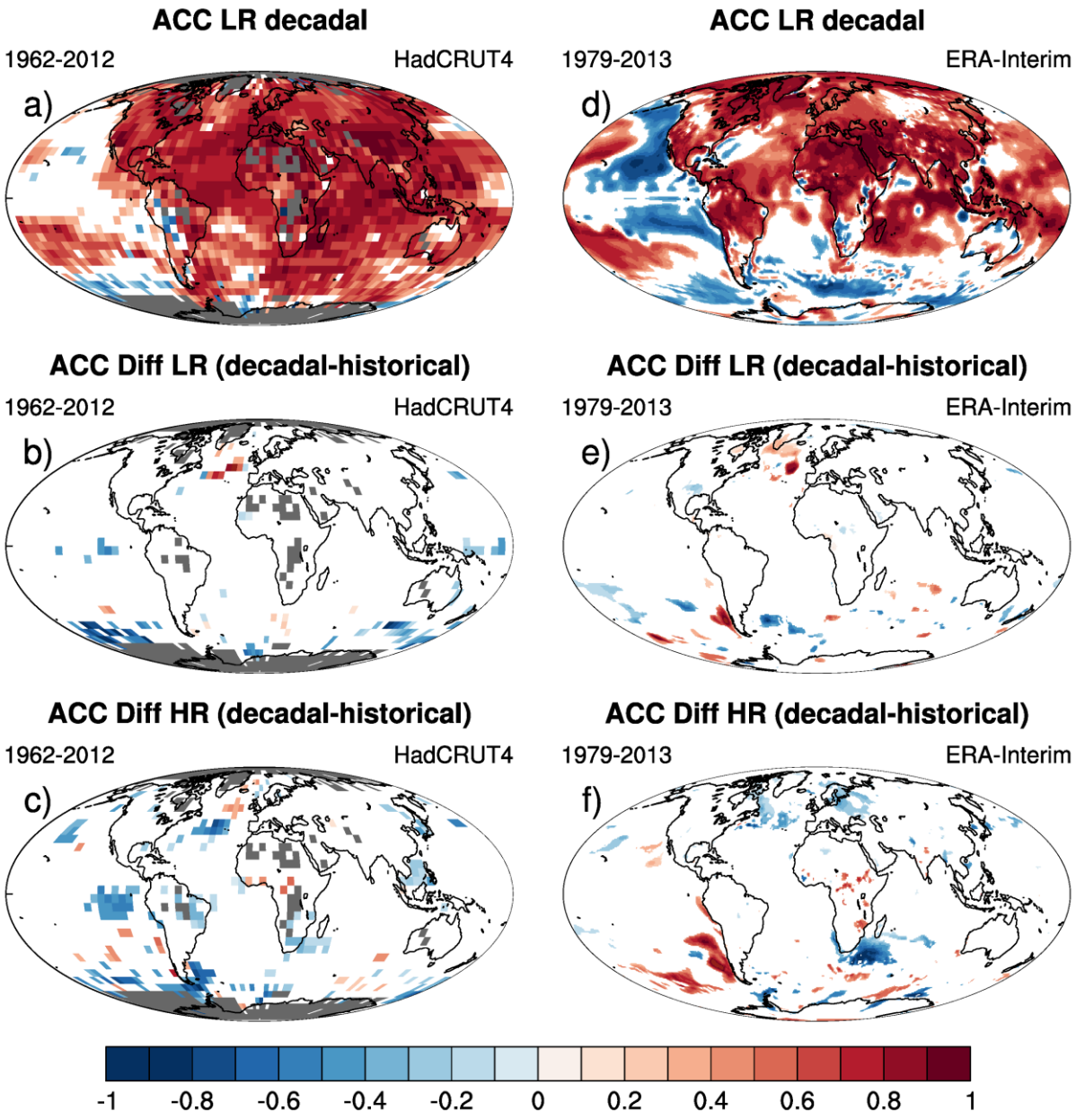

Figure 2.3. Ensemble-mean hindcast skill (ACC) of annual mean near-surface air temperature (TAS) averaged over the lead years 2-5 for MPI-ESM-LR prot decadal hindcasts (a, d) and with the anomaly correlation of MPI-ESM-LR historical simulations subtracted (b, e). Panels c and f are the same as b and e, but for MPI-ESM-HR preop simulations. The ACC is calculated with HadCRUT4 observations (1962-2012; left column, a-c) and ERA-Interim reanalyses (1979-2013; right column, d-f). All values shown are statistically significant at the 95% confidence level according to a t-test for correlation coefficients. Grid cells with a dark gray color denote missing values in the observations.

The ACC skill in the North Atlantic can be improved via initialization (Figure 2.3b, e). Here, the ocean initialization improves prediction skill by increasing local SST skill due to the deep mixed layer in this region (Marotzke et al. 2016). In a predecessor version of the MPI-ESM-LR decadal prediction system, Müller et al. (2017) found significant skill improvements over historical simulations in the representation of the AMOC up to lead year 5, which may be the reason for enhanced TAS skill in the North Atlantic. This improvement is robust against the choice of reference dataset (compare panels b and e in Figure 2.3).

A deviation from the results of Marotzke et al. (2016) is found in the statistical significance of the ACC differences. The two-sided t-test for the correlation coefficients with a Fisher r-to-z transformation applied here gives distinctively fewer statistically significant values than in Marotzke et al. (2016) who used a bootstrap algorithm despite the same confidence level of 95% (compare their Figure 1f to Figure 2.3b of this thesis). The test used in this chapter does not give statistical confidence to the increased hindcast skill south of South America shown in Figure 1f of Marotzke et al. (2016) and shows slightly different values also in the North Atlantic.

On the contrary, in the MPI-ESM-HR *preop* hindcasts (Figure 2.3c, f), no significant improvement in the hindcast skill in the North Atlantic due to initialization is found, independent of the reference dataset used (HadCRUT4, ERA-Interim). A possible reason for that could be the small ensemble size (5 ensemble members instead of 30 members in MPI-ESM-LR *prot*) not covering a large enough spread of different initial conditions. For example, Sienz et al. (2015) recommended at least 10 ensemble members (and even more in regions with a low signal-to-noise ratio), since initial conditions are never known exactly.

2.3.2 Timeseries of pan-Arctic sea ice area

Before assessing hindcast skill in the decadal simulations, possible biases and drifts are identified from the models' climatologies and trends. Figure 2.4 shows the evolution of pan-Arctic (60° - 90° N) SIA between 1979 and 2013 calculated from the two observational NSIDC datasets and simulated by different MPI-ESM runs for different lead years. There is a relatively constant offset between the two NSIDC observations of roughly 1 million km² that results from differences in the retrieval algorithm (see Section 2.2.2). The historical simulations show a smaller bias compared to NSIDC-NT than to NSIDC-BT. This is consistent with results from earlier model versions (Notz et al. 2013). In agreement with both observational datasets, the ensemble means of all MPI-ESM-LR and MPI-ESM-HR simulations show a decline in Arctic summer SIA over this time period, but they underestimate sea ice decline after 2006.

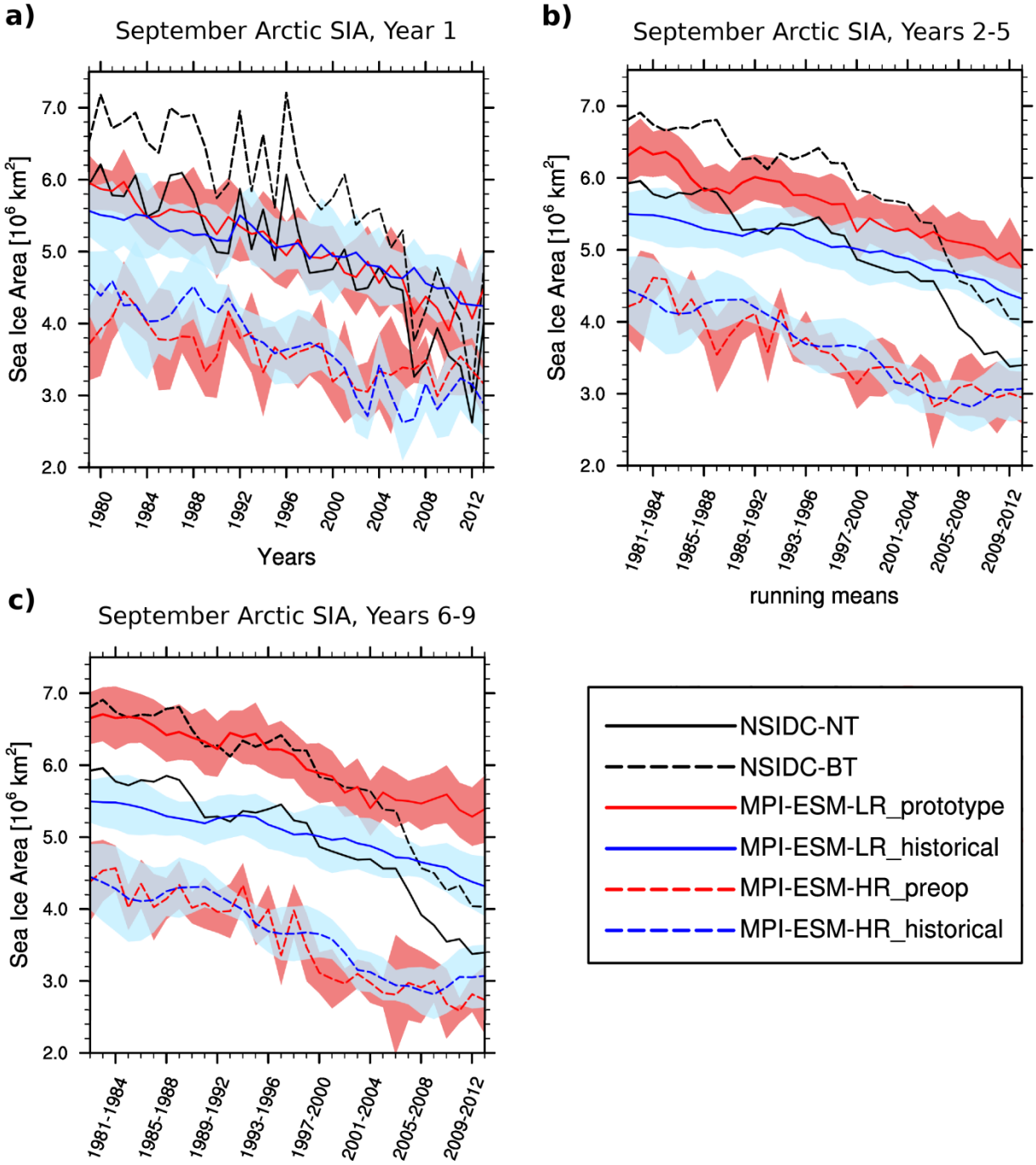

Figure 2.4. Timeseries of September mean Arctic sea ice area (SIA) from 1979 to 2013 from NSIDC-NT and NSIDC-BT observations, and the respective ensemble means of the following model simulations: MPI-ESM-LR historical simulations and *prot* decadal hindcasts, and MPI-ESM-HR historical simulations and *preop* decadal hindcasts. The shaded areas denote the respective ensemble's inter-model standard deviation. The pole hole due to incomplete coverage in the observations has been consistently filled with SIC = 1. (a) lead year 1, (b) lead years 2-5 and (c) lead years 6-9. SIA timeseries are calculated based on the respective native grids of models and observations.

The MPI-ESM-LR *prot* hindcasts show a strong drift in modelled SIA during the ten-year hindcast period: the simulated SIA is closest to NSIDC-NT observations in lead year 1 (Figure 2.4a) and increases during the hindcast time (Figure 2b) reaching a state that is closest to NSIDC-BT in lead years 6-9 (Figure 2.4c). It hereby increasingly deviates from the historical simulations (blue solid line in Figure 2.4), with a maximum offset between the two of around 1 million km² in lead years 6-9. This drift could be caused by the so-called initialization shock: the model that has been nudged to observations returns to its biased equilibrium state immediately after the initialization process when the simulation starts to run freely. Since the assimilation runs are not performed with a coupled model, and since sea ice is not initialized, the initialization shock could be partly due to inconsistencies in the initial conditions. It can cause the model simulation to overshoot its preferred climatology resulting in a larger error than that of the biased equilibrium (Meehl and Goddard 2013). A similar drift has been found in the North Atlantic with the full-field initialization perturbing the overturning circulation, heat transport and associated SST and sea surface salinity in the region of the sub-polar gyre (Kröger et al. 2017). For the analysis presented here, the drift over the hindcast period has been corrected by applying a lead-time dependent cross-validated bias correction (CLIVAR 2011) in the ESMValTool, as mentioned in Section 2.2.3.

In contrast, the MPI-ESM-HR *preop* hindcasts (red dashed line in Figure 2.4) show no drift and remain relatively close to the uninitialized historical simulations (blue dashed line). However, both MPI-ESM-HR hindcasts and historical simulations show a strong negative bias of initially roughly 2 million km² compared to NSIDC-NT and 3 million km² to NSIDC-BT. This bias affects all lead years and strongly decreases with time to less than 1 million km² compared to NSIDC-BT at the end of the assessed time period, indicating a smaller trend than in the observations. It is related to a misrepresentation of the seasonal cycle, as the bias is only present in late summer and fall (i.e., during minimum sea ice conditions), but not in winter (not shown). Such bias might stem from too thin ice in the assimilation run possibly caused by the applied anomaly nudging, as full-field nudging used with an older model version resulted in too thick sea ice and a positive bias in sea ice area in summer (Felix Bunzel, personal communication, May 24th, 2017).

The analysis shown in Figure 2.4 was repeated using sea ice extent instead of SIA (not shown), but the main findings discussed in this section are not sensitive to whether SIE or SIA is used for the analysis.

2.3.3 Hindcast skill in the Arctic

The hypothesis for this chapter is that the enhanced hindcast skill in the North Atlantic in the MPI-ESM-LR *prot* system due to initialization may provide enhanced hindcast skill also in the Arctic for TAS, SST and SIC, which is the focus in the remainder of this section. Since MPI-ESM-HR *preop* does not show an improved skill (see Figure 2.3), from here on only the

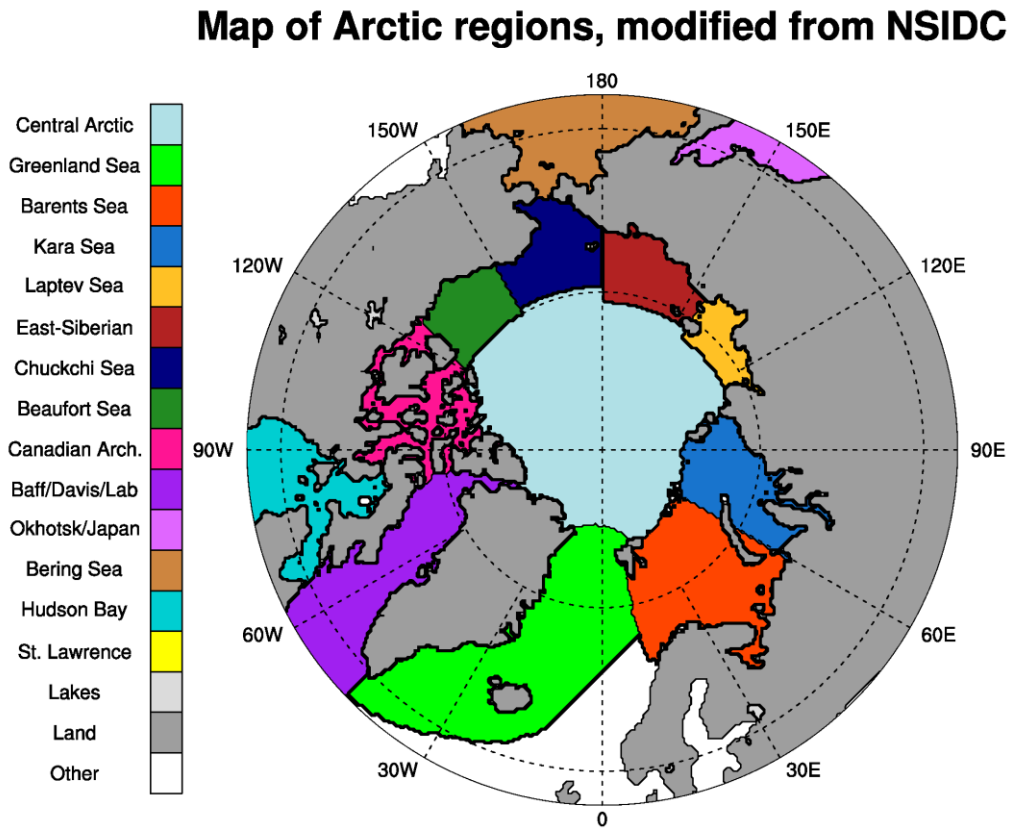

Figure 2.5 The Arctic Ocean and its adjacent seas.

results from the MPI-ESM-LR *prot* system are shown, although the entire analysis has also been done for the MPI-ESM-HR *preop* system to confirm the conclusions. An overview of the different parts of the Arctic ocean, which are referenced in this section, is given in Figure 2.5.

To investigate whether the skill improvement in North Atlantic TAS predictions could result in an improved skill in the Arctic, the analysis of Section 2.3.1 is repeated for TAS and SST in this region (Figure 2.6), focusing on the September and March means (Figure 2.6a-d). ACC for TAS evaluated against ERA-Interim data is generally high (above 0.6) with the exception of the Greenland Sea in March. In terms of improvement due to initialization, however, there is little to no skill gain (Figure 2.6e-h). Solely along the east coast of Greenland and in the Fram Strait there is a statistically significant improvement of ACC skill in March for SST (Figure 2.6d). Note that the skill in SST of the uninitialized MPI-ESM simulations is already high, i.e. any further improvement due to initialization can be expected to be rather small. Indeed, the largest improvement is found in a region where the ACC skill of the decadal hindcasts is not as high as in other regions (compare Figure 2.6d to 2.6h), meaning that largest improvements occur in places where the historical simulations have particularly low skill.

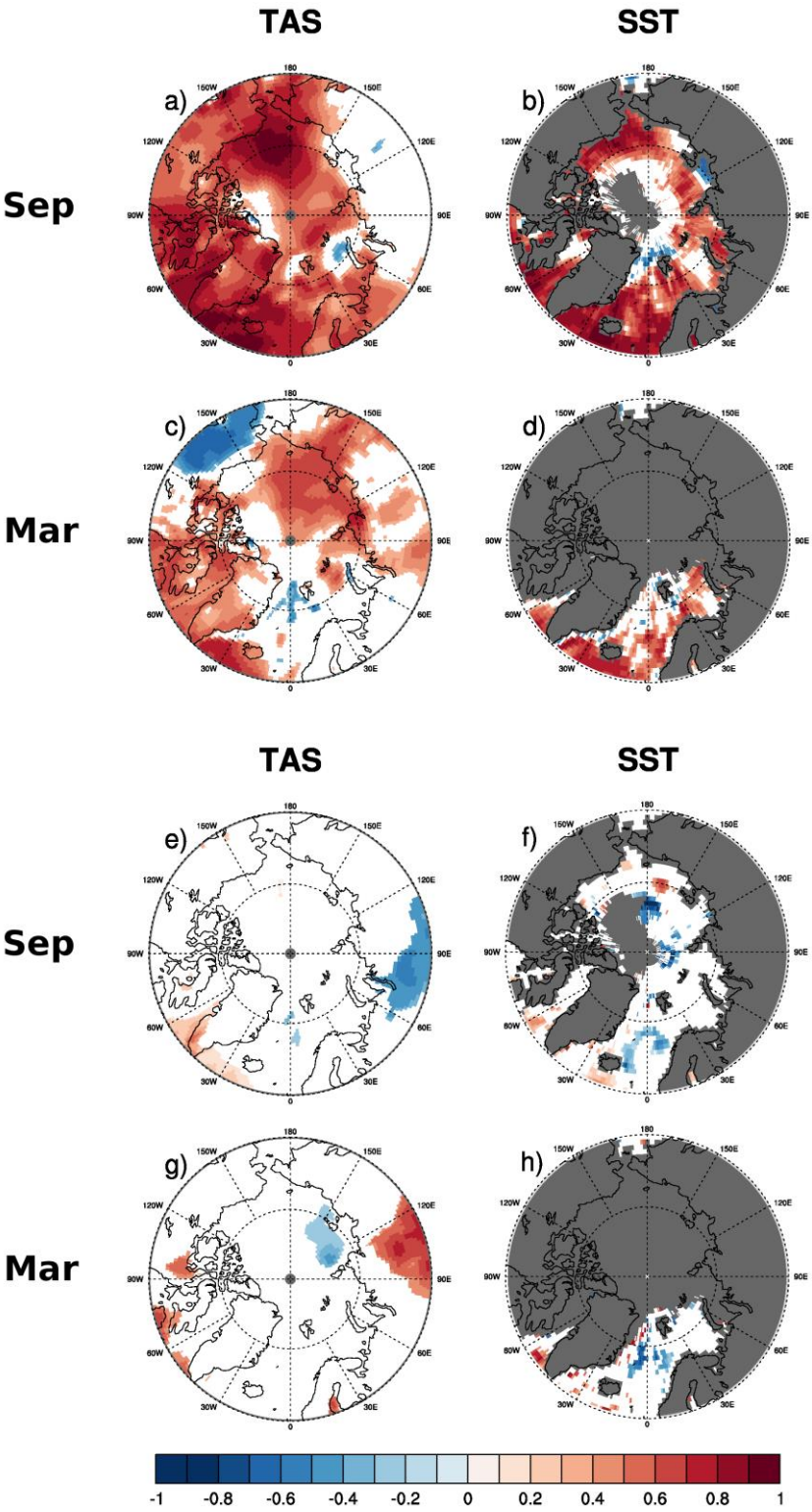

Figure 2.6. a-d: as Figure 1, but for ACC of TAS and SST displayed in polar-stereographic projections in the Arctic and calculated against ERA-Interim and HadISST data, respectively. e-h: ACC difference between MPI-ESM-LR decadal and historical simulations. White grid cells denote values that are not statistically significant at the 95% confidence level. Cells with dark gray color represent missing values that either stem from gaps in the observational data or from constant SSTs due to ice coverage: in grid cells with sea ice, the SSTs in the model are set to a constant value of -1.9 °C.

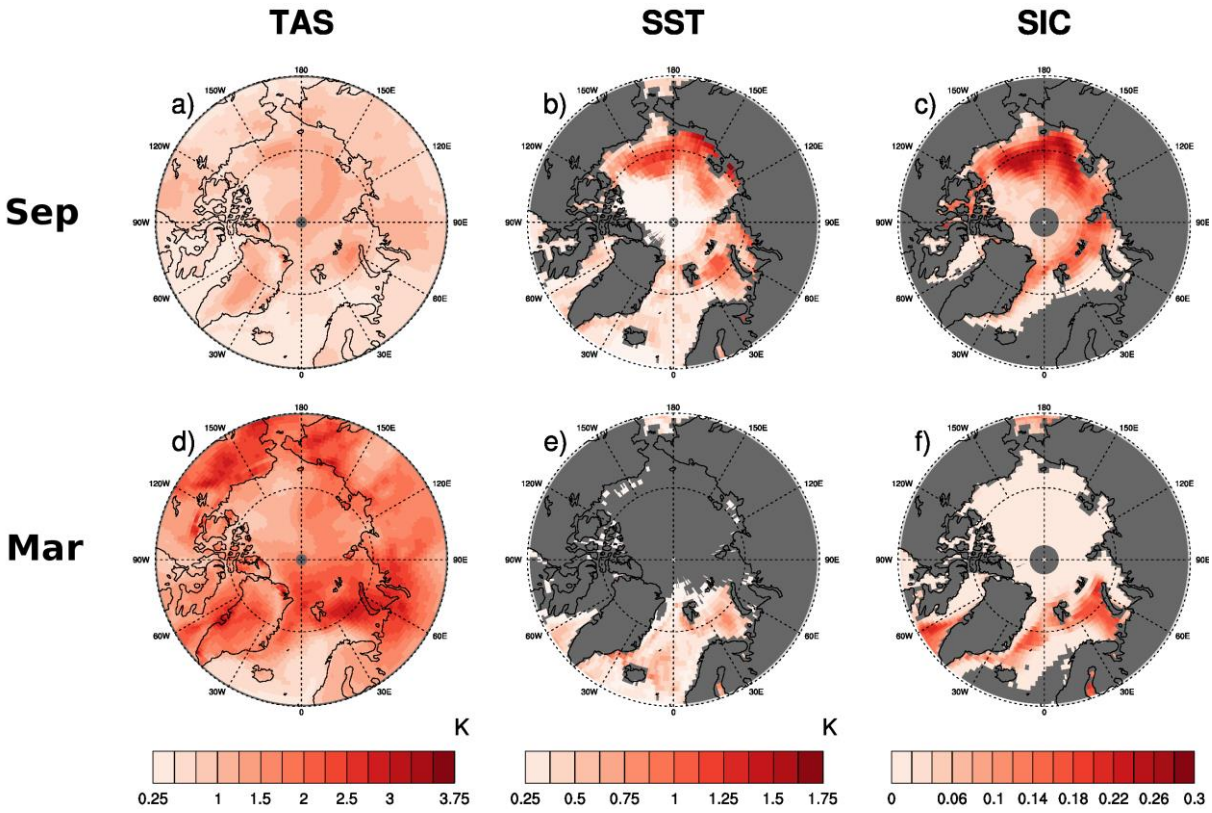

Figure 2.7. Arctic polar-stereographic contour maps of root mean square errors (RMSEs) between anomalies of the 30-member ensemble mean of the MPI-ESM-LR *prot* and respective reference datasets over the time period 1979-2013: ERA-Interim for TAS (a, d), HadISST for SST (b, e), NSIDC-NT for SIC (c, f). Panels a-c and d-f depict September and March means, respectively. All data represent anomalies with respect to their individual climatological mean and are sampled in 4-year running averages, similar to lead years 2-5. Cells with dark gray color represent missing values that either stem from gaps in the observational data or from constant SSTs due to ice coverage: in grid cells with sea ice, the SSTs in the model are set to a constant value of -1.9 °C.

ACC cannot be used to assess the model's hindcast skill in predicting SIC in a meaningful way, since SIC is not normally distributed (Kowalski 1972) but rather follows a bimodal distribution peaking at 0% (no ice) and 100% (fully ice-covered). Since the relative temporal standard deviation of SIC can be very small (for example, in areas with nearly complete ice coverage such as the Central Arctic Ocean), even small errors in the modelled time series can lead to very low ACC values. This results in an unrealistic assessment of the model's performance in reproducing the observed sea ice time series. The RMSE is not affected by this problem and thus provides a more robust estimate of model quality in predicting SIC and has been widely used in several sea ice evaluation studies (e.g., Ahn et al. 2014; Day et al. 2014a; Yang et al. 2017).

Therefore, RMSE is used as a metric and calculated for TAS, SST, and SIC. This is depicted in Figure 2.7 for the MPI-ESM-LR *prot* hindcasts compared to the corresponding reference

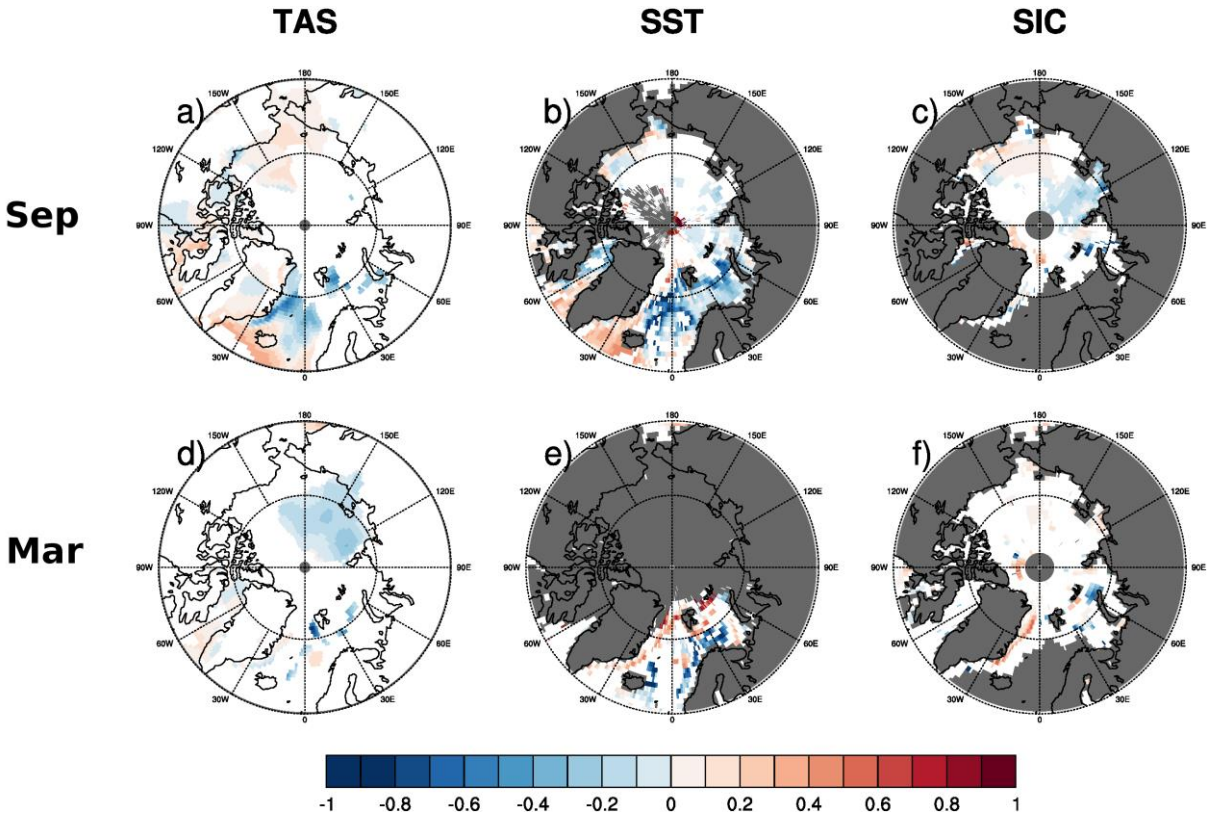

Figure 2.8. Same as Figure 4, but for RMSE skill (Eq. 2.3). Red (blue) colors indicate an improvement (degradation) of the model skill via initialization. All values shown are statistically significant at the 95% confidence level following a significance test including the bootstrap algorithm (see Section 2.3).

datasets. Note that all data represent anomalies to their respective climatological mean, with the hindcasts being lead month-wise bias corrected (see Section 2.2.3). This way, the RMSE is not affected by a constant offset between simulation and reference or a lead-time dependent drift (as seen in Figure 2.4). For TAS, RMSEs are generally higher in March than in September (compare Figure 2.7a to d). Especially in the northern part of the Greenland Sea and in the Kara and Barents Seas in the decadal hindcasts, RMSE values in March show values of up to 3 K compared to values smaller than 1.5 K in September. For both, SST and SIC, highest RMSE values are calculated for the East-Siberian and Beaufort Seas in September of 1.5 K (SST) and up to 0.3 (SIC), and in March in the Atlantic.

The change in model skill via initialization is shown in Figure 5 for the three variables TAS, SST, and SIC (Eq. 2.3). In September, there is a positive RMSE skill score of up to 0.5 (meaning the RMSE was reduced by half) in the North Atlantic in TAS and SST (Figures 2.8a and b), which is the same region for which an increase in TAS ACC skill is found (Figure 2.3). A propagation of this RMSE improvement further northward into the Arctic Ocean is, however, not visible and in general not much improvement is found.

Hindcasts of the Arctic winter SSTs (Figure 2.8e) show a statistically significant improvement in RMSE skill along the east coast of Greenland and in Fram Strait, due to the initialization of ocean temperature and salinity fields. This signal propagates into the decadal winter sea ice predictions decreasing RMSEs of SIC in the same area by about 30% (Figure 2.8f). Note that in the same area the RMSE of SIC decadal hindcasts (Figure 2.7f) is relatively high (around 0.16), which means that the absolute improvement via initialization in that region is also high. The March SST and SIC improvement in RMSE skill along the east coast of Greenland and in Fram Strait is also seen in the ACC (Figure 2.6h). This is the only region in the Arctic with statistically significant improvements by initialization that are robust against different metrics.

Especially in the marginal ice zone of the Atlantic, the decadal hindcasts show deficits in SIC predictions in terms of RMSE skill compared to historical simulations and this in both, summer and winter. This agrees with results from Tietsche et al. (2014), who found largest errors of simulated sea ice concentration in the marginal ice zone and of sea ice thickness along the coasts of the Arctic Ocean in multiple models, including the MPI-ESM-LR. They conclude that spatial patterns of sea ice quantities are more difficult to predict than aggregated quantities like sea ice extent and total sea ice volume. Accordingly, Goessling et al. (2016) find a lower predictability for the Arctic sea ice edge than for sea ice extent, especially in September.

The skill for SST and SIA is further analyzed for different Arctic regions. Only those regions are shown where improvements of hindcast skill could be detected: the Greenland Sea and the Beaufort Sea. In order to investigate the effect of trends on the skill scores, the ACC of detrended SST anomalies and detrended SIA anomalies are assessed for the two regions and for individual lead years (Figure 2.9). In the Greenland Sea, an improvement in skill by initialization is found in lead year 1 in March for SIA. This corresponds to the RMSE reduction in SIC hindcasts seen in Figure 2.8f and may be related to the ACC skill improvement in March SSTs in Fram Strait (Figure 2.6h). No further improvements are found for the Greenland Sea in September, neither for SST, nor for SIA. In the Beaufort Sea, September SST predictions are found to be improved via initialization only in lead year 5. This suggests that the high skill seen in Figure 2.6b mainly stems from the model correctly reproducing the trend. The fact that here, single lead years are analyzed instead of a 4-year average as in Figure 2.6, is another reason for the lower prediction skill shown in Figure 2.9, since multi-year averages are easier to predict than single years (e.g., Goddard et al. 2013). Similar to SST, September SIA predictions are also improved by initialization in lead year 5 (and 6), emphasizing again the strong relationship in prediction skill between the two variables. However, no improvement in ACC is found for the Beaufort Sea in March.

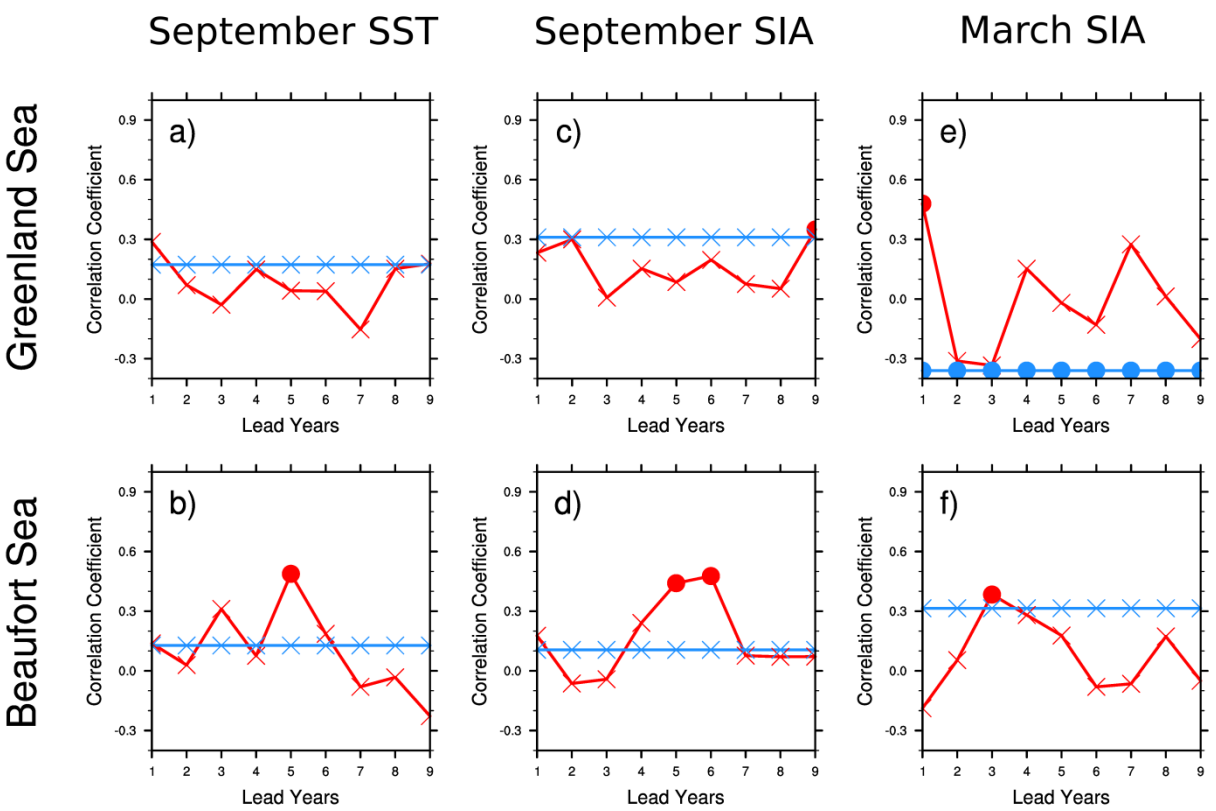

Figure 2.9. ACC against lead years for (a, b) regionally averaged September SST anomalies and (c-f) regionally integrated September and March SIA from MPI-ESM-LR *prot* decadal hindcasts (red) and MPI-ESM-LR historical simulations (blue). All data are detrended by subtracting the least squares linear trend. The two regions are the Greenland Sea (top row) and the Beaufort Sea (bottom row) and follow the definition of NSIDC with modifications. Reference for the correlation calculations are HadISST observations for SSTs and NSIDC-NT observations for SIA. The time period is 1982-2013. Filled circles indicate correlation coefficients that are statistically significant at the 95% significance level according to a t-test for correlation coefficients.

In summary, a robust improvement across different variables and using different metrics could only be found in the Greenland Sea along the east coast of Greenland and in Fram Strait, and only for March. Thus, the skill improvement in the MiKlip prototype and preoperational systems compared to the uninitialized historical experiments in the Arctic is generally weak. Therefore, improved forecasts of the coming decade are not expected. Figure 2.10 shows that the decadal forecasts of MPI-ESM-LR *pr* that were initialized in 2013 simulate on average a 1 million km² larger September Arctic SIE between 2014 and 2023 than the uninitialized RCP 4.5 simulations. The decadal forecasts also show an average bias of 1.5 million km² to the NSIDC-NT observations (that are currently available until 2017), which is not present in the uninitialized simulations.

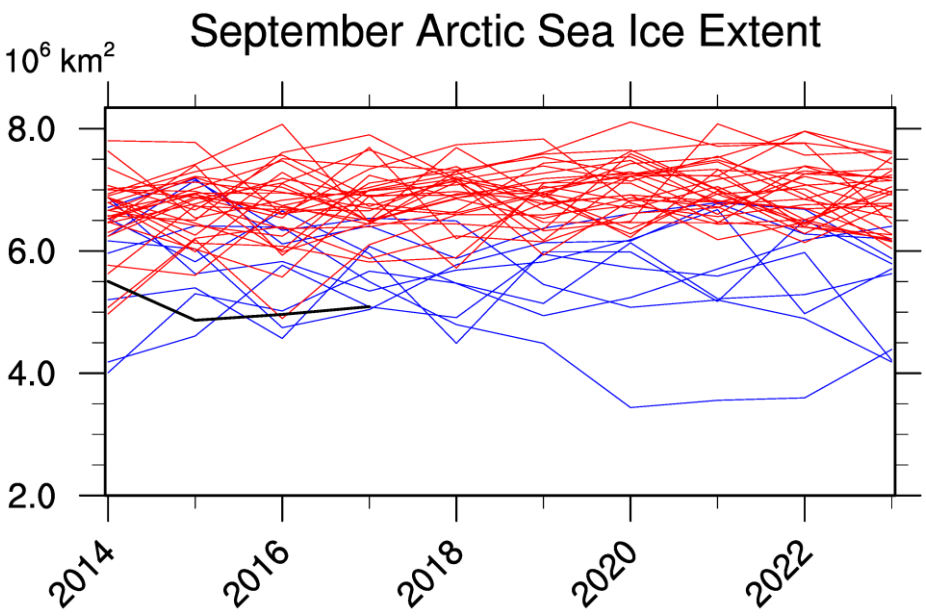

Figure 2.10. Timeseries of September Arctic SIE from MPI-ESM-LR *pr* decadal forecasts (red) and MPI-ESM-LR RCP 4.5 simulations (blue) for the time period 2014-2023. Shown are all available ensemble members of the model results and NSIDC-NT observations for 2014-2017 (black). The decadal forecasts shown here were initialized in 2013.

2.4 Summary and discussion

This chapter assessed the skill in reproducing observed Arctic sea ice, sea surface temperature and near-surface air temperature from the MiKlip prototype decadal prediction system performed with the MPI-ESM-LR by developing and applying the ESMValTool evaluation system for decadal climate predictions. It focused on answering the thesis' first scientific question: *Can model initialization with observations improve Arctic climate predictions?*

A statistically significant improvement in ACC skill for TAS by initialization in the North Atlantic was found, confirming results of previous studies (e.g., Meehl et al. 2009; Müller et al. 2012; Goddard et al. 2013; Marotzke et al. 2016). All of these studies found improved TAS predictions from initialized decadal hindcasts over uninitialized long-term historical simulations in various regions of the globe. These include, among others, the North Atlantic Ocean.

ACC skill scores of Arctic TAS and SST reveal that this improvement in skill in the North Atlantic due to initialization does not propagate into the Arctic and in some parts of it there is even a degradation in skill caused by the so-called initialization shock, i.e. a quick return of the initialized model to its biased equilibrium state after initialization (Meehl and Goddard 2013). The model sometimes overshoots its intrinsic climatology resulting in a larger bias

than without initialization. This can be seen from timeseries of September pan-Arctic SIA showing a strong model drift in the MiKlip prototype system of about 1 million km² during the ten-year simulation periods. Pohlmann et al. (2016) found that the initialization shock in the MPI-ESM-LR decadal hindcasts stems from an overestimated trend in wind stress in the reanalyses used for initialization, leading to a displaced thermocline and large SST anomalies.

The only region of improvement by initialization in the Arctic is found along the east coast of Greenland in March. Due to the initialization of oceanic variables, simulated SSTs are significantly improved in lead years 2-5, which translates to a reduction in winter SIC RMSE in the same region. The SIA integrated over the Greenland Sea correspondingly shows an increase in ACC skill in winter, which, however, lasts only for year one after initialization. This improvement in the Greenland Sea originates from a better representation of the observed year-to-year variability in the decadal hindcasts compared to the historical simulations. In the other Arctic regions virtually no improvement by initialization is found. In agreement with these results, Germe et al. (2014) found a generally weak potential predictability (less than 2 years) for Arctic summer sea ice extent from the CNRM-CM5.1 model. They also found most of the predictability coming from the Atlantic. In the Met Office Hadley Centre Decadal Prediction System (DePreSys), hindcasts of near-surface air temperature and sea surface temperature were improved by initialization in the North Atlantic in lead year 1 and in the Nordic Seas in lead year 2 suggesting a transport of skill northward into the Arctic (Liu et al. 2012). Accordingly, Collins (2002) found the highest potential predictability of near-surface air temperature over the North Atlantic region. These findings support the results in this chapter with high hindcast skill for TAS and SST in the North Atlantic that can translate to improved winter sea ice predictability in that region (Koenigk et al. 2012). Similarly, in the CESM model, Yeager et al. (2015) found significant skill scores in decadal predictions of sea ice extent in the Arctic sector of the Atlantic when including sea ice variables in the initialization process.

Why only certain regions are improved by initialization is still an open topic. In the Greenland Sea, the historical simulations show deficits in realistically simulating the year-to-year variability of sea ice. This is not expected from these long-term climate simulations that aim at predicting the multi-year average climate conditions. The initialization, however, improves the year-to-year variability in that region because the simulations start from the correct phase of the observed natural variability. This directly increases the ACC skill. The underlying reasons for the regionality of improvement need to be further investigated. For example, a sophisticated analysis of the parameterization of relevant processes in the model and a cross-variable assessment of multiple oceanic and atmospheric variables could help to identify the responsible processes.

Several studies suggest that the potential predictability limit estimated from model experiments for Arctic sea ice is higher than the actual prediction skill found here (e.g., Germe et al. 2014; Guemas et al. 2014). Since atmospheric and oceanic variables included in

the initialization process can improve regional temperature predictions, the inclusion of sea ice variables in the initialization process could potentially improve sea ice hindcasts. This has been done in seasonal forecasts with considerable success (e.g., Tietsche et al. 2014; Bunzel et al. 2016; Guemas et al. 2016). Therefore, additional analyses were made with the decadal hindcasts from the MiKlip preoperational decadal prediction system based on the high-resolution model configuration MPI-ESM-HR, where SIC was additionally initialized. However, improvement in TAS hindcast skill over the North Atlantic is not robust against different reference datasets in these simulations. Furthermore, timeseries of September pan-Arctic SIA show a strong negative bias of 1-3 million km² (dependent on reference observations and time) in all MPI-ESM-HR model runs. This bias is not present in March and indicates an unrealistic seasonal cycle of sea ice with too much melting at the end of summer in these simulations. Overall, the additional initialization of SIC brings no improvement in hindcast skill in the Arctic in the preoperational compared to the prototype system. These findings are consistent with several studies in the field of seasonal forecasts suggesting that only initializing SIC but no other sea ice quantity is not enough to improve sea ice forecasts. Only when also sea ice thickness or sea ice age is initialized together with SIC, seasonal forecasts of sea ice extent can be improved (e.g., Day et al. 2014a; Massonnet et al. 2015; Bushuk et al. 2017; Dirkson et al. 2017). This may be due to the fact that sea ice area has higher predictability in regions with thicker ice, with sea ice thickness being generally better predictable than sea ice area (Holland et al. 2010). It is important to note that the MPI-ESM-HR decadal hindcasts available to this study only consist of five ensemble members for each hindcast experiment as opposed to 30 members from MPI-ESM-LR. An increase in ensemble size was shown to improve the quality of decadal predictions (Sienz et al. 2015) by covering a higher variability range and could therefore change the specific conclusions for the MiKlip preoperational system.

Day et al. (2014b) showed that the hindcast quality also depends on the initialization month. Especially for improving the predictions of summer sea ice, initialization of the model in e.g. July instead of January might improve the skill. This is because there is a predictability limit in the melt season that can be overcome by initializing the model in the summer: two predictability re-emergence mechanisms (Blanchard-Wrigglesworth et al. 2011) are thus covered in the initialization: the first mechanism is the re-emergence of correlation occurring when the ice edge is in the same position during melting and freezing and originates from persistence of SST anomalies. The second mechanism is the re-emergence of skill from the persistence of last year's summer sea ice thickness anomalies. By initializing the model in July, the particular atmospheric and oceanic state in summer would be directly incorporated in the model, along with the information of skill re-emergence. In addition, the initialization with more consistent assimilation runs and additional components of the Earth system could further improve prediction skill. For example, in a recent study by He et al. (2017), a four-dimensional variational data assimilation technique (DRP-4Dvar) resulted in more consistent

initial conditions which could significantly reduce the initialization shock. Such techniques should be further investigated.

In conclusion, the short answer to the thesis' first question is that no considerable skill improvement by initialization was found for the Arctic with MPI-ESM. This chapter showed that uncertainties in retrospective predictions of Arctic climate are large. The same is true for the future projections. In the next chapter, the potential to reduce uncertainties in uninitialized long-term projections on the future development of Arctic sea ice is investigated. In addition to MPI-ESM-LR, the results of another 28 CMIP5 Earth system models are used.

3. Constraining uncertainties in projections of Arctic sea ice extent

This chapter is based on Senftleben et al. (2019), see Appendix C.

Observations show that the ongoing warming of the Earth caused the September Arctic SIE to shrink by almost 50% since the 1970s (Stroeve et al. 2012a). Earth system models participating in CMIP5 simulate a further decrease throughout the 21st century in all future RCP scenarios that keep atmospheric greenhouse gas concentrations on the current level or higher. In the RCP 8.5 scenario where atmospheric CO_2 concentration more than double by the end of the 21st century compared to 2000 (Riahi et al. 2011), almost all models project the Arctic to become ice free in summer before 2100. The year of near-disappearance of summer Arctic sea ice (YOD) is defined as the first year of a series of five consecutive years in which the minimum SIE drops below one million km^2 (Wang and Overland 2009). While some models project a YOD beyond 2100, others simulate an ice-free Arctic before 2020. The large spread in the projections stems from three fundamentally different sources of uncertainty: internal (natural) variability, different assumptions and parameterizations in the models (model uncertainty) and the uncertainty in future RCP scenarios (Kay et al. 2011; Stroeve et al. 2012b; Liu et al. 2013; Swart et al. 2015).

The goal of this chapter is to to answer scientific question *Q2*, to quantify and reduce the uncertainties in projections of Arctic SIE, and to give a more precise estimate of YOD than currently available. Therefore, climate projections from a set of 29 Earth system models from 17 different institutes participating in CMIP5 are analyzed. The three different sources of uncertainty in simulated projections of sea ice extent are quantified using the method of Hawkins and Sutton (2009, 2011). For internal variability, this is compared to an estimate resulting from a single model large ensemble similar to Jahn et al. (2016). To narrow model uncertainty, the multiple diagnostic ensemble regression (MDER) method introduced by Karpechko et al. (2013) is used. This method uses statistical relationships between the projected future SIE and historical model performance in terms of different process-oriented diagnostics. Additionally, MDER uses pseudo reality to cross-validate the results and filter spurious relationships. Applying this formal framework to assess the benefits of model weighting compared to using an unweighted multi-model mean is a significant improvement compared to previous studies on Arctic sea ice that use model weighting (see Section 3.1).

The first step in applying MDER is to find observable process-oriented diagnostics that are related to future SIE. Since sea ice is affected by, for example, air and sea-surface temperature and winds, and is a driver of Arctic climate feedbacks, a large variety of

different observable diagnostics that represent these processes is included in MDER (see Table 3.1). This diagnostic selection is unavoidably subjective and relies on expert knowledge. However, in a second step, the most important diagnostics for future SIE are filtered by applying MDER's iterative stepwise regression algorithm that selects the linear combination of diagnostics (i.e., the regression model) that best constrains future SIE. Then, model weights are calculated based on this regression model and the corresponding observations. The weights are used to narrow model uncertainty and constrain the CMIP5 projections of Arctic SIE.

This chapter is structured in the following way: Section 3.1 gives a short overview on the recent literature published on reducing uncertainties in Arctic sea ice projections. Section 3.2 describes the MDER method and how the different types of uncertainty can be quantified. Additionally, the different diagnostics used in this study are introduced and the model simulations and observations are described. In Section 3.3, MDER is applied to constrain SIE projections and narrow model uncertainty. A new estimate of YOD based on the MDER-weighted CMIP5 projections is given for the RCP 8.5 scenario. Section 3.4 closes with a summary and discussion.

3.1. State-of-the-art of uncertainty reduction in Arctic sea ice projections

So far, an MDER based weighting of multi model results has not been used to reduce uncertainties in CMIP5 Arctic sea ice projections. However, different methods have been applied to reduce uncertainties in sea ice projections, and are presented in this section. For example, Massonnet et al. (2012) constrained projections with observations using SIE, sea ice thickness (SIT) and sea ice volume. They select models that successfully simulate the reference observations based on a novel model selection algorithm. With this, they find that future Arctic sea ice loss is linked to present-day sea ice conditions and estimate YOD for the RCP 8.5 scenario to occur between 2041 and 2060. Similarly, Liu et al. (2013) used model selection and the relationship between present and future sea ice conditions to constrain the projected YOD with a statistical fit to reference observations. Their estimate of YOD is even narrower than Massonnet et al. (2012): 2054-2058.

Melia et al. (2015) focused on SIT projections. From different statistical bias correction techniques, they developed a method to constrain the CMIP5 SIT simulations with Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS) reanalysis data (Zhang and Rothrock 2003). They find model uncertainty to contribute most to the total uncertainty and can reduce it by bias-correcting the variance and the mean of the models. With this method, they can narrow the range of YOD to before 2085 and predict it to occur in the 2050s.

To estimate the uncertainty of YOD, Jahn et al. (2016) used large ensemble simulations of one CMIP5 model. With 55 ensemble members, it could be shown that internal variability

alone accounts for an uncertainty range in YOD of 21 years. Using both RCP 4.5 and RCP 8.5 simulations, they find that scenario uncertainty adds another 5 years to this uncertainty range. They conclude that even though the decline of Arctic SIE in the 21st century is not due to internal variability, a precise estimate of YOD is impossible.

Knutti et al. (2017) used model weighting to reduce projection uncertainty in Arctic SIE. They present a weighting scheme that accounts for both model performance and interdependence, i.e., the similarity between simulations of different models that share some of their components or model code. For the weighting scheme, they use multiple sea ice and surface temperature diagnostics like climatological mean September SIE, September SIE trend, and interannual variability of monthly surface temperature. Their weighted CMIP5 model mean shows both a faster decline of Arctic SIE and a smaller model spread than the unweighted mean, but exact values are dependent on the selected diagnostic. A concern raised by the authors is therefore selecting the correct diagnostics for their weighting scheme.

In this chapter, the MDER method is used to formally select the diagnostics that best constrain future Arctic SIE and to filter out diagnostics that exhibit spurious relationships to the target variable. The method is detailed in the following section.

3.2. Methods

The goal of the study presented in this chapter is to reduce uncertainties in multi-model projections of SIE by applying the multiple diagnostic ensemble regression (MDER) method (Karpechko et al. 2013; Wenzel et al. 2016) to simulations from the CMIP5 archive. Different sources of uncertainty (Hawkins and Sutton 2009) in SIE projections are quantified, and the potential of MDER to reduce model uncertainty is demonstrated. This section describes the MDER method, the diagnostics used within MDER, and how the different types of uncertainty are quantified. The section closes with a description of the model experiments and the observational or reanalysis data that were used for this study.

3.2.1. Multiple Diagnostic Ensemble Regression (MDER)

MDER was developed to improve multi-model climate projections and has proven its potential by constraining projections of Antarctic total column ozone (Karpechko et al. 2013) and the position of the summer austral jet stream (Wenzel et al. 2016). The method is based on the correlation between selected process-oriented diagnostics applied to historical or present-day periods for which observations are available, and a future target variable. In an

Table 3.1. Overview of the 15 diagnostics used in this study, including the reanalyses and observations used to constrain the models and their respective references. All diagnostics were calculated over the years 1979-2012.

Acronym	Diagnostic	Reanalysis or observation	Reanalysis or observational value
SIE_c	Climatological mean September Arctic sea ice extent	NSIDC-NT (Cavalieri et al. 1996; Walsh et al. 2015)	$(6.14 \pm 0.16) \times 10^6$ m²
SIE_t	September Arctic sea ice extent trend		$(-0.65 \pm 0.01) \times 10^6$ m² decade^{-1}
SIE_i	Interannual variability of September Arctic sea ice extent		0.95×10^6 m²
SIT_c	Climatological mean September Arctic sea ice thickness	PIOMAS (Zhang and Rothrock 2003)	1.10 ± 0.05 m
SIT_t	September Arctic sea ice thickness trend		-0.254 ± 0.002 m decade^{-1}
SIT_i	Interannual variability of September Arctic sea ice thickness		0.28 m
TAS_c	Climatological mean summer Arctic surface air temperature	ERA-Interim (Dee et al. 2011)	274.9 ± 0.07 K
TAS_t	Summer Arctic surface air temperature trend		0.26 ± 0.01 K decade^{-1}
TAS_i	Interannual variability of summer Arctic surface air temperature		0.40 K
SST_c	Climatological mean summer Arctic sea surface temperature	HadISST (Rayner et al. 2003)	273.79 ± 0.04 K
SST_t	Summer Arctic sea surface temperature trend		0.183 ± 0.003 K
SST_i	Interannual variability of summer Arctic sea surface temperature		0.25 K
PSL_c	Climatological mean September Arctic surface pressure	ERA-Interim (Dee et al. 2011)	1011.20 ± 0.45 hPa
PSL_t	September Arctic surface pressure trend		-0.89 ± 0.04 hPa decade^{-1}
PSL_i	Interannual variability of September Arctic surface pressure		2.66 hPa

iterative step-wise regression algorithm based on von Storch and Zwiers (1999), MDER selects those diagnostics that best explain the future variable while avoiding redundancy. The regression model is then a linear combination of the selected process-oriented diagnostics and is used to constrain the future variable with observations.

The target variable in this chapter is the September Arctic SIE. It is derived from the gridded variable SIC, which describes the fraction of each grid cell that is covered with sea ice. SIE is defined as the total area of all ocean surface grid cells in which SIC ≥ 15%. The month September typically represents minimum sea ice conditions in the Arctic and is commonly used in the literature on sea ice projections, for example in analyses of the timing of an ice-free Arctic (e.g., Massonnet et al. 2012; Jahn et al. 2016; Screen 2018; Sigmond et al. 2018). The target period for future SIE is 2020-2044. This 25-year period is selected to ensure that SIE is larger than zero from all models and at all times, since correlations between historical diagnostics and future SIE may become spurious otherwise. Small noisy fluctuations in sea ice extents close to zero can have a large impact on the correlation coefficient: since it is inversely related to the standard deviation, which is particularly small in near-zero ice conditions, the coefficients can become artificially high due to auto-correlation.

The set of process-oriented diagnostics that was selected by the author as input for the MDER method is listed in Table 3.1 and described in Section 3.2.2, and is based on published literature on processes that determine sea ice concentration. The selection is by no means meant to be an exclusive list. Finding emergent constraints (e.g., Bracegirdle and Stephenson 2012; Borodina et al. 2017) for Arctic sea ice is an ongoing scientific topic and the study can easily be repeated with more or different diagnostics once they become 39vailable. MDER's stepwise regression algorithm takes the set of selected diagnostics as an input and selects a subset of diagnostics used to build the regression model. The selection algorithm iteratively adds and removes diagnostics to and from the regression model until the regression sum of squares is maximized. The stopping criterion is based on an *F* test with a significance level p = 0.1. The final regression model then is the linear combination of the selected diagnostics that best predicts the future SIE and is of the form of

$$y = \beta_0 + \mathbf{X}^{\mathrm{T}}\beta, \tag{3.1}$$

where y is the estimated climate response (here: future SIE), β_0 and $\boldsymbol{\beta}$ are the multiple regression parameters with $\boldsymbol{\beta}$ being a column vector of the size of the number of the selected diagnostics, and $\mathbf{X}$ the matrix of diagnostic values of the selected diagnostics. The values obtained by applying the selected diagnostics to observations are used with the regression model to yield a multi-diagnostic constraint. MDER then calculates model weights based on this constraint following Bracegirdle and Stephenson (2012):

$$\mathbf{W} = [\mathbf{N}^{\mathrm{T}} + (\mathbf{X}_0^{\mathrm{T}} - \overline{\mathbf{X}}^{\mathrm{T}})(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{1}\overline{\mathbf{X}}^{\mathrm{T}})^{-1}(\mathbf{X}^{\mathrm{T}} - \mathbf{X}^{\mathrm{T}}\mathbf{1}\mathbf{N}^{\mathrm{T}})]^{\mathrm{T}}, \quad (3.2)$$

with $\mathbf{N} \equiv (\mathbf{1}^{\mathrm{T}}\mathbf{1})^{-1}\mathbf{1}$ a vector of a size equal to the number of models (n) and the value of all elements equal to n^{-1}, $\mathbf{X}_0$ the vector of observed diagnostics, and $\overline{\mathbf{X}} \equiv (\mathbf{N}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}$ the vector of the multi-model mean diagnostics. The weights are used to calculate a weighted multi-model mean from the CMIP5 model ensemble with the model uncertainty reduced.

The diagnostic selection is one of the key strengths of MDER: compared to simply taking all subjective diagnostics (e.g., Snape and Forster 2014; Knutti et al. 2017; Sanderson et al. 2017), the step-wise regression algorithm reduces redundancy, which is the risk of including multiple diagnostics that effectively describe the same physical process. Another potential danger in using all diagnostics instead of only the MDER-selected diagnostics lies in overfitting caused by spurious relationships between historical diagnostics and the future variable (Bracegirdle and Stephenson 2012). To filter spurious relationships, the MDER results are cross-validated in a pseudo-reality approach: since observations of future variables are naturally unavailable, one model at a time is selected as reference (= pseudo reality) to benchmark the other models by calculating the root mean square error (RMSE) as a measure of prediction uncertainty. Since there is no preferred reference model, each model is taken as pseudo reality once and all of the remaining models are tested against it. The difference in RMSE between the unweighted multi-model mean (uMMM) and the MDER results reveals the potential of MDER to reduce uncertainty in the SIE projections and is a measure of uncertainty of the MDER results.

3.2.2. Diagnostics

The diagnostics used as input for the MDER calculations are process-oriented and cover most of the variables (at least as proxies) that are known to have an impact onArctic sea ice. For each of those variables, three different metrics are calculated: the climatological mean (indicated by _c), the trend (_t) and the interannual (“year-to-year”) variability (_i); see also Table 3.1. Two different sea ice variables are selected to identify biases in SIE predictions (e.g., Laxon et al. 2003; Massonnet et al. 2018): historical SIE and sea ice thickness (SIT). To account for freezing and melting processes, two temperature variables are selected to cover the thermal influence on the ice from above and below (e.g., Zhang et al. 2000; Weeks 2010): near-surface air temperature (TAS) and sea-surface temperature (SST). Atmospheric surface pressure (PSL) is used as a proxy for the influence on ice drift due to atmospheric winds near the surface (e.g., Thorndike and Colony 1982; Spreen et al. 2011). An additional diagnostic to account for the influence of radiation (sea-ice albedo effect (Curry et al. 1995), SI_alb) was originally included, but had to be removed because the required variables were only available from 19 models. As a sensitivity test, study was repeated with the 19 models

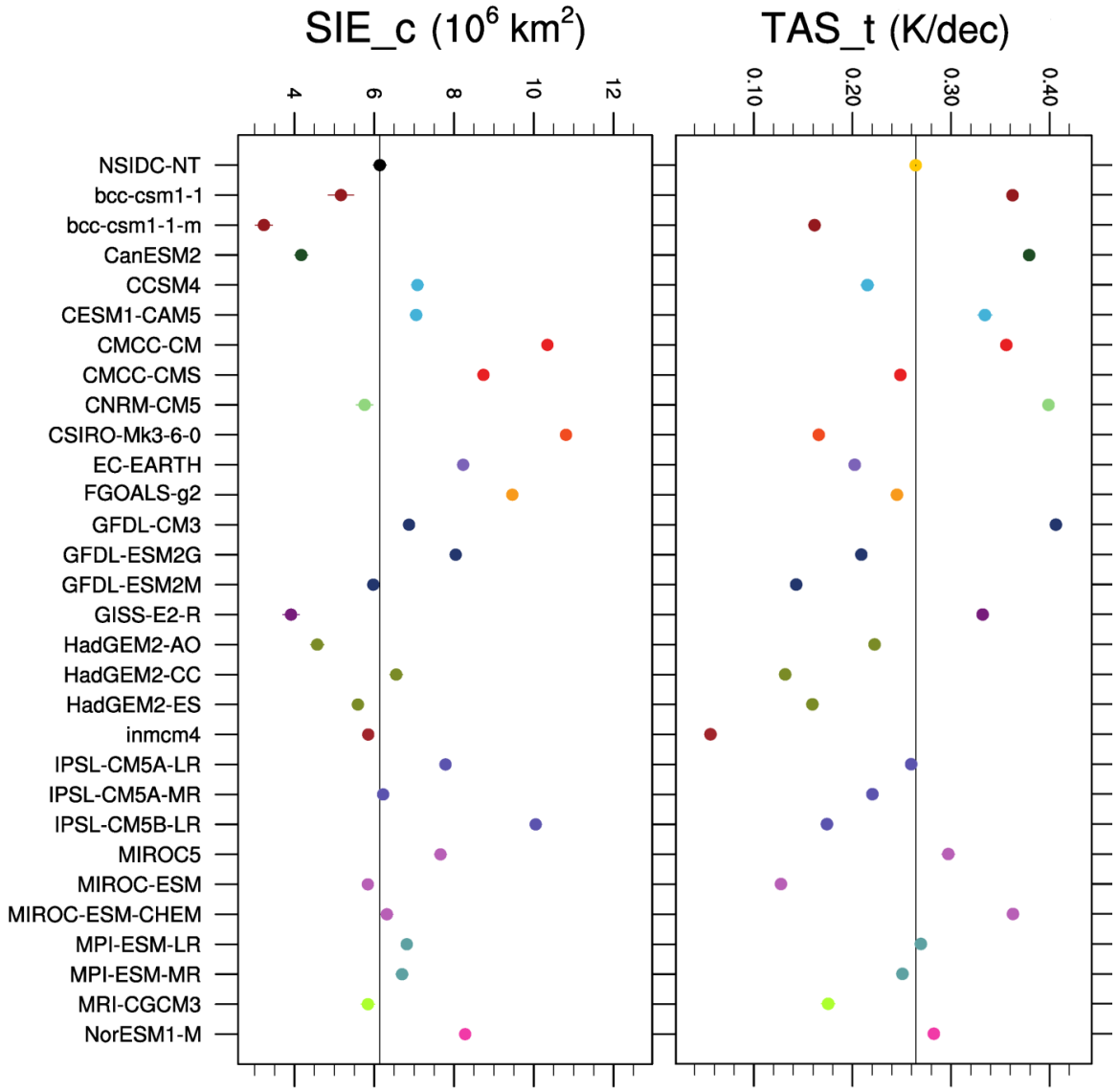

Figure 3.1. Diagnostic values from all CMIP5 models (ensemble means) used in this study (see Table 3.2) for the SIE_c and TAS_t diagnostics (1979-2012). The observational reference value is added in the topmost row, together with a vertical reference line. For details on the diagnostics, see Table 3.1. For the other diagnostics, see Appendix B.

that provided the variables for calculating SI_alb. As it turned out, this diagnostic had the smallest correlation coefficient to future SIE of all diagnostics (r = 0.04, not statistically significant) and was not selected by the stepwise regression algorithm, so the omission of SI_alb is not expected to have an important impact on the results presented in this chapter.

All diagnostics are applied for the same time period, i.e. 1979-2012. This time period was selected so it is as long as possible given that the earliest start is the year 1979 because reliable observations of the required variables are not available for earlier times (see Section 3.2.4). The calculated diagnostic results of each model and observation/reanalysis for two

diagnostics are shown in Figure 3.1 and the values of all diagnostics are in the supplemental information (Appendix B).

3.2.3. Uncertainty estimation

In addition to the MDER method, this chapter applies the method of Hawkins and Sutton (2009, 2011) to quantify different sources of uncertainty in the CMIP5 multi-model projections of the 21st century SIE (2006-2100): internal variability, model uncertainty, and scenario uncertainty. Numerous studies have applied this method to Arctic variables such as CMIP3 summer Arctic sea ice extent (Lique et al. 2016) to investigate Arctic freshwater change, CMIP3 Arctic temperature and precipitation (Hodson et al. 2012), and CMIP5 September Arctic sea ice thickness (Melia et al. 2015). This section describes the sources of uncertainty and how they can be separated, following Hawkins and Sutton (2009, 2011).

To quantify the contribution of internal variability to the overall uncertainty of the model ensemble, each individual projection from all models is individually fit with a fourth-order polynomial over the time period 2006-2100 using the least squares method and creating a smooth fit. The predictions X from each model m for scenario s and year t can be written as

$$X_{m,s,t} = x_{m,s,t} + i_{m,s} + \varepsilon_{m,s,t} \tag{3.3}$$

with i being the observed reference value (year 2000 SIE), x the smooth fit and ε the residual. The internal variability component V is the multi-model mean of the variances of the residuals:

$$V = \frac{1}{N_m} \sum_m var_{s,t}\left(\varepsilon_{m,s,t}\right), \tag{3.4}$$

with $var_{s,t}$ the variance over all scenarios and over time and N_m the number of models. Note that here, V is constant in time by design, which is limitation of this method, especially since interannual variability has been shown to increase as the ice thins (Goosse et al. 2009). The internal variability obtained with this method is compared in Section 3.3.2 to a more recent method by Jahn et al. (2016) who used a large ensemble of a single ESM to estimate internal variability.

Model uncertainty is estimated from the variance in the fits $x_{m,s,t}$ of each scenario. The multi-scenario mean model uncertainty $M(t)$ is calculated as follows:

$$M(t) = \frac{1}{N_s}\sum_s var_m(x_{m,s,t}) \,, \tag{3.5}$$

with N_s the number of scenarios.

The CMIP5 projections are divided into sets of experiments of which each assumes a different amount of external forcing due to the emission of climate-relevant substances. More specifically, the experiments represent four different representative concentration pathways that assume a different radiative forcing in W m^{-2} by the end of 2100: RCP 2.6, RCP 4.5, RCP 6.0, and RCP 8.5. The scenario uncertainty is calculated as the variance of the multi-model means over the RCP scenarios:

$$S(t) = \frac{1}{N_m} var_s\left(\sum_m x_{m,s,t}\right). \tag{3.6}$$

In this chapter, only the scenarios RCP 4.5 and RCP 8.5 were considered, because the variable SIC is not available from all 29 models for the other RCPs. This means that this study considers a medium and a high ("business-as-usual") RCP scenario. However, the study was repeated with only 19 models for which also output for RCP 2.6 was available. The results of this additional analysis are described in Section 3.3.2.

It is important to note that for the analyses in this chapter, the square root of all sources of uncertainty is taken following other studies applying this method (Hodson et al. 2012; Lique et al. 2016). The estimates are therefore based on the standard deviation as opposed to the variance. Furthermore, no ensemble averages are taken in the calculation of the uncertainty estimates.

3.2.4. Models, observations and reanalyses

For this study, only 29 of the around 40 CMIP5 models could be used, because not all models provided the required output. The 29 models are listed in Table 3.2. The historical experiments were extended (here, 1979-2005) with RCP 4.5 results up to the year 2012. The method was also applied to a smaller model ensemble extended with RCP 8.5 which gave similar results (not shown). For each experiment, some models run multiple ensemble members with slightly different initial conditions to estimate internal variability. Only those ensemble members from each model were selected that were available for all experiments and variables. To account for the different ensemble sizes, for each diagnostic an ensemble average is calculated of the diagnostic values which is then used in the analyses. Thus, the multi-model metrics are not biased towards models with many ensemble

Table 3.2. The 29 CMIP5 models used in this study. All available ensemble members (EM) were used for each model, and an ensemble mean was calculated for each model prior to MDER calculations. The numbering starts from 2 for technical reasons (no. 1 is reserved for observations/reanalyses).

No.	Model	Institute	#EM	Reference
02	BCC-CSM1.1	Beijing Climate Center (BCC)	1	Wu et al. (2014)
03	BCC-CSM1.1(m)		1	
04	CanESM2	Canadian Centre for Climate Modelling and Analysis (CCCma)	5	Arora et al. (2011)
05	CCSM4	National Center for Atmospheric Research (NCAR)	5	Gent et al. (2011)
06	CESM1-CAM5		3	
07	CMCC-CM	Centro Euro-Metiterraneo per I Cambiamenti Climatici, Italy	1	Vichi et al. (2011)
08	CMCC-CMS		1	
09	CNRM-CM5	Centre National de Recherches Météorologiques – Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CNRM-CERFACS)	1	Voldoire et al. (2012)
10	CSIRO-Mk3.6.0	Commonwealth Scientific and Industrial Research Organization/Queensland Climate Change Centre of Excellence (CSIRO-QCCCE)	10	Rotstayn et al. (2012)
11	EC-EARTH	European EC-Earth consortium	3	Hazeleger et al. (2010)
12	FGOALS-g2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS	1	Li et al. (2013)
13	GFDL-CM3	National Oceanic and Atmospheric Administration – Geophysical Fluid Dynamics Laboratory (NOAA-GFDL)	3	Donner et al. (2011)
14	GFDL-ESM2G		1	Dunne et al. (2013)
15	GFDL-ESM2M		1	
16	GISS-E2-R	NASA Goddard Institute for Space Studies, USA	2	Schmidt et al. (2014)
17	HadGEM2-AO	Met Office Hadley Centre	1	Martin et al. (2011)
18	HadGEM2-CC		1	
19	HadGEM2-ES		3	
20	INMCM4	Russian Institute for Numerical Mathematics, Russia	1	Volodin et al. (2010)
21	IPSL-CM5A-LR	Institute Pierre-Simon Laplace (ISPL)	4	Dufresne et al. (2013)
22	IPSL-CM5A-MR		1	
23	IPSL-CM5B-LR		1	
24	MIROC5	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, and National Institute for Environmental Studies, Japan	3	Watanabe et al. (2011)
25	MIROC-ESM		1	
26	MIROC-ESM-CHEM		1	
27	MPI-ESM-LR	Max Planck Institute for Meteorology (MPI-	3	Giorgetta et al.

28	MPI-ESM-MR	M)	3	(2013)
29	MRI-CGCM3	Meteorological Research Institute, Japan (MRI)	1	Yukimoto et al. (2012)
30	NorESM1-M	Norwegian Climate Centre (NorClim)	1	Iversen et al. (2012)

members (see also Massonnet et al. 2012). The ensemble average is taken after the diagnostic calculation since calculating the diagnostics on ensemble means would give incorrect estimates of, especially, interannual variability.

The observations and reanalyses for each diagnostic are listed in Table 3.1 and are provided as monthly means. Reliable observations only exist since 1979 when satellite data became routinely available. In the following, the five datasets used in this study are briefly described.

Satellite observations of SIC are used from the National Snow and Ice Data Center (NSIDC, Walsh et al. 2015). The product is available from 1978 to present and was processed with the NASA-Team retrieval algorithm (NT, Cavalieri et al. 1996) from data of Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave sensors. The spatial resolution is 25 x 25 km.

Sea ice thickness reanalyses are taken from the Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS, Zhang and Rothrock 2003), a coupled ice-ocean model forced with National Centers for Environmental Prediction (NCEP) reanalyses and assimilating observations of sea ice concentration and sea surface temperature. Satellite observations of SIT are spatially and temporarily limited (e.g., Kwok et al. 2009; Tilling et al. 2015). Since complete temporal and Arctic-wide coverage is required for the application of the MDER method, the PIOMAS reanalyses are used. PIOMAS was compared to SIT satellite observations and found to be a good estimate of the observed SIT in numerous studies (e.g., Lindsay and Zhang 2006; Schweiger et al. 2011; Laxon et al. 2013; Stroeve et al. 2014). PIOMAS reanalyses are commonly used in studies analyzing SIT (e.g., Melia et al. 2015; Dirkson et al. 2017; Labe et al. 2018).

The Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST, Rayner et al. 2003) provide SST data. It is a global reanalysis product combining data from the Met Office Marine Data Bank (MDB), the Global Telecommunications System (GTS) and the Comprehensive Ocean-Atmosphere Data Set (COADS) and has a spatial resolution of 1° × 1°.

For TAS and PSL, the European Centre for Medium-Range Weather Forecast Re-Analysis Interim data (ERA-Interim, Dee et al., 2011) are used. The data assimilation system is based on the Integrated Forecast System, cycle 31r2 (IFS-Cy31r2) and data are available from 1979 with monthly updates and an approximate horizontal resolution of 80 km.

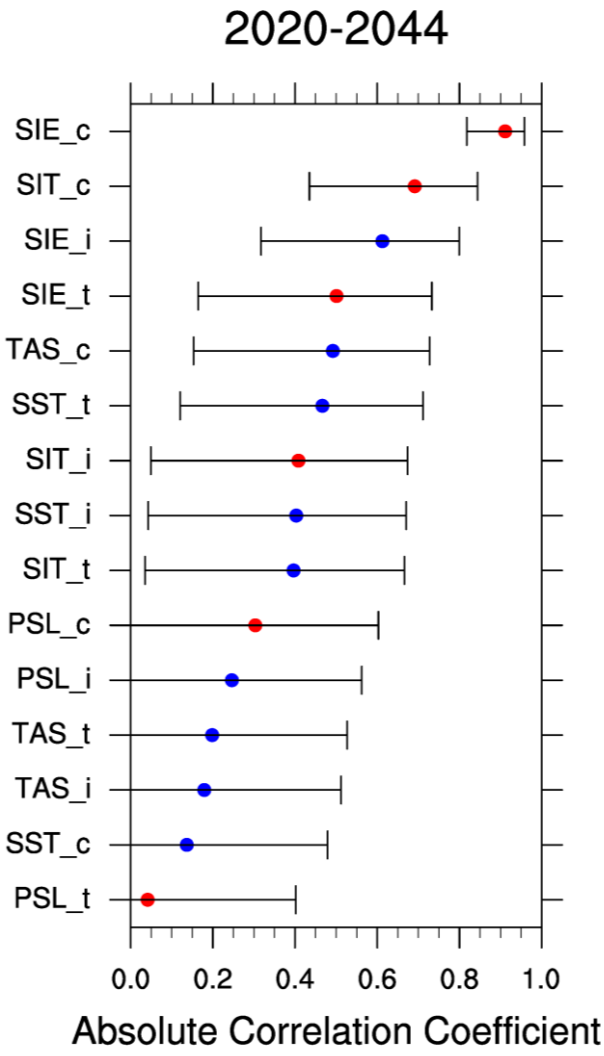

Figure 3.2. Absolute correlation coefficients between the diagnostics (see Table 3.2.1) calculated from historical simulations (1979-2012) and the future SIE (2020-2044) from RCP 8.5 simulations. The correlation coefficients have been calculated for the 29 CMIP5 models given in Table 3.2. Positive coefficients are shown in red and negative coefficients are shown in blue, and error bars indicate the 95% confidence intervals around the correlation coefficients. Correlations with confidence intervals that include negative values (zero) are not statistically significant

3.3. Application of MDER to Arctic sea ice projections

The MDER method presented and applied in Karpechko et al. (2013) and Wenzel et al. (2016) was applied here to projections of September Arctic SIE. The process-oriented historical diagnostics listed in Table 3.1 are candidates for constraining the projections and were calculated using data over the years 1979-2012. The target SIE projections in the MDER regression algorithm were taken from simulations under the RCP 8.5 scenario and calculated over 2020-2044 (Section 3.3.1). The weights produced by MDER were then applied to 21st century SIE projections to narrow model uncertainty (Section 3.3.2) and improve the predictions of YOD (Section 3.3.3). The complete analysis was also done with RCP 4.5 simulations (not shown) giving similar results since differences due to the different forcing scenario in the multi-model projections of SIE are small before 2050 compared to model uncertainty (see also Section 3.3.2). Table 3.2 lists all CMIP5 models that are used in this study.

3.3.1. Diagnostic selection and regression model

Figure 3.2 shows the absolute correlation coefficients of all diagnostics (1979-2012) with future SIE (2020-2044). The climatological mean September Arctic sea ice extent (SIE_c) is the diagnostic with by far the highest correlation coefficient (r = 0.91) and smallest uncertainty range. This means that biases in the historical mean sea ice extent dominate the projections of near-term future sea ice extent: models that simulate a small (large) SIE in the past simulate a small (large) SIE in the near future. Nine out of the 15 diagnostics show a statistically significant correlation with future SIE. After SIE_c, the climatological mean sea ice thickness (SIT_c) shows the largest correlation to future SIE with a correlation coefficient of r = 0.69. Thinner ice is more prone to melting in summer and thus resulting in a smaller future ice extent, but this relationship has a significantly larger uncertainty than SIE_c. The other diagnostics with statistically significant correlations with future SIE are (ordered by decreasing correlation coefficient) SIE_i, SIE_t, TAS_c, SST_t, SIT_i, SST_i, and SIT_t, where i stands for internal variability, t for trend, and c for climatological mean. This means that projections of SIE are in particular influenced by biases in past sea ice conditions and Arctic surface temperatures, but not so much by biases in Arctic wind patterns as estimated by the proxy surface pressure. The uncertainty ranges of all those diagnostics are larger than that of SIE_c by multiple times and their correlation coefficients are smaller with values of around r = 0.5. Near-term future SIE can thus be predominantly improved by correcting for the biases in historical mean SIE. The values for each diagnostic from models and observations are given in Appendix B (Figures B1 through B5) and the scatterplots (Figures B6 and B7) show the correlations between all diagnostics and future SIE. The correlation coefficients were also calculated for a longer target time period (2016-2064, not shown) and were similar for all diagnostics, suggesting that the results are not very sensitive to the selected target period.

It is important to note that internal variability may contribute by 30-50% to SIE_t (Kay et al. 2011; Swart et al. 2015) and should ideally be removed from the trend. To further analyze the CMIP5 SIE trends, trend distributions across the 29 models are analyzed for the historical and future time periods (Figure 3.3). For both time periods and in both scenarios, all models show a negative trend over the complete time periods, ranging between -1.6 million km² per decade and -0.1 km² per decade. As expected, the models simulate a stronger future trend for RCP 8.5 than for RCP 4.5 (compare Figures 3.3b and 3.3c). The observations show a stronger trend (-0.6 million km² per decade) than most models for the historical time period (Figure 3.3a), as can also be seen in Figure 3.1. Swart et al. (2015) also find this model bias (their Figure 2c). They show that internal variability affects the SIE trend distributions for 7-year and 14-year time periods much more than for 35-year periods (see their Figure 2). The historical time period for the diagnostics in this chapter covers 34 years (1979-2012), and the

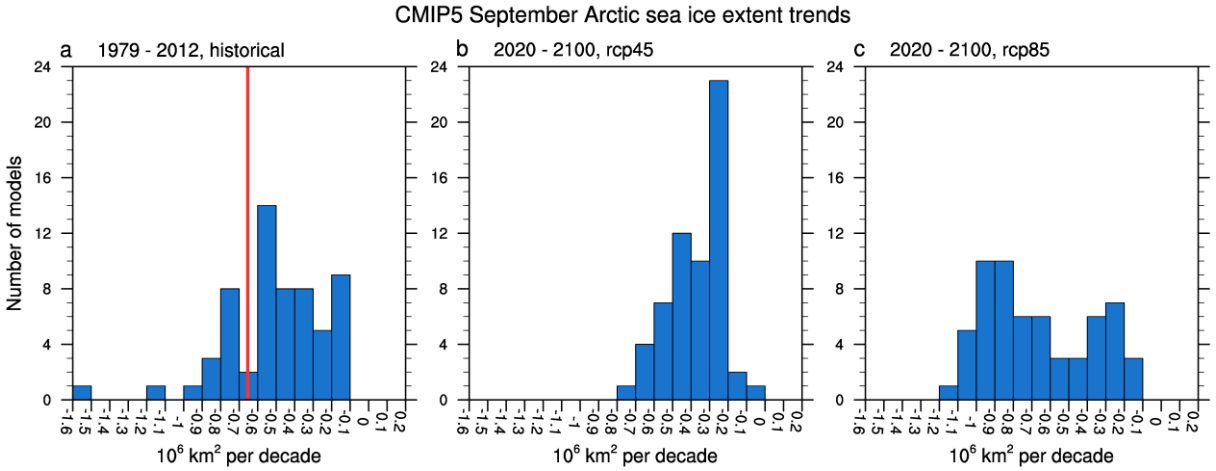

Figure 3.3. Distributions for SIE_t in all ensemble members of the 29 CMIP5 models for (a) the historical simulations (1979-2012) extended with RCP 4.5 after 2005, (b) the RCP 4.5 simulations (2020-2100), and (c) the RCP 8.5 simulations (2020-2100). The red vertical line in Figure 2a represents the trend in NSIDC-NT observations (also 1979-2012).

trend distribution for this time period (Figure 3.3a) is similar to Figure 2c of Swart et al. (2015), which means that internal variability might not play a major role here. As discussed further below, SIE_t is not selected by MDER in this study and its signal is expected to become smaller and therefore even less important if the fraction of internal variability in SIE_t was removed. Hence, the contribution of internal variability to SIE_t was not taken account for here.

The linear combination of multiple diagnostics (parsimonious regression model) that best predicts future SIE is calculated from the pool of diagnostics tested (Table 3.1) by the iterative stepwise regression algorithm within MDER. The scatterplot in Figure 3.4 shows that the two diagnostics selected by MDER to constrain future SIE are SIE_c and TAS_t, with the regression model equation -2.99 + 1.16 × SIE_c – 2.97 × TAS_t. This means that future SIE can be constrained by a linear combination of a bias correction in the historical simulations of SIE and the trend in Arctic surface temperature. Figure 3.1 reveals that the unconstrained values of SIE_c from most models show a positive bias compared to the NSIDC-NT observations (negative constant in the regression model equation). The positive sign of the SIE_c term reflects the positive correlation between historical and future SIE, and the negative sign of the TAS_t term shows that models that simulate a strong positive temperature trend in the Arctic in the historical simulations tend to simulate a smaller SIE in the future. Most models show a negative bias in TAS_t suggesting that the reduction in the constrained SIE is partly due to the TAS_t diagnostic. This is not surprising since Arctic temperatures have risen about twice as much as the global average, largely as a result of sea ice loss (Bellucci et al. 2015). Ding et al. (2017) show that internal atmospheric variability plays a considerable role in recent Arctic change. However, no interannual variability

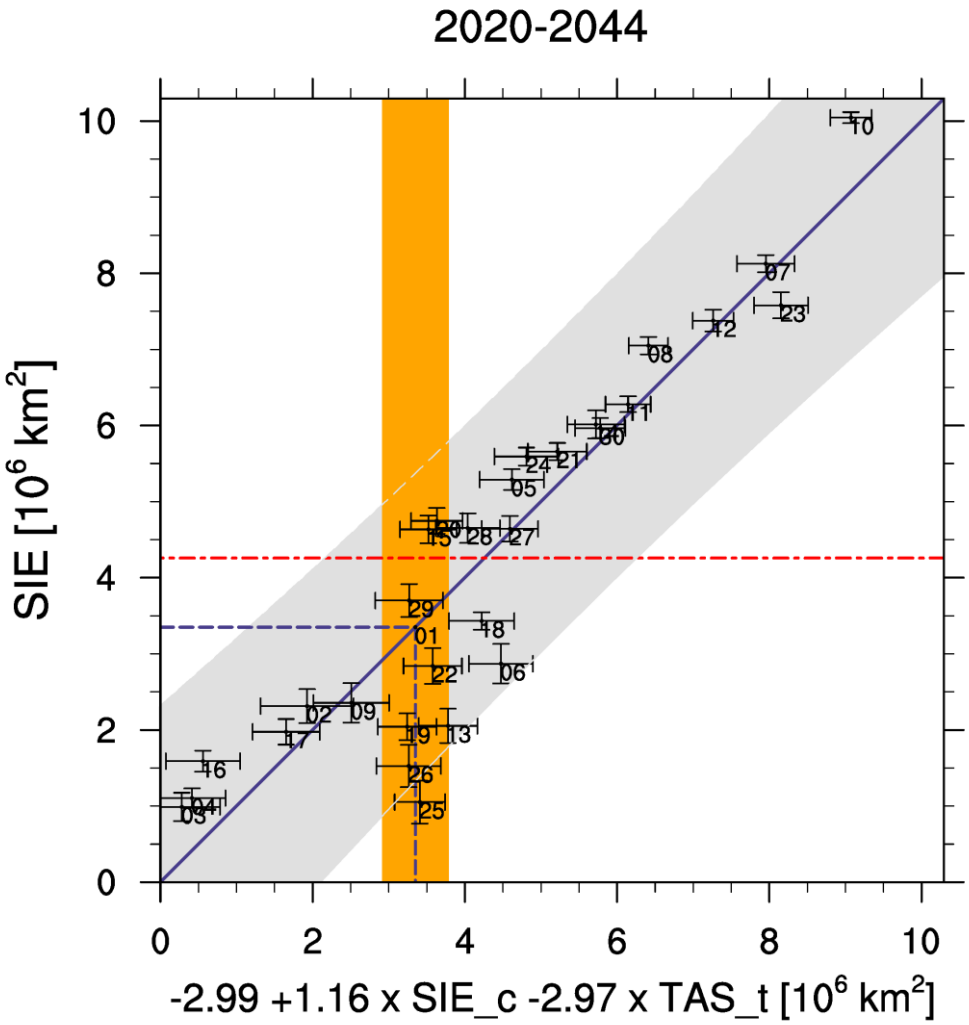

Figure 3.4. Scatterplot showing the relationship between the future climatological mean (2020-2044) September Arctic sea ice extent (RCP 8.5) and the regression model (-2.99 + 1.16 × SIE_c – 2.97 × TAS_t). The models are numbered according to Table 3.2 and error bars show one-standard deviation of the mean values. The solid blue diagonal line is the least squares linear fit to the models and grey shading indicates the 95% prediction interval for the linear regression. The orange vertical bar shows one-standard deviation around the observed climatological mean value (blue dashed vertical line) estimated from NISDC observations and ERA-Interim reanalysis data. The dashed horizontal lines indicate the unconstrained multi-model mean prediction (red) and the constrained MDER prediction (blue).

diagnostic was selected by MDER. The selected diagnostics also show that the linear combination of diagnostics that best predicts the target variable does not necessarily contain the diagnostics with the highest correlation to the target variable, since the correlation coefficient between TAS_t and future SIE is not statistically significant (see Figure 3.2). Applying the regression model equation to the observed SIE_c and TAS_t (blue dashed lines

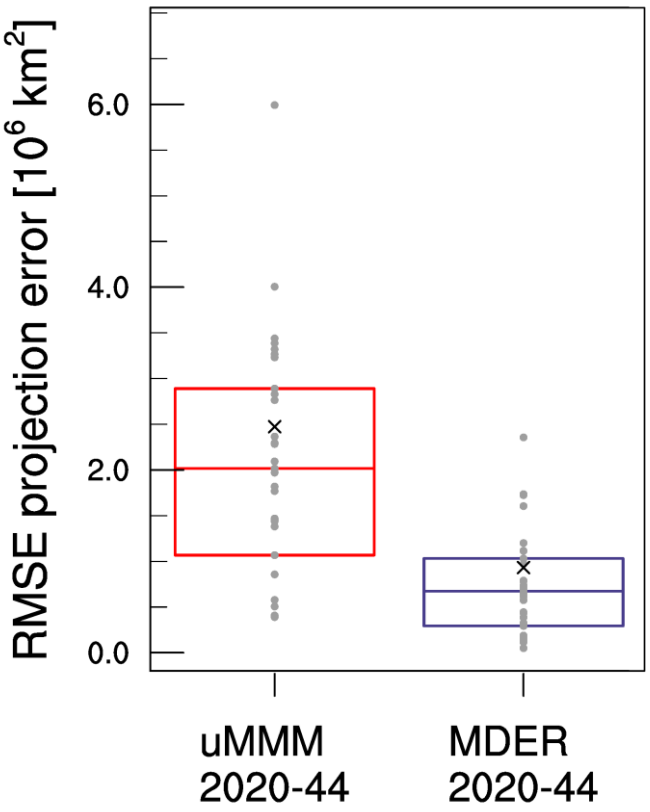

Figure 3.5. RMSE differences between the multi-model mean future SIE and the pseudo reality of future SIE estimated for different pseudo realities (grey dots). The RMSE is calculated for two cases: the unweighted multi-model mean (uMMM) and the MDER method. The crosses show the RMSE for each case and the boxes (red: uMMM; blue: MDER) give the 25th-75th percentiles of the error ensembles estimated from pseudo reality. The horizontal middle line inside each box indicates the median of the model ensemble.

in Figure 3.4), the CMIP5 projections of SIE can be constrained. The constraint reduces the projected multi-model climatological mean SIE (red dashed line) by 0.9 million km², i.e. from 4.25 to 3.35 million km², for the time period 2020 to 2044. This means that applying MDER to SIE results in an over 20% smaller Arctic sea ice extent compared to the unconstrained CMIP5 multi-model mean.

3.3.2. Uncertainty estimation and the potential to narrow model uncertainty

To test if the regression model is suffering from overfitting and in order to investigate whether the MDER method actually gives an estimate of future SIE with a smaller uncertainty than the unweighted multi-model mean, the results are cross-validated using a pseudo reality approach (see Section 3.2.1). Figure 3.5 shows the RMSE for all pseudo realities considered, both for the uMMM and the MDER results. The results show that $RMSE_{MDER}$ (0.93 million km²) is about 62% smaller than $RMSE_{uMMM}$ (2.48 million km²). Similarly, the range spanned by the 25th-75th percentiles of the error ensemble of RSMEs is more than halved from 1.8 million km² (uMMM) to 0.8 million km² (MDER). The large reduction in uncertainty seen in the cross validation can be explained by the fact that MDER explicitly takes historical information into account. The RMSE of the uMMM prediction basically reflects the inter-model spread in the projections of the 2020-2044 mean SIE and is largely influenced by cases where the pseudo reality is an outlier model. In contrast, MDER uses the information of the historical SIE (1979-2012) from the pseudo reality and the other models to estimate the change in mean SIE between 1979-2012 and 2020-2044.

In order to assess the full potential of MDER to reduce the prediction uncertainty, the model weights calculated by MDER are applied to the projections to obtain a weighted multi-model mean. For this, the MDER approach is combined with the method introduced by Hawkins and Sutton (2009) that separates total prediction uncertainty into the three components (see Section 3.2.3) internal (natural) variability, model uncertainty (arising from the spread between the different CMIP5 models), and scenario uncertainty (due to the different RCP scenarios). Note that no ensemble averages are taken in the calculations of the uncertainty estimates, rather each ensemble member from each model is treated separately. Figure 3.6a shows time series of the three sources of uncertainty in unweighted SIE projections. The dominant source of uncertainty throughout the whole time period is model uncertainty, averaging around ± 2.5 million km², which is due to the large inter-model spread in SIE projections among the different CMIP5 models (Kay et al. 2011; Stroeve et al. 2012b; Liu et al. 2013; Swart et al. 2015). Differences in SIE due to different RCP scenarios start to emerge after 2020 and become increasingly larger in the second half of the 21st century. Thus, the scenario uncertainty increases with time and becomes more important than internal variability after 2050. Scenario uncertainty, however, remains smaller than model uncertainty until at least 2100. It is important to note that only simulations using the scenarios RCP 4.5 and RCP 8.5 were used here, since not enough models have performed the RCP 2.6 and RCP 6.0 experiments. The analyses have been repeated with 19 CMIP5 models for which the three scenarios RCP 2.6, RCP 4.5 and RCP 8.5 were available (not shown), and model uncertainty still remained the dominant source over the whole time period.

The internal variability component is assumed constant in time by Hawkins and Sutton (2009), which is a limiting factor (see also Section 2.3). With this method, it has a value of ± 0.6 million km² and is the least important source of uncertainty after 2050. Other studies find a larger contribution of internal variability in other variables (Melia et al. 2015; Jahn et al. 2016; Lique et al. 2016). This raises the question whether the method of Hawkins and Sutton (2009) really captures the internal variability to its full extent. Jahn et al. (2016) estimated internal variability from a large ensemble of the Community Earth System Model (CESM). The assumption is that the spread in large ensembles that only differentiate in their initial conditions (roundoff-level perturbation) represents the internal variability of the climate system within the context of a particular climate model. To compare the results in this chapter to the Jahn et al. (2016) method, this approach was repeated here. Figure 3.7 shows the results for the 38-member large ensemble (CESM LE, grey), which was forced with RCP 8.5, and the 15-member medium ensemble (CESM ME, blue), which was forced with RCP 4.5. The ensemble spreads were calculated as the standard deviation at each time step across all ensemble members. The two ensemble spreads are similar to each other (around 1.5 million km²) until more and more ensemble members of CESM LE reach an SIE of 0, which leads the CESM LE ensemble spread to approach 0 as well. If it is assumed that the ensemble spread represents internal variability, the Jahn et al. (2016) estimate is more than

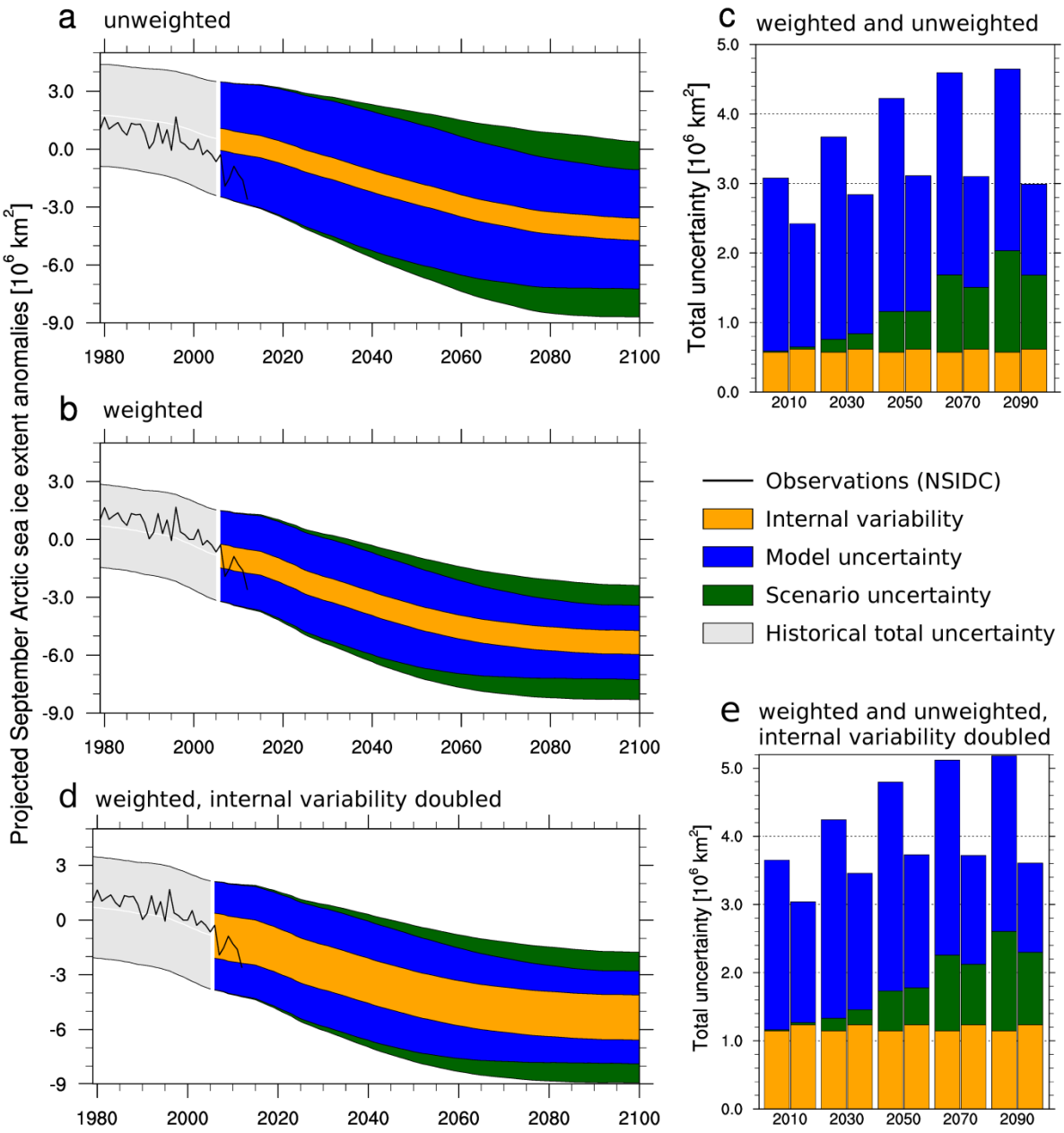

Figure 3.6. Total uncertainty in CMIP5 SIE projections separated into the three fundamentally different sources of uncertainty: internal variability (orange), model uncertainty (blue) and scenario uncertainty (green) estimated from the 29 CMIP5 models listed in Table 2 (RCP 4.5 and RCP 8.5) following Hawkins and Sutton (2009). Panels a and b show time series (1979-2100) of the sources of uncertainty from unweighted SIE projections and from SIE projections weighted with the model weights generated by MDER, respectively. Additionally, the total uncertainty of the historical simulations is shown (grey shading), together with SIE observations (NSIDC, black line). All time series are 10-year running averages and calculated as anomalies with respect to the observed September 2000 Arctic sea ice extent (6.2 million km², NSIDC). Panel c shows the sources of uncertainty in five different time steps both for the unweighted (left) and weighted case (right). Panels d and e are similar to panels b and c, respectively, except for the internal variability component, which is doubled in each time step based on a rough estimate with a single ESM large ensemble.

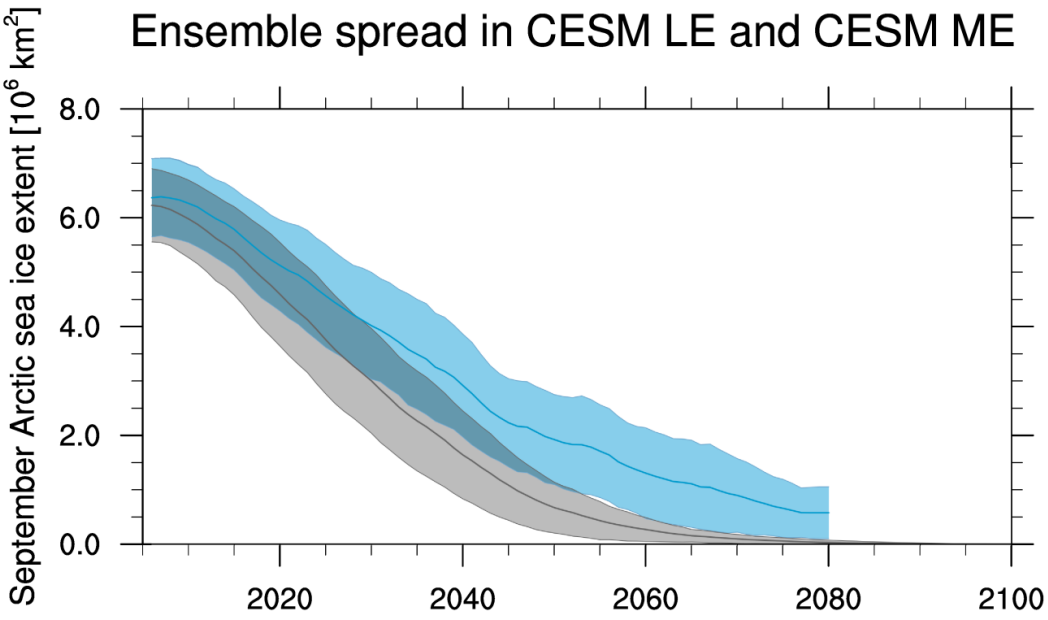

Figure 3.7. Ensemble spread (10-year running mean standard deviation) in September Arctic sea ice extent simulations performed with CESM, for the 38-member large ensemble (CESM LE, 2006-2100) forced with RCP8.5 (grey) and the 15-member medium ensemble (CESM ME, 2006-2080) forced with RCP 4.5 (blue). Thick lines represent the respective ensemble means. For details on CESM LE and CESM ME see Jahn et al. (2016).

twice the internal variability from the Hawkins and Sutton (2009) method. The internal variability estimated by Jahn et al. (2016) quantifies variability as produced by a particular model for a particular scenario. In contrast, the estimate from Hawkins and Sutton (2009) is obtained by performing a statistical fit to the simulated timeseries and quantifying internal variability as the residual from the smooth fit across a number of models and scenarios. These are obviously different metrics and the large difference in estimated values indicates that a true estimate of SIE internal variability that accounts for multiple models is not yet possible. Regardless however, it appears that the Hawkins and Sutton method may underestimate internal variability in the case of September SIE.

Model weighting has the potential to narrow uncertainties in climate model projections (Hawkins and Sutton 2009, 2011; Melia et al. 2015; Knutti et al. 2017; Eyring et al. 2019b). MDER can be used to obtain model weights by calculating the regression of historical diagnostics and future SIE (Section 3.2.1 and Section 3.3.1). Note that these MDER weights are different to the classical performance-based model weights, since they are not necessarily proportional to model biases and can be negative (Bracegirdle and Stephenson 2012). Table B1 lists all models and their weights. The three types of uncertainty are re-calculated weighting the models with the MDER weights. By applying the weights from the target period 2020-2044 to the whole projection period, it is necessary to assume that the same processes selected by MDER are similarly important for SIE projections during the

other periods, because MDER cannot be applied to a longer timer period given the spurious correlations occurring when more and more models reach a sea ice extent close or equal to zero (see Section 3.2.1). The results of the weighting on the types of uncertainty are shown in Figure 3.6b. Compared to the unweighted case (Figure 3.6a) the weighted model uncertainty is considerably smaller, while internal variability is (as expected) not affected by weighting. The bar charts in Figure 3.6c show a quantitative estimate of this uncertainty reduction. The relative reduction of model uncertainty increases with time: from 30% in 2010, to 50% in 2090; or in absolute values, from 0.8 to 1.3 million km². This large uncertainty reduction is mainly due to a bias correction: the MDER weights are calculated using the two diagnostics SIE_c and TAS_t (Section 3.3.1). For SIE_c, the models with the largest biases have the smallest weights, reducing the inter-model spread. Comparing the diagnostic values for SIE_c and TAS_t (Figure 3.1) reveals that the model with the largest positive SIE_c bias (CSIRO-Mk3-6-0) also has an unrealistically small temperature trend (TAS_t) resulting in a small SIE_t and thus gets the smallest weight (Table B1). This example illustrates how model weighting can reduce model uncertainty in multi-model mean results. In the weighted results, the contribution of model uncertainty to the total uncertainty becomes smaller with increasing projection time (from over 80% in 2010 to less than 40% in 2100), but since scenario uncertainty is also slightly reduced by weighting after 2065, model uncertainty remains the most important source of uncertainty in SIE projections throughout the 21st century.

To account for a possible underestimation by the Hawkins and Sutton method (as discussed above), Figures 3.6d and 3.6e show a rough estimate of this uncertainty by doubling the internal variability component, as the Jahn et al. method suggests. Here, internal variability is more important than scenario uncertainty until 2100, accounts for 20-30% of the total uncertainty, and becomes almost equally as important as the weighted model uncertainty towards the end of the 21st century (1.3 million km²). However, even considering a doubled internal variability component, model uncertainty – despite weighting – remains the dominant source of uncertainty throughout this century.

3.3.3. Weighting SIE projections and estimating YOD

In a next step MDER is used to constrain the multi-model projections of future SIE by weighting the models with the MDER weights (Figure 3.8). For both RCP scenarios, the weighted multi-model mean SIE (wMMM) is about 1 million km² smaller than uMMM until YOD is reached (2062 in RCP 8.5) in the wMMM projection of SIE. The smaller SIE from the wMMM suggests a more rapid decline in Arctic sea ice than estimated from uMMM resulting in an earlier disappearance of Arctic sea ice. Two thresholds are calculated to quantify the differences between uMMM and wMMM: the first year in which the multi-model mean SIE

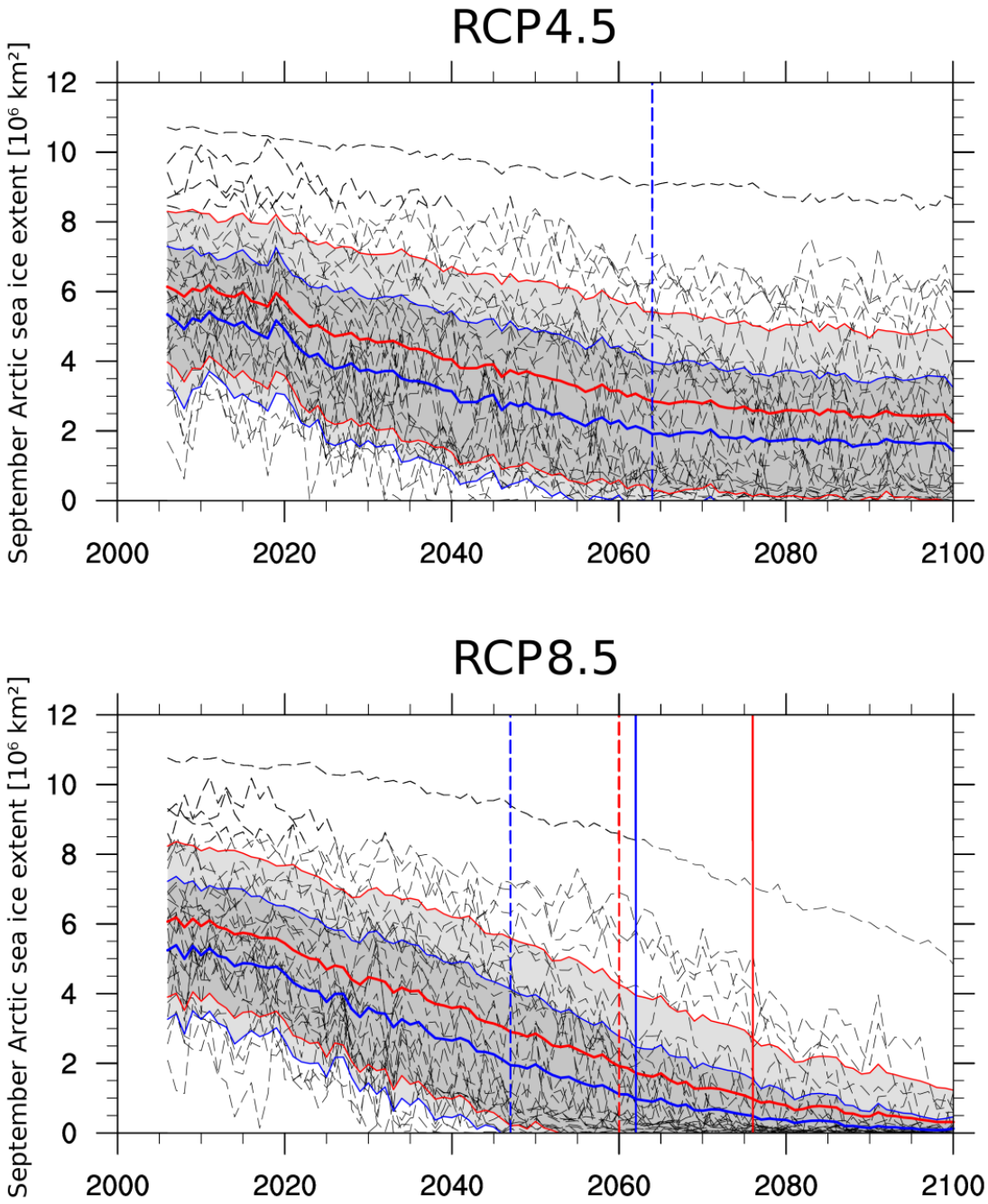

Figure 3.8. Time series (2006-2100) of future SIE assuming the RCP 4.5 scenarios (top) and RCP 8.5 (bottom): ensemble means of the CMIP5 models listed in Table 3.2 (black dashed curves), the uMMM (red curve) and the multi-model mean weighted by MDER (wMMM, blue curve). The grey shading shows the standard deviation of the CMIP5 ensemble, both unweighted (enclosed in thin red lines) and weighted (enclosed in thin blue lines). The vertical lines indicate the first time the multi-model means drop below 2 million km² (YO2, dashed), or the YOD (solid).

drops below 2 million km² (YO2) and the multi-model mean YOD. For RCP 4.5 the YOD occurs after the year 2100 in both cases but YO2 happens much earlier in case of wMMM compared with uMMM: while YO2 happens in uMMM after 2100, YO2 is reached in wMMM in the year 2064 and thus 37 years earlier. In RCP 8.5 both, YOD and YO2, are also earlier in the wMMM: 2047 instead of 2060 in case of YO2 and 2062 instead of 2076 for YOD meaning that when applying MDER to reduce the model uncertainty, nearly ice-free conditions in the summertime Arctic in the RCP 8.5 scenario are reached more than a decade earlier compared with the unweighted CMIP5 multi-model mean.

Weighting affects the multi-model mean similarly to model uncertainty (Section 3.3.2): outlier models are down-weighted and since more models on average show a positive SIE_c bias than negative bias, and since those models also project a larger future SIE than the others, MDER's bias correction predicts a smaller future multi-model mean SIE. Similarly, the weighted standard deviation across the CMIP5 projections is about 0.5 million km² smaller than the unweighted standard deviation until wMMM reaches YO2. Note that the multi-model standard deviation is a different measure than model uncertainty (Equation 3.5) and that the standard deviation is estimated individually for each scenario, which is why the reduction in model uncertainty by weighting (Section 3.3.2) is larger than the reduction in the multi-model standard deviation. Here, the weighted standard deviation is still large amounting up to 5 million km².

Since YOD is clearly scenario-dependent, it provides a measure of the anthropogenic forced response of the climate system. As seen in Section 3.3.2, the contribution of internal variability is too large to precisely predict YOD. However, the results obtained here strongly indicate an earlier near-disappearance of Arctic sea ice than previously assumed.

3.4. Summary and discussion

The aim of this chapter was to answer the second scientific question of this thesis: *Can observation-based constraints improve multi-model projections of 21st-century Arctic sea ice extent?* Therefore, the MDER method was applied to projections of September Arctic sea ice extent (SIE) from 29 CMIP5 models (Table 3.2). From a set of 15 process-oriented diagnostics (Table 3.1) known to influence Arctic sea ice, the stepwise regression algorithm used by MDER selected a linear combination (parsimonious regression model) of two of these diagnostics that best predict the projected future Arctic summer SIE: the climatological mean SIE (SIE_c) and the trend in Arctic near-surface temperature (TAS_t). By applying this regression model to observations, MDER is used to constrain the SIE projections from the models. Compared to unweighted multi-model mean projections, MDER results in smaller ice extents by about 1 million km² and an earlier year of near-disappearance of Arctic sea ice (YOD) by more than a decade in the RCP 8.5 scenario. Cross-validating the MDER results with

a pseudo-reality approach confirms that the regression model has not been overfitted. Note that the predictions of future SIE and YOD provided in this study are affected by unavoidably subjective pre-selection of diagnostics, but can be easily refined once additional emergent constraints become available.

Using the method introduced by Hawkins and Sutton (2009, 2011) to separate the total uncertainty in projections of SIE in its three components (internal variability, model uncertainty, scenario uncertainty), model uncertainty was identified as the largest source of uncertainty in SIE projections throughout the 21st century. It was shown that weighting the models based on MDER's regression model greatly reduces model uncertainty in multi-model projections of SIE by 30 to 50 %. Hodson et al. (2012) found that uncertainties in the 20th century mean state of the Arctic climate are a major source for uncertainties in Arctic climate projections and that observational constraints can greatly improve the precision of Arctic climate projections, which was also demonstrated here. However, the other two sources of uncertainty – internal variability and scenario uncertainty – cannot be significantly reduced by applying the MDER method. By analyzing large ensembles of the CESM model, Jahn et al. (2016) found that these two types of uncertainty make up for a combined prediction uncertainty in YOD of about 25 years. This approach was also used here and the values for internal variability were about twice as high as the ones obtained with the method by Hawkins and Sutton (2009). While the CESM analysis only quantifies internal variability within the context of a single model, the discrepancy with the curve fitting method of Hawkins and Sutton (2009) suggests that this method might underestimate the actual internal variability in SIE. Assuming an internal variability component that is twice as large as the one suggested by Hawkins and Sutton (2009), internal variability becomes almost as large as the weighted model uncertainty at the end of the 21st century (1.3 million km²). Despite the lack of a convincing estimate, internal variability is too large for a precise prediction of YOD. However, MDER strongly indicates an earlier disappearance of Arctic sea ice compared to the unconstrained CMIP5 multi-model mean.

For other sea ice variables, two studies find slightly different contributions of the uncertainty estimates. Melia et al. (2015) found that model uncertainty in September sea ice thickness projections becomes the dominant source of uncertainty after 2022 (before that, the dominant source is internal variability) and that it accounts for maximally 70% of the total uncertainty in sea ice thickness projections. For Arctic SIE, this thesis finds model uncertainty as the dominant source throughout the 21st century, regardless of the method estimating internal variability. For the change in CMIP3 summer (July to September) Arctic SIE, Lique et al. (2016) found internal variability to be more important than model uncertainty until 2020, with a model uncertainty of between 1 and 3 million km². This is similar to the model uncertainty found in this thesis (around 2.5 million km²). However, the estimate of the internal variability component in SIE projections from the Hawkins and Sutton method (0.6 million km²) is smaller than in Lique et al. (2016), which is around 1 million km². This could be

due to the Hawkins and Sutton method not capturing internal variability to its full extent as discussed above, or due to the use of different model results (CMIP3 vs CMIP5).

This chapter demonstrates the potential of MDER to reduce model uncertainty in multi-model projections of Arctic SIE and constrain the prediction of YOD. Its strength is partly based on the strong relationship between past and future climatological mean SIE, with a correlation value of $R^2 = 0.91$ (Figures 3.2 and B6), which is a well-known emergent constraint (e.g., Boé et al. 2009; Bracegirdle et al. 2015; Borodina et al. 2017). Other studies (Massonnet et al. 2012; Liu et al. 2013; Melia et al. 2015; Knutti et al. 2017) also use this relationship to constrain projections of Artic sea ice. Massonnet et al. (2012) used model selection based on four present-day sea ice extent and sea ice volume diagnostics. They obtained an estimate for YOD for RCP 8.5 between 2041 and 2060. Liu et al. (2013) used two approaches: model selection based on model performance in reproducing observed sea ice conditions, and constraining the model biases with observations. With both approaches, they estimate the YOD between 2054 and 2058. The results of this study show that there are still large uncertainties despite weighting. It is surprising therefore somewhat that Liu et al. (2013) find such a narrow estimate for YOD, which might stem from confining their analyses to 5-yr sliding windows. Melia et al. (2015) analyzed bias-corrected SIT projections using a mean and variance correction method. They predict a YOD for the 2050s. Knutti et al. (2017) introduces a weighting scheme that accounts for both model performance and model interdependence using sea ice and surface temperature diagnostics, resulting in a faster decline of SIE than in the unweighted case. They raise the concern of selecting the right diagnostics, but argue that picking unsuitable diagnostics will assign random weights to the models and therefore will not influence the results, as long as the model ensemble is large enough. This is one of the key strengths of the MDER approach: the stepwise regression algorithm filters spurious relationships and only retains the most suitable diagnostics. The potential of the MDER method to constrain the multi-model projections can be further improved by finding additional emergent constraints that can be used as diagnostics.

The results of the different approaches mentioned above are in general agreement with the findings in this chapter. However, despite all the great community efforts to reduce uncertainties in the projections of Arctic climate, model uncertainty remains too large to give a precise estimate of the timing of the first near-disappearance of Arctic sea ice. In fact, considering the large contribution of irreducible internal variability, it is likely that these exact predictions are impossible. Yet, all of these studies, including this thesis, suggest a more rapid decline of Arctic sea ice than estimated from unweighted multi-model mean results. Recent studies found that reducing Arctic warming to 1.5 °C instead of 2 °C by the end of the 21st century can greatly reduce the number of occurrences of an ice-free Arctic (Jahn 2018; Screen 2018; Sigmond et al. 2018). It is therefore imperative to further pursue and enhance global mitigation strategies to limit climate change.

4. Summary and outlook

The Arctic is not only one of the most sensitive regions to global change, but also influences climate worldwide. The fast decline in summer Arctic sea ice extent of about 4 million km² during the past 30 years impacts mid-latitude climate via thermodynamic, radiative and dynamical processes and affects society globally. The future projections of Arctic sea ice made with Earth system models show a large inter-model spread: mid-21st century projections of September Arctic SIE range from 0 to 10 million km², and predictions of the first occurrence of a seasonally ice-free Arctic range from around 2020 to beyond 2100. This thesis contributes to the scientific research effort on understanding and predicting Arctic climate by investigating the use of observations to improve Earth system model simulations in three ways: for model evaluation, for model initialization, and for constraining multi-model projections. Using simulations of both past and future Arctic climate, (1) a specific kind of climate simulations that is initialized with observations was evaluated, and (2) model uncertainties in 21st-century multi-model projections of Arctic sea ice extent were constrained with observations. This chapter summarizes the most important results of this thesis. It also provides a concluding outlook to the likely development of the Arctic sea ice, as well as open questions and next steps in model evaluation of Arctic climate simulations.

4.1. Summary of the results

Part one (Chapter 2) focused on the evaluation of retrospective decadal climate predictions (hindcasts) of Arctic climate parameters and aimed at answering the first scientific question:

Q1: *Can model initialization with observations improve Arctic climate predictions?*

In contrast to long-term historical simulations, decadal hindcasts – that have a forecast horizon of about 10 years – are initialized with observations. This enables the simulations to start from the observed phase of natural variability. Part one of this thesis assessed if the initialization leads to simulations that capture observations better than their uninitialized counterparts. Evaluated were decadal hindcasts from the Earth system model MPI-ESM, and their skill in reproducing the observed Arctic climate between 1979 and 2013 was compared to the skill of historical simulations from the same model. For the evaluation of the model simulations, a verification system for decadal climate predictions has been developed and implemented into the ESMValTool. The evaluation was based two metrics: the anomaly correlation coefficient (ACC) and the root-mean-square error (RMSE). The comparison was

setup in a way so that any skill improvement in the decadal hindcasts solely stems from the initialization. Analyzed were the four variables SST, TAS, SIC, and SIA.

MPI-ESM decadal hindcasts showed improved skill in North Atlantic SST and TAS, which was the motivation to investigate a possibly increased skill in the Arctic. However, it was shown that this increased skill in temperature prediction does not propagate into the Arctic. Only along the east coast of Greenland and only in winter, statistically significant improvements in SST skill were found by initialization, with RMSEs reduced by half. Accordingly, winter RMSE in SIC was reduced in the same region and the integrated SIA over the Greenland Sea showed an improved ACC in winter, but only in the first forecast year after initialization.

It could be shown that the improvements in the winter Greenland Sea stem from the decadal hindcasts better representing the observed year-to-year variability in this region than the historical simulations. This is exactly the aim of decadal climate predictions: whereas long-term climate simulations (such as the historical simulations) are aimed at simulating the average climate conditions over years, decadal hindcasts are initialized with observations in order to start from the correct phase of natural variability and hence, correctly capturing the year-to-year variations of climate. This thesis showed for Arctic simulations with MPI-ESM that this aim could so far only be achieved in the first few years after initialization and only for a certain Arctic region. In fact, some Arctic regions show a degraded skill with initialization such as the marginal ice zone and the Kara and Barents Seas. As it turns out, the main reason for degraded skill from initialized hindcasts is the so-called initialization shock: forcing the climate model to start from the observed climate state may disturb the model in a way that the simulation quickly drifts back to its intrinsic (model) climatological state, often initially overshooting this state. Thus, initialization may introduce temporarily an even larger bias. This model drift could also be seen in MPI-ESM in timeseries of Arctic-wide SIA.

Part two (Chapter 3) focused on the future development of Arctic sea ice by investigating the application of the MDER method to reduce uncertainties in long-term multi-model climate projections of the Arctic:

Q2: *Can observation-based constraints improve multi-model projections of 21st-century Arctic sea ice extent?*

The MDER method relates observable diagnostics to a projected future target variable. For this thesis, it was implemented into the ESMValTool and applied to projections of 21st century Arctic SIE that were calculated assuming the scenarios RCP 4.5 and RCP 8.5. MDER was applied with 15 process-oriented diagnostics calculated with the output of 29 CMIP5 Earth system models. MDER's advantage compared to previous model weighting methods is its iterative stepwise regression algorithm that formally selects those diagnostics from the (subjectively) given set of diagnostics that best predict the target variable. Here, MDER selected the climatological mean SIE and trend in TAS for its regression model, which was

then used to constrain future SIE with observations. The constrained CMIP5 multi-model-mean projection predicts a smaller SIE by about 1 million km² (20%) than the unconstrained multi-model mean throughout the whole time period for RCP 4.5 and until 2060 for RCP 8.5. It also predicts an earlier near-disappearance of summer Arctic sea ice (YOD) by over a decade (from 2076 to 2062) in RCP 8.5.

The application of MDER clearly proved its potential by specifically using the historical information of the two different selected diagnostics and constraining the projections with observations. In addition to MDER, methods to quantify the three fundamentally different sources of projection uncertainties (internal variability, model uncertainty, and scenario uncertainty) were also implemented into the ESMValTool. It was shown that internal variability in Arctic SIE that arises from the chaotic nature of the climate system is still difficult to quantify, since the estimates obtained with two different methods differ by roughly 1 million km². Applying model weights based on the MDER's regression model to the SIE projections, model uncertainty could be reduced by 30-50% (between 0.8 and 1.3 million km²), depending on the projection time. However, despite this strong reduction, model uncertainty remains the largest source of uncertainty in Arctic SIE projections. Together with scenario uncertainty and internal variability (which cannot be reduced), the constrained total SIE uncertainty adds up to at least 3 million km² around mid-21st century. Exact predictions of YOD therefore remain impossible.

4.2. Conclusions and outlook

Evaluating Earth system models with observations is essential to assess the performance or realism of model simulations. However, an actual validation of models is impossible since model simulations are ideally just alternative realizations of natural variability. Furthermore, observations also have uncertainties that originate from statistical errors (noise), measurement uncertainties, and from different assumptions and simplifications made for derived variables. Here, the community is dependent on an estimate of the measurement error provided with some of the observational products. For Arctic SIE, the difference between the available observational datasets can be as large as 1 million km². This is in part due to different measurement methods like ship or buoy measurements and satellite retrievals. However, the difference between the two NSIDC satellite products solely stems from differences in their retrieval algorithms and amounts to roughly 0.5 million km². The different Arctic SIE observations are, however, highly correlated to each other making them well suited to use with metrics based on correlation coefficients, which are not sensitive to constant offsets between data series. Yet, a more formal multi-observational approach for model evaluation is to be developed and should become a routine method in all assessments of model performance.

Using observational datasets to initialize climate models is still a relatively new field. While this thesis could not find large improvements by initialization for decadal hindcasts of Arctic climate, there are several ways for possible improvements. For example, in order to reduce initialization shock, different initialization techniques are being developed and applied in CMIP. Furthermore, the skill of decadal climate predictions in the Arctic may also be improved by including observed sea ice variables in the initialization, such as SIC, SIT, or sea ice age. However, this thesis also showed that including only SIC in the initialization process is not sufficient to improve decadal hindcasts. Additionally, varying the month of initialization – such as in summer instead of at the end of the year – could improve the Arctic hindcasts. Together with the improvements by model initialization found in other regions such as the Atlantic Ocean, all these points strongly motivate further research on decadal climate predictions.

The thesis also demonstrated how observations can be used to reduce uncertainties in climate projections with the MDER method. The strength of the method can be further improved by using better diagnostics, ideally emergent constraints. Emergent constraints are strong process-oriented statistical relationships that can constrain a future (model) variable with observations of the present-day climate. The currently two most promising emergent constraints for Arctic climate are the connection between polar amplification of Arctic warming and present-day mean temperatures, and the relationship between current and future Arctic SIE. In fact, both, the trend in Arctic surface temperature and the mean SIE were selected by MDER to constrain the CMIP5 SIE projections. Finding more and better emergent constraints for Arctic sea ice is a vivid scientific topic and can greatly improve the potential of process-oriented model weighting methods such as MDER.

In summary, the results of this thesis point towards a more pessimistic outlook for Arctic sea ice than previously estimated with CMIP5 multi-model mean results. As shown, this is in accordance with numerous recent studies investigating methods to improve Arctic climate projections. The thesis highlights the deficits in both the understanding of climate processes in the Arctic and the knowledge of the future development of Arctic sea ice. The importance of the Arctic for the world's climate, the deficits in process understanding, increasingly pessimistic projections from more and more studies, and the still large projection uncertainties make Arctic climate research a critical component for society's future. Since current studies show that only restricting global warming to no more than 1.5 °C by 2100 retains a realistic possibility for the survival of summer Arctic sea ice, it is imperative to make the right political choices on global change now.

A. Acronyms

ACC	Anomaly correlation coefficient
AMOC	Atlantic Meridional Overturning Circulation
BMBF	German Federal Ministry of Education (*Bundesministerium für Bildung und Forschung*)
CESM	Community Earth System Model
CMIP	Coupled Model Intercomparison Project
CMIP5	Coupled Model Intercomparison Project Phase 5
DLR	German Aerospace Center (*Deutsches Zentrum für Luft- und Raumfahrt*)
ERA-Interim	European Centre for Medium-Range Weather Forecast Re-Analysis Interim
ESMValTool	Earth System Model Evaluation Tool
EVA	Department Earth System Model Evaluation and Analysis at the Institute of Atmospheric Physics, DLR
GECCO2	German contribution to Estimating the Circulation and Climate of the Ocean
HadCRUT4	Hadley Centre / Climate Research Unit Temperature records 4
HadISST	Hadley Centre Sea Ice and Sea Surface Temperature data set
IPO	Interdecadal Pacific Oscillation
MDER	Multiple diagnostic ensemble regression
MiKlip	BMBF project *Mittelfristige Klimaprognosen* (mid-term climate predictions)
MPI-ESM	Max Planck Institute Earth System Model
MPI-ESM-HR *preop*	MiKlip decadal preoperational prediction system
MPI-ESM-LR *prot*	MiKlip decadal prototype prediction system
NAO	North Atlantic Oscillation
NASA	National Aeronautics and Space Administration

NSIDC	National Snow and Ice Data Center
NSIDC-BT	NSIDC Bootstrap satellite retrieval algorithm
NSIDC-NT	NSIDC NASA-Team satellite retrieval algorithm
ORAS4	Ocean Reanalysis System 4
RCP	Representative Concentration Pathway
RMSE	Root-mean-square error
SIA	Sea ice area
SIC	Sea ice concentration / sea ice area fraction
SIE	Sea ice extent
SIT	Sea ice thickness
SST	Sea-surface temperature
TAS	Near-surface air temperature
uMMM	Unweighted multi-model mean
wMMM	Weighted multi-model mean
YOD	Year of near-disappearance of summer Arctic sea ice

B. Supplemental information on Chapter 3

Table B1. Model weights calculated by MDER for all CMIP5 models used in Chapter 3.

No.	Model	Weight
02	bcc-csm1-1	0.05442876
03	bcc-csm1-1-m	0.05442861
04	CanESM2	0.06936824
05	CCSM4	0.0307124
06	CESM1-CAM5	0.03331346
07	CMCC-CM	0.01811669
08	CMCC-CMS	0.02206415
09	CNRM-CM5	0.05274851
10	CSIRO-Mk3-6-0	0.00190936
11	EC-EARTH	0.01833672
12	FGOALS-g2	0.0168146
13	GFDL-CM3	0.04552687
14	GFDL-ESM2G	0.02425032
15	GFDL-ESM2M	0.03416637
16	GISS-E2-R	0.0612035
17	HadGEM2-AO	0.04922342
18	HadGEM2-CC	0.02943597
19	HadGEM2-ES	0.03303585
20	inmcm4	0.02924009
21	IPSL-CM5A-LR	0.03075209
22	IPSL-CM5A-MR	0.03760888
23	IPSL-CM5B-LR	0.008010222
24	MIROC5	0.03747144
25	MIROC-ESM	0.03407768
26	MIROC-ESM-CHEM	0.04647715
27	MPI-ESM-LR	0.0261536
28	MPI-ESM-MR	0.03634945
29	MRI-CGCM3	0.03726229
30	NorESM1-M	0.02751311

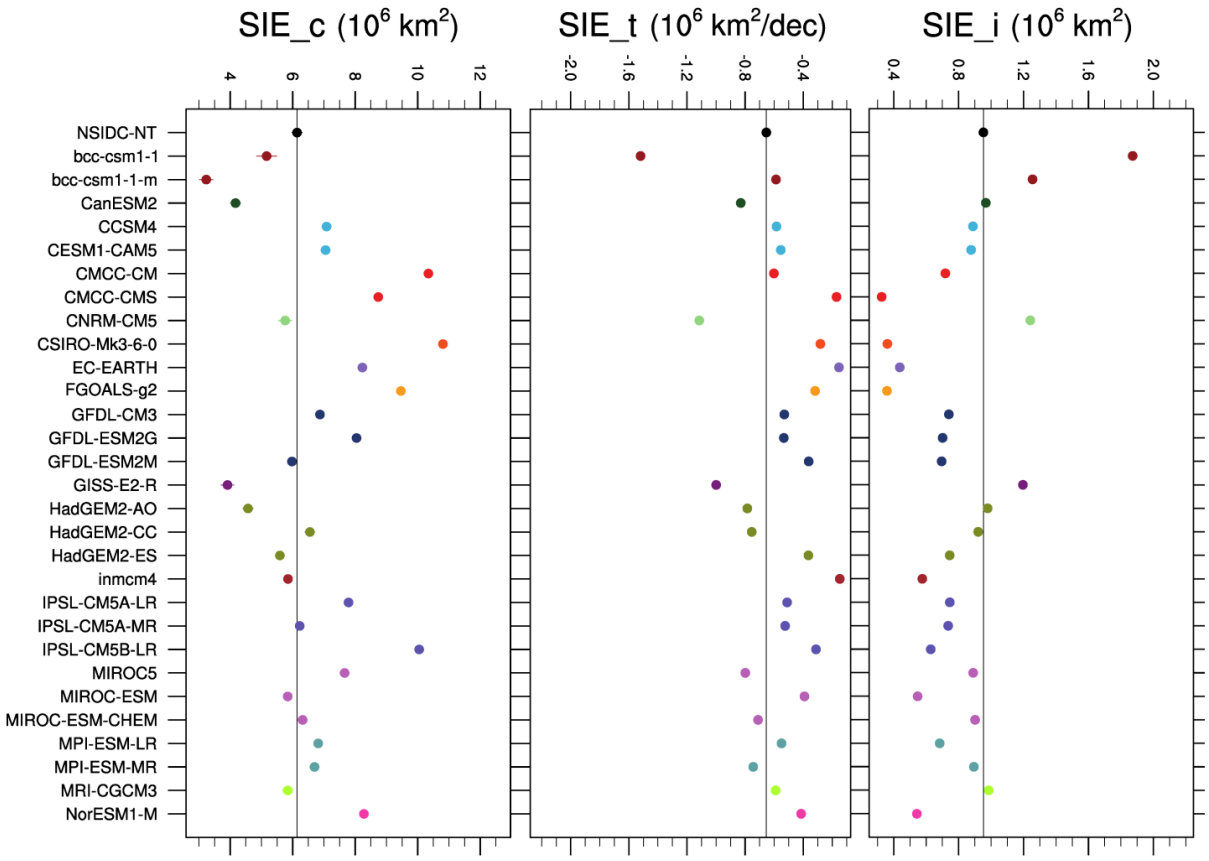

Figure B1. Diagnostic values (averaged for each ensemble) from all CMIP5 models used in this study (see Table 3.2) for the SIE diagnostics (1979-2012). The observational reference value is added in the top position, together with a vertical reference line. For details on the diagnostics, see Table 3.1.

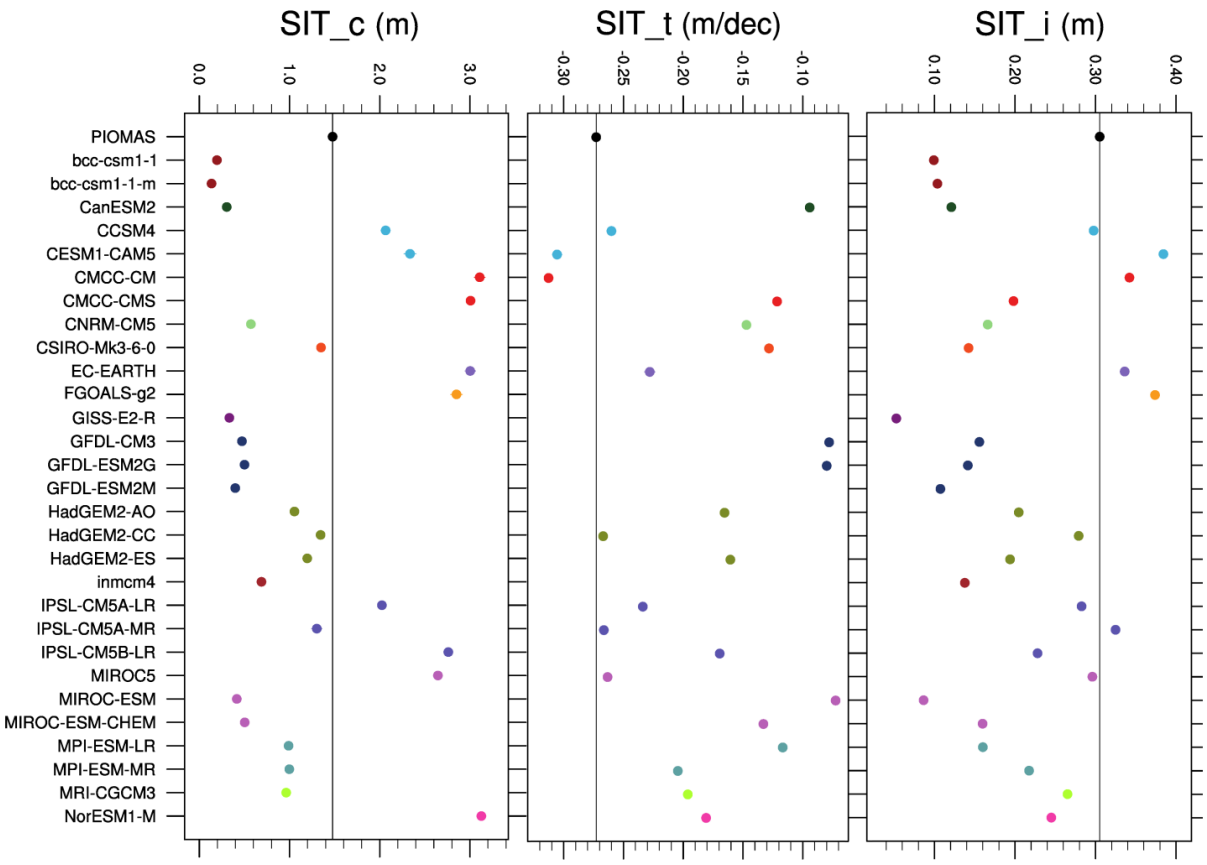

Figure B2. Same as Figure B1, but for SIT diagnostics.

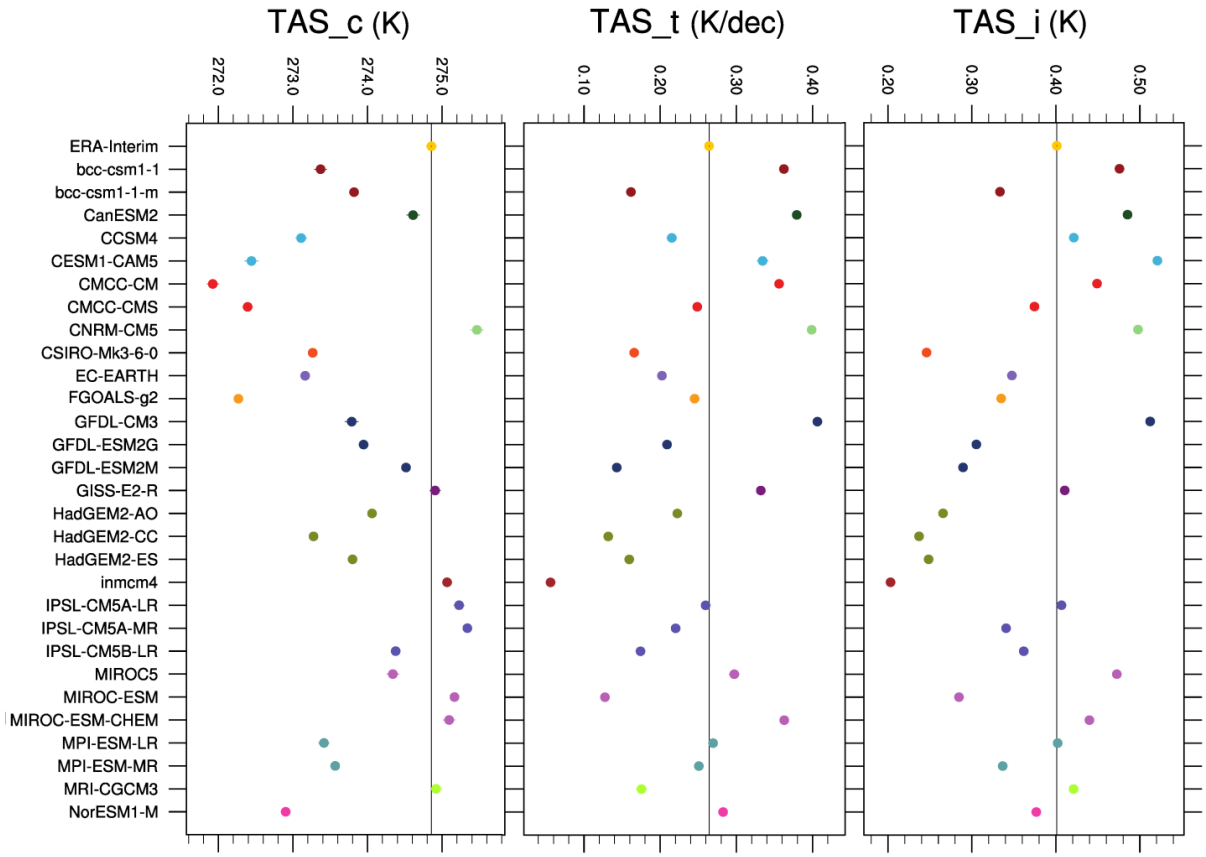

Figure B3. Same as Figure B1, but for TAS diagnostics.

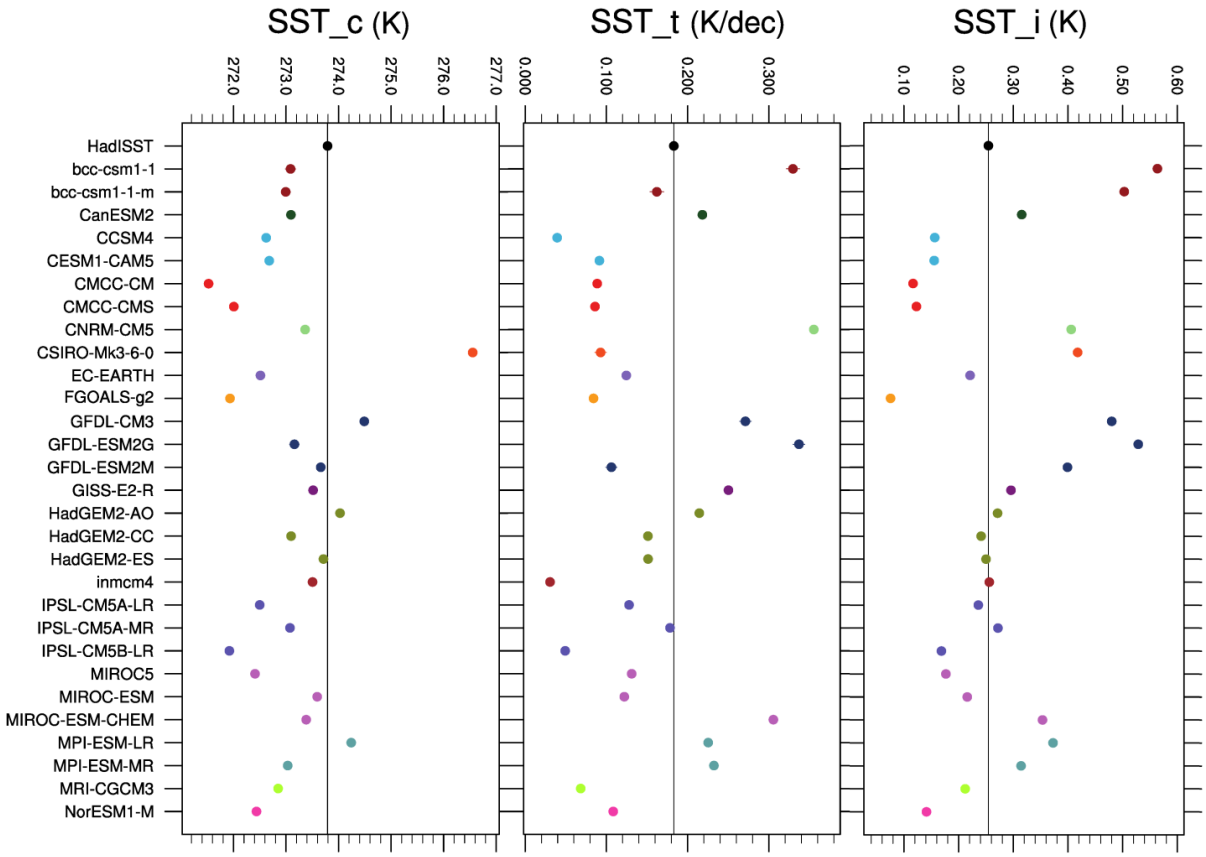

Figure B4. Same as Figure B1, but for SST diagnostics.

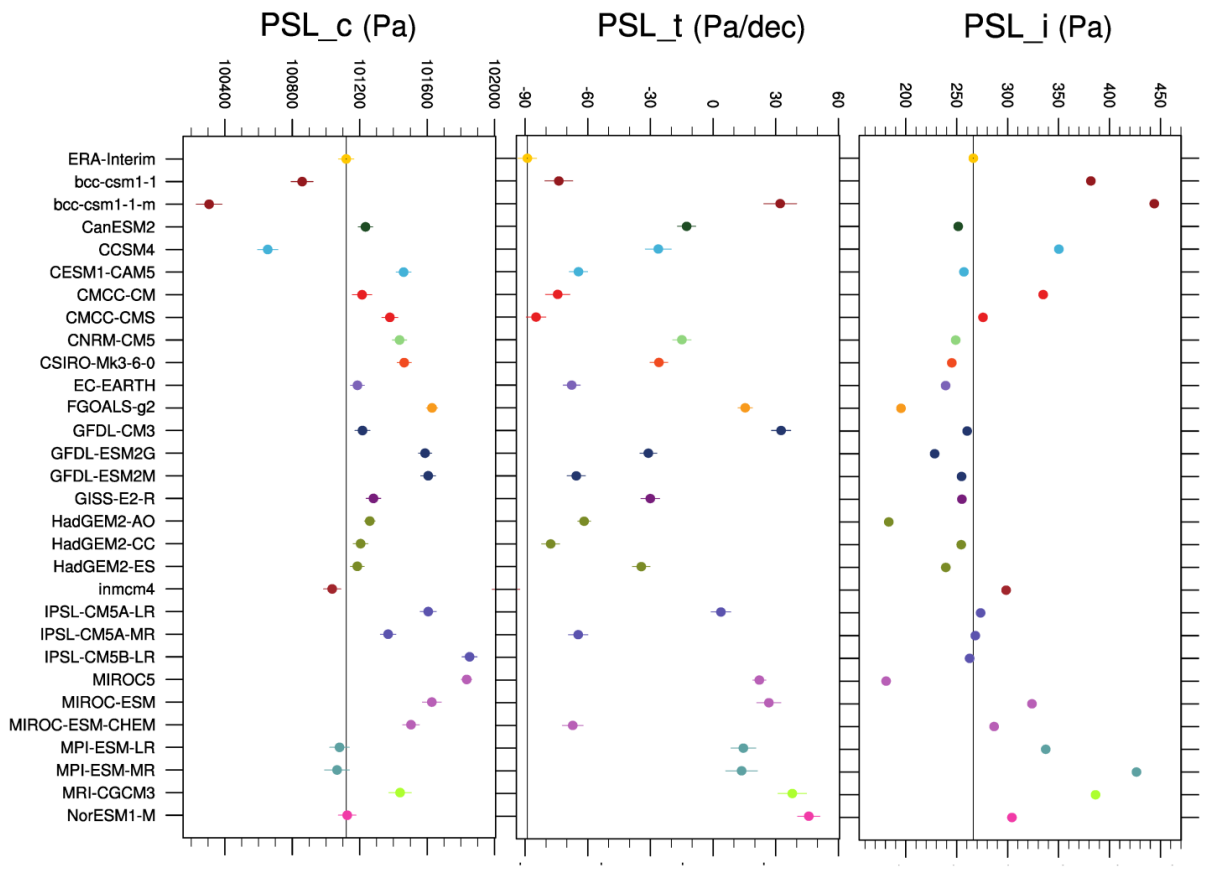

Figure B5. Same as Figure B1, but for PSL diagnostics.

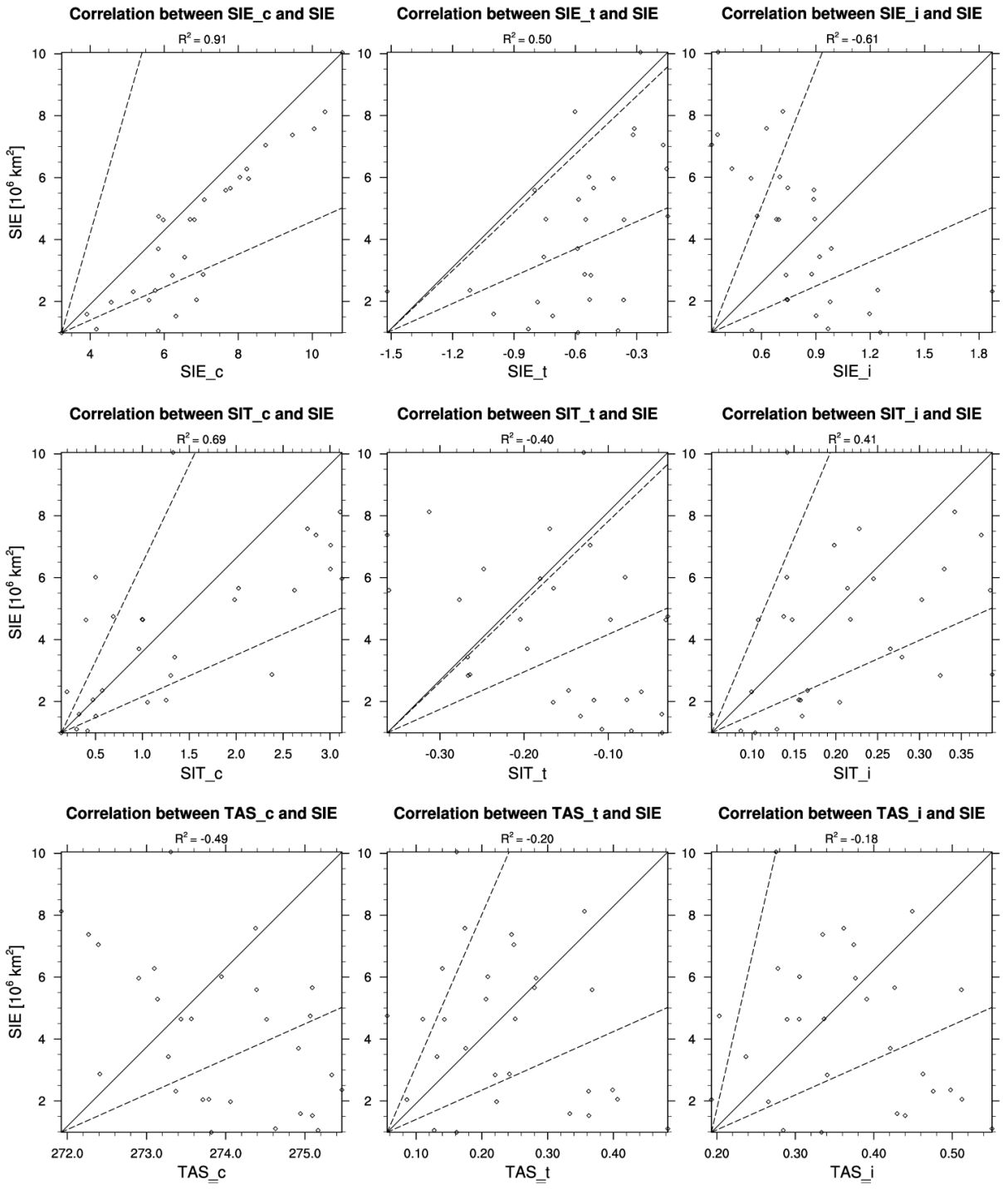

Figure B6. Relationships between each historical diagnostic (1979-2012, see Table 3.1) and the future SIE (2020-2044) in the CMIP5 models (diamonds) referenced in Table 3.2. The correlation coefficients for each diagnostic are indicated as R^2, and the solid line is the reference line for perfect correlation. The dashed lines are the factor-2 reference lines: models between these two lines have a model-to-observation ratio between 0.5 and 2. Shown are the SIE, SIT and TAS diagnostics.

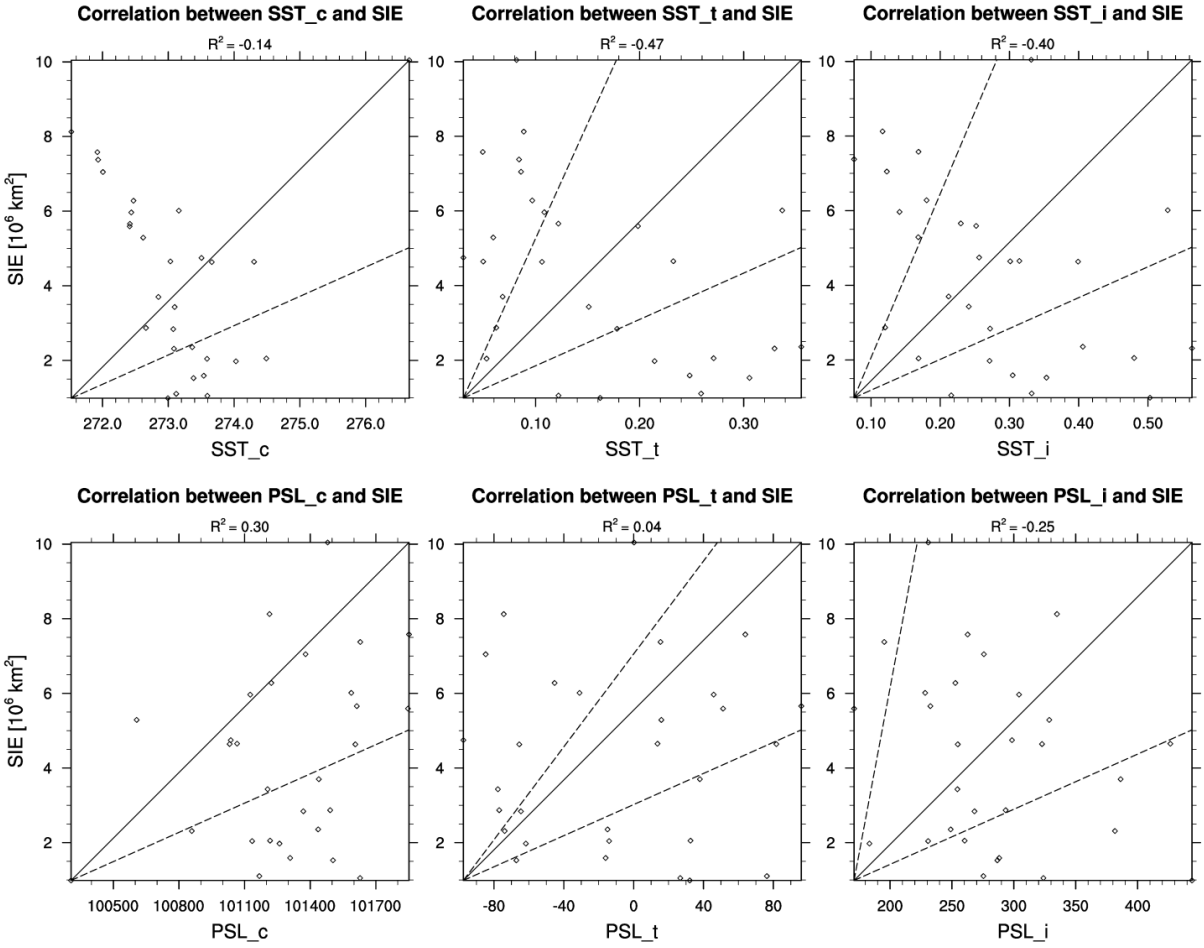

Figure B7. Same as Figure S4, but for SST and PSL diagnostics.

C. Integrated author's references

The following two research articles are part of this thesis (Chapters 2 and 3, respectively) and have been published or are under review by peer-reviewed journals. For both studies, the author of this thesis implemented the according diagnostics into the ESMValTool, analyzed the model and observational data, produced all figures and wrote the text, while the co-authors of the studies helped iterating the concept and the text and interpreting the results. Additionally, V. Eyring supervised both studies. Alterations and enhancements were made to the published text in adapting the studies to the style and structure of this thesis.

Senftleben, D., V. Eyring, A. Lauer, and M. Righi, 2018: Temperature and sea ice hindcast skill of the MiKlip decadal prediction system in the Arctic. *Meteorologische Zeitschrift*, 27, 195-208. doi: 10.1127/metz/2018/0871

Senftleben, D., V. Eyring, A. Lauer, M. M. Holland, and A. Y. Karpechko, 2019: Constraining uncertainties in CMIP5 projections of September Arctic sea ice extent with observations. *Journal of Climate*, under review.

The thesis also contributed to the following peer-reviewed studies by providing sea-ice related text and figures to sections of the articles:

Eyring, V., et al. (incl. **D. Senftleben**), 2016: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP. *Geoscientific Model Development Discussions*, 8, 7541-7661. doi: 10.5194/gmdd-8-7541-2015

Lauer, A., et al. (incl. **D. Senftleben**), 2017: Benchmarking CMIP5 models with a subset of ESA CCI Phase 2 data using the ESMValTool. *Remote Sensing of Environment*. doi: 10.1016/j.rse.2017.01.007

Eyring, V., et al. (incl. **D. Senftleben**), 2019: ESMValTool (v2.0) – an extended set of diagnostics for enhanced evaluation and analysis of CMIP models. *Geoscientific Model Development Discussions*, in prep.

References

Aagaard, K., and E. C. Carmack, 1989: The role of sea ice and other fresh water in the Arctic circulation. *Journal of Geophysical Research*, **94,** 14485. doi: 10.1029/JC094iC10p14485

Ahn, J., S. Hong, J. Cho, et al., 2014: Statistical Modeling of Sea Ice Concentration Using Satellite Imagery and Climate Reanalysis Data in the Barents and Kara Seas, 1979–2012. *Remote Sensing*, **6,** 5520-5540. doi: 10.3390/rs6065520

Archer, D., and B. Buffett, 2005: Time-dependent response of the global ocean clathrate reservoir to climatic and anthropogenic forcing. *Geochemistry, Geophysics, Geosystems*, **6,** n/a-n/a. doi: 10.1029/2004gc000854

Arora, V. K., J. F. Scinocca, G. J. Boer, et al., 2011: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophysical Research Letters*, **38,** n/a-n/a. doi: 10.1029/2010gl046270

Balmaseda, M. A., K. E. Trenberth, and E. Källén, 2013: Distinctive climate signals in reanalysis of global ocean heat content. *Geophysical Research Letters*, **40,** 1754-1759. doi: 10.1002/grl.50382

Bellucci, A., R. Haarsma, N. Bellouin, et al., 2015: Advancements in decadal climate predictability: The role of nonoceanic drivers. *Reviews of Geophysics*, **53,** 165-202. doi: 10.1002/2014RG000473

Blanchard-Wrigglesworth, E., K. C. Armour, C. M. Bitz, and E. DeWeaver, 2011: Persistence and Inherent Predictability of Arctic Sea Ice in a GCM Ensemble and Observations. *Journal of Climate*, **24,** 231-250. doi: 10.1175/2010jcli3775.1

Boé, J., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nature Geoscience*, **2,** 341-343. doi: 10.1038/ngeo467

Boer, G. J., D. M. Smith, C. Cassou, et al., 2016: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, **9,** 3751-3777. doi: 10.5194/gmd-9-3751-2016

Borodina, A., E. M. Fischer, and R. Knutti, 2017: Emergent Constraints in Climate Projections: A Case Study of Changes in High-Latitude Temperature Variability. *Journal of Climate*, **30,** 3655-3670. doi: 10.1175/jcli-d-16-0662.1

Bracegirdle, T. J., and D. B. Stephenson, 2012: Higher precision estimates of regional polar warming by ensemble regression of climate model projections. *Climate Dynamics*, **39,** 2805-2821. doi: 10.1007/s00382-012-1330-3

Bracegirdle, T. J., D. B. Stephenson, J. Turner, and T. Phillips, 2015: The importance of sea ice area biases in 21st century multimodel projections of Antarctic temperature and precipitation. *Geophysical Research Letters*, **42,** 10,832-810,839. doi: 10.1002/2015gl067055

Bunzel, F., D. Notz, J. Baehr, et al., 2016: Observational uncertainty of Arctic sea-ice concentration significantly affects seasonal climate forecasts. *Geophysical Research Letters***,** 1-8. doi: 10.1002/2015GL066928

Bushuk, M., R. Msadek, M. Winton, et al., 2017: Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophysical Research Letters*. doi: 10.1002/2017gl073155

Cavalieri, D. J., C. L. Parkinson, P. Gloersen, and H. Zwally., 1996: Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data.**,** Arctic, full record. doi: 10.5067/8GQ8LZQVL0VL

Church, J. A., N. J. White, R. Coleman, et al., 2004: Estimates of the Regional Distribution of Sea Level Rise over the 1950–2000 Period. *Journal of Climate*, **17,** 2609-2625. doi: 10.1175/1520-0442(2004)017<2609:eotrdo>2.0.co;2

CLIVAR, 2011: Data and bias correction for decadal climate predictions. *CLIVAR Publication Series*, **No. 150,** 6 pp

Collins, M., 2002: Climate predictability on interannual to decadal time scales: the initial value problem. *Climate Dynamics*, **19,** 671-692. doi: 10.1007/s00382-002-0254-8

Comiso, J. C., 2000: Variability and Trends in Antarctic Surface Temperatures from In Situ and Satellite Infrared Measurements. *Journal of Climate*, **13,** 1674-1696. doi: 10.1175/1520-0442(2000)013<1674:VATIAS>2.0.CO;2

Counillon, F., M. Kimmritz, N. Keenlyside, et al., 2017: Seasonal-to-decadal predictability in the Nordic Seas and Arctic with the Norwegian Climate Prediction Model. *EGU General Assembly Conference Abstracts*, **19**

Curry, J. a., J. L. Schramm, and E. E. Ebert, 1995: Sea-Ice Albedo Climate Feedback Mechanism. *Journal of Climate*, 240-247. doi: Doi 10.1175/1520-0442(1995)008<0240:Siacfm>2.0.Co;2

Day, J. J., E. Hawkins, and S. Tietsche, 2014a: Will Arctic sea ice thickness initialization improve seasonal forecast skill? *Geophysical Research Letters*, **41,** 7566-7575. doi: 10.1002/2014GL061694

Day, J. J., S. Tietsche, and E. Hawkins, 2014b: Pan-arctic and regional sea ice predictability: Initialization month dependence. *Journal of Climate*, **27,** 4371-4390. doi: 10.1175/JCLI-D-13-00614.1

Dee, D. P., S. M. Uppala, A. J. Simmons, et al., 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137,** 553-597. doi: 10.1002/qj.828

Ding, Q., A. Schweiger, M. L'Heureux, et al., 2017: Influence of high-latitude atmospheric circulation changes on summertime Arctic sea ice. *Nature Climate Change*, **7,** 289-295. doi: 10.1038/nclimate3241

Dirkson, A., W. J. Merryfield, and A. Monahan, 2017: Impacts of Sea Ice Thickness Initialization on Seasonal Arctic Sea Ice Predictions. *Journal of Climate*, **30,** 1001-1017. doi: 10.1175/jcli-d-16-0437.1

Doblas-Reyes, F. J., I. Andreu-Burillo, Y. Chikamoto, et al., 2013: Initialized near-term regional climate change prediction. *Nature communications*, **4,** 1715. doi: 10.1038/ncomms2704

Donner, L. J., B. L. Wyman, R. S. Hemler, et al., 2011: The Dynamical Core, Physical Parameterizations, and Basic Simulation Characteristics of the Atmospheric Component AM3 of the GFDL Global Coupled Model CM3. *Journal of Climate*, **24,** 3484-3519. doi: 10.1175/2011jcli3955.1

Dufresne, J. L., M. A. Foujols, S. Denvil, et al., 2013: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics*, **40,** 2123-2165. doi: 10.1007/s00382-012-1636-1

Dunne, J. P., J. G. John, E. Shevliakova, et al., 2013: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part II: Carbon System Formulation and Baseline Simulation Characteristics*. *Journal of Climate*, **26,** 2247-2267. doi: 10.1175/jcli-d-12-00150.1

Eade, R., D. Smith, A. Scaife, et al., 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, **41,** 5620-5628. doi: 10.1002/2014GL061146

Eyring, V., L. Bock, A. Lauer, et al., 2019a: ESMValTool (v2.0) – an extended set of diagnostics for enhanced evaluation and analysis of CMIP models. *Geoscientific Model Development Discussions*, **in prep**

Eyring, V., S. Bony, G. A. Meehl, et al., 2016a: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organisation. *Geoscientific Model Development Discussions*, **8,** 10539-10583. doi: 10.5194/gmdd-8-10539-2015

Eyring, V., P. M. Cox, G. M. Flato, et al., 2019b: Taking climate model evaluation to the next level. *Nature Climate Change*. doi: 10.1038/s41558-018-0355-y

Eyring, V., M. Righi, M. Evaldsson, et al., 2016b: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP. *Geoscientific Model Development Discussions*, **8,** 7541-7661. doi: 10.5194/gmdd-8-7541-2015

Flato, G., J. Marotzke, B. Abiodun, et al., 2013: Evaluation of Climate Models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 741-866. doi: 10.1017/CBO9781107415324

Francis, J. A., and S. J. Vavrus, 2015: Evidence for a wavier jet stream in response to rapid Arctic warming. *Environmental Research Letters*, **10,** 014005. doi: 10.1088/1748-9326/10/1/014005

Francis, J. A., S. J. Vavrus, and J. Cohen, 2017: Amplified Arctic warming and mid-latitude weather: new perspectives on emerging connections. *Wiley Interdisciplinary Reviews: Climate Change*, **8,** e474. doi: 10.1002/wcc.474

Gangstø, R., A. P. Weigel, M. A. Liniger, and C. Appenzeller, 2013: Methodological aspects of the validation of decadal predictions. *Climate Research*, **55,** 181-200. doi: 10.3354/cr01135

Gent, P. R., G. Danabasoglu, L. J. Donner, et al., 2011: The Community Climate System Model Version 4. *Journal of Climate*, **24,** 4973-4991. doi: 10.1175/2011jcli4083.1

Germe, A., M. Chevallier, D. Salas y Mélia, et al., 2014: Interannual predictability of Arctic sea ice in a global climate model: regional contrasts and temporal evolution. *Climate Dynamics*, **43,** 2519-2538. doi: 10.1007/s00382-014-2071-2

Giorgetta, M. A., J. H. Jungclaus, C. H. Reick, et al., 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5. *Journal of Advances in Modeling Earth Systems*, **5,** 572-597. doi: 10.1002/jame.20038

Goddard, L., A. Kumar, A. Solomon, et al., 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, **40,** 245-272. doi: 10.1007/s00382-012-1481-2

Goessling, H. F., S. Tietsche, J. J. Day, et al., 2016: Predictability of the Arctic sea ice edge. *Geophysical Research Letters*, **43,** 1-9. doi: 10.1002/2015GL067232

Goosse, H., O. Arzel, C. M. Bitz, et al., 2009: Increased variability of the Arctic summer ice extent in a warmer climate. *Geophysical Research Letters*, **36**. doi: 10.1029/2009gl040546

Guemas, V., M. Chevallier, M. Déqué, et al., 2016: Impact of sea ice initialization on sea ice and atmosphere prediction skill on seasonal timescales. *Geophysical Research Letters*, **43,** 3889-3896. doi: 10.1002/2015gl066626

Guemas, V., E. Blanchard-Wrigglesworth, M. Chevallier, et al., 2014: A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*. doi: 10.1002/qj.2401

Hawkins, E., and R. Sutton, 2009: The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, **90,** 1095-1107. doi: 10.1175/2009bams2607.1

Hawkins, E., and R. Sutton, 2011: The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics*, **37,** 407-418. doi: 10.1007/s00382-010-0810-6

Hawkins, E., S. Tietsche, J. J. Day, et al., 2015: Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems. *Quarterly Journal of the Royal Meteorological Society*. doi: 10.1002/qj.2643

Hazeleger, W., C. Severijns, T. Semmler, et al., 2010: EC-Earth. *Bulletin of the American Meteorological Society*, **91,** 1357-1364. doi: 10.1175/2010bams2877.1

He, Y., B. Wang, M. Liu, et al., 2017: Reduction of initial shock in decadal predictions using a new initialization strategy. *Geophysical Research Letters*, **44,** 8538-8547. doi: 10.1002/2017gl074028

Hibler, W. D., 1979: A Dynamic Thermodynamic Sea Ice Model. *Journal of Physical Oceanography*, **9,** 815-846. doi: 10.1175/1520-0485(1979)009<0815:ADTSIM>2.0.CO;2

Hodson, D. L. R., S. P. E. Keeley, A. West, et al., 2012: Identifying uncertainties in Arctic climate change projections. *Climate Dynamics*, **40,** 2849-2865. doi: 10.1007/s00382-012-1512-z

Holgate, S. J., and P. L. Woodworth, 2004: Evidence for enhanced coastal sea level rise during the 1990s. *Geophysical Research Letters*, **31,** n/a-n/a. doi: 10.1029/2004gl019626

Holland, M. M., and C. M. Bitz, 2003: Polar amplification of climate change in coupled models. *Climate Dynamics*, **21,** 221-232. doi: 10.1007/s00382-003-0332-6

Holland, M. M., D. A. Bailey, and S. Vavrus, 2010: Inherent sea ice predictability in the rapidly changing Arctic environment of the Community Climate System Model, version 3. *Climate Dynamics*, **36,** 1239-1253. doi: 10.1007/s00382-010-0792-4

IPCC, 2013: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, 1535 pp. ISBN: ISBN 978-1-107-66182-0

Iversen, T., M. Bentsen, I. Bethke, et al., 2012: The Norwegian Earth System Model, NorESM1-M – Part 2: Climate response and scenario projections. *Geoscientific Model Development Discussions*, **5,** 2933-2998. doi: 10.5194/gmdd-5-2933-2012

Jahn, A., 2018: Reduced probability of ice-free summers for 1.5 °C compared to 2 °C warming. *Nature Climate Change*, **8,** 409-413. doi: 10.1038/s41558-018-0127-8

Jahn, A., J. E. Kay, M. M. Holland, and D. M. Hall, 2016: How predictable is the timing of a summer ice-free Arctic? *Geophysical Research Letters*, **43,** 9113-9120. doi: 10.1002/2016gl070067

Jevrejeva, S., A. Grinsted, J. C. Moore, and S. Holgate, 2006: Nonlinear trends and multiyear cycles in sea level records. *Journal of Geophysical Research*, **111**. doi: 10.1029/2005jc003229

Jia, L., and T. Delsole, 2013: Multi-year predictability of temperature and precipitation in multiple climate models. *Geophysical Research Letters*, **39,** 1-6. doi: 10.1029/2012GL052778

Jones, P. D., D. H. Lister, T. J. Osborn, et al., 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.*, **117**

Jungclaus, J. H., N. Fischer, H. Haak, et al., 2013: Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model. *Journal of Advances in Modeling Earth Systems*, **5,** 422-446. doi: 10.1002/jame.20023

Karpechko, A. Y., D. Maraun, and V. Eyring, 2013: Improving Antarctic Total Ozone Projections by a Process-Oriented Multiple Diagnostic Ensemble Regression. *Journal of the Atmospheric Sciences*, **70,** 3959-3976. doi: 10.1175/jas-d-13-071.1

Kay, J. E., M. M. Holland, and A. Jahn, 2011: Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters*, **38,** n/a-n/a. doi: 10.1029/2011gl048008

Kennedy, J. J., N. A. Rayner, R. O. Smith, et al., 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research*, **116,** D14104. doi: 10.1029/2010JD015220

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophysical Research Letters*, **39,** n/a-n/a. doi: 10.1029/2012gl051644

Knies, J., P. Cabedo-Sanz, S. T. Belt, et al., 2014: The emergence of modern sea ice cover in the Arctic Ocean. *Nat Commun*, **5,** 5608. doi: 10.1038/ncomms6608

Knutti, R., J. Sedláček, B. M. Sanderson, et al., 2017: A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*. doi: 10.1002/2016gl072012

Koenigk, T., C. K. Beatty, M. Caian, et al., 2012: Potential decadal predictability and its sensitivity to sea ice albedo parameterization in a global coupled model. *Climate Dynamics*, **38,** 2389-2408. doi: 10.1007/s00382-011-1132-z

Köhl, A., 2015: Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic. *Quarterly Journal of the Royal Meteorological Society*, **141,** 166-181. doi: 10.1002/qj.2347

Kowalski, C. J., 1972: On the Effects of Non-normality on the Distribution of the Sample Product-moment Correlation Coefficient. *Journal of the Royal Statistical Society*, **21,** 1-12

Kröger, J., H. Pohlmann, F. Sienz, et al., 2017: Full-field initialized decadal predictions with the MPI earth system model: an initial shock in the North Atlantic. *Climate Dynamics*, **51,** 2593-2608. doi: 10.1007/s00382-017-4030-1

Kumar, S., V. Merwade, J. L. Kinter, and D. Niyogi, 2013: Evaluation of Temperature and Precipitation Trends and Long-Term Persistence in CMIP5 Twentieth-Century Climate Simulations. *Journal of Climate*, **26,** 4168-4185. doi: 10.1175/jcli-d-12-00259.1

Kwok, R., 2018: Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environmental Research Letters*, **13,** 105005. doi: 10.1088/1748-9326/aae3ec

Kwok, R., G. F. Cunningham, M. Wensnahan, et al., 2009: Thinning and volume loss of the Arctic Ocean sea ice cover: 2003–2008. *Journal of Geophysical Research*, **114**. doi: 10.1029/2009jc005312

Labe, Z., G. Magnusdottir, and H. Stern, 2018: Variability of Arctic Sea Ice Thickness Using PIOMAS and the CESM Large Ensemble. *Journal of Climate*, **31,** 3233-3247. doi: 10.1175/jcli-d-17-0436.1

Lamb, G. M., 2008: An icy plunge to save the melting Arctic. *The Christian Science Monitor* [https://www.csmonitor.com/Environment/2008/0110/p14s02-sten.html]

Lashof, D. A., and D. R. Ahuja, 1990: Relative contributions of greenhouse gas emissions to global warming. *Nature*, **344,** 529-531

Lauer, A., V. Eyring, M. Righi, et al., 2017: Benchmarking CMIP5 models with a subset of ESA CCI Phase 2 data using the ESMValTool. *Remote Sensing of Environment*. doi: 10.1016/j.rse.2017.01.007

Laxon, S., N. Peacock, and D. Smith, 2003: High interannual variability of sea ice thickness in the Arctic region. *Nature*, **425,** 947-950. doi: 10.1038/nature02050

Laxon, S. W., K. A. Giles, A. L. Ridout, et al., 2013: CryoSat-2 estimates of Arctic sea ice thickness and volume. *Geophysical Research Letters*, **40,** 732-737. doi: 10.1002/grl.50193

Li, L., P. Lin, Y. Yu, et al., 2013: The flexible global ocean-atmosphere-land system model, Grid-point Version 2: FGOALS-g2. *Advances in Atmospheric Sciences*, **30,** 543-560. doi: 10.1007/s00376-012-2140-6

Lindsay, R. W., and J. Zhang, 2006: Assimilation of Ice Concentration in an Ice–Ocean Model. *Journal of Atmospheric and Oceanic Technology*, **23,** 742-749. doi: 10.1175/jtech1871.1

Lique, C., M. M. Holland, Y. B. Dibike, et al., 2016: Modeling the Arctic freshwater system and its integration in the global system: Lessons learned and future challenges. *Journal of Geophysical Research: Biogeosciences*, **121,** 540-566. doi: 10.1002/2015jg003120

Liu, C., K. Haines, A. Iwi, and D. Smith, 2012: Comparing the UK Met Office Climate Prediction System DePreSys with idealized predictability in the HadCM3 model. *Quarterly Journal of the Royal Meteorological Society*, **138,** 81-90. doi: 10.1002/qj.904

Liu, J., M. Song, R. M. Horton, and Y. Hu, 2013: Reducing spread in climate model projections of a September ice-free Arctic. *Proc Natl Acad Sci U S A*, **110,** 12571-12576. doi: 10.1073/pnas.1219716110

Marotzke, J., W. A. Müller, F. S. E. Vamborg, et al., 2016: MiKlip - a National Research Project on Decadal Climate Prediction. *Bulletin of the American Meteorological Society*

Marsland, S. J., H. Haak, J. H. Jungclaus, et al., 2003: The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modelling*, **5,** 91-127. doi: 10.1016/s1463-5003(02)00015-x

Martin, G. M., N. Bellouin, W. J. Collins, et al., 2011: The HadGEM2 family of Met Office Unified Model climate configurations. *Geoscientific Model Development*, **4,** 723-757. doi: 10.5194/gmd-4-723-2011

Massonnet, F., T. Fichefet, and H. Goosse, 2015: Prospects for improved seasonal Arctic sea ice predictions from multivariate data assimilation. *Ocean Modelling*, **88,** 16-25. doi: 10.1016/j.ocemod.2014.12.013

Massonnet, F., M. Vancoppenolle, H. Goosse, et al., 2018: Arctic sea-ice change tied to its mean state through thermodynamic processes. *Nature Climate Change*, **8,** 599-603. doi: 10.1038/s41558-018-0204-z

Massonnet, F., T. Fichefet, H. Goosse, et al., 2012: Constraining projections of summer Arctic sea ice. *Cryosphere*, **6,** 1383-1394. doi: 10.5194/tc-6-1383-2012

Meehl, G. A., and L. Goddard, 2013: Decadal Climate Prediction: An Update from the Trenches. *Bulletin of the …*, **3,** 1-82. doi: 10.1175/BAMS-D-12-00241.1

Meehl, G. A., C. T. Y. Chung, J. M. Arblaster, et al., 2018: Tropical Decadal Variability and the Rate of Arctic Sea Ice Decrease. *Geophysical Research Letters*. doi: 10.1029/2018gl079989

Meehl, G. A., L. Goddard, J. Murphy, et al., 2009: Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*, **90,** 1467-1485. doi: 10.1175/2009BAMS2778.1

Melia, N., K. Haines, and E. Hawkins, 2015: Improved Arctic sea ice thickness projections using bias-corrected CMIP5 simulations. *The Cryosphere*, **9,** 2237-2251. doi: 10.5194/tc-9-2237-2015

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, **117,** n/a-n/a. doi: 10.1029/2011jd017187

Müller, V., H. Pohlmann, A. Düsterhus, et al., 2017: Hindcast skill for the Atlantic meridional overturning circulation at 26.5°N within two MPI-ESM decadal climate prediction systems. *Climate Dynamics*, **49,** 2975-2990. doi: 10.1007/s00382-016-3482-z

Müller, W. A., C. Appenzeller, and C. Schär, 2004: Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Climate Dynamics*, **24,** 213-226. doi: 10.1007/s00382-004-0492-z

Müller, W. A., J. Baehr, H. Haak, et al., 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophysical Research Letters*, **39,** 1-7. doi: 10.1029/2012GL053326

NASA/Goddard-Space-Flight-Center: 2018 Arctic summertime sea ice minimum extent tied for sixth lowest on record. [Available online at www.sciencedaily.com/releases/2018/09/180927122939.htm.]

Notz, D., 2014: Sea-ice extent and its trend provide limited metrics of model performance. *The Cryosphere*, **8,** 229-243. doi: 10.5194/tc-8-229-2014

Notz, D., 2015: Bedeutung des Meereises für das Weltklima. *Lozán, J. L., H. Grassl, D. Kasang, D. Notz & H. Escher-Vetter(hrsg.). Warnsignal Klima: Das Eis der Erde.*, 189-193. doi: 10.2312/warnsignal.klima.eis-der-erde.28

Notz, D., and J. Marotzke, 2012: Observations reveal external driver for Arctic sea-ice retreat. *Geophysical Research Letters*, **39,** n/a-n/a. doi: 10.1029/2012gl051094

Notz, D., F. A. Haumann, H. Haak, et al., 2013: Arctic sea-ice evolution as modeled by Max Planck Institute for Meteorology's Earth system model. *Journal of Advances in Modeling Earth Systems*, **5,** 173-194. doi: 10.1002/jame.20016

Perovich, D. K., 2002: Seasonal evolution of the albedo of multiyear Arctic sea ice. *Journal of Geophysical Research*, **107**. doi: 10.1029/2000jc000438

Pohlmann, H., J. Kröger, R. J. Greatbatch, and W. A. Müller, 2016: Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific. *Climate Dynamics*. doi: 10.1007/s00382-016-3486-8

Pohlmann, H., J. H. Jungclaus, A. Köhl, et al., 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *Journal of Climate*, **22,** 3926-3938. doi: 10.1175/2009JCLI2535.1

Polyak, L., R. B. Alley, J. T. Andrews, et al., 2010: History of sea ice in the Arctic. *Quaternary Science Reviews*, **29,** 1757-1778. doi: 10.1016/j.quascirev.2010.02.010

Rayner, N. A., D. E. Parker, E. B. Horton, et al., 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, **108,** 4407. doi: 10.1029/2002JD002670

Riahi, K., S. Rao, V. Krey, et al., 2011: RCP 8.5—A scenario of comparatively high greenhouse gas emissions. *Climatic Change*, **109,** 33-57. doi: 10.1007/s10584-011-0149-y

Robson, J., I. Polo, D. L. R. Hodson, et al., 2018: Decadal prediction of the North Atlantic subpolar gyre in the HiGEM high-resolution climate modelc. *Climate Dynamics*, **50,** 921-937

Rotstayn, L. D., S. J. Jeffrey, M. A. Collier, et al., 2012: Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations. *Atmospheric Chemistry and Physics*, **12,** 6377-6404. doi: 10.5194/acp-12-6377-2012

Sanderson, B. M., M. Wehner, and R. Knutti, 2017: Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, **10,** 2379-2395. doi: 10.5194/gmd-10-2379-2017

Schmidt, G. A., M. Kelley, L. Nazarenko, et al., 2014: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, **6,** 141-184. doi: 10.1002/2013ms000265

Schweiger, A., R. Lindsay, J. Zhang, et al., 2011: Uncertainty in modeled Arctic sea ice volume. *Journal of Geophysical Research*, **116**. doi: 10.1029/2011jc007084

Screen, J. A., 2018: Arctic sea ice at 1.5 and 2 °C. *Nature Climate Change*, **8,** 362-363. doi: 10.1038/s41558-018-0137-6

Senftleben, D., V. Eyring, A. Lauer, and M. Righi, 2018: Temperature and sea ice hindcast skill of the MiKlip decadal prediction system in the Arctic. *Meteorologische Zeitschrift*, **27,** 195-208. doi: 10.1127/metz/2018/0871

Senftleben, D., V. Eyring, A. Lauer, et al., 2019: Constraining uncertainties in CMIP5 projections of September Arctic sea ice extent with observations. *Journal of Climate*, **under review**

Shaffrey, L. C., I. Stevens, W. A. Norton, et al., 2009: U.K. HiGEM: The New U.K. High-Resolution Global Environment Model—Model Description and Basic Evaluation. *Journal of Climate*, **22,** 1861-1896. doi: 10.1175/2008jcli2508.1

Shakhova, N., I. Semiletov, A. N. Salyuk, and N. A. Belcheva, 2007: Anomalies of methane in the atmosphere above the East Siberian Arctic Shelf. *Transaction of Russian Academy of Science*, **414,** 819-823

Shimada, K., T. Kamoshida, M. Itoh, et al., 2006: Pacific Ocean inflow: Influence on catastrophic reduction of sea ice cover in the Arctic Ocean. *Geophysical Research Letters*, **33**. doi: 10.1029/2005gl025624

Sienz, F., W. A. Müller, and H. Pohlmann, 2015: Ensemble size impact on the decadal predictive skill assessment. *Meteorol. Z.***,** 1-28. doi: 10.1127/metz/2016/0670

Sigmond, M., J. C. Fyfe, and N. C. Swart, 2018: Ice-free Arctic projections under the Paris Agreement. *Nature Climate Change*, **8,** 404-408. doi: 10.1038/s41558-018-0124-y

Singarayer, J. S., J. L. Bamber, and P. J. Valdes, 2006: Twenty-First-Century Climate Impacts from a Declining Arctic Sea Ice Cover. *Journal of Climate*, **19,** 1109-1125. doi: 10.1175/jcli3649.1

Smith, D. M., R. Eade, N. J. Dunstone, et al., 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nature Geoscience*, **3,** 846-849. doi: 10.1038/ngeo1004

Snape, T. J., and P. M. Forster, 2014: Decline of Arctic sea ice: Evaluation and weighting of CMIP5 projections. *Journal of Geophysical Research: Atmospheres*, **119,** 546-554. doi: 10.1002/2013jd020593

Spreen, G., R. Kwok, and D. Menemenlis, 2011: Trends in Arctic sea ice drift and role of wind forcing: 1992-2009. *Geophysical Research Letters*, **38,** n/a-n/a. doi: 10.1029/2011gl048970

Stevens, B., M. Giorgetta, M. Esch, et al., 2013: Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, **5,** 146-172. doi: 10.1002/jame.20015

Stroeve, J., A. Barrett, M. Serreze, and A. Schweiger, 2014: Using records from submarine, aircraft and satellites to evaluate climate model simulations of Arctic sea ice thickness. *The Cryosphere*, **8,** 1839-1854. doi: 10.5194/tc-8-1839-2014

Stroeve, J. C., M. M. Holland, W. Meier, et al., 2007: Arctic sea ice decline: Faster than forecast. *Geophysical Research Letters*, **34,** 1-5. doi: 10.1029/2007GL029703

Stroeve, J. C., M. C. Serreze, M. M. Holland, et al., 2012a: The Arctic's rapidly shrinking sea ice cover: A research synthesis. *Climatic Change*, **110,** 1005-1027. doi: 10.1007/s10584-011-0101-1

Stroeve, J. C., V. Kattsov, A. Barrett, et al., 2012b: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters*, **39,** 1-7. doi: 10.1029/2012GL052676

Swart, N. C., J. C. Fyfe, E. Hawkins, et al., 2015: Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, **5,** 86-89. doi: 10.1038/nclimate2483

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, **93,** 485-498. doi: 10.1175/BAMS-D-11-00094.1

Thomson, A. M., K. V. Calvin, S. J. Smith, et al., 2011: RCP4.5: a pathway for stabilization of radiative forcing by 2100. *Climatic Change*, **109,** 77-94. doi: 10.1007/s10584-011-0151-4

Thorndike, A. S., and R. Colony, 1982: Sea ice motion in response to geostrophic winds. *Journal of Geophysical Research*, **87,** 5845. doi: 10.1029/JC087iC08p05845

Tietsche, S., J. J. Day, V. Guemas, et al., 2014: Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophysical Research Letters*, **41,** 1035-1043. doi: 10.1002/2013GL058755

Tilling, R. L., A. Ridout, A. Shepherd, and D. J. Wingham, 2015: Increased Arctic sea ice volume after anomalously low melting in 2013. *Nature Geoscience*, **8,** 643-646. doi: 10.1038/ngeo2489

Tsamados, M., D. Feltham, A. Petty, et al., 2015: Processes controlling surface, bottom and lateral melt of Arctic sea ice in a state of the art sea ice model. *Philos Trans A Math Phys Eng Sci*, **373**. doi: 10.1098/rsta.2014.0167

Uppala, S. M., P. W. KÅllberg, A. J. Simmons, et al., 2005: The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, **131,** 2961-3012. doi: 10.1256/qj.04.176

Våge, K., R. S. Pickart, V. Thierry, et al., 2008: Surprising return of deep convection to the subpolar North Atlantic Ocean in winter 2007–2008. *Nature Geoscience*, **2,** 67-72. doi: 10.1038/ngeo382

van Vuuren, D. P., J. Edmonds, M. Kainuma, et al., 2011: The representative concentration pathways: an overview. *Climatic Change*, **109,** 5-31. doi: 10.1007/s10584-011-0148-z

Vichi, M., E. Manzini, P. G. Fogli, et al., 2011: Global and regional ocean carbon uptake and climate change: sensitivity to a substantial mitigation scenario. *Climate Dynamics*, **37,** 1929-1947. doi: 10.1007/s00382-011-1079-0

Vihma, T., 2014: Effects of Arctic Sea Ice Decline on Weather and Climate: A Review. *Surveys in Geophysics*, **35,** 1175-1214. doi: 10.1007/s10712-014-9284-0

Voldoire, A., E. Sanchez-Gomez, D. Salas y Mélia, et al., 2012: The CNRM-CM5.1 global climate model: description and basic evaluation. *Climate Dynamics*, **40,** 2091-2121. doi: 10.1007/s00382-011-1259-y

Volodin, E. M., N. A. Dianskii, and A. V. Gusev, 2010: Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. *Izvestiya, Atmospheric and Oceanic Physics*, **46,** 414-431. doi: 10.1134/s000143381004002x

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 494 pp pp

Walsh, J. E., W. L. Chapman, and F. Fetterer, 2015: Gridded Monthly Sea Ice Extent and Concentration, 1850 Onward, Version 1. Boulder, Colorado USA.

Wang, M., and J. E. Overland, 2009: A sea ice free summer Arctic within 30 years? *Geophysical Research Letters*, **36,** n/a-n/a. doi: 10.1029/2009gl037820

Wang, Y., F. Counillon, N. Keenlyside, et al., 2018: Seasonal-to-decadal prediction with the Norwegian Climate Prediction Model. *Geophysical Research Abstracts*, **20**

Warner, J., J. A. Screen, and A. Scaife, 2017: Role of Arctic sea ice in skilful predictions of the North Atlantic Oscillation. *EGU General Assembly Conference Abstracts*, **19**

Watanabe, S., T. Hajima, K. Sudo, et al., 2011: MIROC-ESM: model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development Discussions*, **4,** 1063-1128. doi: 10.5194/gmdd-4-1063-2011

Weeks, W. F., 2010: *On sea ice.* University of Alaska Press. ISBN: 978-1-60223-079-8

Wenzel, S., V. Eyring, E. P. Gerber, and A. Y. Karpechko, 2016: Constraining Future Summer Austral Jet Stream Positions in the CMIP5 Ensemble by Process-Oriented Multiple Diagnostic Regression*. *Journal of Climate*, **29,** 673-687. doi: 10.1175/jcli-d-15-0412.1

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. *International Geophysics Series*, **102**. doi: 10.1198/jasa.2007.s163

Wu, T., L. Song, W. Li, et al., 2014: An overview of BCC climate system model development and application for climate change studies. *Acta Meteorologica Sinica*, **28,** 34-56. doi: 10.1007/s13351-014-3041-7

Yang, Q., S. N. Losa, M. Losch, et al., 2017: Assimilating summer sea-ice concentration into a coupled ice–ocean model using a LSEIK filter. *Annals of Glaciology*, **56,** 38-44. doi: 10.3189/2015AoG69A740

Yeager, S. G., A. R. Karspeck, and G. Danabasoglu, 2015: Predicted slowdown in the rate of Atlantic sea ice loss. doi: 10.1002/2015GL065364.Received

Yukimoto, S., Y. Adachi, M. Hosaka, et al., 2012: A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3 ^|^mdash;Model Description and Basic Performance^|^mdash. *Journal of the Meteorological Society of Japan*, **90A,** 23-64. doi: 10.2151/jmsj.2012-A02

Zhang, J., and D. A. Rothrock, 2003: Modeling Global Sea Ice with a Thickness and Enthalpy Distribution Model in Generalized Curvilinear Coordinates. *Monthly Weather Review*, **131,** 845-861. doi: 10.1175/1520-0493(2003)131<0845:mgsiwa>2.0.co;2

Zhang, J. L., D. Rothrock, and M. Steele, 2000: Recent changes in Arctic sea ice: The interplay between ice dynamics and thermodynamics. *Journal of Climate*, **13,** 3099-3114. doi: Doi 10.1175/1520-0442(2000)013<3099:Rciasi>2.0.Co;2

Acknowledgments

This work was funded by the BMBF MiKlip Climate Model Validation by confronting globally Essential Climate Variables from models with observations (ClimVal) project and by the Advanced Earth System Model Evaluation for CMIP (EVal4CMIP) project funded by the Helmholtz Association of German Research Centers.

First and foremost, I would like to thank Professor Veronika Eyring for giving me the opportunity to do my PhD studies at her department Earth System Model Evaluation and Analysis (EVA) at the German Aerospace Center (DLR), for being my first assessor, and for supervising both my Master's and PhD Theses. Her tremendous support and scientific experience have helped me substantially. Furthermore, I thank her for giving me the opportunity to actively participate in international climate conferences, such as the Polar Prediction Workshops in Bremerhaven and New York, and meeting many inspiring climate scientists from all over the world.

I thank Professor George Craig for being my second assessor and providing advice throughout my time as PhD student.

Moreover, I would like to thank Dr. Marika Holland, Professor Alexey Karpechko, and Dr. François Massonnet for insightful discussions and their contributions to my second study.

Special thanks are due to Professor Andreas Burkert and Professor Fritz Schmielau for their encouragements and helpful second opinions from outside the field of climate research.

A tremendous thank you goes to all of my colleagues of department EVA for the great time. To Dr. Mattia Righi for providing technical support on NCL and the ESMValTool and for his great contributions to my first study. To Manuel Schlund and Dr. Björn Brötz, who always shared their ingenuity in Python and Git. To Tina Gier for her helpful comments on my manuscripts. To Sabrina Zechlau for sharing her PhD experience and giving great advice. To Dr. Franziska Winterstein who always had the answers to my questions right there in her desk drawer. To Ricarda Rettinger for the many talks and laughs during long office hours. To Dr. Birgit Hassler for being such an awesome person (and for always providing chocolate). And last but certainly not least, to Dr. Axel Lauer – for simply everything.

A very special thank you to Josh and Katya for the much needed distractions and for always being there for me.

And finally, and most importantly, the biggest thank you ever to my family for their never ending and invaluable support in all areas. DANKE.