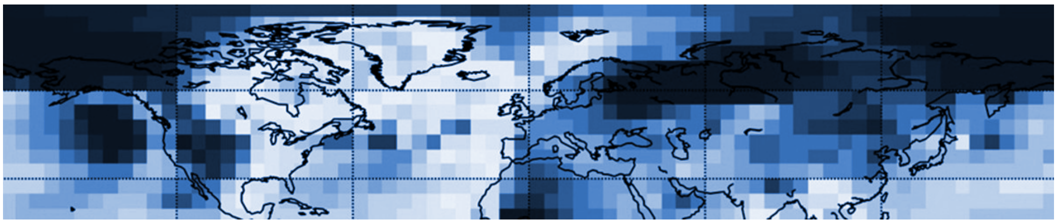
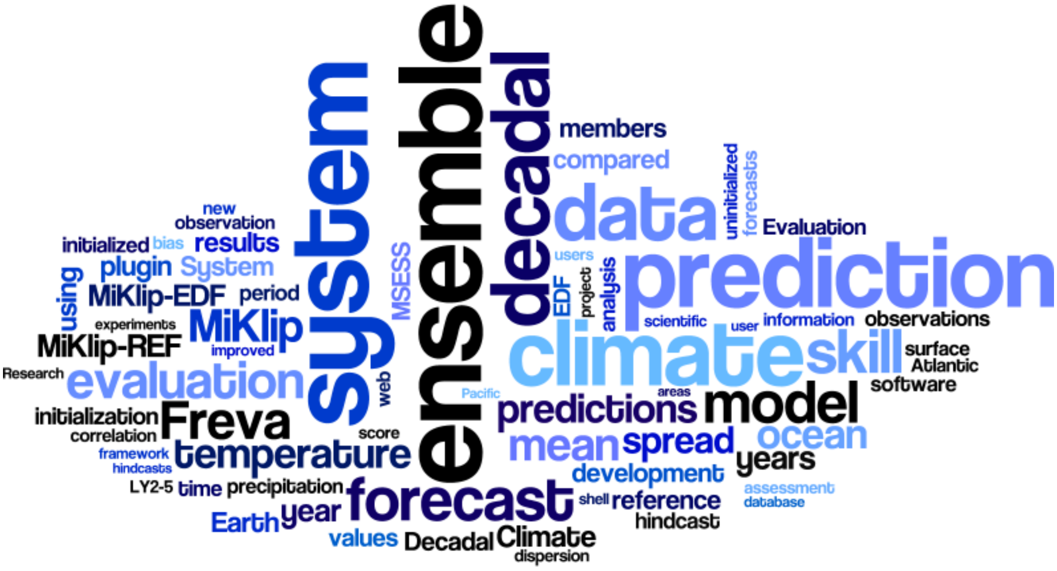


Improving decadal climate predictions by ocean ensemble dispersion filtering and an efficient systematic evaluation

Christopher Kadow

Cumulative Dissertation at the
Institute of Meteorology of the
Freie Universität Berlin
October 2018



Cover Figure: First ever stated annual 2m-temperature forecast by MiKlip. This forecast for the year 2014 was released by the PhD candidate in 2013 and published by the Max-Planck-Institute for Meteorology[1] and Kadow et al., 2016. Re-colored for the cover stylesheet.

[1] https://www.mpimet.mpg.de/en/communication/news/focus-on-overview/decadal-climate-predictions

Freie Universität Berlin

Department of Earth Sciences
Institute of Meteorology
Climate Modeling Group

Dissertation
zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften
am Fachbereich Geowissenschaften der Freien Universität Berlin

PhD Thesis

Improving decadal climate predictions by ocean ensemble dispersion filtering and an efficient systematic evaluation

Christopher Kadow

1. Reviewer	Univ.-Prof. Dr. Ulrich Cubasch Institute of Meteorology Freie Universität Berlin
2. Reviewer	Univ.-Prof. Dr. Uwe Ulbrich Institute of Meteorology Freie Universität Berlin

October 22th, 2018
Revised January 18th, 2019

Christopher Kadow
Improving decadal climate predictions by
ocean ensemble dispersion filtering and
an efficient systematic evaluation
Thesis (Dissertation), October 22th, 2018
Defense (Disputation), January 11th, 2019
Reviewer: Univ.-Prof. Dr. Ulrich Cubasch
and Univ.-Prof. Dr. Uwe Ulbrich
Freie Universität Berlin
Climate Modeling Group
Institute of Meteorology
Department of Earth Sciences
Carl-Heinrich-Becker-Weg 6-10
12165 Berlin

Für Mairhi und Logan

Publications of this Cumulative Dissertation

This is work is presented as a thesis by publications and comprises the following research articles that were published and submitted to international peer-review ISI-index journals:

Chapter 2:
Kadow, C., S. Illing, O. Kunst, T. Schartner, I. Kirchner, J. Grieger, M. Schuster, A. Richling, H.W. Rust, U. Cubasch, and U. Ulbrich, Freva - Free Evaluation System Framework for Earth System Modeling, *Journal of Open Research Software*, E-ISSN: 2049-9647 (2018, in review)

Chapter 3:
Kadow, C., S. Illing, O. Kunst, H. W. Rust, H. Pohlmann, W. A. Müller, and U. Cubasch (2016), Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system, *Meteorologische Zeitschrift - Open Access Schweizerbart Science Publisher*, Volume 25, Pages 631-643, `https://doi.org/10.1127/metz/2015/0639`.

Chapter 4:
Kadow, C., S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch (2017),
Decadal climate predictions improved by ocean ensemble dispersion filtering,
Journal of Advances in Modeling Earth Systems - An Open Access AGU Journal,
9, 1138–1149, `https://doi.org/10.1002/2016MS000787`.

Abstract

Decadal climate predictions have the objective to predict the development of the climate for the following years to decades. Numerical Earth system models are initialized with observational values, similar to the methodology applied in weather forecasting. Additionally, they are forced by boundary conditions, like greenhouse gas scenarios, to project the long term development. This thesis investigates decadal climate predictions with Earth system models and their further improvement.

A decadal prediction system is evaluated and investigated for sources of potential skill. Hence a systematic evaluation strategy is developed. It contains the assessment of accuracy of the ensemble mean and the ensemble spread, and compares decadal experiments with climatology, observations, and climate projections. This initialized reference system leads to good predictive skill in temperature and precipitation forecasts. The evaluation shows that the decadal prediction is scientifically sound, but it also has potential for improvement. The initialization with observed ocean data and the prediction with the ensemble mean of a larger ensemble size turn out to be sources of skill for decadal predictions. The entire assessment is performed within a novel evaluation system called Freva. This system is designed to complement climate modeling by a systematic and efficient assessment. Freva serves as a resource-efficient process framework between the data generation and its evaluation, to detect decadal climate prediction potential.

A new prediction technique called 'Ensemble Dispersion Filter' is developed. It exploits two important climate prediction paradigms: the ocean's heat capacity and the advantage of the ensemble mean. The Ensemble Dispersion Filter averages the ocean temperatures of the ensemble members every three months, uses this ensemble mean as a restart condition for each member, and further executes the prediction. The evaluation by the new verification framework shows that the Ensemble Dispersion Filter results in a significant improvement in the predictive skill compared to the unfiltered reference system. Even in comparison with prediction systems of a larger ensemble size and higher resolution, the Ensemble Dispersion Filter system performs better. In particular, the prediction of the global average temperature of the forecast years 2 to 5 shows a significant skill improvement. Compared to the observational climatology forecast, the Ensemble Dispersion Filter experiment has a Mean Squared Error Skill Score of 0.83, while the unfiltered reference system exceeds only 0.68. With major improvements over the Pacific and North Atlantic, the regional distribution of the Ensemble Dispersion Filter experiment is more accurate than the reference. In precipitation forecasts, improvements are seen over the continents. The prediction of the cyclone frequencies improves over the key region of the North Atlantic. Consequently, the thesis demonstrates a substantial advance in research on decadal climate predictions.

Zusammenfassung

Dekadische Klimavorhersagen haben das Ziel, die Entwicklung des Klimas in den kommenden Jahren und Jahrzehnten vorherzusagen. Hierfür werden numerische Erdsystemmodelle mit Beobachtungsdaten ähnlich zur Methodik der Wettervorhersagen initialisiert. Zusätzlich werden die Modelle für die langfristige Entwicklung wie bei Klimaprojektionen mit Randbedingungen wie Treibhausgasen angetrieben. Diese Dissertation untersucht dekadische Klimavorhersagen mittels Erdsystemmodellen und deren weitere Verbesserung.

Ein dekadisches Vorhersagesystem wird ausgewertet und auf Quellen der Vorhersagegüte hin untersucht. Hierfür wird eine Auswertestrategie entwickelt. Diese beinhaltet die Genauigkeit des Ensemblemittels und die Ensemblestreuung, und vergleicht dekadische Experimente mit Klimatologien, Beobachtungen und Klimaprojektionen. Das initialisierte Referenz-Vorhersagesystem besitzt eine Vorhersagegüte in Temperatur- und Niederschlagsvorhersagen. Die dekadische Vorhersage funktioniert bereits, weist jedoch noch Verbesserungspotenzial auf. Die Initialisierung mit Ozeandaten und eine Vorhersage mit dem Ensemblemittel erweisen sich als Quellen der Vorhersagegüte. Die gesamte Analyse findet in dem eigens entwickelten Evaluierungssystem Freva statt. Es wurde zur Erweiterung der Klimamodellierung mittels systematischer Verifikation konzipiert. Freva dient als ressourceneffiziente Schaltzentrale zwischen den Modelldaten und den Auswerteverfahren, um Potenziale der dekadischen Klimavorhersage zu erkennen.

Eine neue Vorhersagemethode namens 'Ensemble-Dispersionsfilter' wird entwickelt. Diese nutzt zwei Klimavorhersage-Paradigmen: die Wärmekapazität des Ozeans und den Vorteil des Ensemblemittels. Der Ensemble-Dispersionsfilter mittelt nach jeweils drei Monaten die Ozeantemperaturen des Ensembles und benutzt dieses Ensemblemittel als Neustartbedingung für jedes Ensemblemitglied. Die Auswertung mit dem neuen Evaluierungssystem zeigt eine signifikante Verbesserung, verglichen mit dem ungefilterten Referenzsystem. Das Ensemble-Dispersionsfilter Experiment ist sogar besser als Systeme mit höherer Auflösung oder mehr Ensemblemitgliedern. Verglichen mit der Vorhersage der Klimatologie, erreicht die Vorhersagegüte der globalen Mitteltemperatur in den Vorhersagejahren 2 bis 5 im Gütemaß der mittleren quadratischen Abweichung 0.83 beim Ensemble Dispersionsfilter, das ungefilterte Referenzvorhersagesystem hingegen nur 0.63. Mit Verbesserungen über dem Zentral-Pazifik und Nord-Atlantik zeigt das Ensemble-Dispersionsfilter Experiment auch im regionalen Vergleich die bessere Güte. Beim Niederschlag zeigen sich regionale Verbesserungen vor allem über den Kontinenten. In der Vorhersage der Zyklonenhäufigkeit ist ein signifikanter Fortschritt über der entscheidenden Nord-Atlantik Region zu verzeichnen. Folglich zeigt diese Arbeit eine substantielle Weiterentwicklung für die Forschung der dekadischen Klimavorhersagen.

Contents

1 Introduction and Research Agenda

> *The paradigm of physics - with its interplay of data, theory, and prediction - is the most powerful in science.*
>
> — **Geoffrey West**
> (Physicist)

The prediction of the atmospheric development has always been important for many sectors of society, industry, and economy. The weather prediction on the scale of several days, as well as the climate projection on the scale of centuries are common tools used in management and planning. In the recent decades, the attention of scientists has been focused on the development of multi-model systems for seasonal forecasts. In the recent years, the extension to a seamless prediction from scales of seasons to one hundred years has become the ultimate goal.

The ’Decadal Climate Prediction’ is still a young branch within climate science. For this forecast horizon, there is a growing demand for a broad range of accurate climate information for medium-term planning activities like the design of power stations, or water management. The evolution of the climate in the near term is the combination of climate variability and climate change (Figure 1.1). Changes in natural variability are large enough to temporarily reduce or intensify climate trends (Easterling and Wehner, 2009). Research aims to combine short-term forecasts (initial value) and long-term projection (boundary condition) approaches as described in the following.

The long-term climate projections up to centuries ahead pose a boundary condition problem (Figure 1.1 - right) —e.g. the increase of greenhouse gases— and examine the long-term climate development (Meehl et al., 2009; Mehta et al., 2011). Therefore, the climate model needs external information like the concentration of greenhouse gases and volcanic aerosols, or solar radiation. These boundary conditions influence the development of the climate system. Climate projections extent to a century ahead and represent the mean path of the anthropogenic forced climate evolution within an envelope of uncertainty. For boundary conditions of the past, observations can be used for hindcasts or historical simulations. For simulations which extend into the future, projected boundary conditions are applied. Representative Concentration Pathways (RCPs) show scenarios for the future evolution (Vuuren et al., 2011). For the Intergovernmental Panel on Climate Change (IPCC[1]) four RCPs were selected and defined by their total radiative forcing pathway and level by the year 2100. The RCPs were chosen to represent a broad range of climate outcomes. Each RCP could result from different combinations of economic, technological, demographic, policy, and institutional futures. Decadal climate predictions also need the RCPs scenarios for their forecast ability. For decadal climate predictions usually the RCP4.5 scenario is chosen for its future boundary condition. The RCP4.5 scenario represents the stabilization without overshoot pathway to 4.5 W/m^2 at stabilization after 2100[1]. However, these climate projections alone do rarely account for the actual, initial state-dependent evolution of climate in the near term.

Short to medium-range weather forecasts represent an initial value problem at the beginning of a forecast (Figure 1.1 - left). The more we know about the current atmospheric state, the better the initialized forecast becomes. For a perfect forecast one would have an observed value for all its grid-boxes and all its variables at the start of a simulation. However, observations do not exist for each point in time and on Earth, so that initial conditions usually consist of a blend of mean observational values and coupled model simulations such as (re-)analyses. Dynamical processes of the climate system act on different time-scales. Especially the atmosphere has a short-term memory of only a few weeks (Meehl et al., 2009), which is important for weather forecasts. The longer the forecasts lasts - from months, over seasons, to several years - the more important the role of the ocean gets. Large scale changes in the ocean not only have an influence on the atmosphere for a couple of days, but possibly over several years. This multi-year ’memory’ of the climate system can be used for better predictions beyond the short-term.

For the time of seasons to decades, there is a need for the prediction of the combined impact of externally forced and internally generated variability. Addressing this need can only be achieved by predictions initialized with the current climate state. Initialization is the key-word for decadal predictions which developed out of climate projections, which are uninitialized free simulations with forcing parameters. The variability of these simulations are not in phase with the actual climate state of the real world. Therefore, the climate model needs the information as starting parameters - it needs to be initialized. The weather forecasts using and developing these initialization and assimilation strategies within the last decades for their atmosphere models. As stated already, the decadal time scale is dominated by the inertia of the oceans (Meehl et al., 2009). Therefore, a usable numerical decadal prediction is only possible through the coupling of atmospheric and oceanic numerical models (Figure 1.1 - top and bottom).

[1] www.ipcc.ch

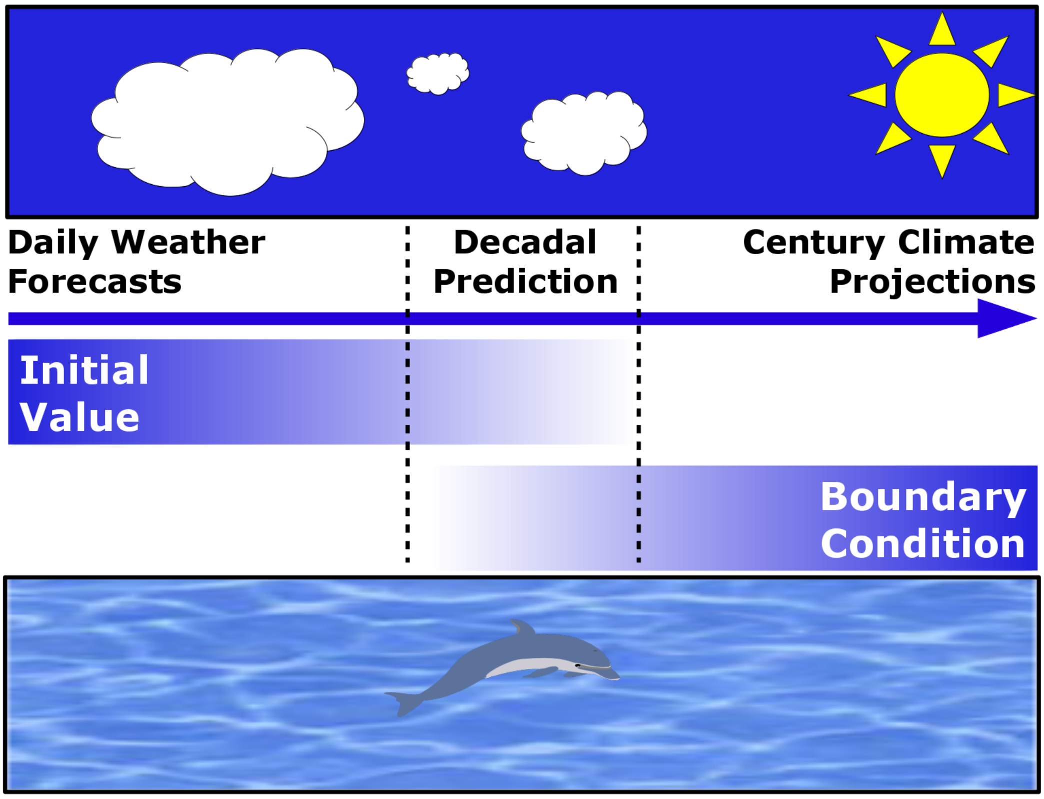


Fig. 1.1: Scheme of climatic predictions on different scales with different challenges from initial value problems with daily weather forecasts at one end, and century projections as a forced boundary condition problem at the other, with decadal prediction in between. The importance of atmospheric-ocean (model) interaction is highlighted as well. Adapted from Meehl et al., 2009.

This was achieved with designing modern fully coupled Earth system models within the last decade. Still, there was the need to develop strategies for the ocean model initialization to start the model system with the correct climate state [Smith et al., 2007, Keenlyside et al., 2008, Pohlmann et al., 2009]. Additionally, initialization strategies with observational data [e.g. Matei et al., 2012, Smith et al., 2013b, Meehl et al., 2014] have been adopted for these coupled models. The observations or reanalyses usually get nudged or assimilated into one model simulation. Hereby, specific and selected observational variables are adjusted or exchanged with the model values. The selection of variables takes their impact on the decadal scale into account. Therefore, oceanic variables (e.g. temperatures, salinity, currents, etc.) and atmospheric variables which drive the ocean (e.g. wind-stress, 2m temperature, pressure, etc.) are typically involved. However, the strategies on initialization variables differ between research institutions.

The same is valid for the two main methods exchanging the variables: Full-field and anomaly. The full-field initialization uses observational or reanalysis data as it is. As there is no correction applied, the model forecast is close to the actual and absolute values of the applied observations or reanalysis in the first few months. However, the longer the forecast lasts, the more the climate model starts to adjust the values to the models own climate. This results in potential drifts within the model, reacting to the so called initial-shock (Pohlmann et al., 2017). Potential bias or drift correction methods can correct the results to a certain level [e.g. Kruschke et al., 2015, Smith et al., 2013b, Ferro, 2007]. For the anomaly initialization the observational values are corrected before their application. Anomalies of the observational values are calculated and added to the model values regarding the climate model's climatology. The anomaly method does not have a strong shift away from the model's climate. The danger of a potential drift of the model system is much smaller. However, with the anomaly initialization the model only produces correct absolute values, if the model bias is small compared to the observation. This is usually not the case. Therefore, only anomaly forecasts are possible with the anomaly initialization. As climate predictions are anomaly forecasts with respect to a certain climate period, it is neither a argument against or for the technique.

Beside the main research on developments of initialization techniques, the science community focused its research on the increase of spatial resolution [e.g. Pohlmann et al., 2013, Menary et al., 2015, Shaffrey et al., 2017] and the generation of ensemble forecasts [e.g. Goddard et al., 2013, Boer et al., 2016, Sienz et al., 2016]. Both procedures showed promising results and encouraged the decadal science communities to consider these development strategies. To evaluate the skill of the prediction of the future climate states, one performs retrospective predictions of the past - so-called hindcasts. At present, hindcasts are set up annually to produce sufficient data for a meaningful statistical evaluation (Boer et al., 2016). With all these requirements (coupling of atmosphere and ocean, initialization and assimiliation techniques, high resolution models, large number of ensembles, annually launched hindcast sets, etc.) even modern high-performance computers rapidly reach their limits when calculating and evaluating decadal predictions. This leads to a demand for new and more efficient prediction and evaluation methods that provide improvements of the forecasts without the need for additional computer resources.

Concepts for carrying out decadal prediction experiments have been developed by the World Climate Research Programme (WCRP) within the framework of the fifth phase of the Coupled Model Intercomparison Project [CMIP[2] - CMIP5 (e.g. Taylor et al., 2012)] and have entered the recent IPCC[3] report. Many research groups around the world are developing decadal prediction capabilities. The UK Met Office has taken on the task to coordinate the exchange of near-real time decadal predictions presenting actual forecasts of several climate models, which is intended to facilitate research and collaboration on this topic (Smith et al., 2013). Within

[2] www.wcrp-climate.org/wgcm-cmip
[3] www.ipcc.ch

CMIP6, (Eyring et al., 2016) the Decadal Climate Prediction Project (DCPP - Boer et al., 2016) has its own Model Intercomparison Project (MIP). DCPP coordinates the international comparison, which includes experiment designs and evaluation strategies. The World Climate Research Programme (WCRP) defined the rules for the DCPP. It categorizes 'Near-Term Climate Prediction' as one of its 'Grand Challenges'. It coordinates the international research and development to improve multi-year to decadal climate predictions and their application by stake holders and decision makers. The climate predictions ranging from several years to decades require a much higher technical effort then e.g. climate projections (50 model years for each ensemble member for a projection from 1960 to 2010), because hindcast sets consist of decadal predictions annually setup over the last five decades (500 model years for each ensemble for a prediction set from 1960 to 2010). Decadal predictions also not have reached a production level yet, which is necessary for an operational application.

1.1 Thesis Structure and Research Tasks

At present, the decadal prediction research is rapidly developing and remain a lot of potential to improve the predictions. The following three research tasks (RT) are open for further investigation:

RT1) Develop and implement an evaluation system for Earth system modeling and decadal climate prediction to verify enhancements of skill in different development stages, with the full flexibility of model and observational data comparison in a sophisticated, reproducible, and efficient way.

RT2) Formulate and incorporate a systematic and comprehensible statistical framework for decadal climate prediction into the evaluation system and fully assess a prediction system to reveal scientific plausibility, prediction skill, and sources of potential skill.

RT3) Having a fully assessed and skillful decadal prediction system at hand, exploit detected sources of potential skill to further improve the decadal prediction system.

This thesis addresses these research tasks and aims at scientific improvements of decadal climate prediction and evaluation systems. The study focuses on the development of a novel forecast technique as well as the establishment of a systematic evaluation strategy to verify the new method. The new forecast technique combines the research of the open tasks to reveal and apply left potential of decadal predictions:

In **Chapter 2** the base for the next Chapters starts with a root cause analysis in the data-driven climate science and introduces the development of a new scientific evaluation software system for Earth system models to allow efficient qualitative and quantitative validations in decadal climate research.

In **Chapter 3** a verification framework to determine accuracy and ensemble spread of hindcasts is developed and implemented into the evaluation system of Chapter 2, delivers a comprehensive skill assessment of a reference decadal prediction system, and detects sources of potential skill.

In **Chapter 4** the newly developed Ensemble Dispersion Filter (EDF) forecast technique is derived from fundamental climate science paradigms: the memory of the ocean heat capacity and the advantage of the ensemble mean. The EDF is introduced and applied within the reference decadal prediction system, and evaluated with the verification framework and evaluation system of the earlier Chapters.

In **Chapter 5** the thesis closes with an overall summary, a discussion of results, an outlook for follow-on studies, and a unifying and interpreting conclusion.

Each of the Chapters 2 to 4 was published or submitted as an article within/to an ISI-index journal. The papers of Chapters 3 and 4 were published in peer-reviewed international open access journals. Chapter 2 is under review as a publication, which will be published open-access in an international journal after the peer-review is completed. More information is available within the Chapters. The thesis is embedded within the 'Mittelfristige Klimaprognosen' (MiKlip[4]) project in Germany. MiKlip develops a decadal climate prediction and evaluation system that will be transferred to the German meteorological service DWD for operational use (Marotzke et al., 2016).

[4] www.fona-miklip.de

2 Freva - Free Evaluation System Framework for Earth System Modeling

Abstract

In this study, we present the Free Evaluation System Framework (Freva). Freva is an all-in-one solution to efficiently handle evaluation and validation systems of research projects, institutes or universities in the Earth system and climate modeling community. It is a scientific software framework for high performance computing and provides all available features in both, the shell and web environment. The main system design is equipped with the common and standardized model database, programming interface, and history of evaluations. Freva's interface to the model database satisfies the international data standards provided by the Earth System Grid Federation and the World Climate Research Programme. Therefore, Freva indexes different data projects into one common search environment by storing the meta data information of the model, reanalysis and observational data sets in a database. This implemented meta data system with its advanced but easy-to-handle search tool supports scientists and their plugins to retrieve the required information of the database. A generic application programming interface allows scientific developers to connect their analysis tools with the evaluation system independently of the programming language. Users of the evaluation techniques benefit from the common interface of the evaluation system without any need to understand the different scripting languages. The history and configuration sub-system stores every analysis performed with the evaluation system in a database. Configurations and results of the tools can be shared among scientists via shell or web system. Research groups benefit from scientific transparency and reproducibility. Furthermore, if saved configurations match while starting an evaluation plugin, the system suggests to use results already produced by other users – saving CPU/h, I/O, disk space and time. Freva's efficient interaction between different technologies improves the Earth system modeling science.

The following chapter (pre-print) consists of the publication submitted to the *Journal of Open Research Software (JORS)*. This paper will be published after the review process.

Kadow, C., S. Illing, O. Kunst, T. Schartner, J. Grieger, M. Schuster, A. Richling, I. Kirchner, H.W. Rust, U. Cubasch, and U. Ulbrich, Freva - Free Evaluation System Framework for Earth System Modeling, *Journal of Open Research Software (JORS)*, E-ISSN: 2049-9647 (2018, in review)

Freva is published as open source software and accessible for the Earth system modeling community via GitHub: https://github.com/FREVA-CLINT and citable:

Kadow, C. and S. Illing. (2018, August 1). FREVA-CLINT/Freva v1.0-beta (Version v1.0-beta). *Zenodo*. http://doi.org/10.5281/zenodo.1325148

2.1 Introduction

State of the Art Challenge in Earth System Modeling

The Earth system modeling community nowadays uses information technology, data, and software as an indispensable support for science. Scientists use climate models as their main tools to simulate and research the past, present, and future climate. The Intergovernmental Panel on Climate Change (IPCC[1]) urges that 'it is crucial therefore to evaluate the performance of these models'. A growing variety of research software and the increase in computer power allows scientists to study a steadily increasing amount of data. The ongoing production of data and model development stages need to be evaluated in a sustainable way. Therefore, scientists develop evaluation and verification software with the code of best practice in mind. However, usually scientists are not software engineers. Scientists have to invest a lot of time in their software development skills. It is also common, that scientists develop software routines about topics, which were developed already many times by other scientists - probably unintended, because they are unaware of existing ones. This leads to a huge amount of partly redundant results and software development history.

It is difficult to accomplish reproducible, transparent, and efficient scientific results. Thus, there is a demand of software and community frameworks supporting scientists to overcome technical hurdles and concentrate on the research. With the growing amount of research data there is also a risk of losing track of research possibilities. Several model intercomparison projects (MIPs) were started in recent past to make climate modeling activities comparable. This was only achievable by using common international data standards and granting international data availability through the Earth System Grid Federation (ESGF[2]). These projects facilitated data standardization, validation, model comparisons, and multi-model assessment.

The ESGF database is a huge collection of Earth system modeling data. However, scientists still need to find ways of detecting and incorporating these amount of data in their science. There is also the need to incorporate other sets of observations, reanalysis, or model data, because research gets turnarounds during evaluation. Flexibility and efficiency are therefore important in data relevant research. With that being said, there is a growing need for common scientific infrastructures in the Earth system modeling community.

[1] www.ipcc.ch

[2] https://esgf.llnl.gov

Origin

The Free Evaluation System Framework (Freva[3]) has been developed for climate modeling research of decadal climate prediction within the 'Mittelfristige Klimaprognosen' (MiKlip) major project funded by the Federal Ministry of Education and Research in Germany (BMBF). Within MiKlip, the Freva framework hosts the MiKlip Central Evaluation System (CES) (Marotzke et al., 2016) on a high performance computer (HPC) at the German Climate Computing Centre (DKRZ[4]).

Exemplary Research Group

Marotzke et al., 2016 state: *'The MiKlip hub furthermore provides a central evaluation system. The evaluation system, the necessary observational data, and the entire set of MiKlip prediction results conform to the CMIP5 data standards (Taylor et al. 2012) and reside on a dedicated data server. The MiKlip server makes the prediction results and evaluation system immediately accessible to the entire MiKlip community, thereby providing a crucial interface between production on the one hand and research and evaluation on the other hand. [...] The central evaluation system is constantly expanded with contributions from the MiKlip evaluation module and, together with its reference data pool for verification, resides on the same data server as the entire MiKlip prediction output. The analyses are collected into a database ensuring reproducibility and transparency. Providing the central evaluation system to the entire MiKlip project is also an effective training tool, especially for those researchers who have only recently joined the rapidly expanding field of decadal prediction.'*

Target Group

Freva is a research software environment, hosting verification routines and observational, reanalysis, and model data in customized central evaluation systems of research groups like described in the MiKlip project. The potential user of Freva can be an institute, university, research center, project (like MiKlip), or simply an individual scientist. To address potential user classes with one term, we call it research group hereafter. Freva gives full control of the scientific tool development and improves science through efficient tool application, distinct data access, and integration into a central system. This combination requires a fluent interplay and user guides - which will be in addition to this paper. Freva as a framework is designed for three different user groups who will be addressed in this study and in their

[3] Developed at the Freie Universität Berlin, Freva's naming is a wordplay of a free software and the German name 'Freie', which means 'free and independent'.

[4] https://www-miklip.dkrz.de

individual user guides. All three groups are scientists in the field of Earth system modeling. First, there are the *users* of the research group's evaluation system that look for help in the basic user guide (BUG). The second group are plugin *developers* who fill Freva with scientific applications and retrieve documentation by using the basic developer guide (BDG). Of course, the *developers* are *users* as well. Last but not least, the *admins* of the research group host the Freva instance as a scientific infrastructure for *users* and *developers*. The *admins* may resort to the basic admin guide (BAG).

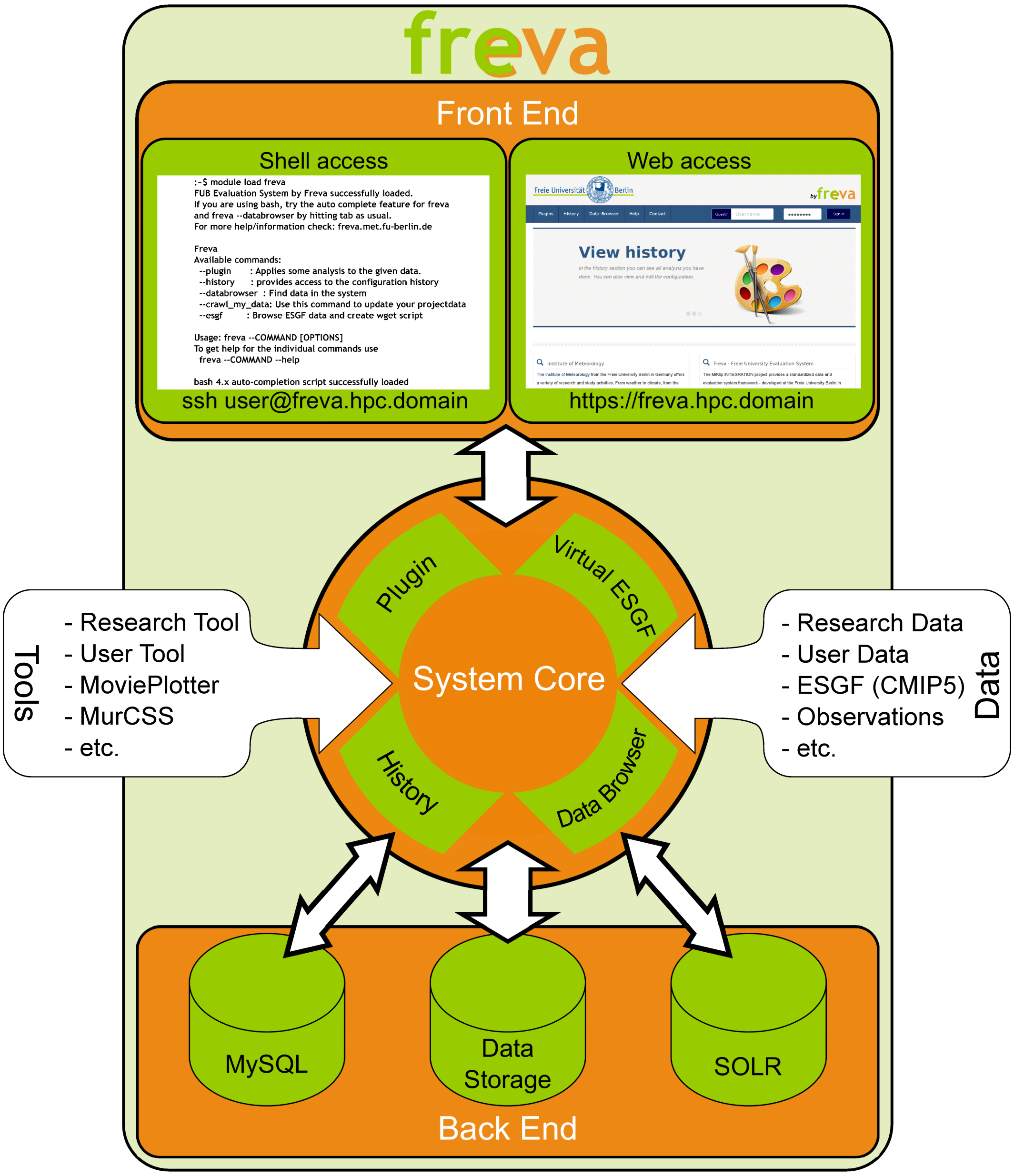
Research Agenda

In this study, we present the system design of Freva, its main features, and its combination of different software technologies (Fig. 2.1). Freva is a combination of a well-defined software plugin management, Earth system model data retrieval, and a backup of all analyses within a portal including a web and a shell front end on a high performance computer (HPC). The system offers a balance between usability and flexibility but being presupposed by transparency and reproducibility (Sect. 2.2). The main use and features of Freva offer a single program solution (Sect. 2.3). We then discuss the advantages of a hybrid evaluation system making use of big data HPCs in climate science and Earth system modeling (Sect. 2.4).

As a picture is worth a thousand words, hands on a software is way more intuitive, then reading about it in a paper. Readers are invited to go to *freva.met.fu-berlin.de*, click on 'Guest?', login, and compare the following sections with the live evaluation system while getting inside views.

2.2 Framework System Design - General Concept

Freva is an evaluation system framework for scientific validation data and software, and it runs as a hybrid system in the web and shell (Fig. 2.1). In this section the concept is explained addressing the general purpose of the system. Freva's integrated front ends fulfill an optimum usage and well-defined interaction between the users and the evaluation system (Sect. 2.2.1). The System Core of Freva consists of software components, the wrapping of the plugin interface, the history database, and the model data browser (Sect. 2.2.2). The combination of different open source technologies into the main framework allows the evaluation system to be generated by one software solution - details of the general software lineup of Freva can be found in the Appendix (Sect. 2.5).



```
:~$ module load freva
FUB Evaluation System by Freva successfully loaded.
If you are using bash, try the auto complete feature for freva
and freva --databrowser by hitting tab as usual.
For more help/information check: freva.met.fu-berlin.de

Freva
Available commands:
  --plugin       : Applies some analysis to the given data.
  --history      : provides access to the configuration history
  --databrowser  : Find data in the system
  --crawl_my_data: Use this command to update your projectdata
  --esgf         : Browse ESGF data and create wget script

Usage: freva --COMMAND [OPTIONS]
To get help for the individual commands use
  freva --COMMAND --help

bash 4.x auto-completion script successfully loaded
```

In the history section you can see all analysis you have done. You can also view and edit the configuration.

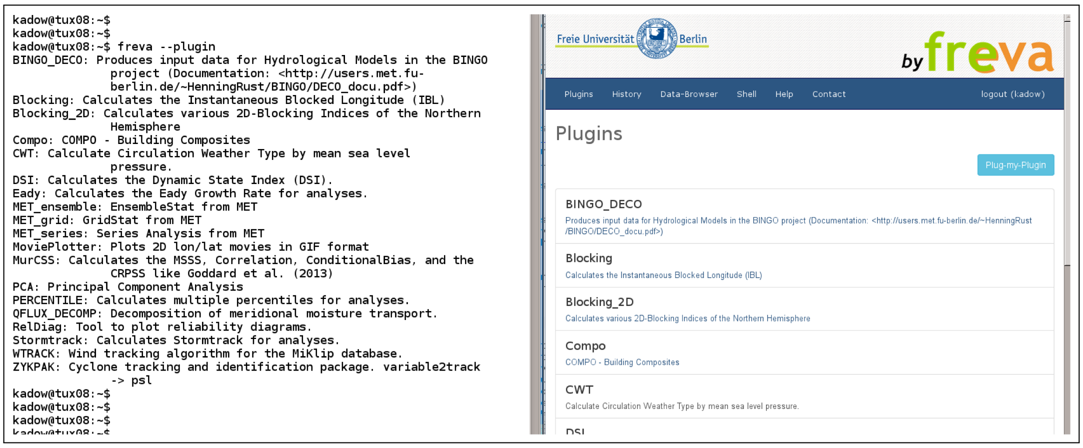
The Institute of Meteorology from the Freie University Berlin in Germany offers a variety of research and study activities. From weather to climate, from the

The MiKlip INTEGRATION project provides a standardized data and evaluation system framework - developed at the Freie University Berlin in

Fig. 2.1: Freva - The Free Evaluation System Framework and its design combining several following technologies into one common software solution. The System Core contains the plugin API handling tools, the history saving configurations, the data-browser where to find data. The scheme represents the basic structure of this study including its subsections.

2.2.1 Front ends of Freva - Usability and Flexibility

The front ends of Freva give users and plugin developers access to the resources of the System Core and the back end databases. Both web and shell front ends connect the scientists with the application system as they represent the interface of the core commands *plugin* (Sect. 2.2.2), *history* (Sect. 2.2.2), and *databrowser* (Sect. 2.2.2). The two interfaces connect the scientists with the application system. The scientists can decide, which degree of freedom they like in using the shell and web by starting,



```
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$ freva --plugin
BINGO_DECO: Produces input data for Hydrological Models in the BINGO
              project (Documentation: <http://users.met.fu-
              berlin.de/~HenningRust/BINGO/DECO_docu.pdf>)
Blocking: Calculates the Instantaneous Blocked Longitude (IBL)
Blocking_2D: Calculates various 2D-Blocking Indices of the Northern
              Hemisphere
Compo: COMPO - Building Composites
CWT: Calculate Circulation Weather Type by mean sea level
              pressure.
DSI: Calculates the Dynamic State Index (DSI).
Eady: Calculates the Eady Growth Rate for analyses.
MET_ensemble: EnsembleStat from MET
MET_grid: GridStat from MET
MET_series: Series Analysis from MET
MoviePlotter: Plots 2D lon/lat movies in GIF format
MurCSS: Calculates the MSSS, Correlation, ConditionalBias, and the
              CRPSS like Goddard et al. (2013)
PCA: Principal Component Analysis
PERCENTILE: Calculates multiple percentiles for analyses.
QFLUX_DECOMP: Decomposition of meridional moisture transport.
RelDiag: Tool to plot reliability diagrams.
Stormtrack: Calculates Stormtrack for analyses.
WTRACK: Wind tracking algorithm for the MiKlip database.
ZYKPAK: Cyclone tracking and identification package. variable2track
              -> psl
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
```

Plugins

BINGO_DECO
Produces input data for Hydrological Models in the BINGO project (Documentation: <http://users.met.fu-berlin.de/~HenningRust /BINGO/DECO_docu.pdf>)

Blocking
Calculates the Instantaneous Blocked Longitude (IBL)

Blocking_2D
Calculates various 2D-Blocking Indices of the Northern Hemisphere

Compo
COMPO - Building Composites

CWT
Calculate Circulation Weather Type by mean sea level pressure.

DSI

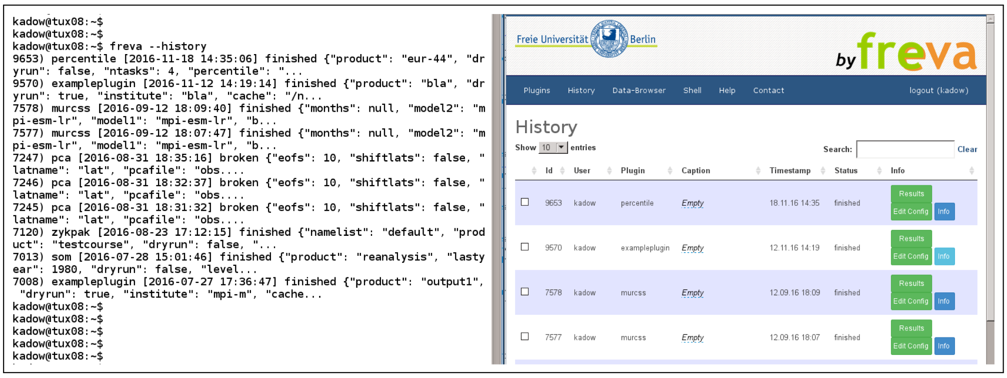
Fig. 2.2: The plugin list in the shell and web interface (snapshot).

adjusting, and operationalizing evaluation procedures as described in the following.

Shell Interface

The shell interface is the most useful when accessing an HPC environment in climate science. The command-line approach allows the development of adjustable Unix-based routines. It grants fast and flexible data access using efficient climate data processing tools. The opportunity to write code within the software applications running within Freva for example with regular expression and basic bash commands improves software and data handling. In that way Freva can for example be started and monitored regularly by Cron jobs. Even big evaluation routines by Freva can be started within Bash loops.

In the following list, we explain the three main features (see Sect. 2.2.2 for details) of the shell interface applying Freva's core-commands:
The **--plugin** (Fig. 2.2) section holds all plugged-in tools and helps the user to start one. When the user forgets a mandatory option of a --*plugin*, Freva gives the name of the name of the missing option. When the user mistypes an option of a plugin, Freva suggests the right one (see also Fig. 2.6
The **--history** (Fig. 2.3) command gives direct access to all analysis and their result directories. Distinct IDs are utilized to sort all results and show their respective history entries. Furthermore, the history holds all configurations and starting commands, which are editable and restartable.
The **--databrowser** (Fig. 2.4) interface efficiently searches the model database. The integrated bash-completion automatically fills the data browser search facets by simply tabbing, thus leading the user easily to the needed dataset or given overview of the database.



```
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$ freva --history
9653) percentile [2016-11-18 14:35:06] finished {"product": "eur-44", "dr
yrun": false, "ntasks": 4, "percentile": "...
9570) exampleplugin [2016-11-12 14:19:14] finished {"product": "bla", "dr
yrun": true, "institute": "bla", "cache": "/n...
7578) murcss [2016-09-12 18:09:40] finished {"months": null, "model2": "m
pi-esm-lr", "model1": "mpi-esm-lr", "b...
7577) murcss [2016-09-12 18:07:47] finished {"months": null, "model2": "m
pi-esm-lr", "model1": "mpi-esm-lr", "b...
7247) pca [2016-08-31 18:35:16] broken {"eofs": 10, "shiftlats": false, "
latname": "lat", "pcafile": "obs....
7246) pca [2016-08-31 18:32:37] broken {"eofs": 10, "shiftlats": false, "
latname": "lat", "pcafile": "obs....
7245) pca [2016-08-31 18:31:32] broken {"eofs": 10, "shiftlats": false, "
latname": "lat", "pcafile": "obs....
7120) zykpak [2016-08-23 17:12:15] finished {"namelist": "default", "prod
uct": "testcourse", "dryrun": false, "...
7013) som [2016-07-28 15:01:46] finished {"product": "reanalysis", "lasty
ear": 1980, "dryrun": false, "level...
7008) exampleplugin [2016-07-27 17:36:47] finished {"product": "output1",
 "dryrun": true, "institute": "mpi-m", "cache...
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
```

History

Show 10 entries Search: Clear

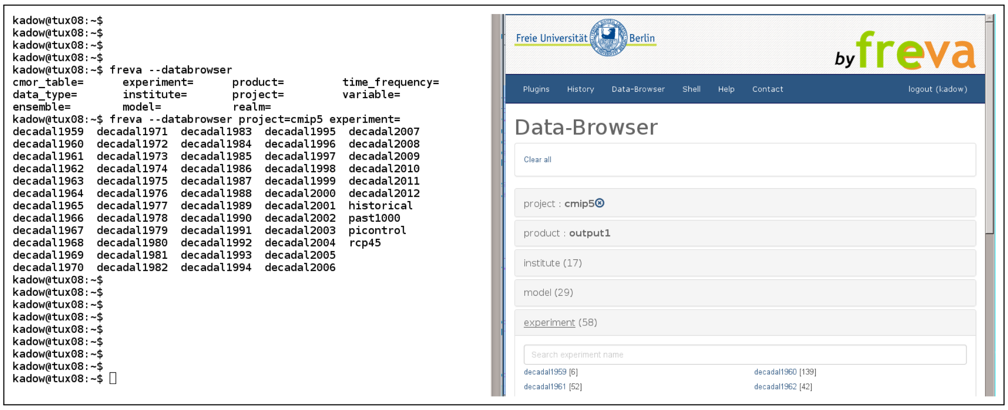
	Id	User	Plugin	Caption	Timestamp	Status	Info
□	9653	kadow	percentile	*Empty*	18.11.16 14:35	finished	Results Edit Config Info
□	9570	kadow	exampleplugin	*Empty*	12.11.16 14:19	finished	Results Edit Config Info
□	7578	kadow	murcss	*Empty*	12.09.16 18:09	finished	Results Edit Config Info
□	7577	kadow	murcss	*Empty*	12.09.16 18:07	finished	Results Edit Config Info

Fig. 2.3: The history in the shell and web interface (snapshot).

Beside these main options, there are assisting side commands only available in the shell:
The **--help** always gives detailed information about Freva, its subcommands, and plugins.
The **--esgf** helps users to download data from the ESGF, establishes a connection to the ESGF and generates the necessary WGET script using the standardized attributes and facets.
The **--crawl_my_data** subcommand offers the opportunity to implement additional standardized datasets. Users can compare their data sets against the ones of the research group, the ESGF projects, or data from other users.

Web Interface

The web interface works similar to the shell interface (Sect. 2.2.1). However, it advances Freva's usability. Usually on HPC environments there is no comfortable way to find or process data and even view results. The web interface introduces easy entrance points for beginners and experts. The three main features (see Sect. 2.2.2 for details) stay the same - plugin, history, databrowser. In the following the advantages of these three features in the web interface are explained.
The **Plugin** section (Fig. 2.2) gives access to plugins, an overview of their options, and assists the user during the individual starting procedure with pre-filled facets. When a user forgets to set a mandatory option, the web interface points to the missing plugin option. There are two ways of accessing the HPCs database. It is possible to point to a specific file by browsing the user's main directories of the user or project, or even use the databrowser to search for some file to analyze. Plugins can apply the more advanced CMOR options to search the whole database of the research project or a virtual ESGF project - option by option (e.g. project, experiment, variable, etc.). This built-in databrowser search is increasing efficiency by decreasing



```
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$ freva --databrowser
cmor_table=       experiment=       product=          time_frequency=
data_type=        institute=        project=          variable=
ensemble=         model=            realm=
kadow@tux08:~$ freva --databrowser project=cmip5 experiment=
decadal1959  decadal1971  decadal1983  decadal1995  decadal2007
decadal1960  decadal1972  decadal1984  decadal1996  decadal2008
decadal1961  decadal1973  decadal1985  decadal1997  decadal2009
decadal1962  decadal1974  decadal1986  decadal1998  decadal2010
decadal1963  decadal1975  decadal1987  decadal1999  decadal2011
decadal1964  decadal1976  decadal1988  decadal2000  decadal2012
decadal1965  decadal1977  decadal1989  decadal2001  historical
decadal1966  decadal1978  decadal1990  decadal2002  past1000
decadal1967  decadal1979  decadal1991  decadal2003  picontrol
decadal1968  decadal1980  decadal1992  decadal2004  rcp45
decadal1969  decadal1981  decadal1993  decadal2005
decadal1970  decadal1982  decadal1994  decadal2006
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$
kadow@tux08:~$ ▯
```

Data-Browser

project : **cmip5**

product : **output1**

institute (17)

model (29)

experiment (58)

decadal1959 [6]
decadal1960 [139]
decadal1961 [52]
decadal1962 [42]

Fig. 2.4: The databrowser in the shell and web interface (snapshot).

the number of CMOR facets with every selection made and only showing remaining possible combinations.

The **History** (Fig. 2.3) shows the completed, scheduled, or running evaluations. All configurations, including the GIT (Hamano, Torvalds, et al., 2015) versioning information, can be retrieved. It is also possible to restart a finished evaluation (*Edit Configuration*). To organize their results, the user is allowed to set a *caption* or delete them from the history section. The *Search* bar allows to search within the configurations started with Freva and filter for used options and e.g. CMOR options.

The **Data-Browser** (Fig. 2.4) gives a convenient way of finding data in the database of the research group. By just clicking through the given standardized (DRS, CMOR, CORDEX, ANA4MIPS, etc. - see ESGF[5]) facets, the user finds data sets and data directories. The web front end provides even more meta information of the search facets, like *variable*, *model*, or *institute*, to explain the meaning of the abbreviations and help to find the right data sets or see what is available. Furthermore, the web part allows to stream the meta data of a specific file by starting ncdump from the NetCDF package.

Besides the main options, there are some extras on the web:

The **Help** section hosts information about the evaluation system build with Freva. A web tour explains the usage of the web page. The scientists find documentation of the research project and developed plugins. Guidelines are also available in the Help section.

The **Shell** section within the web interface also allows the command-line access to the high performance computer of the research group. Applying the shell-in-a-box enables the users to directly start Freva from the bash through the web.

[5] https://esgf.llnl.gov

2.2.2 System Core including Backend

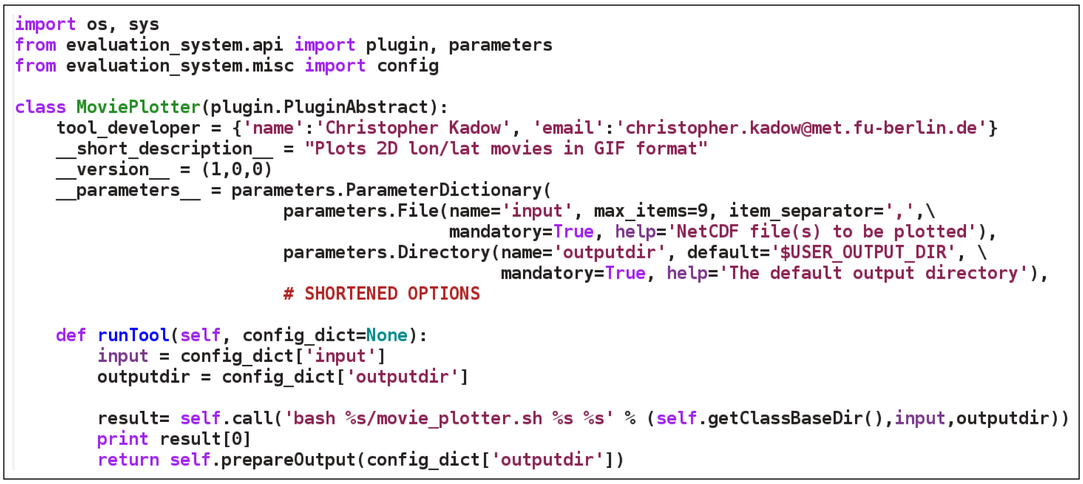
The System Core is the main part of every evaluation system build with Freva (Fig. 2.1). It is an efficient combination of the following technologies and their communication before, during, and after the analysis of the evaluation system. Its plugin interface manages the incorporation of software tools and their common application in the front end (See *Plugin - Application Programming Interface* Sect. 2.2.2). All configurations and information of the executed plugins and analyzed data sets are saved to satisfy the commitment to transparency and reproducibility (see *History - Transparency and Reproducibility* Sect. 2.2.2). In order to keep track and to overview the database, Freva can implement standardized interfaces to model, reanalysis, and observational data sets or even data incorporated by the users (see *Databrowser - Standardized Model Data Access* Sect. 2.2.2).

Plugin - Application Programming Interface

The expertise of scientific evaluation in Earth system modeling usually resides with experts of the field. These experts also take care of the translation of their research field into scientific software. Not every scientist is also an expert in software development. Freva serves as a development interface to assist scientists to fulfill the code of best practice in terms of developing scientific software. The next paragraph will give some insight in the technical details.

The plugin framework of Freva handles the connectivity of stand-alone tools to the evaluation system of the research group through an application programming interface (API). The plugin API, written in Python, is well structured to assist tool developers during the process of plugging-in a tool. Every tool gets an *api.py* wrapper to realize the exchange of options between the Freva system and the plugin. The API transmits all necessary options to Freva and to the tool. The following minimum code requirements guide the plugin developer to structure the tool by providing meta information of the plugin.

A simple implementation of a plugin is shown in Figure 2.5 with an example of the MoviePlotter plugin. The class is derived from the *PluginAbstract* base class and implements some mandatory meta information like *tool_developer*, *short_description*, *long_description*, and the plugin *version*. The *parameters* section automatically collects the tool options by *name* and the corresponding *default*, *mandatory* and *help* information by *ParameterType* and defines the plugin interface to the user. The arguments get parsed from the plugin, retrieving not only the options set by the user but also the default values if parameters are unset. The plugin transforms the



```
import os, sys
from evaluation_system.api import plugin, parameters
from evaluation_system.misc import config

class MoviePlotter(plugin.PluginAbstract):
    tool_developer = {'name':'Christopher Kadow', 'email':'christopher.kadow@met.fu-berlin.de'}
    __short_description__ = "Plots 2D lon/lat movies in GIF format"
    __version__ = (1,0,0)
    __parameters__ = parameters.ParameterDictionary(
                          parameters.File(name='input', max_items=9, item_separator=',',\
                                          mandatory=True, help='NetCDF file(s) to be plotted'),
                          parameters.Directory(name='outputdir', default='$USER_OUTPUT_DIR', \
                                               mandatory=True, help='The default output directory'),
                          # SHORTENED OPTIONS

    def runTool(self, config_dict=None):
        input = config_dict['input']
        outputdir = config_dict['outputdir']

        result= self.call('bash %s/movie_plotter.sh %s %s' % (self.getClassBaseDir(),input,outputdir))
        print result[0]
        return self.prepareOutput(config_dict['outputdir'])
```

Fig. 2.5: The basic plugin.py as example of the MoviePlotter plugin, with condensed option list for display reasons.

incoming strings into Freva options, and the parameter classes validate them by type (e.g. string, integer, bool). Next to these ordinary *string*, *integer* or *bool* fields, the *data-browser fields* in the plugin API communicates with Freva's Solr server (see Sec. 2.2.2) and can be interpreted by the web interface. The plugin API offers some system variables set up by the admin in the configuration of Freva. System variables are for example the default user output directory, plot directory, or cache directory, which can be used for a clear organization of the plugin results.

Software developments need flexibility without interferences between the groups - users want to use plugins; developers want to design or re-design plugins. The publicly available plugins are defined in the main configuration file of Freva, and the actual loading is handled by the *PluginManager*. The *PluginManager* controls the upload into the evaluation system and gives access to the plugins as a central registration. Freva offers developers the possibility to connect their new plugins or temporarily redirect the link to the plugin used by Freva to their own version - independently of the main systems plugins. The overwritten plugin is only applicable by the developer. The system tells the user which version, i.e. the one from the main system or their own linked version is used when the plugin is started. This is especially useful during development stages because developers can test new features or completely new software without disturbing the production system. The *PluginManager* is parsing the incoming command and generates a configuration as *configDict* each time a plugin is started. The *PluginManager* is able to start the plugged-in tool using the *runTool* interactively in shell or via the available *batch mode*.

History - Transparency and Reproducibility

Transparency and reproducibility are important qualities in science. For a scientist it is significant work to take care of the traceability of his research. In that sense, Freva also serves as research recording clerk. The scientific development stages are recorded, easily reviewable, and restartable. The next paragraph will give some insight in the technical details.

All information about performed analyses with Freva is saved in a MySQL database. When a plugin is started, the System Core sets certain information through the *PluginManager*. Each evaluation receives a unique identification number (ID), which is then combined with the user's ID, the plugin name, a time stamp, and status. The configuration parameters of the plugin, including possible data retrieval options (e.g., *Solr fields*), are stored in MySQL. Furthermore, Freva is saving all GIT versioning information, including repository directory and internal version number of the plugin and the Freva version itself, for each analysis. Thus, Freva is flexible enough to guarantee a full recovery of the whole system or just one particular plugin whenever it may be necessary to reproduce old evaluations. In most cases, it is not necessary to set back the system or plugin. Usually it is enough to browse the history of the respective experiment, retrieve the plugin command via shell or web, and rerun the plugin possibly after slight modification, e.g. outputdir, time ranges, etc. To provide a better overview to the user and help them find old configurations and results, they have the opportunity to entitle each analysis with a *caption*. The history also contains the plugin's interactive standard output. The history class of the System Core establishes several statuses, permissions, and result types of each analysis, which can be retrieved by the front ends (Sect. 2.2.1).

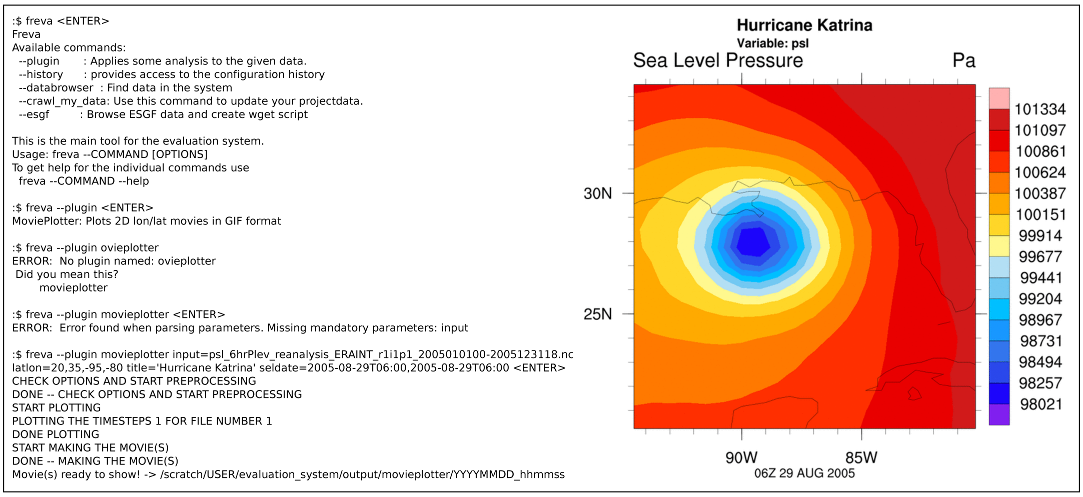
The history-database in MySQL gets monitored for all evaluations done by Freva. Admins of the research group evaluation system have the possibility to view these. Freva saves the status of the started plugins, for example, *finished* or *broken*. This is an advantage over stand-alone tools and decentralized usage. Because, this monitoring helps to reveal data discrepancies and software bugs, as users are not always reporting problems. Freva helps to inform so that users can adjust their *broken* analysis and inform them about how to proceed. If users keep utilizing the system and do not step away after some failed attempts the evaluation system and the research around it improves.

Databrowser - Standardized Model Data Access

The data browser of Freva is more than a search engine. It is a joint commitment to a common Earth system model data output standard within in a research group. It was a step change in development when the climate communities first agreed on a specific data structure for model intercomparison projects (MIPs). As a consequence, nowadays, there are many opportunities to evaluate different models e.g. by the same tool. This meant, that software no longer needed to be adjusted to the model data to be analyzed. The next paragraph will give some insight in the technical details.

Freva's main data standard is the Data Reference Syntax (DRS) of CMIP5 which is publicly available at the ESGF. The DRS has distinct meta data requirements, including the Climate and Forecast (CF) meta data convention which uses NetCDF and the even more restrictive Climate Model Output Rewriter (CMOR) guidelines to bring meta data information into the directory structure of the model output database. This basic approach of using the CMOR options allows to set up a common and easy to understand model database for a research group. This database can be easily extended at later stage e.g. by model data of upcoming development stages of the research group or even the model data of users. Due to the fact that in the ESGF several data standards exist, Freva even gives the possibility to set up several different databases with different data standards, e.g. obs4mips, ana4mips, CORDEX, etc. at the same time. However, for a distinct plugin development, using these meta data directories as options to retrieve data sets, it is recommended to use just one standard or at least imbedded standards like DRS or CMOR. Therefore, Freva also ships with some example scripts to standardize and re-standardize datasets. These scripts also help users to bring their own model results into the required standard format to ultimately incorporate the data into the system.

Freva indexes these output directory structures (model, reanalyzes, observations, etc.) of the research group and saves this meta data information in a Solr database. Solr has a faceting component which is part of the standard request handler which allows a faceted navigation. Therefore, Freva applies the Solr faceted search on the data directories and datasets using, for example, the DRS. All files of a chosen directory get registered or 'crawled', and thereafter all model datasets and their locations get ingested into the Solr server. The stand-alone Solr server is started via Java (see Sect. 2.5) and allows http requests. The System Core of Freva has a python class called *solr core* to encapsulate these requests to the Solr server. This way Freva retrieves the locations of the ingested model data sets via its meta data. This allows the assignment of the datasets to multiple categories. The scientific developer benefit from these categories to precisely different model data sets and exchange them easily. Plugins can use the databrowser to identify the model data needed for evaluation. The plugin interface of the *System Core* allows developers to clearly define which options in the *Solr fields* will be set by the users and which are pre-set by default values. If the data base contains a versioning of the database like e.g. DRS of CMIP5 does - which is recommended - Freva helps to keep track



```
:$ freva <ENTER>
Freva
Available commands:
  --plugin       : Applies some analysis to the given data.
  --history      : provides access to the configuration history
  --databrowser  : Find data in the system
  --crawl_my_data: Use this command to update your projectdata.
  --esgf         : Browse ESGF data and create wget script

This is the main tool for the evaluation system.
Usage: freva --COMMAND [OPTIONS]
To get help for the individual commands use
  freva --COMMAND --help

:$ freva --plugin <ENTER>
MoviePlotter: Plots 2D lon/lat movies in GIF format

:$ freva --plugin ovieplotter
ERROR:  No plugin named: ovieplotter
 Did you mean this?
        movieplotter

:$ freva --plugin movieplotter <ENTER>
ERROR:  Error found when parsing parameters. Missing mandatory parameters: input

:$ freva --plugin movieplotter input=psl_6hrPlev_reanalysis_ERAINT_r1i1p1_2005010100-2005123118.nc
latlon=20,35,-95,-80 title='Hurricane Katrina' seldate=2005-08-29T06:00,2005-08-29T06:00 <ENTER>
CHECK OPTIONS AND START PREPROCESSING
DONE -- CHECK OPTIONS AND START PREPROCESSING
START PLOTTING
PLOTTING THE TIMESTEPS 1 FOR FILE NUMBER 1
DONE PLOTTING
START MAKING THE MOVIE(S)
DONE -- MAKING THE MOVIE(S)
Movie(s) ready to show! -> /scratch/USER/evaluation_system/output/movieplotter/YYYYMMDD_hhmmss
```

Fig. 2.6: The basic usage of Freva in the shell environment including help, listing of plugins, user mistypings and missing informations, and Freva suggestions to guide the user to the final result (on the right). Applying a plugin named MoviePlotter for a quick view at the sea level pressure in Pascal of ERA-Interim reanalysis around New Orleans (USA) while hurricane Katrina was hitting its coast.

with the newest versions without unnecessary extra options. Per default the data browser lists the latest published data of an updated experiment set, but the search can be extended by all accessible versions. This is especially useful for reproduction of research results.

The Virtual ESGF (see Fig. 2.1) is an add-on to the Databrowser. However it is still under development. Therefore, its description can be found in the Appendix (2.5).

2.3 Scientific Application of Freva

Earth system models are important tools for climate science. While the models underwent major computational development stages in the last decades, verification systems are behind the state-of-the-art technologies. However, evaluation system frameworks for the verification equations can be what Earth system model frameworks are for the primitive equations: A systematic computationally efficient tool to research the climate. We examine the importance of on state-of-the-art evaluation system application and address its scientific development for Earth system modeling with the example application of decadal climate prediction.

Based on the corresponding plugin API in Figure 2.5, a simple sample application of the usage is shown in Figure 2.6. It shows an easy way of plugging a stand-alone tool into Freva. The automatic help during the progress supports its application. The MoviePlotter applies the *parameters.File* in the plugin directing the software to one file: The reanalysis of the mean sea level pressure of the ERA Interim (Dee et al., 2011). The figure shows a quick analysis of Hurricane Katrina in the Gulf of Mexico. The application can be efficiently changed to a different variable, different reanalysis, different time range, etc. But with the basic idea of a very simple application, which only needs one input parameter to be used by other plugins for their plotting procedures, the MoviePlotter is just the first step in the evaluation complexity in decadal climate prediction science.

A more complex approach, using the CMOR facets directly in the plugin, is shown by the MurCSS tool from Illing et al. (2014) for decadal climate prediction research. They include two independent CMOR option parts communicating with the Solr server (see Sec. 2.2.2). Thereafter, it is possible to compare two different model versions (e.g. Pohlmann et al., 2013) or even two different experiment setups (e.g. Kadow et al., 2016) against observations or reanalysis data. The development of this efficient basic validation tool for decadal evaluation in MiKlip (see Sect. 2.1 and Marotzke et al., 2016) framed by Freva, which ensures usability and reproducibility is a huge step forward in climate data verification. The research group may detect improvements in the research field 'decadal prediction' much faster and is able to share this knowledge between scientists. Freva was applied in decadal prediction research for example in the assessment of a future volcano eruption on forecasts (Illing et al., 2018), the development of novel forecast techniques (Kadow et al., 2017), the investigation of the East Asian Monsoon (Huang et al., 2018), the assessment of the initial shock (Kröger et al., 2017), the vertical skill evaluation compared to radiosondes (Pattantyús-Ábrahám et al., 2016), the effect of a wind-stress initialization method (Thoma et al., 2015), the decadal skill due to volcano eruptions (Timmreck et al., 2015), the re-calibration of decadal predictions using observations (Pasternack et al., 2018), and the general research of the development stages in MiKlip (Marotzke et al., 2016) - to name a few. Many plugins[6] with different expertise have been developed and shared by Freva within the MiKlip research group.

Hosting an evaluation system of a research group via Freva rather than using stand-alone tools has even more advantages. It is not only scientific developers that can share knowledge by usable plugins; users can also share configurations or even results. This can be done actively via saving the configuration in the shell or by *share results* in the web with colleagues of the research group. But this could be also done passively using big data approaches by Freva. While filling out the web form of a plugin, Freva automatically scans the history database and looks for similar configurations. Even before the plugin is started, the web interface suggests to use results of previously performed experiments, maybe even by other users. This is possible, because Freva is an open system, and all results are accessible by the entire research group. On the research side, this improves the research group's connectivity

[6] www-miklip.dkrz.de/plugins

and saves time for the users. New ideas can be developed as researches can be more productive. From the HPCs point of view, this saves CPU/h, I/O, disk space, and energy.

Evaluation systems framed by Freva can be found at the Freie Universität Berlin[7] for research and teaching, at the DKRZ for the MiKlip[8] project for decadal climate prediction research and CMIP6[9] project for scientific applications and evaluations done by the ESMVal (Eyring et al., 2016), at the Research Applications Laboratory (RAL) of the National Center for Atmospheric Research (NCAR) for MET tools applications[10], and at the German Weather Service (DWD) for interdisciplinary meteorological analysis and visualizations[11].

2.4 Discussion and Conclusion

This paper introduced a complex and efficient framework for the evaluation of data in the context of Earth system modeling, the Freva system. The simple yet powerful concept of the collective commitment to a common data standard (CMOR) and the applicable provision of knowledge on Earth system model science offers the potential to improve the efficiency of research groups. Freva as a host respects the fact that scientists need their scope for development to detect scientific findings. Freva emphasizes transparency and reproducibility of open science in a research project. Plugged-in tools and experiments are reviewable, editable, and repeatable. Although it is desirable to exclusively use the most efficient programming language as the common language in a project, Freva allows to plugin stand-alone tools in a variety of programming languages. Freva enables the utilization of a multitude of software plugins by acquainting only one common framework. The combination of the easy use with the flexibility of incorporating user specific data sets in agreement with research group's standardization of model data, reanalysis, observations, or even ESGF data, is a huge advantage.

Furthermore, Freva supports research groups in terms of sustainability. The full control of a constructed evaluation system by including user specific data and plugging-in individual interfaces and the group's version control is mandatory for a software system in science. Due to the commitment of a research group to work together in a central system like Freva, there is a need for an efficient and convenient communication. For the growth and quality of the system it is also important to invite and convince scientists to be part of the common framework. Therefore, Freva

[7] freva.met.fu-berlin.de

[8] www-miklip.dkrz.de

[9] cmip-eval.dkrz.de

[10] freva.rap.ucar.edu - behind firewall

[11] mavis.dwd.de - behind firewall

addresses three types of clients: The User, the Developer, and the Admin. All of them are usually scientists with different research aims. Because of the comprehensible web platform, users usually get started right away. Over time the user's requirements getting more and more complex. After that, users sometimes jump to cloning, adapting, and re-plugin of a versioned plugin. Freva is at its most impressive when users become developers, and scientists start to cooperate on scientific tasks. Freva guides scientists over technical hurdles and allows them to concentrate on science itself. Another well known issue in science is the fluctuation of scientists in research groups. A clear infrastructure set up with Freva can help to sustain and pass on the knowledge and keep experience in the research group - even when developing scientists leave the field.

A central system can host Earth system model data for the whole research group. By handing out example standardization scripts to the users allows them to complement missing data, standardize them, and file them in the users' data archive or let the admins archive them in the main data structure of their research project. The better the users understand the interfaces, the better the system becomes.

A major issue of the data structure is the change of standards from time to time, e.g. the progression from CMIP5 to CMIP6. Freva tackles this challenge easily, as it is fully adjustable in terms of data standards, and new standards can be included anytime. However, for one plugin it is difficult to deal with different data standards having different attributes. Setting one common data standard and re-standardizing other data sets according to it is the most efficient way for the plugins. Freva is flexible enough, that the data standard could be also set to a completely independent version of the research group - aside from standards in the ESGF.

An automatic application of the CMOR data standard interface to the output of plugins is the next step in the evolution of evaluation systems. The ability to link or register data results to the database of the user will allow its immediate evaluation by, e.g. statistical approaches. Furthermore, developers are able to connect different plugins with each other to pre- or post-process data - from plugin to plugin. This facilitates a common development of a multitude of basic plugins for a kit of evaluations without the necessity to 're-invent the wheel' every time. In addition, the facilitation of the provision and scientific usage of software and climate data automatically increases the number of scientists working with the data sets and identifying discrepancies.

A publication of a software package like this one is always just a snapshot of what has been developed up to that very moment. The software design may have changed over time but the main system framework idea has remained the same ever since we started the development of Freva in 2011. Clear interfaces in terms of tools and data have been established. A well-structured and stable model database was set up, which is flexible to adapt to the research group's needs. Freva offers automated reproducibility and transparency while increasing the usability of tools by different programming languages in shell and web on a HPC. The share of knowledge can be advanced by developing plugins together and by providing Earth system model data. In addition it is possible to produce, share and discuss results of the evaluation system within the research group. Retrospectively, the MiKlip project and Freva have been mutually beneficial for one another. Many plugins have been developed and shared, and a huge model database has been produced within the MiKlip Central Evaluation System for decadal climate prediction as seen in Section 2.3. The MiKlip project is a perfect example of a nationwide project with a special focus and plenty of scientists jointly working on one HPC. Freva, as a central infrastructure, organized MiKlip's tool development and data retrieval. The efficient interaction between different technologies and the increased efficiency of evaluation frameworks next to modeling frameworks improves the Earth system modeling research.

Code and data availability

Freva is open source and accessible for the Earth system modeling community via GitHub: https://github.com/FREVA-CLINT citable through

Kadow, C. and S. Illing. (2018, August 1). FREVA-CLINT/Freva v1.0-beta (Version v1.0-beta). Zenodo. http://doi.org/10.5281/zenodo.1325148

Acknowledgements

The authors would like to acknowledge funding from the Federal Ministry of Education and Research in Germany (BMBF) through the research program MiKlip II (FKZ: 01LP1519A, 01LP1519B, 01LP1520A) and the CMIP6-DICAD project (FKZ: 01LP1605D), and from CoreLogic SARL Paris. All MiKlip project members need to be acknowledged. The authors thank Estanislao Gonzalez for his valuable contribution for Freva within the first months of development. The authors declare that they have no conflict of interest.

2.5 Appendices to Chapter 2

General Software Lineup - Technical Details

Freva is designed to be implemented on IT platforms like Linux (GNU/Linux Community, 2015) for scientists in a research group including user account, compute resources, and storage. The main framework, including shell executables and the web-interface, is written in Python (Python Software Foundation, 2015) using several third party packages. The whole system including the plugged-in tools are version controlled with GIT (Hamano, Torvalds, et al., 2015). In the shell front end, Freva is meant to be loaded by Modules (Environment Modules Project, 2015) or sourced using preferably Bourne-again shell (BASH, Fox and GNU Project, 2015), thus allowing users to stay in the general work environment of, for example, an HPC. In the web front end, which is built using Django (Django Software Foundation, 2015), the users can log in via existing user accounts. Per default Freva is sourcing all user information via Lightweight Directory Access Protocol (LDAP, Carter, 2003), granting/not granting access via group permissions. Therefore, it is not necessary to build an extra user database.

All communications between the web front end and the HPC are realized via Secure Shell (SSH, OpenSSH project, 2015) using the user account. Started plugins via web are handled by the Freva batch mode using a job scheduler, the Simple Linux Utility for Resource Management (SLURM, Slurm Commercial Support and Development, 2015). The database of all produced results is accounted to the user in a structure that is configurable and reachable from all processing hardware. Only the central databases stay within the central evaluation system, e.g. like the plot preview section for the web page. This add-on keeps the preview graphics for the web small and available for the research project. These previews are produced by ImageMagick's convert (ImageMagick Studio LLC, 2015) command. The results in the preview section are connected to the research results and plugin configurations of the history section and are stored in a database like MySQL (Oracle Corporation, 2015b).

The processed standardized Earth system model data can be found using the faceted search via the indexing Solr (Apache, 2015) server running in Java (Oracle Corporation, 2015a) as described in Section 2.2.2. Due to the fact that the Earth system model community, including the ESGF, mainly uses NetCDF data format, helpful accessory software are NetCDF libraries (University Corporation for Atmospheric Research, 2015), NetCDF Operators (NCO, The NCO project, 2015), and Climate Data Operators (CDO, Max-Planck-Institute for Meteorology, 2015). Thus, ncdump of NetCDF is used to retrieve meta data for the web application. The virtual database of the ESGF is hosted by the Filesystem in Userspace (FUSE, Szeredi, 2015), taking care of the bridge between incoming dataset requests, their download, and virtual database caching.

All software setups are described in one configuration file of Freva, coordinating the combination of necessary programs, ports, and communicators.

Virtual ESGF - Evaluation Data Extension

Nowadays, most of the computational data handling can be done via the internet. Cloud and Grid computing services offer fast IT solutions. However, Earth system modeling is still on the edge of possible or practical ways for scientists. Network processing of aggregated data (like yearly global means) is easily possible but an analysis based on high spatial and temporal resolution data is extremely computationally expensive and time consuming. A long term hosting of several terabytes of external model data as explained in Section 2.2.2 is not a practical solution. A database of a research project is usually increasing with time, e.g. the data amount of CMIP6 is estimated to be 20 times larger than that of CMIP5 (Eyring et al., 2016). Therefore, Freva offers a beta-version of a virtual database especially designed for the integration of ESGF projects into the databrowser.The next paragraph will give some insight in the technical details.

The virtual ESGF maps a project like CMIP5 onto the respective data structure of the research project using (FUSE, Szeredi, 2015) as described in the following. For this purpose, we use Freva's ESGF API which addresses the ESGF via attributes and search facets. A listener script is running on the IT platform waiting for requests. Whenever a user or a plugin of Freva asks to access virtual datasets through the databrowser, only these are downloaded into a temporary cache. This cache is adjustable in a way that, for example, one month old unused data will be deleted automatically. During this time frame, the downloaded data is physically reachable. The virtual ESGF allows flexible adjustments while streaming them into the data browser. It is possible to map an ESGF project from the available standard into the research group's chosen standard. In addition, the research group can manipulate the data via NCO or CDO when known issues of data sets of the ESGF, e.g. wrong missing values need to be fixed.

The increased data resources through the virtual ESGF extend the evaluation possibilities for the research group without a restriction in the usability. The virtual ESGF can map several ESGF projects like CMIP5, CORDEX, obs4MIPs, etc. into Freva. An external dependency - the ESGF itself - is restricting the data accessibility and therefore the stability of Freva which needs to be communicated inside the research group when using this powerful feature. Due to ESGF network availability, we recommend a clear separation of these virtual data sets from the local ones, customizable through the databrowser.

The topic 'virtual datasets' is still work in progress. While the design of the virtual ESGF is already fully developed, the practical implementation suffers from sporadic connectivity gaps to the ESGF.

Related Software Developments

There is a growing need for common scientific infrastructures in the Earth system modeling community. However, several big geoscientific software communities arguing about their preferred programming language and their history of software development. Thus, the attempt to migrate to one common software in a research project can be challenging in practice. Several research groups developed and provided their own software packages during the last decades (e.g. PCMDI metrics package[12], Global Marine[13], RCMES tool[14], ESMVal[15]). In the majority of cases, these packages focus on one specific research topic without aiming to be open for a broader audience. Usually, these software packages are provided as scripts which need to be adapted in the programming language they are written in. While this way of providing tools is very flexible because it is possible to adapt the tool completely to one's own project needs, these scripting formats lack usability. In order to improve usability, a few research centers in the recent years have developed websites which present some pre-calculated research galleries (e.g. the Decadal Predictability Working Group[16]). Even less research centers also provide a dynamic calculation of results depending on the chosen options (e.g. Climate Explorer[17], BirdHouse[18]). Often these sites do not offer the possibility to adapt the tool or to use own software and datasets. With this restriction of flexibility, but interactive production of graphics by the users, it is at least possible for users to produce pre-defined evaluations. Furthermore, these platforms provide no opportunities to build specific portals needed by research groups that want to work within a self-contained environment.

Information

Information about the candidate's and co-authors' work on that paper:

The **candidate** developed the scientific idea of the paper -namely the evaluation system framework Freva; the software design of the shell and web work space including standardized data and tool solutions; the data embedding, software engineering of the framework for high-performance computer as well as climate research software;

[12] http://doi.org/10.5281/zenodo.809463
[13] https://www.nceas.ucsb.edu/globalmarine
[14] https://rcmes.jpl.nasa.gov
[15] https://www.esmvaltool.org/
[16] http://clivar-dpwg.iri.columbia.edu
[17] https://climexp.knmi.nl
[18] https://github.com/bird-house

analyzed and discussed the results; drafted the paper.

The **co-authors** provided software engineering development, re-development, and enhancements of the Freva framework in shell and web; developed climate research software in terms of plugins; helped to improve the text of the paper by numerous comments and took care of the funding.

3

Evaluation of Forecasts by Accuracy and Spread in the MiKlip Decadal Climate Prediction System

Abstract

We present the evaluation of temperature and precipitation forecasts obtained with the MiKlip decadal climate prediction system. These decadal hindcast experiments are verified with respect to the accuracy of the ensemble mean and the ensemble spread as a representative for the forecast uncertainty. The skill assessment follows the verification framework already used by the decadal prediction community, but enhanced with additional evaluation techniques like the logarithmic ensemble spread score. The core of the MiKlip system is the coupled Max Planck Institute Earth System Model. An ensemble of 10 members is initialized annually with ocean and atmosphere reanalyses of the European Centre for Medium-Range Weather Forecasts. For assessing the effect of the initialization, we compare these predictions to uninitialized climate projections with the same model system. Initialization improves the accuracy of temperature and precipitation forecasts in year 1, particularly in the Pacific region. The ensemble spread well represents the forecast uncertainty in lead year 1, except in the tropics. This estimate of prediction skill creates confidence in the respective 2014 forecasts, which depict less precipitation in the tropics and a warming almost everywhere. However, large cooling patterns appear in the Northern Hemisphere, the Pacific South America and the Southern Ocean. Forecasts for 2015 to 2022 show even warmer temperatures than for 2014, especially over the continents. The evaluation of lead years 2 to 9 for temperature shows skill globally with the exception of the eastern Pacific. The ensemble spread can again be used as an estimate of the forecast uncertainty in many regions: It improves over the tropics compared to lead year 1. Due to a reduction of the conditional bias, the decadal predictions of the initialized system gain skill in the accuracy compared to the uninitialized simulations in the lead years 2 to 9. Furthermore, we show that increasing the ensemble size improves the MiKlip decadal climate prediction system for all lead years.

The following chapter consists of the main publication and the supporting information published in the open access journal *Meteorologische Zeitschrift (MetZ)* an international journal of the meteorological societies of Germany, Austria, and Switzerland (Schweizerbart Science Publishers). This paper is part of the *Special issue: Verification and process oriented validation of the MiKlip decadal prediction system*.

Kadow, C., S. Illing, O. Kunst, H. W. Rust, H. Pohlmann, W. A. Müller, and U. Cubasch (2016), Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system, *Meteorologische Zeitschrift*, 25, 631-643, `https://doi.org/10.1127/metz/2015/0639`.

3.1 Introduction

Decadal climate prediction research gains progressively more attention in climate science as well as in society, industry and economy. The research aims to close the gap between short term forecasts and long term projections. Numerical weather predictions focus on an initial value problem in the beginning of a forecast. On the other hand, climate projections as a boundary condition problem examine the long-term development (Meehl et al., 2009) , Mehta2011. In order to accommodate the demand for reliable informations on near-term climate variability on the crucial timescales of a year up to a decade, different national and international initiatives have been launched. The Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al., 2012) offers a platform to approach decadal predictions on a common basis via hindcast experiments in the ’observation’ period from 1960 to 2010.

The ’Mittelfristige Klimaprognosen’ (MiKlip) project (Marotzke et al., 2016), funded by the Federal Ministry of Education and Research in Germany (BMBF), is based on CMIP5 and currently develops a decadal forecast system using the Max Planck Institute Earth System Model (MPI-ESM). With the improvements made through initialization techniques using ocean and atmosphere reanalyses in a coupled initialization (Pohlmann et al., 2013), the MiKlip model version outperforms the CMIP5 complement (Müller et al., 2012), especially in the tropics.

In this study we present the forecasts and the skill assessment of the MiKlip decadal climate prediction system following the verification framework for interannual-to-decadal prediction experiments recommended by Goddard et al. (2013). For this purpose, we employ the decadal evaluation tool ’MurCSS’ (Illing et al., 2014) as part of the MiKlip Central Evaluation System (see Marotzke et al., 2016). We point out the importance of a detailed evaluation by combining initialized decadal climate predictions with their prediction skill using the MiKlip system. In section 2 we present the statistical methods used to evaluate the accuracy and the spread of the ensemble hindcast experiment. We present decadal forecasts and their prediction skill for near surface air temperature and precipitation for lead year 1 and lead years 2 to 9, as well as the improvement due to increased ensemble size in section 3. In section 4, we discuss the combination of predictions and the prediction skill of the MiKlip system.

3.2 Data and Methods

The MiKlip decadal forecasts and hindcasts (Baseline1, see also Pohlmann et al., 2013) used in this study were conducted with the earth system model from the Max-Planck-Institute in the low resolution version (MPI-ESM-LR). It is a coupled atmosphere-ocean system triggered by two different initialization techniques. The ocean component MPI-OM (Jungclaus et al., 2013) with the resolution of 1.5○/L40 was initialized with temperature and salinity anomaly fields from the European Centre for Medium-Range Weather Forecasts (ECMWF) ocean reanalysis system 4 (ORAS4 - Balmaseda et al., 2013). The atmospheric component ECHAM6 (Stevens et al., 2013) with the resolution of T63L47 was obtained by a full-field initialization with ECMWF atmosphere reanalyses, including fields of temperature, vorticity, divergence, and surface pressure (ERA40 in 1960-1989 and ERA-Interim in 1990-2013, Uppala et al. (2005) and Dee et al. (2011) respectively). The simulations were started annually for the period 1961 to 2013, each initialization simulating a decade and consisting of 10 ensemble members.

Uninitialized runs with the same model configuration and in the same time period serve as references (Goddard et al., 2013; Matei et al., 2012), disclosing the effect of the initialization and its potential gain of skill. The uninitialized simulations equate to the 'historical' experiment performed during CMIP5 using observed external forcings. Due to the fact that the 'historical' experiment ends in 2005, the reference run was extended by the CMIP5 'rcp45' experiment consisting of the projected RCP4.5 scenario (Taylor et al., 2012). A 10 member experiment of uninitialized runs was conducted to have an equivalent ensemble size to the initialized runs.

We compare near surface air temperature to the HadCRUT3v (Brohan et al., 2006) dataset from the Hadley Centre and Climatic Research Unit for the period 1961 to 2012. This commonly used anomaly data set is chosen to maintain the comparability to other decadal prediction studies (Pohlmann et al., 2013; Goddard et al., 2013; Matei et al., 2012). To enable a global comparison of precipitation with observation over land and ocean, a shorter time period was selected, focussing on the era of satellite data. The Global Precipitation Climatology Project Satellite-Gauge (GPCP-SG) dataset (Adler et al., 2003) was used for the period from 1979 to 2012. However, for full comparability with the decadal prediction community and the evaluation over the longer timescale, we also present the evaluation of precipitation with the Global Precipitation Climatology Centre (GPCC) Full Data Reanalysis Version 6 dataset (Schneider et al., 2011; Becker et al., 2013) over land in the supplementary material of this publication. For both evaluated variables, anomalies are considered for comparison with the model data to ensure that no general bias is influencing the results like differences in the height of model and observation.

The anomaly real-time forecasts for temperature and precipitation are available for the year 2014 and the time period of the years 2015 to 2022. The reference period is 1981 to 2010. The uninitialized simulations are used as reference datasets for the anomaly calculations.

The following skill assessment – based on the decadal climate prediction verification framework (Goddard et al., 2013) – includes spatial averaging on a 5x5 degree grid, temporal aggregation and lead-time dependent bias adjustment in a cross validated manner (ICPO, 2011). The lead year 1 hindcast continues the observed initial conditions in the first prediction year. For the lead years 2 to 9, the representation of the decadal-scale climate predictions excludes the skill of lead year 1. Significance of the verification scores was estimated using a non-parametric bootstrap approach (Wilks, 2011; Mason and Mimmack, 1992) taking auto-correlation into account (Goddard et al., 2013). First, we investigate the gain of accuracy in the ensemble mean due to the initial conditions compared to uninitialized climate change projections. In a second step, we analyze whether the ensemble spread is an appropriate representation of the forecast uncertainty on average.

3.2.1 Accuracy of the ensemble mean

The mean squared error skill score (MSESS) compares the accuracy of two predictions (Murphy, 1988) of the past, so called hindcasts. The initialized hindcasts H_{ij} consist of their ensemble members i=1,...,m and the start times j=1,...,n. The mean squared error (MSE) between the hindcast ensemble mean H_j and the observations O_j over j=1,...,n start times can be expressed as

$$\mathrm{MSE}_H = \frac{1}{n}\sum_{j=1}^{n}(H_j - O_j)^2. \tag{3.1}$$

Compared to some reference prediction, such as the climatological forecast $MSE_{\bar{O}} = \frac{1}{n}\sum_{j=1}^{n}(\bar{O} - O_j)^2$, the skill can be determined by the

$$\mathrm{MSESS}(H,\bar{O},O) = 1 - \frac{\mathrm{MSE}_H}{\mathrm{MSE}_{\bar{O}}}. \tag{3.2}$$

Applying the Murphy-Epstein decomposition and using anomalies, the MSESS for the climatological forecast can be written as:

$$\mathrm{MSESS}(H,\bar{O},O) = r_{HO}^2 - \left[r_{HO} - \frac{s_H}{s_O}\right]^2 \tag{3.3}$$

with r_{HO} being the sample correlation coefficient between the hindcasts and the observations, and the sample variance of the hindcasts s_H^2 and observations s_O^2 (Murphy, 1988; Murphy and Epstein, 1989). This decomposition allows to differentiate between the correlation coefficient and the conditional prediction bias (second term on the right hand side of Eq. 3.3). When comparing the initialized hindcasts H with the uninitialized reference R, the MSESS can be written as

$$\mathrm{MSESS}(H,R,O) = \frac{\mathrm{MSESS}_H - \mathrm{MSESS}_R}{1 - \mathrm{MSESS}_R} \tag{3.4}$$

to assess the change of skill from the uninitialized to the initialized prediction system.

The MSESS represents the improvement in the accuracy of the hindcasts H over the climatology $\bar{O}$ or a reference forecast R with respect to the observations O, where - $\infty < \mathrm{MSESS} \leq 1$. A positive value suggests an improved accuracy of the hindcast ensemble mean compared to the reference, and a negative value indicates the opposite.

The correlation coefficient $-1 \leq \ r \leq 1$ as the potential skill of a prediction system represents the linear relationship between a hindcast and the observation. For assessing the change in the correlation coefficient of the hindcast against a reference prediction, the difference of r_{HO} and r_{RO} is presented, with values ranging from -2 to 2.

The conditional bias $-\infty < r_{HO} - \frac{s_H}{s_O} < \infty$ is the difference of the correlation and the ratio of standard deviation from a prediction and observation - it is zero at its best. The gain of the conditional bias against a reference prediction is calculated by subtracting the absolute values $|r_{RO} - \frac{s_R}{s_O}| - |r_{HO} - \frac{s_H}{s_O}|$. Positive values represent a decrease of bias or, in the sense of the MSESS, a gain of skill and vice versa.

3.2.2 Ensemble spread as forecast uncertainty

The spread of an ensemble forecast (ensemble variance) is meant to be an estimate of the forecast uncertainty due to uncertainty in the initial conditions. If the mean squared deviation of the observations from the ensemble mean (MSE) corresponds to the ensemble variance, the latter is a good estimate of the forecast uncertainty. Is the ensemble variance smaller than the MSE the ensemble is said to be under-dispersive (overconfident); an ensemble variance larger than the MSE indicates an over-dispersive (underconfident) ensemble. This answers the question, if the ensemble spread can be used as reference for the forecast uncertainty. Following Goddard et al. (2013), the ensemble spread is compared to the forecast uncertainty using a particular version of the continuous ranked probability skill score (CRPSS). The CRPSS is based on the continuous ranked probability score (Matheson and Winkler, 1976)

$$\mathrm{CRPS}(H_{ij}, O_j) = \int_{-\infty}^{\infty} (F_{H_j}(y) - \mathcal{H}(y - O_j))^2 dy, \tag{3.5}$$

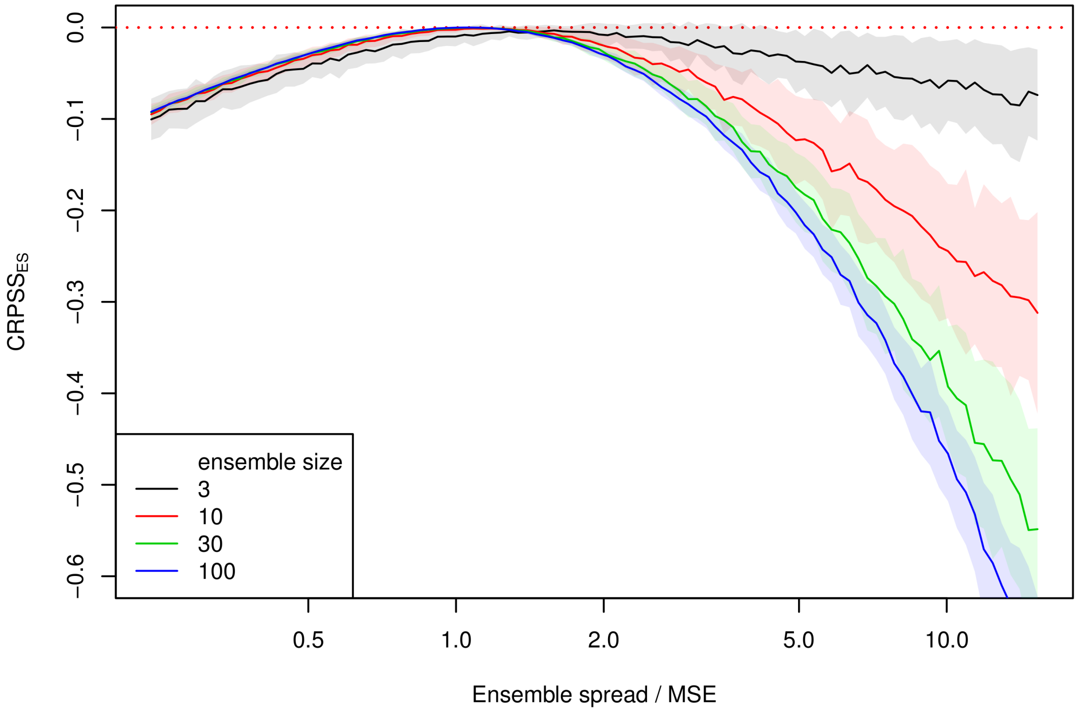


Fig. 3.1: The $\mathrm{CRPSS_{ES}}$ as function of the ratio between ensemble spread ($\overline{\sigma^2_{\hat{H}}}$) and MSE ($\sigma^2_R$) for different ensemble sizes. When the given ratio is one, the $\mathrm{CRPSS_{ES}}$ reaches its maximum value of zero.

which integrates the squared difference between the probability distribution F_{H_j} of the ensemble forecast and the observation for a given instance $j = 1, \dots, n$ in probability space over the predictand y. The Heaviside function $\mathcal{H}(y - O_j)$ is the associate cumulative distribution function for the single observation. Gneiting and Raftery (2007) suggested to use a normal distribution with mean H_j and variance σ^2_H for the forecast probability density function $F_{H_j} = \mathcal{N}(H_j, \sigma^2_{H_j})$. The CRPS can be expressed with the standard normal probability density and cumulative distribution function φ and ϕ, respectively

$$CRPS(\mathcal{N}(H_j, \sigma^2_{H_j}), O_j) = \sigma_{H_j} \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{O_j - H_j}{\sigma_{H_j}}\right) - \frac{O_j - H_j}{\sigma_{H_j}} \left(2\phi\left(\frac{O_j - H_j}{\sigma_{H_j}}\right) - 1 \right) \right]. \quad (3.6)$$

To quantify the ensemble spread against the standard error, we use the average ensemble spread

$$\overline{\sigma^2_{\hat{H}}} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{m-1}\sum_{i=1}^{m}(\hat{H}_{ij}-\hat{H}_j)^2 \tag{3.7}$$

with the ensemble members $\hat{H}_{ij}$ and the ensemble mean $\hat{H}_j$ corrected for mean and conditional bias. The reference prediction has the same mean, but its variance is replaced by the MSE

$$\sigma^2_R = \frac{1}{n-2}\sum_{j=1}^{n}(\hat{H}_j - O_j)^2. \tag{3.8}$$

Using these hindcast and reference distributions in the continuous ranked probability skill score for the assessment of the ensemble spread, the resulting $\mathrm{CRPSS}_{\mathrm{ES}}$ reads

$$\mathrm{CRPSS}_{\mathrm{ES}} = 1 - \frac{\sum_j \mathrm{CRPS}_H(\mathcal{N}(\hat{H}_j, \overline{\sigma^2_{\hat{H}}}), O_j)}{\sum_j \mathrm{CRPS}_R(\mathcal{N}(\hat{H}_j, \sigma^2_R), O_j)}. \tag{3.9}$$

The reference CRPS_R using the MSE represents the forecast uncertainty and thus defines the desired value for the CRPS_H, therefore $\mathrm{CRPSS}_{\mathrm{ES}} \leq 0$. The optimum $\mathrm{CRPSS}_{\mathrm{ES}} = 0$ is attained for $\overline{\sigma^2_{\hat{H}}} = \sigma^2_R$, and $\overline{\sigma^2_{\hat{H}}} \neq \sigma^2_R$ leads to a negative $\mathrm{CRPSS}_{\mathrm{ES}}$. The respective simulation study with varying ensemble size is utilized in Fig. 3.1. This behavior does not allow to determine whether the ensemble spread is over- or underestimating the forecast uncertainty (MSE). To add this missing information, we consider the spread score (see Palmer et al., 2006; Keller et al., 2008), with a log-transform to obtain the logarithmic ensemble spread score

$$\mathrm{LESS} = \ln\left(\frac{\overline{\sigma^2_{\hat{H}}}}{\sigma^2_R}\right). \tag{3.10}$$

The LESS shows negative (positive) values for under-dispersive (over-dispersive) forecasts. A meaningful combination of the $\mathrm{CRPSS}_{\mathrm{ES}}$ and the LESS depicts the skill and the sign of dispersion. This addresses the question whether the ensemble spread is an adequate representation of the forecast uncertainty on average posed by Goddard et al. (2013). In this study, we define a skill score based on the LESS to compare model development stages. Different sized ensembles of the model system can be evaluated with respect to spread development.

$$\mathrm{LESSS} = 1 - \frac{\mathrm{LESS}^2_{\mathrm{pred}}}{\mathrm{LESS}^2_{\mathrm{ref}}} \in (-\infty, 1] \tag{3.11}$$

The LESSS answers the question if the prediction system improves this ratio between the average ensemble spread and the mean squared error compared to the reference prediction.

3.3 Results

3.3.1 Forecasts and skill assessment of temperature

In general, the anomaly forecast of near surface air temperature for the year 2014 with the MiKlip system shows rather warming than cooling signals in the different regions of the world (Fig. 3.2a). However, there are regions with strong negative and positive signals. The North-East Pacific, the western part of North America including Alaska, Central and Southern Africa as well as Russia show distinct hot spots with anomalies over 1.5K. There are cooling patterns as well, mainly over the north-eastern North-America, India and southern China, the Antarctic Circumpolar Current and the northern North Atlantic. The forecast for the eastern Pacific points to a cooling in the ENSO region and positive anomalies in the surrounding. Over Europe, the forecast shows a warming of around 0.75K. The climate forecast for the years 2015 to 2022 predicts a clear warming signal on the Northern Hemisphere from 60°N northwards with values over 1.5K, beside the cooling spot in the northern North-Atlantic (Fig. 3.3a). The forecast shows also a cooling area in the Pacific-Antarctic Basin, e.g. over the Amundsen Sea. All continents show a warming signal of around 1K, as do the equatorial eastern Pacific, the eastern Atlantic, and the western Indian Ocean.

The analysis of the near surface air temperature in lead year 1 indicates an improvement from the uninitialized projections to the initialized hindcasts (Fig. 3.2g,h,i). Combining the effect of increased correlation and reduced conditional bias, the MSESS exhibits significant positive values over the ocean, most likely due to the ocean initialization. The North Pacific in particular benefits from the initialization (Fig. 3.2g). The North Atlantic provides a contrast: while there is at least some improvement in correlation compared to the uninitialized runs (Fig. 3.2h), it is accompanied by a decrease in the conditional bias (Fig. 3.2i). The initialized hindcast experiments (Fig. 3.2) of lead year 1 add confidence to the forecast of surface temperature in Figure 3.2a.

For lead years 2 to 9 (see Fig. 3.3), the initialized and uninitialized experiments perform similarly. Due to catching the long-term trend of the climate system, the correlation coefficients for surface temperature are significantly high. Apart from the ENSO-related tropical Pacific, this is comparable to Goddard et al. (2013) and Müller et al. (2012). Little correlation is lost almost over the whole globe in the initialized

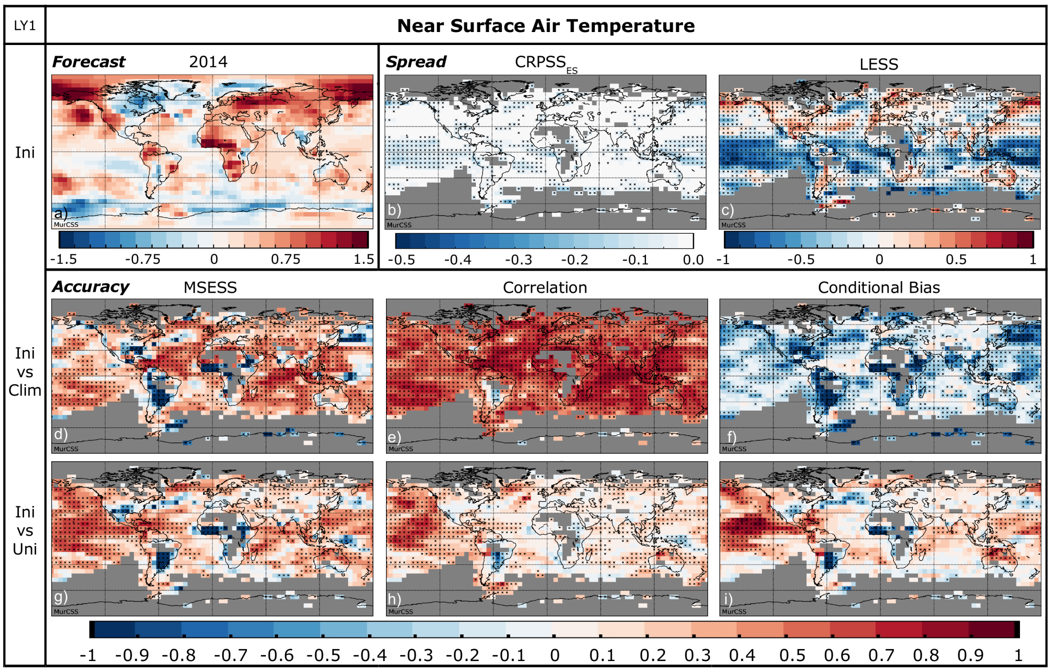


Fig. 3.2: Anomaly forecast of the MiKlip decadal prediction system for near surface air temperature in Kelvin for the year 2014 (a). Anomalies are calculated relative to the years 1981 to 2010 from the uninitialized (historical and rcp45) simulations and interpolated on the 5x5 grid for skill assessment. The evaluation of the ensemble spread is to the right of the forecast with the continous ranked probability skill score of the ensemble spread vs the reference error (CRPSS$_{ES}$ in b) and the logarithmic ensemble spread score (LESS in c). The ensemble mean hindcast skill is shown in the middle and bottom row - mean squared error skill score (MSESS - left column) and its decomposition in correlation (middle column) and conditional bias (right column) of near surface air temperature averaged over the first prediction year against observation from HadCRUT3v over the period 1961-2012. It shows the skill of the initialized decadal experiments against a climatological forecast (middle row) including the MSESS (d), correlation (e) and the conditional bias (f). The lower row uses the uninitialized simulations (historical, extended with rcp45 to year 2012) as the reference prediction in the MSESS (g), the correlation differences (h) and depicting the change in magnitude of the conditional bias (i). Colorbars in the accuracy section are scaled to -1 to 1. Crosses denote values significantly different from zero exceeding at a 5% level applying 1000 bootstraps. Gray areas mark missing values with less than 90% data consistency in the observation.

runs compared to the historical runs (Fig. 3.3h). However, small areas of positive gain in correlation can be found in the North Atlantic (Fig. 3.3h). The conditional bias (Fig. 3.3f) improved in the initialized runs, leading to an overall positive skill (Fig. 3.3i). The MSESS in the initialized runs against uninitialized hindcast for the surface temperature increases significantly in the tropics (Fig. 3.3g). It decreases over areas such as northern Asia and suffers from an increased conditional bias and negative correlation.

The $\mathrm{CRPSS}_{\mathrm{ES}}$ in Fig. 3.2b, 3.3b shows that the ensemble spread can represent forecast uncertainty in various regions. This is not the case in the central Pacific for

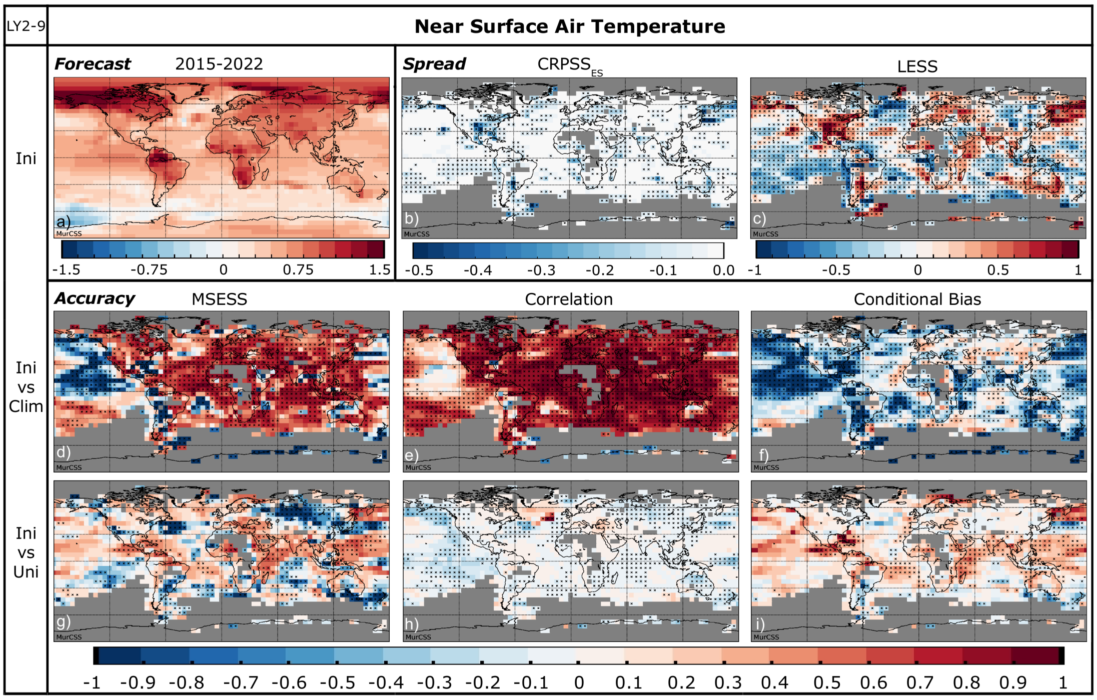


Fig. 3.3: As in Figure 3.2 but for the forecast of 2015-2022 and evaluation of lead years 2 to 9 over the period 1962-2012.

lead year 1 (Fig. 3.2b). The LESS in Figure 3.2c reveals that the spread is too small in the tropics and the Southern Hemisphere; this improves slightly for years 2 to 9 (Fig. 3.3c). Variabilities around the North Atlantic as well as the North Pacific in lead year 1 (Fig. 3.2c) show patterns with over- and under-dispersive spreads next to each other. The ensemble is over-dispersive for North America, the North Atlantic, Europe as well as around the Kuroshio, which means the ensemble spread is too large compared to the reference error (Fig. 3.2c, 3.3c).

The model system used in this study also participates in a multi-model comparison project as accomplished by Smith et al. (2013). However, a different initialization strategy is applied, when comparing the real-time forecasts. The anomaly initialization in the ocean was conducted through a NCEP forced assimilation run, so called MiKlip Baseline0 simulation (see Matei et al., 2012; Müller et al., 2012). The MiKlip system as analyzed in this study (Baseline1) is closer to the multi-model average as shown in Smith et al. (2013) than Baseline0 (not shown). In general a more uniform warming (less regions with cooling) is predicted with Baseline1 compared to Baseline0 on the longer timescales beyond lead year 1.

3.3.2 Forecasts and skill assessment of precipitation

The prediction of precipitation is more challenging, and consequently results are more dispersive than for temperature. The forecasts feature strong anomalies in the

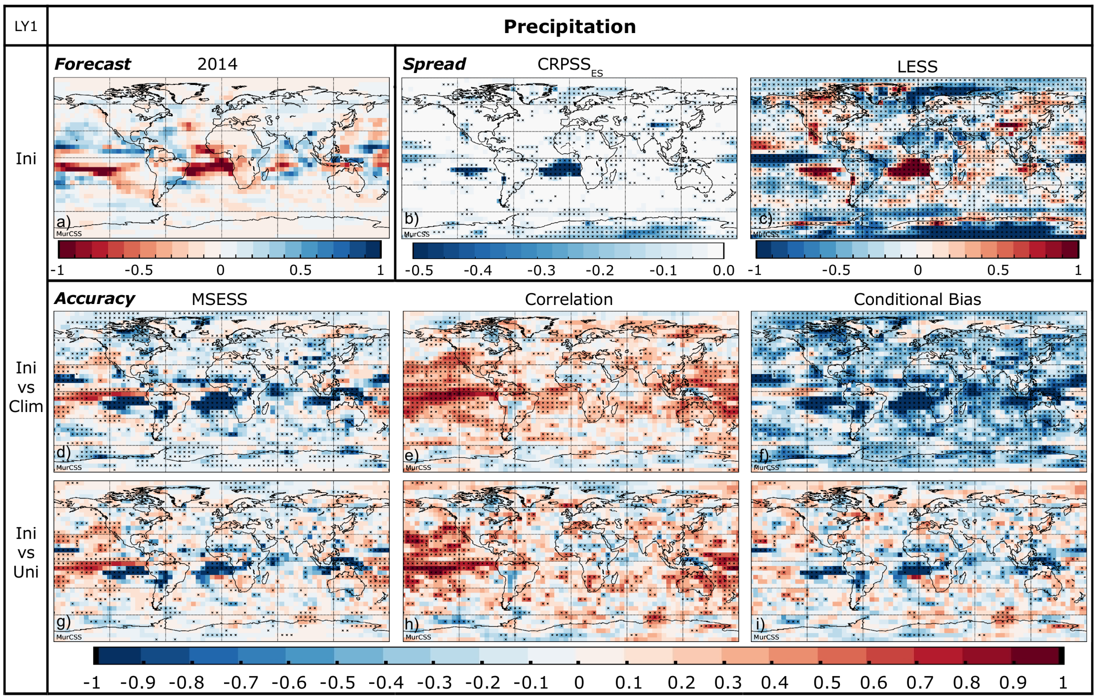


Fig. 3.4: As in Figure 3.2 but for precipitation in mm/day and using the observation from GPCP-SG over the period 1979-2012 for skill assessment.

tropics and over the oceans (Fig. 3.4a, 3.5a). The anomaly forecast in Figure 3.4a shows an increase in precipitation for the year 2014 in the northern West Pacific, East Atlantic and Indian Ocean. Precipitation is decreasing in the southern equatorial Pacific and Atlantic. The forecasts over Africa and northern South America predict an overall drying, while Central America and India show a wetter signal. For the next 2 to 9 years (2015 - 2022) precipitation rates decrease over the northern equatorial Atlantic as well as south of the equator in the Indian Ocean and increase in the tropical Pacific. The latter shows El Niño like structures (Fig. 3.5a). In general, the continents in the northern hemisphere show an increase, whereas the southern continents including Africa rather indicate a decrease.

The evaluation of lead year 1 shows a significant gain in correlation for the initialized over the uninitialized experiment (Fig. 3.4h). Significant positive correlation between the decadal hindcasts and the observations from GPCP-SG (Fig. 3.4e) is present mainly in the tropical Pacific, but can also be detected in the equatorial Atlantic and the Indian ocean. Conditional biases for initialized (Fig. 3.4f) and uninitialized (Fig. 3.4i) simulations are large and negative over the whole globe compared to GPCP-SG. In the tropics in particular, the model has difficulties to reproduce precipitation variability. For the initialized run the performance is worse compared to the climatological forecast. However, the combined MSESS still shows some skill (Fig. 3.4d, g), which can be traced back to the strong improved correlation compared to the uninitialized simulations.

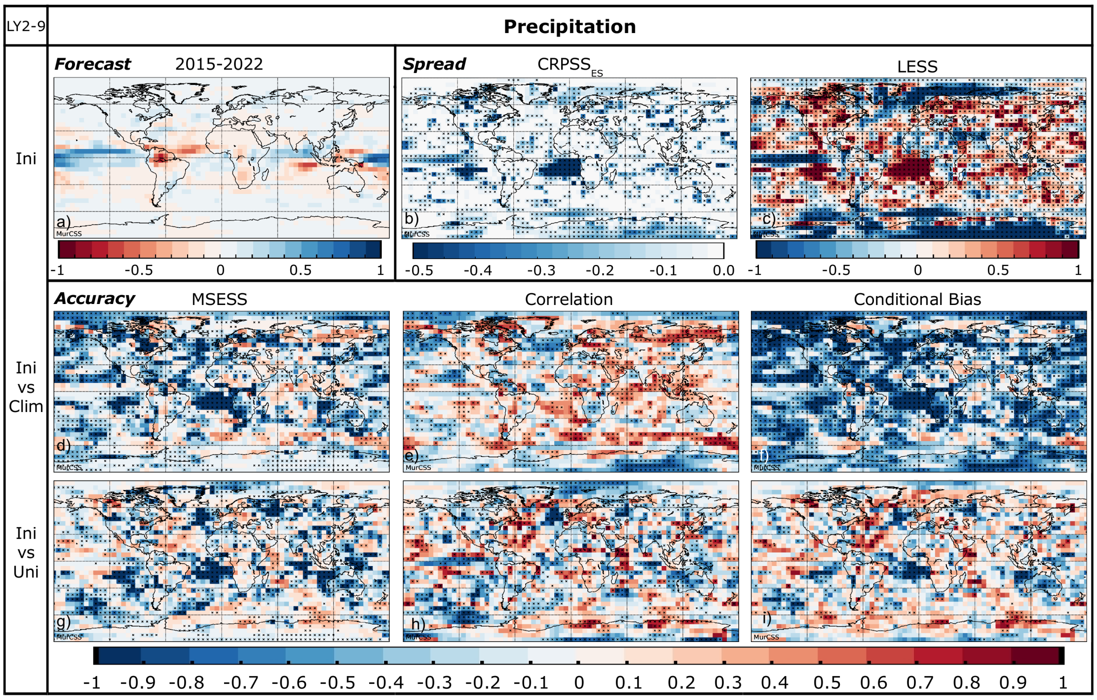


Fig. 3.5: As in Figure 3.4 but for the lead years 2 to 9 over the period 1980-2012.

The various skill scores (Fig. 3.5) become noisy for the lead years 2 to 9. However, we present these results as well—for consistency and comparability with other international studies (Goddard et al., 2013; Smith et al., 2013). Some continental areas like Europe, the Middle East and North-East Asia, as well as the Indian Ocean, show some positive correlation in the decadal hindcasts compared to the climatological forecast (Fig. 3.5e). The decadal hindcasts improve over Europe when compared to the uninitialized simulations (Fig. 3.5h). This comes along with an improved temperature and therefore energy budget over Europe when compared to the uninitialized hindcast for the lead years 2 to 9. This gets more obvious, when the initialized system clearly outperforms the uninitialized system in the detrended temperature analysis of the MSESS and correlation in the leadyears 2 to 9 (Fig. 3.9). This is because annual precipitation is not that trend related (Kumar et al., 2013), especially in Europe (Cubasch and Kadow, 2011) and the North Atlantic is shown to be the source of skill over Europe (Ghosh et al., 2015). But, due to the loss of correlation for precipitation in most of the other regions by contrast with the uninitialized runs and the negative conditional bias in the North Atlantic, as well as the same difficulties as experienced for lead year 1 at the equatorial regions in the conditional bias (Fig. 3.5f, i), the MSESS comparison from initialization runs versus uninitialized simulations (Fig. 3.5d, g) shows almost no skill for precipitation.

For lead year 1 the ensemble spread is an adequate estimate for the forecast uncertainty for most regions (Fig. 3.4b). This is no longer valid for lead years 2 to 9 (Fig. 3.5b), with only some small areas left over the ocean with the spread being close

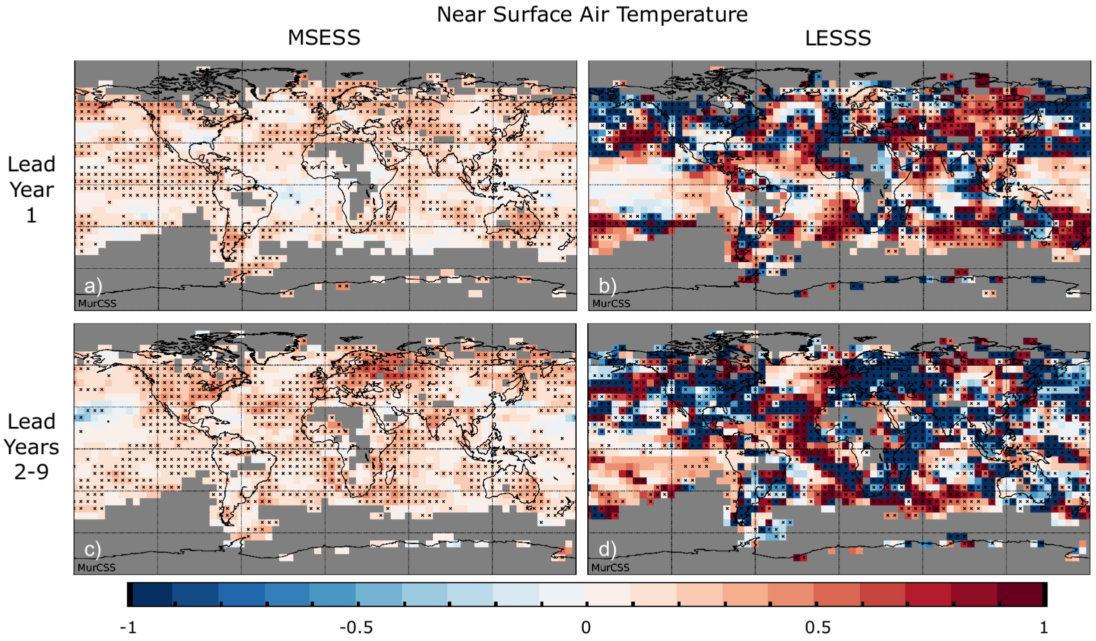


Fig. 3.6: Comparison of the hindcast skill of different sized ensemble model versions (10 member vs 3 member). MSESS and LESSS for near surface air temperature over the period 1961-2012 against HadCRUT3v for the lead year 1 (upper row) and lead years 2 to 9 (lower row). The MSESS shows the improvement made in the hindcast ensemble mean prediction and the LESSS exhibit the improvement in the ensemble spread as an adequate representation of the forecast uncertainty. Crosses denote values significantly different from zero exceeding at a 5% level applying 1000 bootstraps. Gray areas mark missing values with with less than 90% data consistency in the observation.

to the reference error. The $\mathrm{CRPSS}_{\mathrm{ES}}$ highlights the areas in the tropical Pacific and Atlantic showing no skill. The LESS demonstrates the over-dispersion (Fig. 3.4c, 3.5c) in these regions. Here, the precipitation rates suffer from positive temperature biases in the ocean in these areas (not shown), which leads to more convective activity and variability. Furthermore, the LESS reveals that areas of small and large ensemble spreads are next to each other in the central Pacific and Atlantic. This points to problems in the correct representation of small scale processes on these time scales in the spread of the ensemble. Variabilities in convective and large scale precipitation processes in climate models are difficult to represent. The standard error of satellite instruments is also relatively high in regions with little precipitation, especially in the first years of the GPCP-SG dataset (Adler et al., 2003). The short observational period of the satellite observations is problematic too, when analyzing the lead years 2 to 9.

3.3.3 Ensemble Size

The CMIP5 (Taylor et al., 2012) decadal experimental design with initializations every 5 years led to an unreliable skill assessment (Goddard et al., 2013). Since then, most of the prediction systems are initialized annually. The small ensemble size of these experiments is another known issue, particularly for comparing different prediction systems (Smith et al., 2013). Pohlmann et al. (2013) analyze only 3 ensemble members of the MiKlip system in order to have a clean comparison with the results of the 3 available members in the CMIP5 system. Kruschke et al. (2014) use a bias corrected RPSS to compensate for different ensemble sizes. A comprehensive study on the effect of the ensemble size on decadal prediction is given in Sienz et al. (2016). To fill the gap between the MiKlip system analyzed in Pohlmann et al. (2013) and the results shown in this study, we present the change of skill for lead years 1 and 2 to 9 by increasing the ensemble size from 3 to 10 ensemble members.

The MSESS in Figure 3.6 shows a significant gain of prediction skill for surface temperature. Besides the Central Atlantic, the temperature prediction skill for lead year 1 increases for the whole globe - not significant everywhere (Fig. 3.6a). But on the long run, the forecast for the Central Atlantic benefits from the larger ensemble for lead years 2 to 9 (Fig. 3.6c). The LESSS for temperature shown in Figures 3.6b) and 3.6d) improves in the tropics where the $\mathrm{CRPSS_{ES}}$ reveals significant negative skill (Fig. 3.2b, 3.3b) and the LESS (Fig. 3.2c, 3.3c) depicts an under-dispersion. Therefore, the decreasing under-dispersion due to the increased ensemble size leads to a slightly better representation of the uncertainty by the ensemble spread. Precipitation shows an improvement in the MSESS in lead years 1 and 2 to 9 (not shown). The LESSS improves only in local areas in the development of the ensemble spread as an adequate forecast uncertainty in the comparison of the 10 to the 3 ensemble member system for precipitation (not shown).

3.4 Discussion and conclusions

Combining forecasts and detailed evaluation for the MiKlip system for near surface air temperature and precipitation provides a comprehensive assessment of the decadal climate predictions. With a strong impact in lead year 1, initialization techniques improve the prediction system in comparison to an uninitialized system. Both atmospheric parameters benefit from an initialization with an oceanic reanalysis. Mainly the Pacific region temperature forecast improves, which causes an improved convection, triggering precipitation fluxes. The equatorial regions suffer from an under-dispersive ensemble in temperature and an over-dispersion of precipitation in regions of western South America over the Pacific and western Central Africa over the Atlantic. Both variables exhibit a large negative conditional bias in lead year 1. The largest temperature anomalies for year 2014 are forecasted in areas where the performance of the model system is less satisfying, e.g. a warming of 3 Kelvin in West Africa or a cooling of 2 Kelvin in a small region in the North Atlantic. Regions with few data for validation like the southern Pacific can not be reliably evaluated using observational reconstructions.

As the initialized system drifts towards the same state as the uninitialized model, the lead years 2 to 9 produce similarly performances for the initialized and uninitialized experiments. The improvement of the initialized prediction system on these timescales stems from the decreased conditional bias in combination with an increased ensemble size, at least for temperature. The conditional bias exists, when a climate model e.g. over-responds to increasing greenhouse gases (Goddard et al., 2013). This can result in an overestimation of temperature anomalies. In this respect, the initialized MiKlip prediction system performs better in the MSESS than the uninitialized due to matching the climate trend much better. But it is difficult to differentiate between a model drift of the initialized system towards a warmer state of the uninitialized system and a possible predicted warming after the hiatus (Meehl et al., 2011; Kosaka and Xie, 2013). Analyzing a decadal prediction system being between an initial and boundary condition problem leads to several factors for potential skill. The correct initial condition in the beginning of the forecast improves the forecast on the seasonal to the interannual timescale. The memory of the ocean plays a big role on interannual to decadal timescale, when running a coupled model. But the trend due to increased greenhouse gases has even more influence on the long-term development. Analyzing the time range of 2 to 9 years mixes these potentials of skill and the uninitialized system improves on the long run. Therefore the uninitialized can outperform the initialized system in correlation like shown in this study. But, filtering the trend in the temperature hindcasts and observations showed that the initialized system beats the uninitialized simulations in terms of correlation on these timescales. However, the long-term temperature trend belongs to the 2 to 9 year forecast. This cannot be adjusted, when presenting decadal predictions.

The comparison of the 10 ensemble member system against the 3 ensemble member system (used in Pohlmann et al., 2013), shows clear improvements in the MSESS over the whole globe. Even for regions of overestimated precipitation, the forecasts improved for lead years 2 to 9. The analysis of the LESSS also shows a slight improvement in ensemble spread in the tropics, comparing two different ensemble sizes. In most of the regions the ensemble spread is an adequate representation for the uncertainty of this system and it is much closer to the reference error (MSE) than for other decadal prediction systems (Goddard et al., 2013).

Including the LESS and the LESSS to the set of skill assessment for decadal prediction allows to distinguish between an over- or under-dispersive ensemble and detect improvements made when aiming at larger ensemble sizes (Sienz et al., 2016). The LESSS could also be used to evaluate different ensemble generation methods of the same model system to assess their possible improvement. After the development stages and accomplished improvements (Müller et al., 2012; Pohlmann et al., 2013; Kruschke et al., 2014; Stolzenberger et al., 2015; Spangehl et al., 2015), the next step in the ongoing MiKlip project is to switch from the anomaly initialization in the ocean with ORAS4 to full-field multi-reanalysis initialization with ORAS4 and GECCO2 (Köhl, 2014). A first study on these combined predictions is given by Kruschke et al. (2015). The coming 30 member prediction system will allow a more robust assessment. It will be possible to involve other scores to this combined prediction system, e.g. the error spread score (Christensen et al., 2014), which needs ensemble sizes larger than available in this study.

The decadal skill assessment used in this study is an operational part of the central evaluation in MiKlip. It is available to the climate science community (Illing et al., 2014) and is planned to be deployed in the next stages of the MiKlip project development.

Acknowledgments

We acknowledge funding from the Federal Ministry of Education and Research in Germany (BMBF) through the research programme MiKlip[1], projects INTEGRATION (FKZ: 01LP1160A), ENSDIVAL (FKZ: 01LP1104A), and FLEXFORDEC (FKZ: 01LP1144A). We have used observational data provided by the Hadley Centre[2], Global Precipitation Climatology Project[3], Global Precipitation Climatology Centre[4] at Deutscher Wetterdienst and reanalyses from the ECMWF[5]. The climate simulations were performed at the German Climate Computing Centre (DKRZ[6]). The skill assessment is part of the Central Evaluation System[7] of MiKlip. The authors thank Mareike Schuster, Tim Kruschke, Madlen Fischer and Martina Baltkalne whose comments helped to improve this paper.

[1] http://www.fona-miklip.de/en
[2] http://www.metoffice.gov.uk/hadobs
[3] http://www.gewex.org/gpcpdata.htm
[4] http://gpcc.dwd.de/
[5] http://www.ecmwf.int
[6] https://www.dkrz.de
[7] https://www-miklip.dkrz.de

3.5 Appendices to Chapter 3

Variable or Equation	Explanation
i=1,...,m	ensemble members
j=1,...,n	start or initialization times of experiments
H_{ij}	initialized hindcasts
H_j	ensemble mean of hindcasts
O_j	observations
$\mathrm{MSE}_H = \frac{1}{n}\sum_{j=1}^{n}(H_j - O_j)^2$	mean squared error of the hindcast (against observation)
$\mathrm{MSE}_{\bar{O}} = \frac{1}{n}\sum_{j=1}^{n}(\bar{O} - O_j)^2$	mean squared error of the climatological forecast (against observation)
$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}}$	general expression of a skill score (A value for accuracy measure, A_{perf} the value for perfect prediction and A_{ref} the value for a reference forecast system
$\mathrm{MSESS}(H, \bar{O}, O) = \frac{\mathrm{MSE}_H - \mathrm{MSE}_{\bar{O}}}{0 - \mathrm{MSE}_{\bar{O}}}$ $= 1 - \frac{\mathrm{MSE}_H}{\mathrm{MSE}_{\bar{O}}}$	mean squared error skill score of the hindcast H vs the climatological forecast $\bar{O}$ (with $\mathrm{MSE}_{perf} =$ 0)
$\mathrm{MSESS}(H, \bar{O}, O) = r_{HO}^2 - \left[r_{HO} - \frac{s_H}{s_O}\right]^2$	Murphy-Epstein decomposition of the MSESS
$\mathrm{MSESS}(H, R, O) = 1 - \frac{\mathrm{MSE}_H}{\mathrm{MSE}_R}$ $= \frac{\mathrm{MSESS}_H - \mathrm{MSESS}_R}{1 - \mathrm{MSESS}_R}$	mean squared error skill score of the hindcast H vs a reference prediction R
r_{HO}	sample correlation coefficient between hindcasts (H) and observations (O)
s_H^2 and s_O^2	sample variance of the hindcasts and observations
$r_{HO} - \frac{s_H}{s_O}$	conditional bias of hindcasts (H) compared to observations (O)

Tab. 3.1: Overview table of used variable names and equations.

Variable or Equation	**Explanation**
$\mathrm{CRPS}(H_{ij}, O_j) = \int_{-\infty}^{\infty} (F_{H_j}(y) - \mathcal{H}(y - O_j))^2 dy$	continuous ranked probability score
$\mathrm{CRPS}(\mathcal{N}(H_j, \sigma^2_{H_j}), O_j) = \sigma_{H_j}\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{O_j - H_j}{\sigma_{H_j}}\right) - \frac{O_j - H_j}{\sigma_{H_j}}\left(2\phi\left(\frac{O_j - H_j}{\sigma_{H_j}}\right) - 1\right)\right]$	continuous ranked probability score expressed with the standard normal probability density (φ) and cumulative distribution function (ϕ)
$\mathcal{H}(y - O_j) = \begin{cases} 1, & \text{if } y \geq O_j \\ 0, & \text{if } y < O_j \end{cases}$	Heaviside function as the associate cumulative distribution function for the single observation
$F_{H_j} = \mathcal{N}(H_j, \sigma^2_{H_j})$	probability distribution of the ensemble forecast
φ and ϕ	standard normal probability density (pdf) and cumulative distribution function (cdf)
$\hat{H}_{ij}$ and $\hat{H}_j$	ensemble members and ensemble mean corrected for mean and conditional bias
$\overline{\sigma^2_{\hat{H}}} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{m-1}\sum_{i=1}^{m}(\hat{H}_{ij} - \hat{H}_j)^2$	average ensemble spread
$\sigma^2_R = \frac{1}{n-2}\sum_{j=1}^{n}(\hat{H}_j - O_j)^2$	mean squared error (MSE)
$\mathrm{CRPSS}_{\mathrm{ES}} = 1 - \frac{\sum_j \mathrm{CRPS}_H(\mathcal{N}(\hat{H}_j, \overline{\sigma^2_{\hat{H}}}), O_j)}{\sum_j \mathrm{CRPS}_R(\mathcal{N}(\hat{H}_j, \sigma^2_R), O_j)}$	continous ranked probability skill score for the assessment of the ensemble spread
$\mathrm{LESS} = \ln\left(\frac{\overline{\sigma^2_{\hat{H}}}}{\sigma^2_R}\right)$	logarithmic ensemble spread score
$\mathrm{LESSS} = 1 - \frac{\mathrm{LESS}^2_{\mathrm{pred}}}{\mathrm{LESS}^2_{\mathrm{ref}}} \in (-\infty, 1]$	logarithmic ensemble spread skill score

Tab. 3.2: Overview table of used variable names and equations.

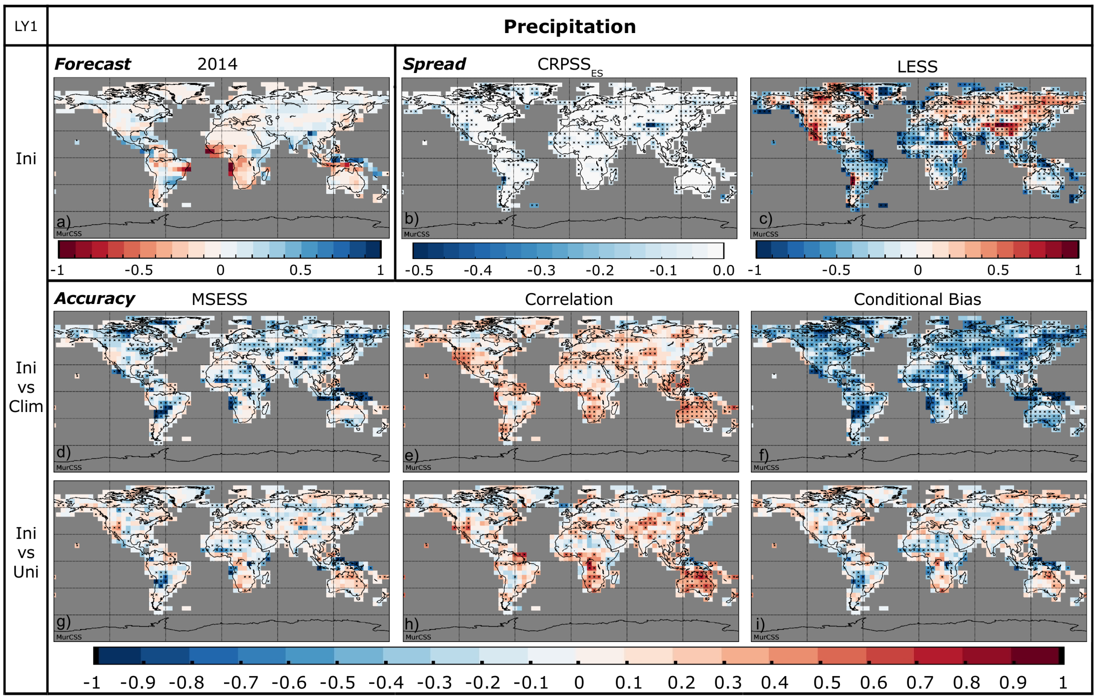


Fig. 3.7: As in Figure 3.4 but using the observation from GPCC over the period 1961-2012 for skill assessment.

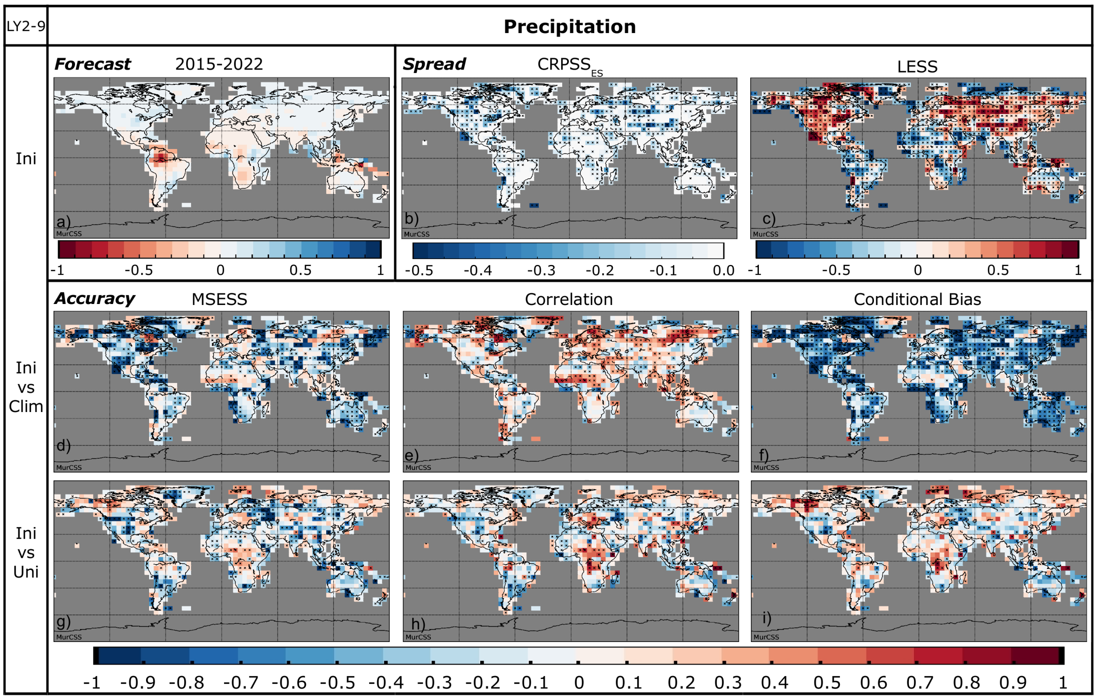


Fig. 3.8: As in Figure 3.5 but using the observation from GPCC over the period 1962-2012 for skill assessment.

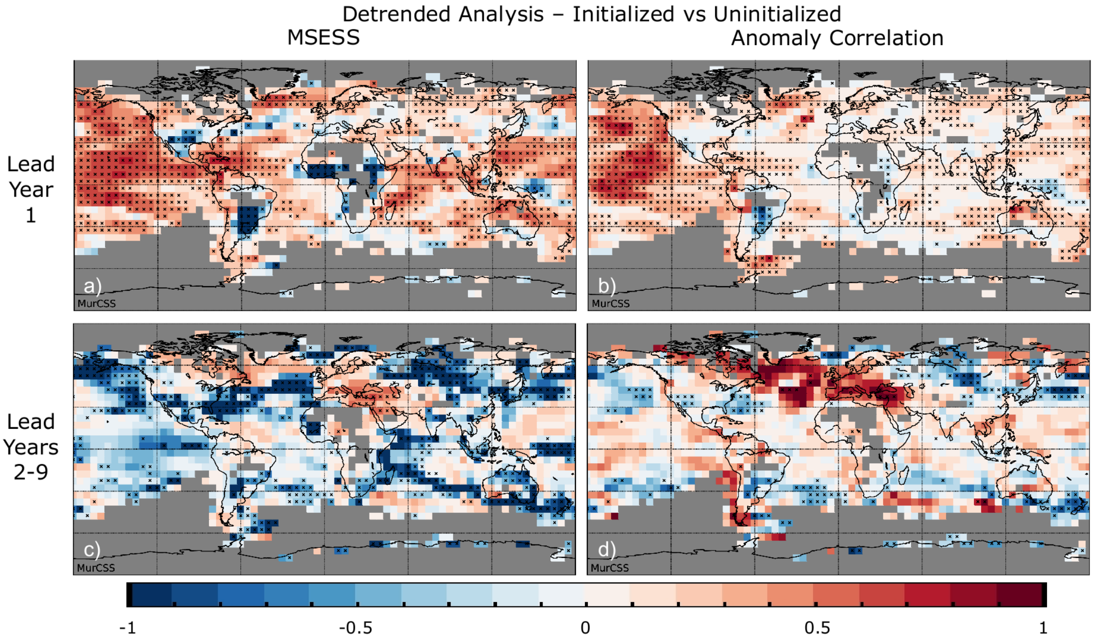


Fig. 3.9: Comparison of the detrended analyses from initialized vs uninitialized simulations. Anomaly correlation and the Mean Squared Error Skill Score (MSESS) for near surface air temperature over the period 1961-2012 against HadCRUT3v for the lead year 1 (upper row) and lead years 2 to 9 (lower row). The anomaly correlation and MSESS shows the added value of the initialization made in the hindcast ensemble mean prediction when neglecting the linear climate trend. Crosses denote values significantly different from zero exceeding at a 5% level applying 1000 bootstraps. Gray areas mark missing values with with less than 90% data consistency in the observation.

Information

Information about the candidate's and co-authors' work on that paper:

The **candidate** developed the scientific idea of the paper -namely the evaluation strategy for decadal prediction-, which includes the combination of an anomaly forecast, accuracy (MSESS, correlation, conditional bias) and spread (CRPSS, LESS) assessment with climatological and uninitialized references; enhancements and redevelopment of existing methods; development of a novel skill score (LESSS) for the spread assessment; added assessment of ensemble size; analyzed, combined, and discussed the results; revealed and confirmed sources of skill, namely the ocean model initialization and the enlargement of ensemble members; drafted the paper.

The **co-authors** provided data, software, and manpower to evaluate the results; added a re-adjustment of the CRPSS; added the logarithmic scale to the Ensemble Spread Score; and helped to improve the text of the paper by numerous comments and took care of the funding.

4

Decadal climate predictions improved by ocean ensemble dispersion filtering

Abstract

Decadal predictions by Earth system models aim to capture the state and phase of the climate several years in advance. Atmosphere-ocean interaction plays an important role for such climate forecasts. While short-term weather forecasts represent an initial value problem and long-term climate projections represent a boundary condition problem, the decadal climate prediction falls in-between these two timescales. In recent years, more precise initialization techniques of coupled Earth system models and increased ensemble sizes have improved decadal predictions. However, climate models in general start losing the initialized signal and its predictive skill from one forecast year to the next. Here we show that the climate prediction skill of an Earth system model can be improved by a shift of the ocean state towards the ensemble mean of its individual members at seasonal intervals. We found that this procedure, called ensemble dispersion filter, results in more accurate results than the standard decadal prediction. Global mean and regional temperature, precipitation, and winter cyclone predictions show an increased skill up to 5 years ahead. Furthermore, the novel technique outperforms predictions with larger ensembles and higher resolution. Our results demonstrate how decadal climate predictions benefit from ocean ensemble dispersion filtering towards the ensemble mean.

The following chapter consists of the main publication and the supporting information published in the open access *Journal of Advances in Modeling Earth Systems (JAMES)* of the American Geophysical Union (AGU). Not subject to U.S. copyright. This chapter has been published as:

Kadow, C., S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch (2017), Decadal climate predictions improved by ocean ensemble dispersion filtering, *Journal of Advances in Modeling Earth Systems*, 9, 1138–1149, `https://doi.org/10.1002/2016MS000787`.

In addition, the publication possesses open access data within the *World Data Center PANGAEA* for transparency and reproducibility of this study.

Kadow, C., S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch (2017), Earth system model results by the MPI-ESM-LR of the MiKlip Decadal climate prediction experiment improved by ocean ensemble dispersion filtering, links to NetCDF files, *PANGAEA*, `https://10.1594/PANGAEA.874231`.

4.1 Introduction

Climate prediction and climate predictability using comprehensive Earth system models have become an important contribution of the climate science community (Meehl et al., 2014) to society. The seamless prediction - ranging from weather forecasts, over seasonal to decadal prediction, to century projections - conducted with one model system is the ultimate goal. However, the field of decadal prediction has several challenges. The research aims to bridge the gap between short term forecasts and long term projections. Short to medium-range weather forecasts focus on an initial value problem in the beginning of a forecast. On the other side, climate projections as a boundary condition problem - like greenhouse gases - examine the long-term climate development (Meehl et al., 2009; Mehta et al., 2011). Climate change projections are good indicators for the trend of the climate system. The natural variability of the climate around this trend is the real challenge. Lately, considerable progress has been made by initializing a decadal prediction system using ocean reanalyses [e.g. Marotzke et al., 2016]. Fitting the actual state of the Earth's climate system into a model allows it to capture the phase of current large scale variability (Smith et al., 2007). While the atmospheric processes act on a daily to sub-seasonal scale, the ocean processes dominate the inter-annual to decadal time scale. As sea surface temperatures of ocean basins are key factors determining the atmospheric global mean temperature (Meehl et al., 2013; Meehl et al., 2014; Kosaka and Xie, 2013), predicting the ocean can be considered as the main key to decadal predictability in our climate system (Keenlyside et al., 2008). As climate projections do not deal with actual states of the ocean, they cannot be predictors for multi-annual changes of the climate.

Several techniques including the ocean evolved by setting up retrospective forecasts or so called hindcasts. Adding or nudging anomalies of atmospheric or ocean observations into the model system is called anomaly initialization (Keenlyside et al., 2008; Pohlmann et al., 2009; Matei et al., 2012). Putting the actual state of the observations or usually of some reanalysis product into the model system is called full-field initialization (Yeager et al., 2012; Fyfe. et al., 2012). Recent studies discussed these methods causing errors in the prediction system in terms of drift and initial shocks (Smith et al., 2013; Kharin et al., 2013; Marotzke et al., 2016). Within the Coupled Model Intercomparison Project 5 [CMIP5 - Taylor et al., 2012] the decadal experiments started to be investigated in a community effort. Several modeling groups were involved. The following Decadal Climate Prediction Project [DCPP – Boer et al., 2016] within CMIP6 [Eyring et al., 2016] set up a more detailed protocol. The evaluation strategy of DCPP involves a setup of a common framework to evaluate and compare their hindcast sets focusing on accuracy and ensemble spread (Goddard et al., 2013). In recent years, additional efforts investigated in probabilistic measurements and forecast reliability [e.g. Weisheimer and Palmer, 2014; Kruschke et al., 2015; Stolzenberger et al., 2015]. The application of an ensemble approach is essential for a decadal prediction system (Sienz et al., 2016) – in many ways. Due to non-linear filtering of errors the ensemble average is closer to the truth (Kumar and Hoerling, 2000; Kalnay et al., 2006). Therefore, the evaluation of the accuracy of a model system with an ensemble mean is likely to be more skillful than using any of its individual members (Eade et al., 2014).

The innovation discussed in this paper consists in the combination of these two just mentioned scientific findings leading to an improvement of forecasts: (1) the ocean and its initialization plays a crucial role on the decadal time-scales of climate predictions, and (2) the ensemble mean of a forecast is generally more accurate than any of its individual members. We give detailed information on the experimental set-up of a new decadal forecast procedure, its evaluation methods, and the observational data used to validate the hindcast sets (see Sect. 4.2). We show results of the global mean and regional temperature, precipitation, and winter cyclone predictions with the new method and its reference (see Sect. 4.3), before we discuss and conclude this study (see Sect. 4.4).

4.2 Modeling, Methods, and Data

4.2.1 Common Base – Model and Prediction System

The decadal prediction system of MiKlip (Marotzke et al., 2016) is based on the Max-Planck-Institute Earth System Model (Stevens et al., 2013; Jungclaus et al., 2013). The low resolution version (MPI-ESM-LR) of the Max-Planck-Institute Earth system model is the coupled climate model applied in this study. The atmospheric component ECHAM6 (Stevens et al., 2013) has a resolution of T63L47 and the oceanic component MPI-OM (Jungclaus et al., 2013) has a resolution of 1.5°/L40. It is a high computational effort to produce a yearly initialized decadal hindcast set in a lead-time-dependent way. In the decadal component of CMIP5 (Taylor et al., 2012) most of the decadal prediction hindcast sets reached no more than 3 ensemble members or just initialization every fifth year.

The MiKlip setup (Pohlmann et al., 2013; Kadow et al., 2016) is used as reference prediction hindcast set, hereafter called MiKlip-REF. The configuration used in this study follows the protocol of the Decadal Climate Prediction Project [DCPP - Boer et al., 2016] within the Coupled Model Intercomparison Project (CMIP). Following the DCPP protocol, the MiKlip-REF system consisting of 5 ensemble members is initialized every year on the 1st of January. The individual ensemble member of the full model system is started on different start days following the 1st of January to spread the ensemble - called lag day initialization. The set-up covers the decadal experiments from 1974 to 2012. Each initialization simulated a pentad. This time range is used to be able to evaluate the lead years (LY) 1 to 5 in the same time frame from 1979 to 2013 by shifting the experiments (e.g. LY1 uses experiments initialized in 1978 to 2012, LY2 in 1977 to 2011, and so on) as suggested in the DCPP protocol. An “assimilation run” was set-up to guide the MPI-ESM-LR model system towards an observational state. This reanalysis-like model run was used to start the prediction from. Therefore, the following reanalyses data was used. The ocean model was anomaly initialized by the Ocean ReAnalysis System 4 (ORAS4) (Balmaseda et al., 2013) from the European Centre for Medium-Range Weather Forecasts (ECMWF). Oceanic temperature and salinity anomalies were nudged to MPI-OM. The atmosphere full-field initialization comes from ECMWF ERA40 (Uppala et al., 2005) for the period 1974-1989 and from ERA-Interim (Dee et al., 2011) for 1990-2012. The actual values of temperature, surface pressure, vorticity, and divergence of the reanalysis replaced the ECHAM6 values.

MiKlip-REF is the base for the new development explained in the next sub-section and therefore the most important reference when assessing the skills. However, other interesting comparisons to different approaches within MiKlip can be done. The MiKlip-REF-10 is an extension of MiKlip-REF from 5 to 10 ensemble members (Kadow et al., 2016). This approach stands for the idea of increasing the ensemble size. The MiKlip-REF-MR uses the mixed resolution version (Pohlmann et al., 2013) of MPI-ESM. Its data set consists of a higher ocean resolution (0.4°L40) and more vertical levels in the atmosphere (T63L95). This reference reflects the idea of increasing the model resolution. The MiKlip-REF-FF is part of the newer Prototype (Marotzke et al., 2016) system of MiKlip. It uses full-field initialization, this means actual values of the oceanic variables by ORAS4 instead of anomalies as used in MiKlip-REF. Uninitialized runs of the MPI-ESM-LR serve as references as well (Kadow et al., 2016), which are called MiKlip-REF-UN. This reference is usually taken to determine the trend and the added value of initialization procedures. MiKlip-REF-UN equates to a mixture of the ‘historical’ and ‘rcp45’ experiments according to CMIP5 (Taylor et al., 2012) using observed (historical) and projected (rcp45) external forcing.

4.2.2 Ensemble Dispersion Filter – Setup and Details

In this study, we present a new forecasting technique using and name it an ensemble dispersion filter (EDF) to retrieve the initialized climate signal more precisely. Producing an ensemble for climate predictions is common practice. Small perturbations of the model system lead to different variations of the models climate system [e.g. Lorenz, 1963. Using its ensemble mean helps to reduce errors and increase accuracy of predictions (Kalnay et al., 2006). This is usually done after model runs. In

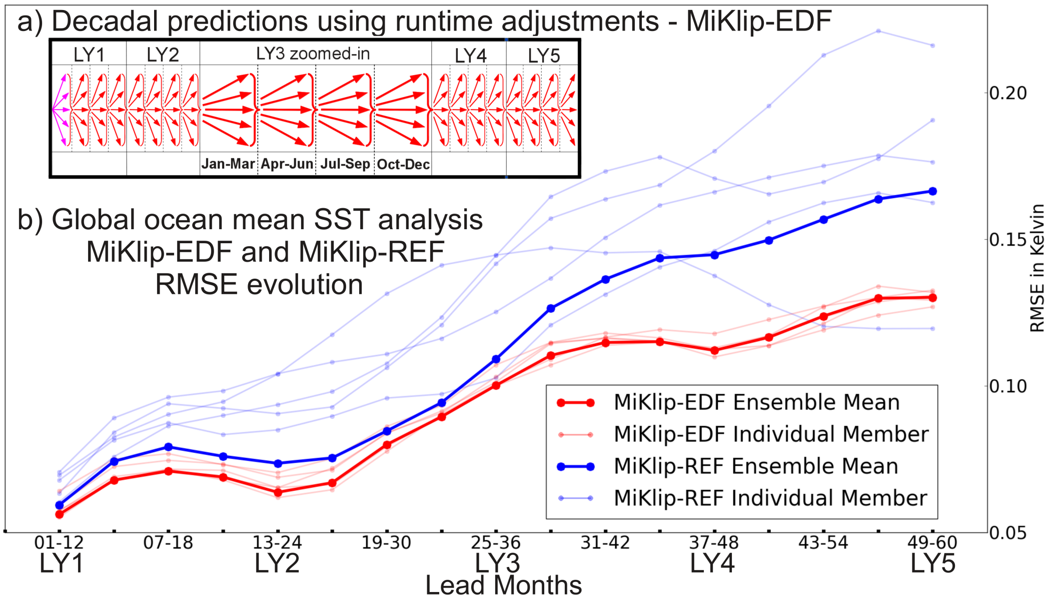


Fig. 4.1: (a) (top) Schematic decadal climate prediction hindcast experiment setup of MiKlip-EDF in red. MiKlip-EDF consists of five ensemble members and 5 year integrations. The first 3 months of every experiment and ensemble member from MiKlip-EDF and MiKlip-REF are identical (magenta). Time frame of decadal experiments from 1974 to 2012 to cover analysis years from 1979 to 2013 for all 5 lead years. (b) (bottom) Global ocean mean sea surface temperature (SST) analysis of MiKlip-REF in blue and MiKlip-EDF in red. The ensemble mean in dark colors and the individual members in light colors. Shown is the development of the root-mean-squared error (RMSE) in comparison to the HadSST3 observation in differences in Kelvin over lead months in 12 months (yearly) chunks every 3 months. The analysis covers the years from 1979 to 2013.

machine learning Bayesian model averaging or to be more specific Bootstrap Aggregation (Bagging) on unstable procedures smooth out variance and reduce mean squared error leading to improved predictions (Hastie et al., 2013). If perturbing the learning set can cause significant changes in the predictor constructed, then Bagging can improve accuracy (Breiman, 1996). Applying the ensemble mean during the model run of a perturbed (lag-day initialized) decadal climate prediction could lead to much more distinct signals of the prediction system. It benefits from the ensemble within its prediction process applying the EDF. The ensemble dispersion filter approach in this study uses ocean and surface temperatures of the initialized decadal prediction system to improve its performance on the first pentad.

We started MiKlip-EDF as MiKlip-REF from the “assimilation” run consisting of a MPI-ESM-LR model run with observational (reanalyses) information. While MiKlip-REF simulates the climate for the next years as an independent run after initialization, MiKlip-EDF was stopped after 3 months. Thereafter, the model’s restart files of the 5 ensemble member were processed with the help of the NetCDF Operators (Zender, 2008). Due to the fact that the sea surface temperature is part of the atmospheric component of the MPI-ESM, there was the need to modify the MPI-OM and the

ECHAM. The ensemble mean of the ocean temperatures (MPI-OM code 2 - variable THO) and the (land and sea) surface temperature fields (ECHAM code 169 – variable tsurf) were calculated (see also text 4.5). Every level of the ocean temperature was used to maintain the memory of the deep ocean. The surface temperature was used to allow an atmosphere-ocean interaction of this forecast technique. We added some spread for the development of its ocean temperatures by using only 4 of the 5 ensemble members when calculating the ensemble mean. Therefore, we have 5 combinations of 4 members leaving out one member at every step calculating the ensemble means. These in-run perturbations or leave-one-out cross bootstraps keep the idea of building ensemble means during the prediction alive. This was done for every 3 month period until the 5 year forecast of one decadal experiment was finished. This whole hindcast setup was used for every decadal experiment between 1974 and 2012 (Fig. 4.1a).

4.2.3 Evaluation Strategy for Decadal Climate Predictions

We evaluated the decadal prediction system using the published software package MurCSS (Illing et al., 2014) applied and developed within the Central Evaluation System of MiKlip (Marotzke et al., 2016). It follows the evaluation strategy (Goddard et al., 2013) for decadal prediction systems by analyzing them in a lead-year manner in terms of accuracy and spread compared to observations. The lead years (LY) are the forecasted years of all decadal experiments in the hindcast. We combine the first forecast year of all experiments into a LY1 time series, to verify the skill of the first year prediction by evaluating the hindcast in its first lead year. Accordingly this is been done for all lead years. In this study, we focus on the accuracy of the prediction by evaluating the ensemble mean, but investigating partly into ensemble spread and forecast reliability. As suggested by the DCPP, we analyze the yearly initialized experiments in the same time frame for all lead years. Analyzing the time frame 1979 to 2013 is a typical range in decadal climate prediction, focusing on the most certain observational period. All methods and formulas are identical to those applied and written down in its open access predecessor study as given by Kadow et al., 2016 evaluating the MiKlip system, here the reference system MiKlip-REF. The mean squared error skill score ($\infty < \mathrm{MSESS} \leq 1$) compares the accuracy of two predictions (Murphy, 1988) of the past, so called hindcasts. Applying the Murphy-Epstein decomposition, the MSESS for the hindcast H vs the observational climatology $\bar{O}$ compared to the observation O can be written as:

$$\mathrm{MSESS}(H, \bar{O}, O) = 1 - \frac{\mathrm{MSE}_H}{\mathrm{MSE}_{\bar{O}}} = r_{HO}^2 - \left[r_{HO} - \frac{s_H}{s_O} \right]^2 \rightarrow 1 = \text{perfect skill score} \tag{4.1}$$

with r being the sample correlation coefficient between the hindcasts and the observations, and the sample variance s of the hindcasts and observations (Murphy, 1988;

Murphy and Epstein, 1989). When comparing the hindcast H with some reference hindcast set R, the MSESS can be written as

$$\mathrm{MSESS}(H,R,O) = \frac{\mathrm{MSESS}_H - \mathrm{MSESS}_R}{1 - \mathrm{MSESS}_R} \rightarrow 1 = \text{perfect skill score} \tag{4.2}$$

to assess for example the change of skill comparing two development steps of a prediction system. It represents the improvement in the accuracy of the hindcast H over the climatology $\bar{O}$ or a reference hindcast R with respect to the observations O. A positive value suggests an improved accuracy of the hindcast ensemble mean compared to the reference, and vice versa. The correlation coefficient (-1 $\leq r \leq$ 1) as the potential skill of a prediction system represents the linear relationship between a hindcast and the observation. The evaluation of the ensemble spread and reliability helps to determine the forecast uncertainty. For ensemble spread we consider the spread score (see Palmer et al., 2006; Keller et al., 2008, with a log-transform to obtain the logarithmic ensemble spread score (Kadow et al., 2016) which is symmetric around zero. To quantify the ensemble spread against the standard error, we use the average ensemble spread $\sigma^2_{\hat{H}}$ and the reference MSE_H

$$\mathrm{LESS} = \ln\left(\frac{\overline{\sigma^2_{\hat{H}}}}{\mathrm{MSE}_\mathrm{H}}\right). \tag{4.3}$$

If the MSE corresponds to the ensemble variance, the latter is a good estimate of the forecast uncertainty. Is the ensemble variance smaller than the MSE the ensemble is said to be under-dispersive (overconfident). An ensemble variance larger than the MSE indicates an over-dispersive (underconfident) ensemble. For the forecast reliability we parametrize the slope within the reliability diagram (Hsu and Murphy, 1986) with four categories. Reliability diagrams are graphical tools to investigate the correspondence of forecast probabilities of dichotomous events and the observed frequency given the forecast (Wilks, 2011). A weighted linear regression of all forecast probabilities and relative observed frequency pairs results in a reliability line of which the slope - including its uncertainty range - can be used as indicator of reliability (Weisheimer and Palmer, 2014; Stolzenberger et al., 2015). Categories of reliability are defined following Weisheimer and Palmer, 2014 combining their lowest two:

$$\text{Reliability Classification:} = (\text{perfect} \mid \text{still useful} \mid \text{marginally useful} \mid \text{not useful}) \tag{4.4}$$

The binary event is defined as the exceedance of the climatological median at every grid point. To increase sample size of the estimations the nearest neighbors of each grid point are taken into account leading to a smoothed field of reliability. Observational and model data were spatially interpolated into a common 5°x5° grid, and temporally averaged to yearly anomalies using the evaluation period for climatology, and a cross-validated and lead-time-dependent bias adjustment (ICPO, 2011). The lead-time-dependent bias adjustment uses the temporal mean of a specific lead year to calculate anomalies to account for potential lead-time dependent drifts of the model system. The cross validation leaves out the year which is corrected within the temporal mean for bias correction. Annual averaged climate values are normally distributed or will be at least approximately Gaussian (Wilks, 2011), which is important for the applied statistics (Kadow et al., 2016). Significance of the verification scores was estimated using a non-parametric bootstrap (1,000 fold) approach (Wilks, 2011) taking auto correlation into account (Goddard et al., 2013). We focused on the LY2-5 period because it is the typical time frame to look at in a decadal prediction as suggested by the DCPP.

4.2.4 Observational Data Sets

In this study we will evaluate global mean and regional temperature, precipitation, and winter cyclone hindcasts to assess the skill of the prediction systems. For the evaluation of the near-surface temperature we compared the model simulations with the observational anomaly data set, which technically speaking is the median of HadCRUT4 (“Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set”). It is a collaborative product with the ocean component HadSST3 (“Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization”) of the Met Office Hadley Centre and the land component CRUTEM4 of the Climatic Research Unit at the University of East Anglia (“Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010”). The evaluation of precipitation was carried out using the Global Precipitation Climatology Centre (GPCC) Full Data Reanalysis (V7) (Becker et al., 2013) operated by the German Weather Service (DWD) under the auspices of the World Meteorological Organization (WMO). We assessed the cyclone track densities after post-processing mean sea level pressure (PSL) of the ERA-Interim reanalysis by the ECMWF. The cyclone tracking uses the Laplacian of the PSL to identify cyclones, and afterward the track densities are calculated (Pinto et al., 2005; Murray and Simmonds, 1991). This method was applied to MiKlip-EDF and MiKlip-REF as well. For a clean assessment of the track density and tracking, the PSL of ERA-Interim was interpolated onto the grid of the MPI-ESM. We used the Met Office HadISST (“Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century”) data set as a reference in the supplementary information section to evaluate the sea surface temperature biases of MiKlip-EDF and MiKlip-REF in comparison to the observations.

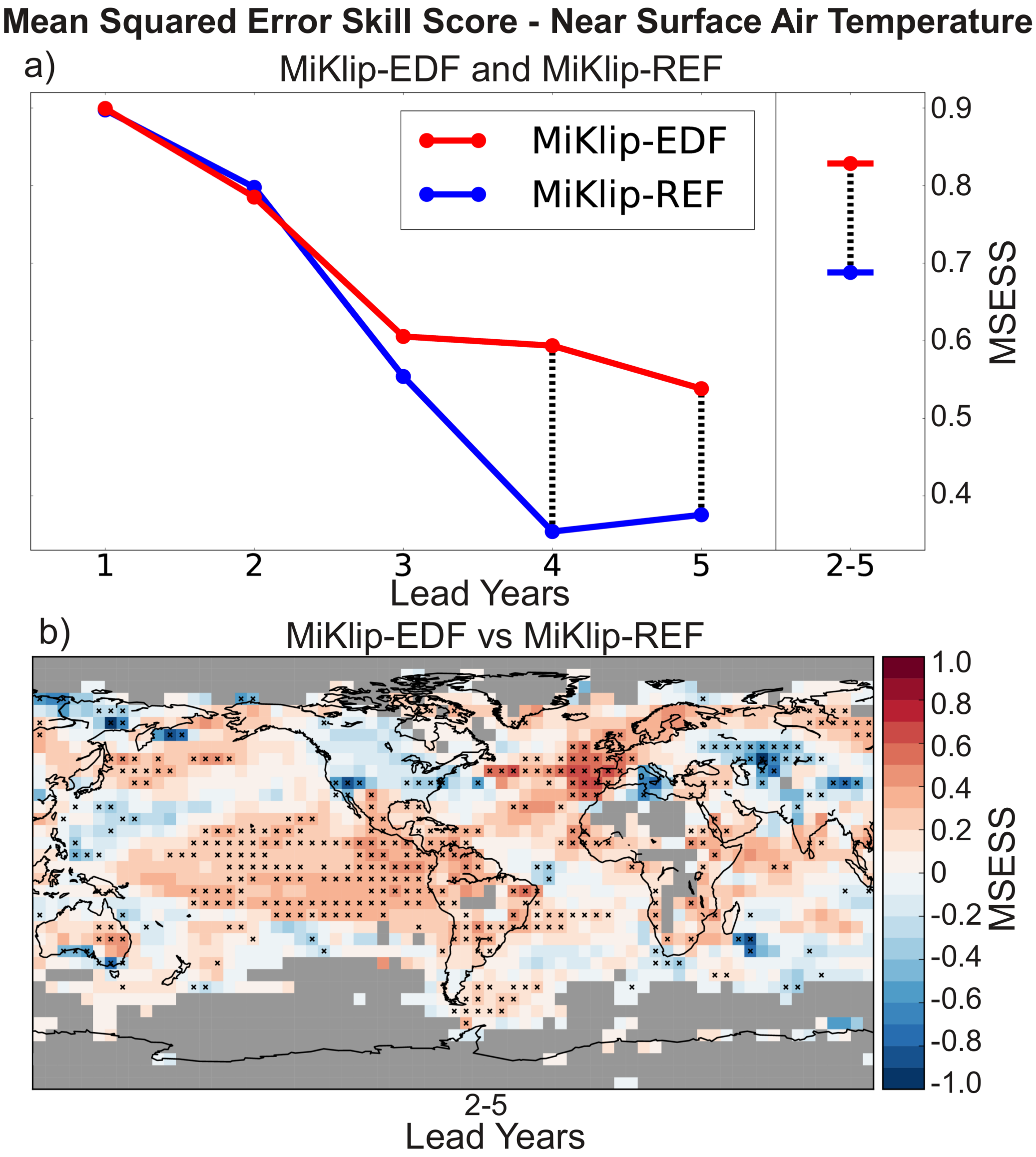


Fig. 4.2: (a) Mean squared error skill score (MSESS) of the global mean temperature ensemble mean of MiKlip-REF (blue) and MiKlip-EDF (red) for LY1 to LY5 and LY2-5 compared to HadCrut4 with climatology as a reference prediction on the top. Significant differences of MiKlip- EDF to its reference prediction MiKlip-REF in the lead year skill are marked by black dashed lines. (b) The corresponding regional analysis of the LY2-5 MSESS shows the improvement of MiKlip-EDF compared to its reference prediction MiKlip-REF with observations of HadCrut4 on the bottom. Crosses denote values significantly different from zero exceeding at a 5% level applying 1000 bootstraps. Gray areas indicate missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013.

4.3 Results

The temporal development of the global ocean mean sea surface temperature as the RMSE in Kelvin compared to observations (“Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization") already indicates the benefits of the novel technique (Fig. 4.1b). First of all it confirms that in MiKlip-REF and MiKlip-EDF, the ensemble mean is in most cases closer to the observed development of the climate than any of its individual members. Additionally, the ensemble mean of MiKlip-EDF is closer to the observation than the ensemble mean of MiKlip-REF. This effect gets even larger with increasing lead time. There are large differences between individual members of MiKlip-REF, and they grow with lead time. This effect cannot be found within MiKlip-EDF. With an improved estimate of the ocean state, we show now that the resulting atmospheric climate variables are more accurate as well.

The comparison of MiKlip-EDF with its reference system MiKlip-REF with respect to global surface temperature reveals the benefits of the novel forecasting technique (Fig. 4.2). Here, the MSESS compares MiKlip-EDF and MiKlip-REF against forecasting the climatological mean (Fig. 4.2a). It confirms that the initialization effect is strongest in the first lead year (LY1), decreasing thereafter in both hindcast sets. Differences in the first 3 lead years between MiKlip-EDF and MiKlip-REF are small and statistically not significant. A significant improvement of MiKlip-EDF is found in the LY4 and LY5 where it maintains skill longer than MiKlip-REF. This results in a more accurate and significantly better forecast of the LY2-5 global mean temperature by MiKlip-EDF in reference to MiKlip-REF as well.

In addition to the improvement in the global mean temperature predictions, an enhancement of the skill on the regional scale can be seen (Fig. 4.2b). Here the MSESS of MiKlip-EDF uses MiKlip-REF as a reference in the LY2-5 analysis. The near-surface temperature prediction reveals large regions of significant improvement. The strongest effect is located in the North Atlantic and Western Europe. Also the tropics, including the high impact ENSO region in the Central and North Pacific, as well as South America, Africa, and Australia show more accurate predictions. A few regions with a significant loss of skill, like Central Asia, the Central-West Pacific, and the Mediterranean Sea, can be found as well.

By definition of this new technique, MiKlip-EDF decreases the ensemble spread in the near surface temperature compared to MiKlip-REF. This can be determined in Fig. 4.3a and 4.3b. The LESS reveals the obvious small ensemble spread of MiKlip-EDF compared to the MSE especially over the ocean. Over the continents the ensemble spread in MiKlip-EDF is closer to the MSE than in MiKlip-REF. This results in a MiKlip-EDF ensemble spread which is better over continents, but tends to be overconfident over ocean. However, the forecast reliability analysis next to the spread evaluation shows no significant difference (Fig. 4.3c, d). In both forecast systems the reliability patterns over the whole globe are quite similar. Thus, even if we lose ensemble spread, the EDF is not reducing the forecast reliability compared

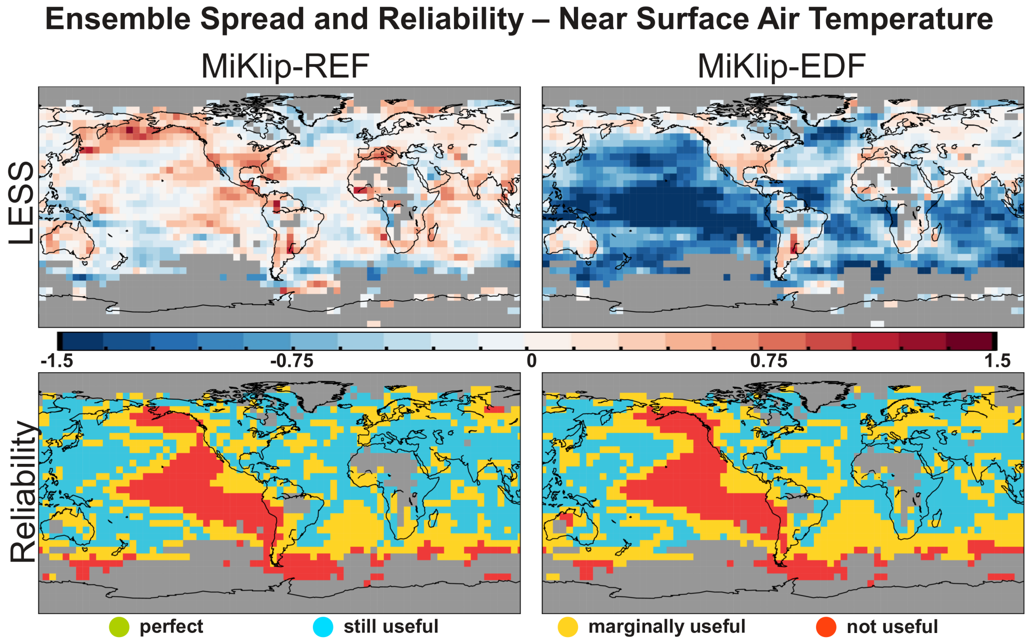


Fig. 4.3: (top) Logarithmic ensemble spread score and (bottom) forecast reliability for near-surface air temperature in (left) MiKlip-REF and (right) MiKlip-EDF. Analyses are done for LY2-5 compared to HadCrut4. Gray areas indicate missing values with less than 90in the observation. The analyses cover the time period from 1979 to 2013.

to the free reference run. An additional and future approach could be bundling several independent 5 member EDF systems. The general spread would increase. In addition we would have a spread of the ensemble means which could be a valuable information around this technique. Besides determining the spread and reliability, it is worth analyzing the mean bias of sea surface temperature as well. We note for example a reduced North Atlantic cold bias (Supplementary Figure 4.6).

For a more comprehensive temperature assessment, we also compared the new MiKlip-EDF dataset to more recently developed sets of MiKlip experiments (Marotzke et al., 2016) representing different decadal prediction strategies with the same model system (see Table 4.1). Figure 4.4 shows the comparison of MiKlip-EDF and MiKlip-REF to be directly compared with Figure 4.2. MiKlip-EDF outperforms all of them in the most important time frame of LY2-5. MiKlip-EDF shows significant improvements in the global mean analysis as well as large regional patterns. MiKlip-REF is worse than the other more sophisticated systems. However, these results especially in the global mean LY2-5 are not significant. The comparison against the uninitialized runs (Fig. 4.4 bottom) is a general way to evaluate the added value of initialized decadal predictions. Here, MiKlip-EDF shows clear signs of a significant added value in contrast to MiKlip-REF.

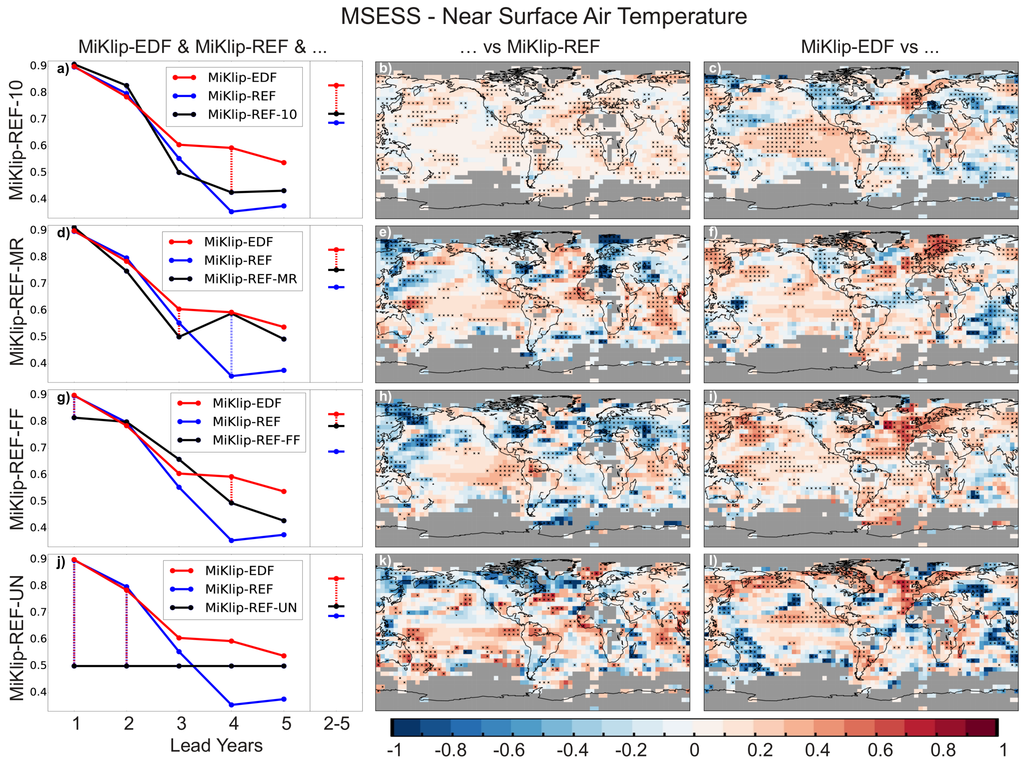


Fig. 4.4: Mean squared error skill score (MSESS) of the global mean temperature with climatology as a reference prediction for (left column) LY1 to LY5 and LY2-5 of MiKlip-REF (blue), MiKlip-EDF (red), and other MiKlip reference data sets (black): (first row) MiKlip-REF-10 includes 10 instead of 5 ensemble members, (second row) MiKlip-REF-MR uses the MPI-ESM-MR with higher resolution instead of MPI-ESM-LR, (third row) MiKlip-REF-FF uses full-field instead of anomaly initialization in the ocean, and (fourth row) MiKlip-REF-UN is the uninitialized mix of the historical and rcp45 experiments. (middle column) The corresponding regional analysis of MiKlip-REF as a reference for the other MiKlip-REF-XX prediction system in the LY2- 5. (right column) The MiKlip-EDF analysis versus another MiKlip-REF-XX prediction system as a reference in the LY2-5. Significant differences are marked (left) by dashed lines or (middle and right) by crosses. Gray areas mark missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013. The figures are constructed to be compared with the main results in Figure 4.2.

Hindcast System	Ens. Size	Eva. Period	Dec. Exp.	Atmos Ini and Res	Ocean Ini and Res	EDF Freq.
MiKlip-EDF	5	1979-2013	1974-2012	FF/T63L49	AN/1.5°L40	**3 mon**
MiKlip-REF	5	1979-2013	1974-2012	FF/T63L49	AN/1.5°L40	-
MiKlip-REF-10	**10**	1979-2013	1974-2012	FF/T63L49	AN/1.5°L40	-
MiKlip-REF-MR	5	1979-2013	1974-2012	FF/T63**L95**	AN/**0.4**°L40	-
MiKlip-REF-FF	5	1979-2013	1974-2012	FF/T63L49	**FF**/1.5°L40	-
MiKlip-REF-UN	5	1979-2013	**–**	**–**/T63L49	**–**/1.5°L40	-

Tab. 4.1: Overview of hindcast systems used in this study. Information about the ensemble size, the evaluation period, the yearly initialized decadal experiments, the MPI-ESM atmosphere initialization (Ini) technique and resolution (Res), the MPI-ESM ocean initialization (Ini) technique and (Res) resolution, and the frequency of applying the ensemble dispersion filter. Initialization techniques are full-field (FF) and anomaly (AN). Highlighted in bold are the main differences to the basic reference system MiKlip-REF.

Forecasting long-term precipitation changes and multi-annual variations of rain is a challenge in climate science (Goddard et al., 2013). Both, MiKlip-REF and MiKlip-EDF show rather small, if any, skill (Supplementary Fig. 4.7) in predicting large scale anomalies on the global scale, which is in line with results of other studies (Kadow et al., 2016). We focus on the Northern Hemisphere and LY2-5 for a more detailed analysis of regional precipitation skill (Fig. 4.5a, b). With respect to observations (Fig. 4.5a), the correlation map of MiKlip-REF shows large and significant positive patterns over the north of East Asia and the Middle East. The MiKlip-EDF shows large significant positive patterns over North America, the Middle East, northern Central Europe, and smaller ones around Iceland and Greenland (Fig. 4.5b). Regions influenced by the ocean state and by atmospheric wind systems like cyclone tracks (see next paragraph and Fig. 4.5c, d), like Central Europe, show signs of significant improvements. This indicates that more accurate sea surface temperatures of the global prediction system lead to improvements in regional precipitation prediction. However, the precipitation patterns are far more local than temperature and some regions show negative developments as well. The evaluation of the ensemble spread and the forecast reliability shows that there is no appreciable difference between MiKlip-EDF and MiKlip-REF (Supplementary Fig. 4.8). In fact, the reliability in MiKlip-EDF is slightly better in most regions like Northwest America in comparison to MiKlip-REF. The ensemble dispersion filter applied on ocean temperatures has no negative effect on precipitation in terms of these ensemble metrics. A more comprehensive investigation on the prediction of large scale and convective rain as well as differentiation between seasons is beyond the scope of this study (see also Supplementary Fig. 4.7).

Prediction of extra-tropical cyclones is another benchmark for decadal climate forecast systems (Kruschke et al., 2014). We show the correlation of the lead winter (DJF) 2-5 cyclone track densities of MiKlip-REF and MiKlip-EDF compared to the ERA-Interim reanalysis in the Northern Hemisphere (Fig. 4.5c, d). MiKlip-REF reveals no significant predictability (Fig. 4.5c). MiKlip-EDF, however, shows large areas of significant positive correlation, especially over the North Atlantic and Europe, and along the North Atlantic storm track (Fig. 4.5d). As there is a strong connection between SST patterns and cyclone track density in climate models (Zappa et al., 2013) like the MPI-ESM (Kruschke et al., 2014), the improvement in the SST prediction leads to a significant step forward in predicting winter cyclones a pentad in advance. The evaluation of the ensemble spread and the forecast reliability underpins this finding. Besides an improvement in the North Atlantic Stormtrack region and Europe in the forecast reliability there is no appreciable difference between MiKlip-EDF and MiKlip-REF (Supplementary Fig. 4.9).

4.4 Conclusion

The novel forecast technique presented here improves the multi-annual temperature, precipitation, and winter cyclone prediction in comparison to the predictions obtained by the standard forecast technique. This is possible without a considerable increase of computational power, which would be necessary in the case of increasing the ensemble size or the model resolution. Even experiments with the MiKlip model system employing larger ensemble size and higher model resolution are outperformed by MiKlip-EDF as well – especially on the most important LY2-5 time scale. Skill is preserved much longer in MiKlip-EDF than in MiKlip-REF. This can be understood from the general forecast rule that the observed state is likely to be closer to the ensemble mean than to any individual ensemble member. Smoothing out variance and reducing the error in the perturbed signal of the initialization improves the forecast close to the ideas of machine or statistical learning. Especially in the improved North Atlantic region, MiKlip-EDF and its atmospheric model component responded to a different and more accurate sea surface temperatures. Usually the model atmosphere is not constrained strongly enough by the relevant drivers of predictability (Eade et al., 2014). The re-centering of the forecast ensemble improves skill by reducing the growth rate of model biases as well, by e.g. reducing the North Atlantic cold bias.

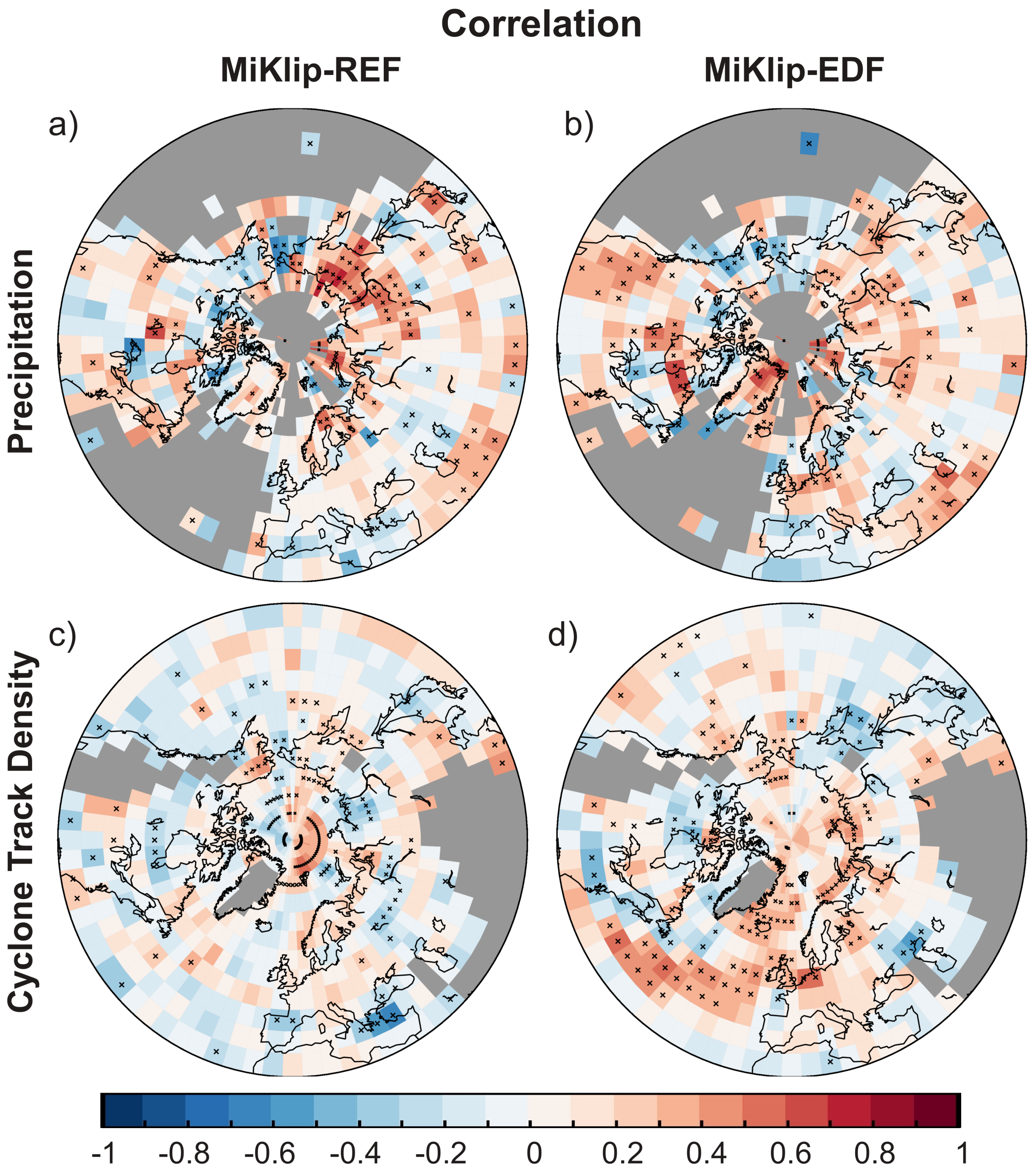


Fig. 4.5: Correlations of (left) MiKlip-REF and (right) MiKlip-EDF (top) for precipitation compared to GPCC observations and (bottom) for DJF cyclone track density compared to ERA-Interim for the LY2-5 hindcasts sets over the period from 1979 to 2013. Significance is marked by black crosses. Gray shading indicates missing data of (top) GPCC and regions higher than 1 km in the cyclone track density analysis (bottom).

However, more research on this forecast technique is necessary. For example, other meteorological variables or other restart time frequencies should be explored within the ensemble dispersion filter. The reduction in ensemble spread in the applied method within this study could be problematic for other research scenarios especially over or within the ocean. ENSO as well the North Atlantic sub-polar gyre state could lose potential information especially on seasonal time-scales when applying the EDF every 3 months. Therefore, an increase of the ensemble size should be beneficial as well. Connecting distinct members and building independent bundles would add more degrees of freedom to the analysis. This would introduce a new kind of ensemble spread, which should increase the temperature spread and amend its forecast uncertainty. The method itself is very much dependent on the initialized signal, because the EDF strengthen this, no matter if it is a good or bad initialized signal. In machine learning, it is known that Bagging a good classifier can make it better, but Bagging a bad classifier can make it worse (Hastie et al., 2013). Therefore, an investigation of a full-field initialization in the ocean would be an interesting addition. If the initialized observational signal would stay longer in the model system in addition to a reduction of the error growth rate, not just decadal prediction, but seasonal prediction - usually applying full-field - could improve as well. Slowing down the drift of full-field predictions should get investigated then as well.

A more advanced way of fostering the ensemble memory by using Ensemble Kalman Filter (Evensen, 2003) instead of simply using ensemble means should be explored as well. Next to other model development ideas like the 'Supermodel' in "Dynamically combining climate models to "supermodel" the tropical Pacific", the main ideas of the EDF could live up in other techniques like combining neural networks with numerical models. The synchronization of members in terms of information exchange could be a valuable add-on. In general, the approach should work with all numerical model systems producing a decadal prediction system. This study opens new possibilities for other ensemble forecasting disciplines in science and especially Earth system research, which could benefit from these or similar 'forecasting from forecasted mean state' methods using ensemble dispersion filter.

Acknowledgements

We thank M. Schuster and M. Baltkalne for critical discussions and reading the manuscript; H. Pohlmann and the MiKlip-Team providing the original MiKlip-REF (Baseline1) prediction system data; MiKlip (www.fona-miklip.de) is funded by the German Federal Ministry for Education and Research (BMBF), projects MiKlip-INTEGRATION (FKZ: 01LP1160A), and MiKlip-II INTEGRATION (FKZ: 01LP1519B); all simulations were carried out at the German Climate Computing Centre (DKRZ), all evaluations were carried out with the MiKlip Central Evaluation System (www-

miklip.dkrz.de) on the MiKlip server of the DKRZ, which also provided all major data services. Method and used software are cited and can be obtained via open access. The MiKlip-EDF model output is available through PANGAEA data publication: Kadow et al. [2017] (doi:10.1594/PANGAEA.874231). The source code for the model used in this study, the MPI-ESM, is available at http://www.mpimet.mpg.de/-en/science/models/mpi-esm. Both the data and input files necessary to reproduce the experiments with MPI-ESM are available from the authors upon request (Christopher.Kadow@met.fu-berlin. de). The raw model data are archived at the German Climate Computing Center (DKRZ). Correspondence and requests for materials should be addressed to Christopher Kadow (Christopher. Kadow@met.fu-berlin.de). The authors also acknowledge the anonymous reviewers whose contributions greatly improved the manuscript.

4.5 Appendices to Chapter 4

The supplementary information consist of additional data sets analyzed, a mean state analysis of MiKlip-REF and MiKlip-EDF, scientific side aspects, and technical difficulties during the production process.

Mean State Analysis

Supplementary Figure 4.6 shows the differences between the temporal mean of the sea surface temperatures of MiKlip-EDF and MiKlip-REF for the period from 1979 to 2013. In general, MiKlip-EDF produces warmer high latitudes and colder tropics. The most pronounced temperature difference can be found in the Sub-polar Gyres of the Pacific and Atlantic. MiKlip-EDF is up to 2K warmer in the central North Atlantic than MiKlip-REF. Earth system models, including the MPI-ESM-LR, produce too low temperatures in the North Atlantic in comparison to observations. MiKlip-REF already reduces the so called North Atlantic cold bias in contrast to the uninitialized runs by up to 0.5K (not shown). The forecasting technique suggested here leads to an even stronger reduction of this SST cold bias. The cold bias is not completely eliminated and other biases remain stable. However, this corrected inner-model shift in the energy budget of the North Atlantic is followed by a more accurate prediction not only of the near-surface air temperature but also of associated variables. Experience in numerical weather and seasonal forecasts showed, that skill can be considerably improved by reducing model systematic error (Keenlyside et al., 2008), which can be confirmed by this study as well.

Scientific Side Aspects

Changing just one parameter in a physical consistent model could cause inconsistency. The differences in the ensemble members are small enough (particularly in the deep ocean) that the ocean model re-adjusts to temperature shifts, similar, as it does with nudging or initialization procedures. An assimilation strategy to adjust the salinity and sea ice cover as well, were computationally too expensive in the current experimental set-up, but could be considered in future experiments.

Technical difficulties and study adjustments

The ocean component MPI-OM of the MPI-ESM in its current and applied version is not able to provide output when stopped within a year because of default output of yearly means. This is causing run-time errors. Therefore, there was a need to switch off the MPI-OM output. Accordingly, we did not had a chance to analyze ocean data beyond surface. The focus in the study was the atmospheric climate system driven by the ocean. In future experiments a new version of the MPI-OM or an experimental re-design is necessary to analyze the deep ocean behavior as well.

The introduction of the sub-ensemble mean temperature dataset during the ensemble dispersion filtering of MiKlip-EDF may lead to model instabilities. This happened in some rare cases (14 out of 3800) of the ensemble mean building and its following restarts, mainly in the atmospheric component ECHAM6. As a simple solution the failing member was not replaced and run for 6 months instead of 3.

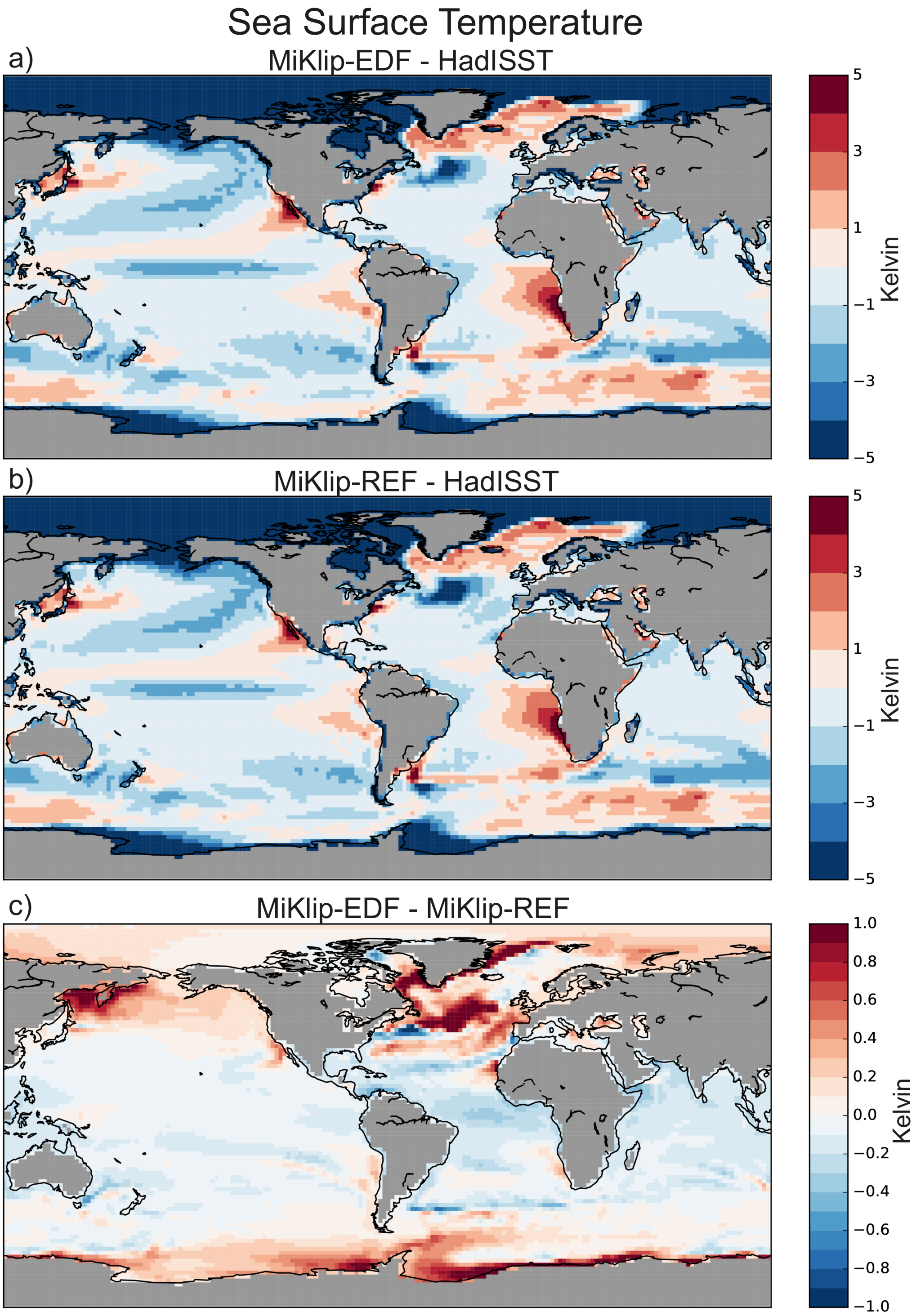


Fig. 4.6: Sea surface temperature differences in Kelvin by showing to show the mean bias. MiKlip-EDF (a) and MiKlip-REF (b) minus HadISST as well as the difference of MiKlip-EDF and MiKlip-REF. The LY2-5 analysis covers the period from 1979 to 2013 and the HadISST temporal mean of 1979 to 2013 is interpolated onto the MPI-ESM-LR grid.

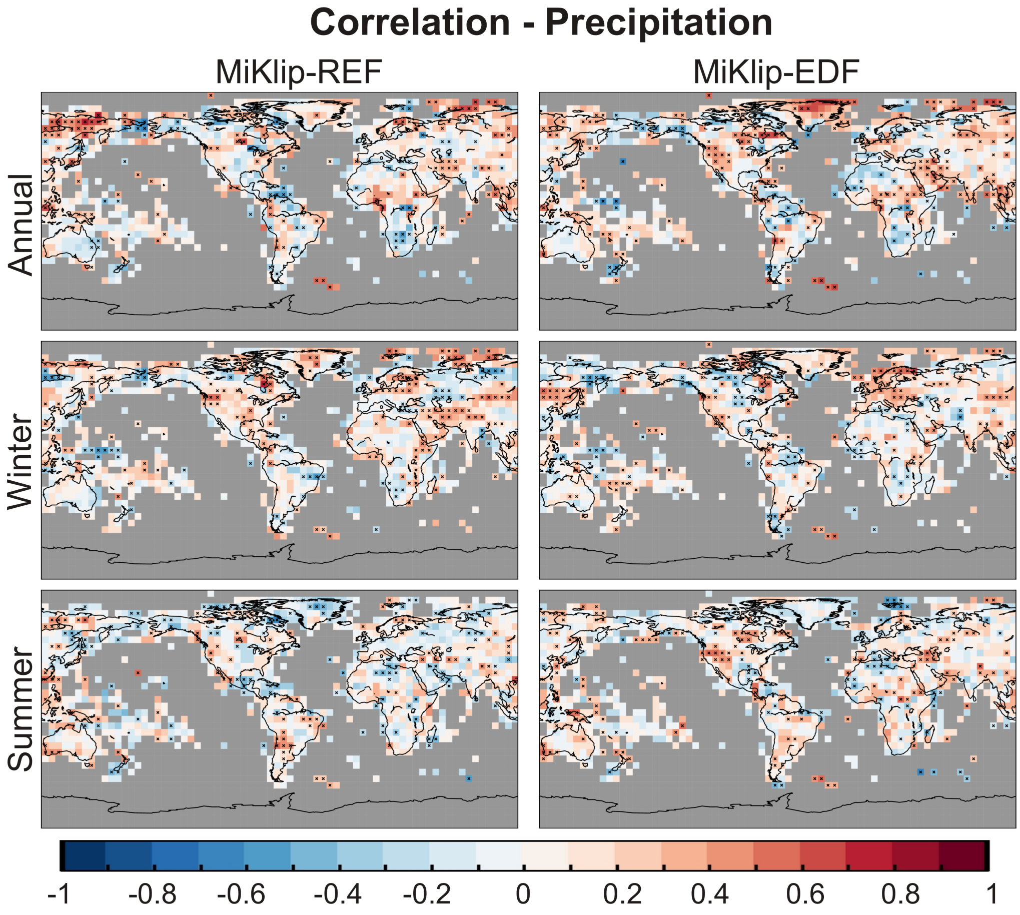


Fig. 4.7: Correlations of MiKlip-REF (left) and MiKlip-EDF (right) for precipitation compared to GPCC observations for the annual LY2-5 (top – comparable with Figure 4.5a, b), the winter (DJF) LY2-5 (middle), and the summer (JJA) LY2-5 (bottom) over the period from 1979 to 2013. Significance is marked by black crosses. Grey shading indicates missing data of GPCC.

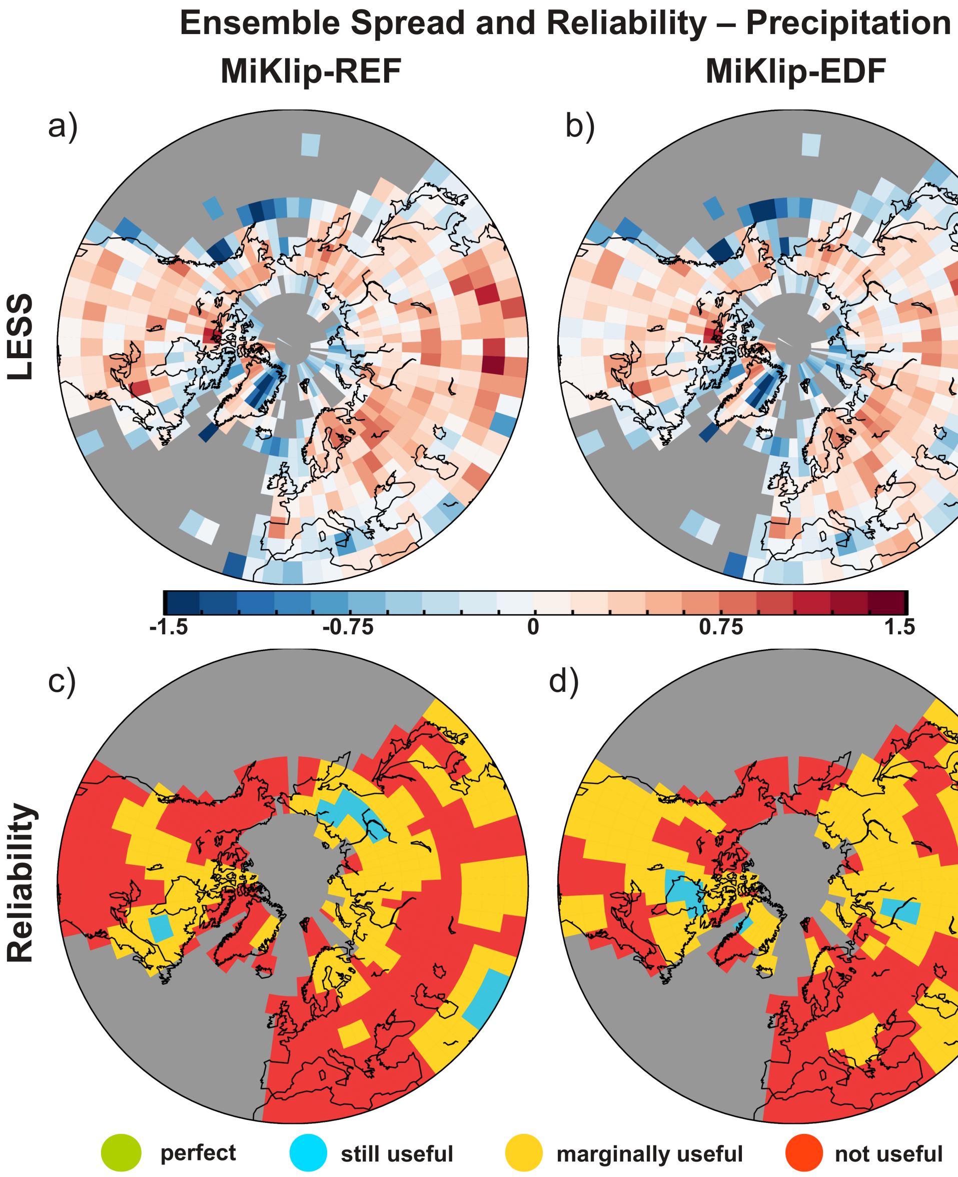


Fig. 4.8: Logarithmic Ensemble Spread Score (top) and Forecast Reliability (bottom) for Precipitation in MiKlip-REF (left) and MiKlip-EDF (right). Analyses are done for LY2-5 compared to GPCC. Gray areas indicate missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013.

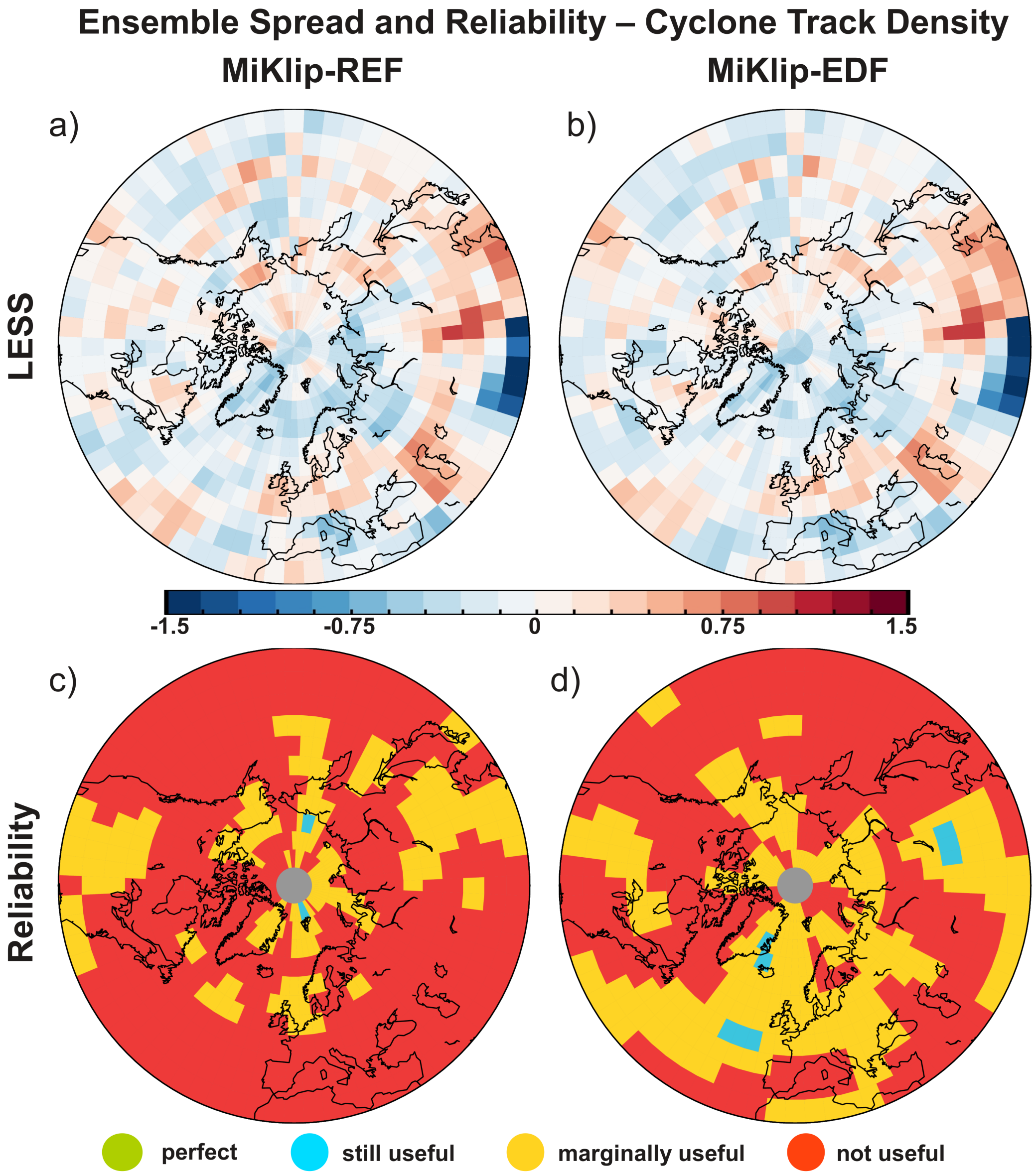


Fig. 4.9: Logarithmic Ensemble Spread Score (top) and Forecast Reliability (bottom) for Cyclone Track Density (DJF) in MiKlip-REF (left) and MiKlip-EDF (right). Analyses are done for LY2-5 compared to ERA-Interim. Gray areas indicate missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013.

Information

Information about the candidate's and co-authors' work on that paper:

The **candidate** developed the scientific idea of the paper -namely the Ensemble Dispersion Filter for decadal predictions-; designed the research question for exploiting sources of skill to improve predictions, implemented the technical adjustments into the Earth system model; installed and applied the climate model in the new setup; evaluated the EDF and its reference system; analyzed the results; drafted the paper.

The **co-authors** provided software to evaluate the results, and helped to improve the text of the paper by numerous comments and took care of the funding.

5 Summary, Discussion, Outlook, and Conclusion

5.1 Summary and General Achievements

The work at hand focused on the analysis, evaluation, and the development of decadal climate prediction systems in order to contribute to their fundamental understanding as well as to the improvement of their actual forecast capability.

The research of this thesis started with a development of an efficient computational and systematic approach in climate science verification shown in Chapter 2. An evaluation framework called Free Evaluation System Framework for Earth System Modeling ('Freva') was designed, developed, and applied. The main system design of Freva features the common and standardized model database of climate modeling and in particular decadal prediction hindcast sets. A programming framework for efficient verification routines, and a history system of evaluations to keep track of the research was established. Freva constituted the foundation to use the growing knowledge around decadal prediction in the MiKlip Central Evaluation System to detect improvements for further developments in combination with the evaluation strategy of Chapter 3.

The research in Chapter 3 showed that a decadal forecast with the numerical Earth system model of MiKlip is possible and scientifically sound. There was a need to set up a statistical framework for this young research branch and the MiKlip project. The combination of accuracy and spread metrics and the addition of new verification metrics made precise evaluations of the decadal prediction system possible. The investigation exposed the system's sources of potential skill (initial values, boundary conditions, assimilation strategy, ensemble size, etc.). The importance of the memory in the ocean (model) was confirmed as well. The investigation also showed the difference in skill on the short (first year) and long term (up to a decade) prediction. The prediction skill of the first lead years is much higher than the long term prediction skill several years ahead. However, the high skill in the first year is promising for the multi-annual forecast. The ability to synchronize the climate model with observations confirms the capability of the numerical models, which is evident in the early years of the decadal forecast. The potential that this synchronization is extendable to subsequent years was investigated in Chapter 4.

With the help of the systematic evaluation (Chapter 3) and the efficient application (Chapter 2), scientific findings around the MiKlip decadal prediction system led to a novel forecast technique. Two aspects that are important in climate predictions inspired this approach. [1] The ocean memory due to its heat capacity holds a large potential of forecast skill - applicable with a fully coupled Earth system model. [2] Instead of applying one single member prediction, exploiting the whole ensemble and its ensemble mean, does improve a prediction system. The combination of both effects forms this new approach - named Ensemble Dispersion Filter (EDF). The climate prediction skill of the MiKlip ensemble prediction system was improved by a shift of the model's ocean state towards the ensemble mean of its individual members at seasonal intervals. Applied as an add-on to the reference prediction system MiKlip-REF, the EDF led to even more accurate results. Global mean and regional temperature, precipitation, and winter cyclone predictions showed increased skill up to 5 years ahead. Especially the later lead years (LY4 and LY5) benefited from the EDF. This forecast technique was a new effort to combine sources of skill to improve a decadal prediction system during model run-time.

In the following sections, each of the research tasks stated in the beginning will be revisited (Sect. 1.1). The corresponding contributions from Chapters 2, 3, and 4 will be summarized, discussed, and interlinked. Alongside a Discussion (Sect. 5.2) with subsequent research questions regarding the EDF, which provide a deeper insight into the applied new methodologies, the Outlook section (Sect. 5.3) provides follow-on ideas. Concluding remarks (Sect. 5.4) set the results into a broader context and their scientific relevance is highlighted.

5.1.1 Research for scientific efficiency and reproducibility (RT1)

Scientific progress is usually really fast, when research breaks new ground. This was also true for decadal climate prediction. Within a decade after the first publications about decadal prediction [Smith et al., 2007, Keenlyside et al., 2008, Pohlmann et al., 2009], research evolved fast and diverse. Even the MiKlip project investigated several and different areas of potential improvements and it had several development stages (Marotzke et al., 2016). Additionally, decadal prediction consists of large and complex (e.g. lead year) data structures, which was a data challenge on top of the scientific challenge. There was a fundamental need for an efficient way to keep track of the scientific developments, implementations of scientific software of statistical frameworks (see above), organization of research data and its exchangeability, and scientific reproducibility and traceability. This issue has been addressed by the first research task:

Development and implementation of an evaluation system for decadal climate prediction to verify enhancements of skill in different development stages with the full flexibility of *model and observational data comparison in a sophisticated, reproducible, and efficient way.*

The Chapter 2 showed in combination with Chapter 3 and 4 that the climate modeling science and as an example decadal prediction research can be enhanced with the appropriate software to faster detect and accomplish scientific improvements. A root cause analysis of data-driven climate science showed challenges and potential solutions for the efficient evaluation in Earth system modeling. The new scientific software development Freva incorporates several data sets and verification routines into a common software and data framework. The ingested knowledge and toolset overcomes technical hurdles to efficiently test new prediction systems on high performance computers. It also combines the knowledge of the scientists into one easily applicable evaluation system to connect the scientists through their work. The software efficiency on HPCs with web and shell front ends leads to faster and easier applications and results, which enhances the science by building bridges from HPCs to the scientists.

In combination with the other Chapters 3 and 4 and other studies which applied Freva (Sect. 2.3), the publication showed that climate verification benefits from explicit software designs on HPCs. The times are over, where climate model data production is performed on modern HPCs and the rest of the work including verification takes place on the PC of the scientist (Ranilla et al., 2014). Evaluation systems filled with scientific routines and framed by Freva can be raised to a similar level as climate models on HPCs. Modern software engines like Freva give climate scientists a common and efficient developer's base.

5.1.2 Statistical framework for decadal scale evaluation (RT2)

Whenever a new research branch - as in this case decadal climate prediction - is born, there is a fundamental need to evaluate its scientific results. As decadal prediction is the interface between short-range climate forecasts and long-term century climate projections, it benefits and suffers from both sides in terms of evaluation. It needs to take into account that the initialization and its potential obstacles take a major role in the beginning of the forecasts. The same accounts for the boundary conditions of the forcing on the later stage of a forecast. As evaluation strategies were already developed for short and long term, there was a need to combine both evaluation types for a systematic evaluation strategy for decadal prediction. This is in particular important to keep track with the decadal prediction developments - to evaluate its development stages. This issue has been addressed by the second research task:

Formulate and incorporate a systematic and comprehensible statistical framework for decadal climate prediction into the evaluation system and fully assess a prediction system to reveal scientific plausibility, prediction skill, and sources of potential skill.

In Chapter 3 a detailed statistical framework for a decadal scale climate prediction system was set up and applied with the MiKlip prediction system. An actual decadal forecast is combined with evaluation metrics regarding the forecast system's hindcast set. The accuracy assessment of the prediction system is combined with its ensemble spread assessment. The evaluation takes the observed climate development into account. The decadal prediction system skill is assessed by comparing with climatological, uninitialized, different sized ensembles, and detrended hindcasts in different variables. The systematic evaluation of different variables -in terms of temperature and precipitation forecasts on different lead times- is applied in one common framework. Therefore it is reaching comparability and comprehensibility across target values.

While the accuracy assessment of the decadal climate prediction community (Goddard et al., 2013) is quite robust, the spread assessment turned out to be quite raw (Sect. 3.2.2). It is shown that the proposed (Goddard et al., 2013) skill measurement for spread is not satisfying (Fig. 3.1). With the developments of the Logarithmic Ensemble Spread Score (LESS) and Logarhytmic Ensemble Spread Skill Score (LESSS) in Chapter 3 missing information about the spread behavior could be revealed. Too large or too small spreads of decadal forecast systems can be detected with the LESS. If a decadal climate prediction system is for example improving due to a larger ensemble size, this can be revealed by the skill score of the LESS, namely the LESSS (Fig. 3.6).

The study showed an actual climate forecast next to the evaluation strategy. This enabled a regional validation of the forecast - whether it may or may not be trustworthy. This new verification framework became the basic evaluation for forecast verification within the MiKlip project and led to its main forecast evaluation strategy of the MiKlip project[1].

This statistical framework, developed for decadal predictions, showed that the MiKlip prediction system improved, compared to the climatological forecast and the uninitialized climate projections. The applied ocean initialization showed a forecast enhancement compared to the uninitialized system. The extensive standardized assessment revealed that the MiKlip reference system and decadal predictions in general leave room for improvements especially beyond leadyear 1. The skill for the first year is substantially better than for the subsequent years. This means that the

[1] www.fona-miklip.de/decadal-forecast

MiKlip decadal prediction system loses skill pretty fast, which was investigated in the following research task.

5.1.3 Exploit sources of skill to improve decadal prediction (RT3)

Decadal prediction research already underwent several development stages. The investigations on how to improve this research is an ongoing process. It is closely linked to the decadal predictability research focusing on what is possible, what are the boundaries, and what are sources of potential skill. The research of the prior Chapters of this thesis helped to identify obstacles and possibilities of the MiKlip decadal prediction system. In particular, the ocean 'memory' on the decadal time scale, as well as the advancement of the prediction using a larger ensemble got confirmed. This thesis took these two main aspects into further investigation. These have been addressed by the third research task:

Having a fully assessed and skillful decadal prediction system at hand, exploit detected sources of potential skill to further improve the decadal prediction system.

The earlier chapters showed that we know, we can identify, and we can confirm sources of skill for decadal scale prediction of the climate system. In Chapter 4 the newly developed forecast technique, the EDF, was the first effort to actually combine and exploit the following sources for a decadal prediction system: the ocean memory and the ensemble mean. The ensemble mean application on the 3-dimensional ocean temperature (heat capacity) evolution during the run-time of the model (every 3 months) significantly improved the forecast of the MiKlip prediction system on the time-scales up to 5 years ahead.

Many standard improvement strategies like initialization techniques, higher resolution, different ensembles sizes, etc. were clearly outperformed by the new EDF forecast technique (Fig. 4.4). No other MiKlip forecast system, using the same model system and setup of the MiKlip reference system Baseline1, outperformed Baseline1 like the EDF did in this study. The decadal prediction community within and outside of the MiKlip project put a lot of effort into an optimization of initialization methods to exploit observational initial values - another source of potential skill improvement. New model and prediction systems arose after the publication of the Chapter 4. As the EDF is an add-on technique being applied during the forecast, the EDF is not in competition with initialization (e.g. ensemble Kalman filter) or ensemble generation (e.g. breeding) techniques, which are applied before or at the start of the forecast. In the future, it is conceivable that a smart combination of methods applied before, during, and after the forecast will be used in decadal prediction research. Therefore, many scientific and technical factors will be brought together and are hard to separate when analyzing the skill. The new and fast methods of the prior

Chapters 2 and 3 around Freva help to improve the efficiency of the evaluation of prediction development stages as they did with the evolution of the EDF prediction system.

The EDF demonstrated that it is worth to study methods which go beyond standard procedures. It is possible to exploit sources of skill to improve the prediction skill significantly. Related ideas -similar to the EDF- which exploit sources of skill should be investigated as well and are discussed in the following Section 5.2.2. New scientific and technical methods should be developed (see Outlook 5.3), and scientists should be encouraged to explore novel, unconventional forecast methods.

5.2 Discussion

The thesis showed the development of a new forecast technique within decadal climate science. The Ensemble Dispersion Filter (EDF) improved the prediction of several climate parameters. However, this method leads to additional questions, which will be discussed and investigated in the following:

5.2.1 Effect of the Ensemble Dispersion Filter after its first application

The EDF mainly improved the skill in the later forecast years - lead years 4 to 5 (LY4-5). Is the decadal prediction of the EDF further along the line (LY4 and LY5) more accurate, because the EDF already improves the prediction in its early stages maybe even after its first application?

The main idea behind the EDF is to improve each single ensemble member by making use of the ensemble mean, and in turn to improve the whole prediction system in the long run. The analysis of the lead months (LM) 3 and 4 of the ensemble mean and the single members shows that the EDF already improves the system at the beginning of the forecast (Fig. 5.1). By design MiKlip-REF and MiKlip-EDF are identical in LM3 in the ensemble mean and for each member (as in Fig. 4.1a). However, in the first month after applying the EDF on the ocean temperatures (LM4), it is already apparent that there is an improvement in the temperature prediction in each single member compared to its un-filtered counterpart MiKlip-REF (Fig. 5.1d, f, h, j, l). This finding supports the main idea of improving each member with the common ensemble to improve the whole ensemble in the long run. The improvement is not yet seen in the ensemble mean of MiKlip-EDF compared to MiKlip-REF (Fig. 5.1b). Indeed, the ensemble mean of the atmospheric state in MiKlip-REF in LM4 is still

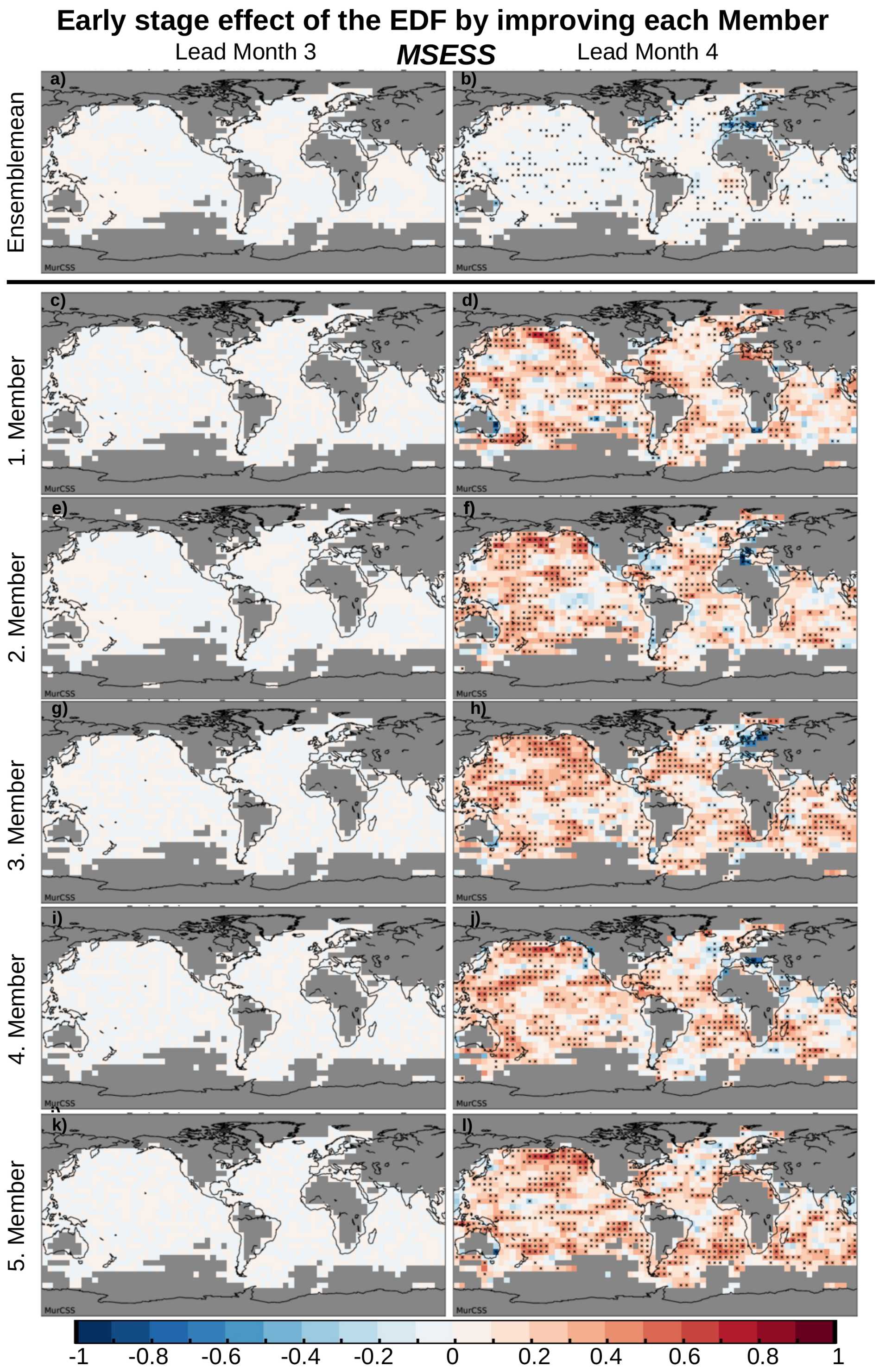


Fig. 5.1: Early stage effect of the EDF in the first month after application: The MSESS shows the improvement of MiKlip-EDF compared to its reference prediction MiKlip-REF with observations of HadCrut4 in the ensemble mean (top row) and each member before the EDF is applied (LM3 - left) and after (LM4 - right). Figure is constructed as Fig. 4.2 to be comparable with - established by the common evaluation system.

very close to the actual development, i.e. the observations. However, this effect gets lost along the forecast as seen in Chapter 4, which favors the application of the EDF (Figure 4.1). In the long run it could still be an additional positive effect, that a low pass filtering reduces noise, through the averaging. This should be investigated in the future. Therefore it is out of the scope of this thesis.

5.2.2 Effect of regions and forecast developments in predictions

If the early time period is important for the EDF, how important is the early (e.g. first year's) development of a forecast for its later stage? Are specific regions important for a decadal climate prediction?

A larger independent ensemble than the MiKlip-EDF or MiKlip-REF needs to be applied in order to answer that question. This was tested in the following additional study set-up. The most comprehensive prediction system currently available is the MiKlip Prototype system. It consists of 15 members using the same initialization data sets as MiKlip-REF, but applies full-field initialization in the ocean (MiKlip-REF-FF in Chapter 4).

Method Assuming we know the development of the climate in the first year, we select the best (worst) 5 members of MiKlip Prototype which are closest (furthest) from the sea surface temperature observation by using the root mean squared error (RMSE) and correlation in the specific field mean of the regions: Global, ENSO, PDO, SPDO, NA (Figure 5.2 and Table 5.1). With the *RMSE-Selection* (Fig. 5.3) we capture the smallest (largest) distance and with the *Correlation-Selection* (Fig. 5.4) we capture the most (dis-)similar variability of the investigated regions within the first prediction year. The evaluation of LY2-5 is performed with Freva from Chapter 2 and its implemented evaluation strategy from Chapter 3.

Result The RMSE selection of the regions Global (Fig. 5.3a) and PDO (Fig. 5.3e) show the highest added value in the MSESS, especially in the Pacific region. The selection of the RMSE in the ENSO region show some added value in the LY2-5 correlation (Fig. 5.3d). The NA region shows some improvement in the ENSO region, but more importantly an added value over Europe (Fig. 5.3i). The effects on the SPDO region are negligible in the RMSE-Selection comparison (Fig. 5.3g). However, the SPDO shows the most promising results in the Correlation-Selection (Fig. 5.4g, h). The correlation is significantly better in the regions ENSO, SPDO, and eastern Pacific. The most interesting result can be seen in the NA region correlation selected LY2-5 analysis (Fig. 5.4i, j). Teleconnection patterns between North Atlantic and Pacific emerge. Figure 5.4j shows that the ENSO region is much better predicted in LY2-5, if the evolution of the NA region is better predicted in the first year.

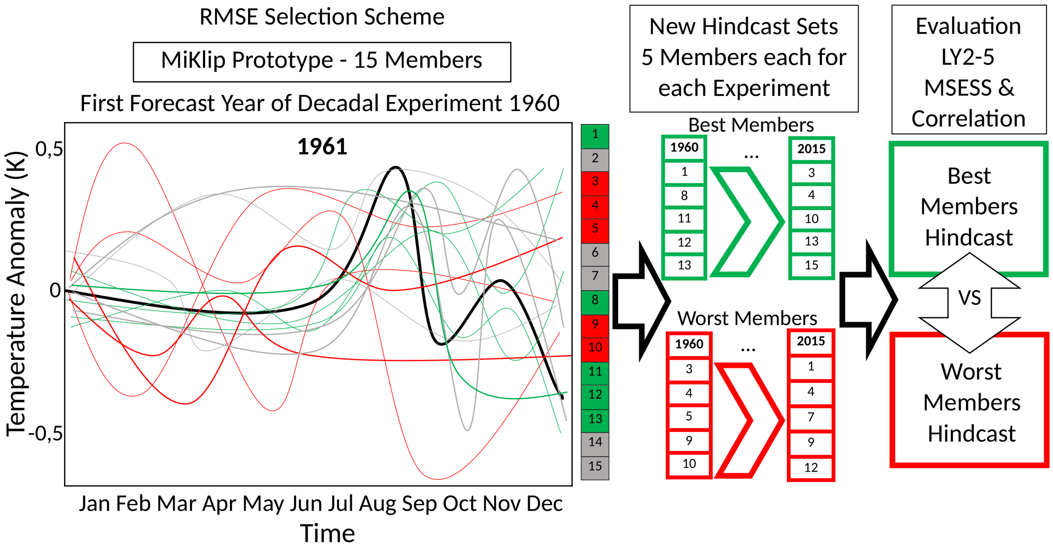


Fig. 5.2: Schematic experiment setup for the RMSE selection (similar approach for the Correlation selection). The decadal hindcast experiment 1960 starts in 1961. The 15 member of the MiKlip Prototype system (thin lines) are compared against an observation (thick black line) in a regional/field (Global, ENSO, PDO, SPDO, NA) mean. The 5 members with the smallest (green) and biggest (red) RMSE get selected. Unselected members in gray. This procedure is done for all decadal experiments between 1960 and 2015. The selected members form two hindcast sets: Best (Worst) Members in green (red) which represent the hindcast with the closest (farthest) development compared with observations in the 1st forecast year. These two hindcast sets get evaluated in the lead years 2 to 5 in terms of MSESS and Correlation compared to observations.

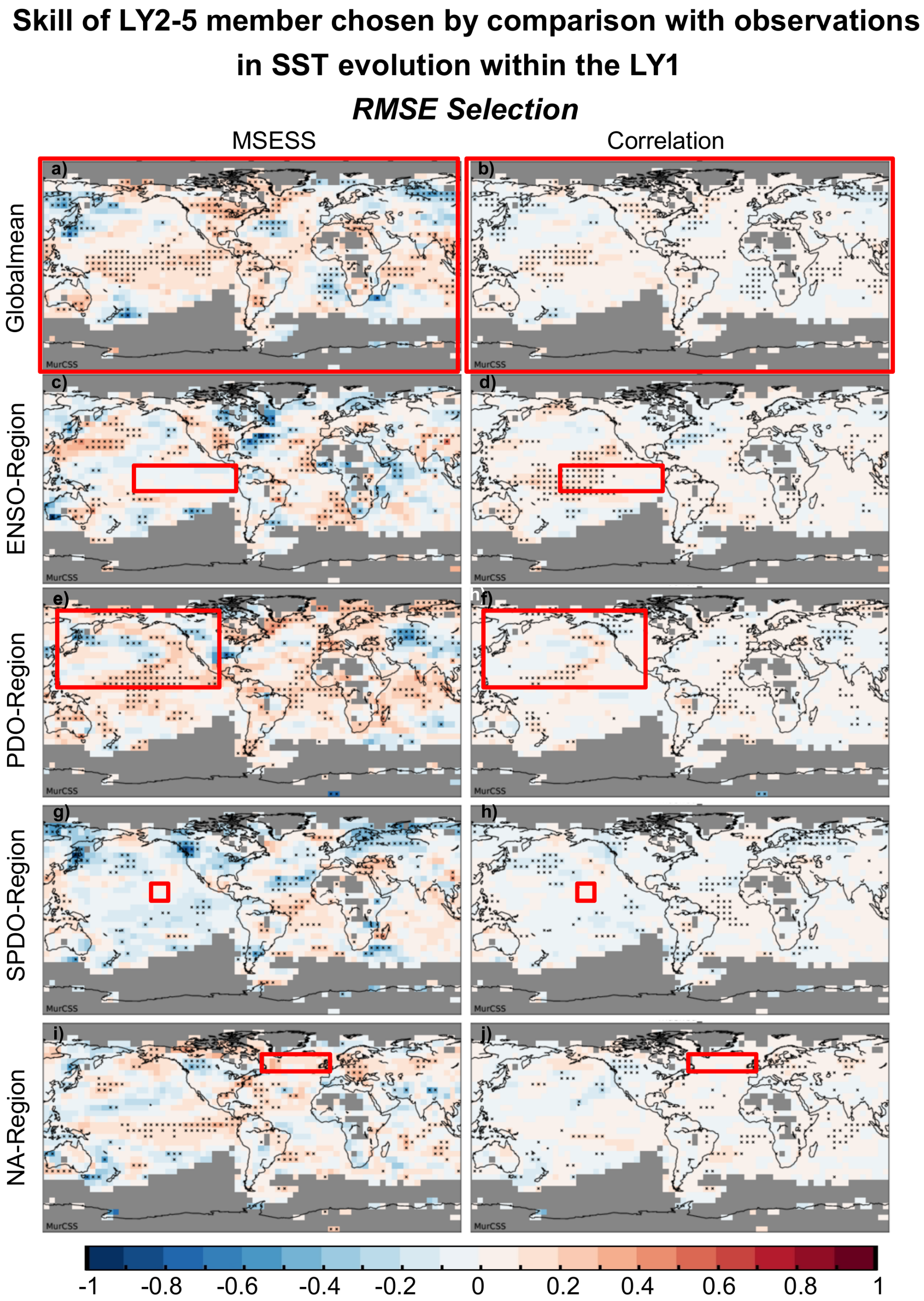


Fig. 5.3: Skill analysis of the LY2-5 MSESS in special sub-set of the MiKlip Prototype system. Selection of 5 member being (not) close to the first year evolution (12 month) to the observations by the (biggest) smallest RMSE in a specific regional field mean shown in red boxes. Crosses denote values significantly different from zero exceeding at a 5% level applying 1000 bootstraps. Gray areas indicate missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013.

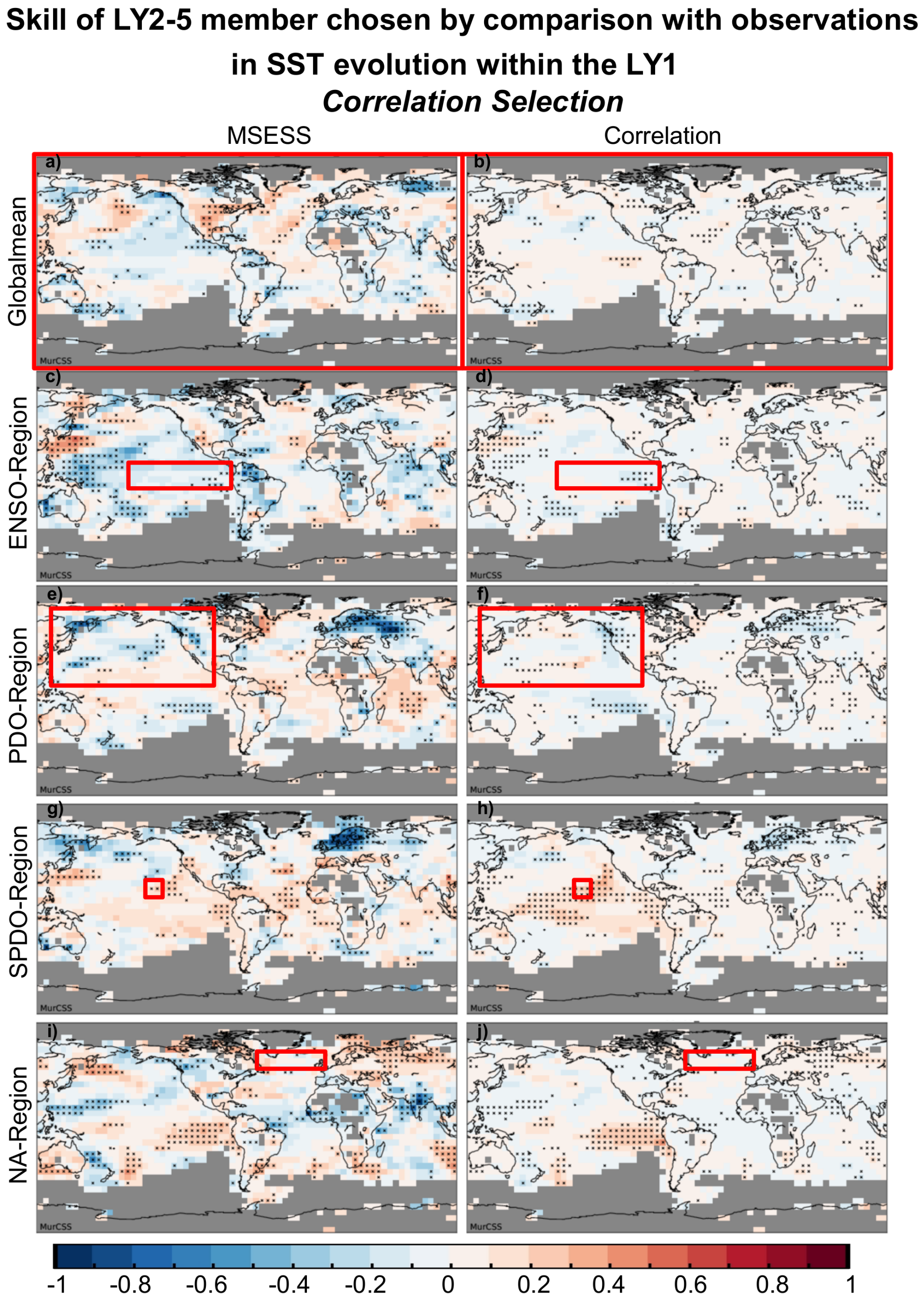


Fig. 5.4: Skill analysis of the LY2-5 MSESS in special sub-set of the MiKlip Prototype system. Selection of 5 member being (not) close to the first year evolution (12 month) to the observations by the (lowest) highest correlation in a specific region field mean shown in red boxes. Crosses denote values significantly different from zero exceeding at a 5% level applying 1000 bootstraps. Gray areas indicate missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013.

Conclusion The skill in key areas is important within the first year (LY1) of the prediction, when forecasting several years ahead (LY2-5). However, this relation needs further investigation. More independent ensemble members are necessary to clearly distinguish between 'close' and 'not close' to the observations. The 15 members system (all members of MiKlip Prototype) is still better than the 5 selected ones which show positive effects (not shown). Also, it is not clear what would be necessary to transfer this idea to a decadal prediction system, besides the fact that we need to improve the seasonal system as well. The basic idea and the method shown here is set in a relation to the EDF technique, and is discussed in the Outlook section (see also 5.3 - key word *Meta-Selection*). In particular, the 5 member EDF system is still significantly better than the best 5 members of MiKlip prototype analyzed in this section (not shown).

A regional detection experiment, which allows to investigate the effects on the global scale, is already part of the DCPP as part of CMIP6. The DCPP suggests an experiment with observed SSTs in key regions in a historical (uninitialized) run as in Kosaka and Xie, 2013. This would show the effect of a specific region in a numerical model to the climate development compared with observations. The results shown in the paragraph support these additional experiments from another standpoint, especially the teleconnectivity should be explored (Fig. 5.4j). However, these sensitivity experiments are outside the scope of the thesis, but should be considered in follow-up studies.

Name	**Explanation of Region for Field Mean**	**Region Lon/Lat**
Global	Whole Globe	180E 180W / 90S 90N
ENSO	'El Nino Southern Oscillation Index' Reg	180W 70W / 20S 20N
PDO	'Pacific Decadal Oscillation Index' Reg	160E 120W / 20N 60N
SPDO	(S)mall part of the PDO Region	160W 150W / 15N 25N
NA	North Atlantic Region	60W 0 / 50N 65N

Tab. 5.1: Overview table of used region for the field-mean Correlation- and RMSE-Selection - red boxes in Figures 5.3 and 5.4.

5.3 Outlook

The original purpose of the EDF was its application to decadal predictions. Therefore, the selection of a variable to be filtered by the EDF led to ocean temperatures. However, other variables could be tested as well. The adaptation of ocean salinity is an obvious counterpart of ocean temperatures. Also, other variables, e.g. land variables, sea ice area fraction, or the quasi-biennial oscillation should be examined to study the seasonal or multi-annual implications of the EDF. Other time intervals than the 3-months when applying the EDF, are worthwhile to explore as well. Monthly as well as yearly iterations will grant more insights to this technique.

Another focus should be on the amount of ensemble members and how they interact with each other. As mentioned in Chapter 4, the development of independent ensemble bundles of the EDF (like 5 members each) would be an interesting add-on. In addition, the mixture of several bundles could open new insights. Such a mixture setup would use a 10 member bundle A, where only 5 of its members exchange ensemble mean information with 5 members of a different 10 member bundle B. This could answer the question, if we could keep a larger ensemble spread in the EDF experiments.

This thesis analyzed only forecasts with 5 years into the future. What is the effect of the EDF after 5 years on the second pentad? The research on a longer than the so far analyzed five years' time scale as well as the full range of state-of-the-art decadal prediction systems, from the 1960s onwards, should be investigated with a new setup of model runs.

The EDF takes the value of each independent member to compute the ensemble mean. It improves each member itself (Fig. 5.1) and also the whole prediction system (Fig. 4.2). With the results shown of the first year evolution and the importance of regions (Fig. 5.3, 5.4), two alternative approaches are closely related to the EDF: (1) A majority vote of members could get weighted towards the most probable climate development. (2) A meta-selector could use a meta function to select good and poor predictors of the climate system. These two methods have the potential to be transformed into a climate simulation approach similar to the EDF. They are discussed in the following, not been done in this thesis, but could be seen as ideas for follow-on studies.

The majority vote idea could be called 'Ensemble Majority Vote' (EMV) of members. It is an interesting approach within a decadal prediction system. An ensemble runs for one forecast year and then the majority of members is selected, which point into a certain direction of the climate system. The selection checks the first year similar to the analysis shown before (Sect. 5.2.2), without comparing it to the observation, but the ensemble members itself. Next to the analysis shown in temperature, it could check for other important climate indicators like sea ice, ocean heat content, the Atlantic overturning circulation, etc. After the selection one ignores the minority members and restarts only the majority members for the next year(s) and so on. Applying slightly perturbed versions of the majority members to increase the member size again to its original size, should be applicable. However, the ensemble size is still the biggest problem when considering EMV–it is by far too small when thinking about a majority vote. A size of 100 members is probably the smallest ensemble which should be applied, to infer a clear percentage vote from distribution (Wilks, 2005). With increasing computing power in the future, the application of this methodology to numerical models in the climate prediction research seems feasible and promising.

The meta selection idea could be called ’Ensemble Meta Selection’ (EMS) of members. It is similar to EMV, but would use a meta level for a weight function or selection of members. This meta level selects between ’good’ and ’bad’ members. ’Good’ and ’bad’ needs to be defined beforehand and on a science basis. In climate science, this would be as the analysis applied before, by selecting the first year’s evolution of climate indicator regions [see Section 5.2.2]. The alternative would be, that we use ’last’ year’s forecast to predict the next decade and evaluate the forecast in its first year. This is possible, but probably inefficient. At present the first forecast year is always the one with the best skill (Fig. 4.2). The second forecast year needs to outperform this first forecast year, and the third needs to outperform the second, etc. Another meta level selection could be the suspension of unusual extremes, which are clearly disproportionate. Finding the right meta selection for the climate evolution could be challenging. The weight of members and the decision about which region, variable, tendency, etc. is the most important and beats the others in a selection ranking is not straightforward to be made. Already the selection of regions and selection method (RMSE- and Correlation-Selection in the first year) as described in Discussion Section 5.2.2 is challenging. At least EMS does not need as many ensemble members as EMV.

The ideas and methods discussed could get investigated as well. Could the EMV or EMS outperform the EDF? One advantage of EMV and EMS is due to the fact, that each ensemble member runs untouched as a ’physical’ solution of the model. Each of the described methods is worth to be followed-up in subsequent studies. The application of the EDF approach already shows very promising results. At present, the EDF approach is the one which is the most feasible to apply. Simply due to the fact, that it does not need so much computing power as EMV or EMS. Therefore, the next steps should be the enhanced exploration of the EDF including more ensemble members, independent bundles, longer time ranges of up to ten years, etc. in decadal prediction. The three mentioned techniques (EDF, EMV, EMS) open new ways of establishing forecasts. These techniques are related to approaches in machine learning (Bagging, Boosting, Stacking - Hastie et al., 2013). It should be beneficial for climate science and machine learning to learn from each other.

The in-depth evaluation strategy as shown in Chapter 3 should be the base of evaluations in future studies of decadal climate prediction. However, the evaluation strategy itself must be evaluated from time to time and maybe extended in the future by probabilistic and reliability measurements as already indicated in assessment of the EDF in Chapter 4. The evaluation system framework Freva as shown in Chapter 2 should be a sustainable add-on to climate studies. Freva could be a beneficial tool for other applications beyond decadal climate prediction.

5.4 Conclusion

To conclude this study, the development of a new forecast technique –namely the Ensemble Dispersion Filter– improves the decadal climate prediction research. Other configurations or slightly different ideas close to the EDF have already been discussed in the outlook, which could potentially even improve the results shown in this thesis. The advancement of prediction skill in the important climate parameters temperature, precipitation, and cyclones encourages to conduct further investigations. Other forecasting disciplines, such as seasonal prediction or even fields outside of climate research, for example weather or space science, might benefit as well from this method. Novel forecast techniques highly benefit from efficient evaluation systems. The creation of a verification strategy which includes new metrics in decadal climate prediction have become an essential element of decadal climate prediction. The detailed analysis of the MiKlip reference prediction system and its first scientific decadal forecast provide the base for subsequent development steps. The development of an efficient evaluation system tool –namely Freva– for model data and evaluation procedures enhanced the research around the EDF. As numerical climate models and verification software are executed on modern high-performance computers, there is a scientific need for climate research infrastructures like Freva to conduct evaluations in a reproducible but efficient way. The evaluation strategy improved the research and was used in many other studies (see Bibliography). In summary, this thesis showed the synthesis of scientific improvements in decadal climate predictions, due to enhancements in climate research, climate modeling, and climate evaluation.

Bibliography

Research of the publications by this cumulative dissertation influenced several other studies. Here, related co-authored publications of the PhD candidate within the field of decadal climate prediction are presented:

MiKlip: A National Research Project on Decadal Climate Prediction,
J Marotzke, WA Mueller, FSE Vamborg, ... **C Kadow**, ..., U Ulbrich, Bulletin of the American Meteorological Society 97 (12), 2379-2394 9, 2016 DOI: 10.1175/BAMS-D-15-00184.1

MurCSS: A tool for standardized evaluation of decadal hindcast systems,
S Illing, **C Kadow**, K Oliver, U Cubasch, Journal of Open Research Software 2 (1) 9, 2014 DOI: 10.5334/jors.136

Improved forecast skill in the tropics in the new MiKlip decadal climate predictions,
H Pohlmann, WA Mueller, K Kulkarni, ..., **C Kadow**, . . . , J Marotzke, Geophysical Research Letters 40 (21), 5798-5802 44, 2013 DOI: 10.3402/tellusa.v66.22830

Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, T Kruschke, HW Rust, **C Kadow**, WA Mueller, H Pohlmann, GC Leckebusch, Meteorologische Zeitschrift 25 (6), 721-738 17, 2016 DOI: 10.1127/metz/2015/0641

Decadal hindcasts initialized using observed surface wind stress: evaluation and prediction out to 2024, M Thoma, RJ Greatbatch, **C Kadow**, R Gerdes, Geophysical Research Letters 42 (15), 6454-6461 15, 2015 DOI: 10.1002/2015GL064833

The impact of stratospheric volcanic aerosol on decadal-scale climate predictions,
C Timmreck, H Pohlmann, S Illing, **C Kadow**, Geophysical Research Letters 43 (2), 834-842 10, 2016 DOI: 10.1002/2015GL067431

Evaluating decadal predictions of northern hemispheric cyclone frequencies,
T Kruschke, HW Rust, **C Kadow**, GC Leckebusch, U Ulbrich, Tellus A: Dynamic Meteorology and Oceanography 66 (1), 22830 10, 2014 DOI: 10.3402/tellusa.v66.22830

Bias and Drift of the Medium-Range Decadal Climate Prediction System (MiKlip) validated by European Radiosonde Data, M Pattantyus-Abraham, **C Kadow**, S Illing, ..., Meteorologische Zeitschrift 2, 2016 DOI: 10.1127/metz/2016/0803

Seasonal prediction skill of East Asian summer monsoon in CMIP5-Models,
B Huang, U Cubasch, **C Kadow**, Earth System Dynamics, 9.3, pp.985-997, 2018
DOI: 10.5194/esd-9-985-2018

Assessing the impact of a future volcanic eruption on decadal predictions,
S Illing, **C Kadow**, H Pohlmann, and C Timmreck: Earth Syst. Dynam., 9, 701-715, 2018
DOI: 10.5194/esd-9-701-2018

References

Adler, RF, GJ Huffman, A Chang, et al. (2003). “The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present)”. English. In: *Journal Of Hydrometeorology* 4.6, 1147–1167. DOI: 10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2 (cit. on pp. 32, 42).

Apache (2015). “Solr - open source enterprise search platform”. In: Apache Lucene Project, License: Apache License 2.0 (cit. on p. 25).

Balmaseda, Magdalena Alonso, Kristian Mogensen, and Anthony T. Weaver (2013). “Evaluation of the ECMWF ocean reanalysis system ORAS4”. English. In: *Quarterly Journal Of The Royal Meteorological Society* 139.674, A, 1132–1161. DOI: {10.1002/qj.2063} (cit. on pp. 32, 55).

Becker, A., P. Finger, A. Meyer-Christoffer, et al. (2013). “A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present”. In: *Earth System Science Data* 5.1, pp. 71–99. DOI: 10.5194/essd-5-71-2013 (cit. on pp. 32, 59).

Boer, G. J., D. M. Smith, C. Cassou, et al. (2016). “The Decadal Climate Prediction Project (DCPP) contribution to CMIP6”. In: *Geoscientific Model Development* 9.10, pp. 3751–3777. DOI: 10.5194/gmd-9-3751-2016 (cit. on pp. 4, 5, 53, 54).

Breiman, Leo (1996). “Bagging predictors”. In: *Machine Learning* 24.2, pp. 123–140. DOI: 10.1007/BF00058655 (cit. on p. 56).

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006). “Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850”. English. In: *Journal Of Geophysical Research-atmospheres* 111.D12. DOI: {10.1029/2005JD006548} (cit. on p. 32).

Carter, G. (2003). *LDAP System Administration*. ISBN 1-56592-491-6). O’Reilly (cit. on p. 25).

Christensen, HM, IM Moroz, and TN Palmer (2014). “Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts”. In: *Quarterly Journal of the Royal Meteorological Society* (cit. on p. 45).

Cubasch, U. and C. Kadow (2011). “Global Climate Change and Aspects of Regional Climate Change in the Berlin-Brandenburg Region”. In: *DIE ERDE* 142, pp. 3–20 (cit. on p. 41).

Dee, D. P., S. M. Uppala, A. J. Simmons, et al. (2011). “The ERA-Interim reanalysis: configuration and performance of the data assimilation system”. English. In: *Quarterly Journal Of The Royal Meteorological Society* 137.656, A, 553–597. DOI: {10.1002/qj.828} (cit. on pp. 20, 32, 55).

Django Software Foundation (2015). “Django Web Framework”. In: 1.8. Lawrence Journal-World, License: 3-clause BSD (cit. on p. 25).

Eade, Rosie, Doug Smith, Adam Scaife, et al. (2014). “Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?” In: *Geophysical Research Letters* 41.15. 2014GL061146, pp. 5620–5628. DOI: 10.1002/2014GL061146 (cit. on pp. 54, 65).

Easterling, David R. and Michael F. Wehner (2009). “Is the climate warming or cooling?” In: *Geophysical Research Letters* 36.8. DOI: 10.1029/2009GL037810. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2009GL037810 (cit. on p. 1).

Environment Modules Project (2015). “Environment Modules”. In: "John L. Furlani, "Modules: Providing a Flexible User Environment", Proceedings of the Fifth Large Installation Systems Administration Conference (LISA V), pp. 141-152, San Diego, CA, September 30 - October 3, 1991.", License: GNU General Public License (version 2) (cit. on p. 25).

Evensen, Geir (2003). “The Ensemble Kalman Filter: theoretical formulation and practical implementation”. In: *Ocean Dynamics* 53.4, pp. 343–367. DOI: 10.1007/s10236-003-0036-9 (cit. on p. 67).

Eyring, V., S. Bony, G. A. Meehl, et al. (2016). “Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization”. In: *Geoscientific Model Development* 9.5, pp. 1937–1958. DOI: 10.5194/gmd-9-1937-2016 (cit. on pp. 5, 22, 26, 53).

Ferro, Christopher A. T. (2007). “Comparing Probabilistic Forecasting Systems with the Brier Score”. In: *Weather and Forecasting* 22.5, pp. 1076–1088. DOI: 10.1175/WAF1034.1. eprint: http://dx.doi.org/10.1175/WAF1034.1 (cit. on p. 4).

Fox, B. and GNU Project (2015). “BASH - Unix Shell and Command Language”. In: License: GNU GPL v3+ (cit. on p. 25).

Fyfe., J. C., Merryfield W. J., Kharin V., et al. “Skillful predictions of decadal trends in global mean surface temperature”. In: *Geophysical Research Letters* 38.22. DOI: 10.1029/2011GL049508. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011GL049508 (cit. on p. 53).

Ghosh, R., W. A. Müller, J. Baehr, and J. Bader (2015). “Impact of observed North Atlantic multi-decadal variations to European summer climate: A quasi-geostrophic pathway”. In: *in preparation for Climate Dynamics* (cit. on p. 41).

Gneiting, Tilmann and Adrian E. Raftery (2007). “Strictly proper scoring rules, prediction, and estimation”. English. In: *Journal Of The American Statistical Association* 102.477, 359–378. DOI: {10.1198/016214506000001437} (cit. on p. 35).

GNU/Linux Community (2015). “Linux”. In: Linus Torvalds (cit. on p. 24).

Goddard, L., A. Kumar, A. Solomon, et al. (2013). “A verification framework for interannual-to-decadal predictions experiments”. English. In: *Climate Dynamics* 40.1-2, 245–272. DOI: {10.1007/s00382-012-1481-2} (cit. on pp. 4, 31–34, 36, 37, 41, 42, 44, 53, 57, 59, 64, 78).

Hamano, J., L. Torvalds, et al. (2015). “GIT - Version Control System”. In: License: GNU General Public License (cit. on pp. 15, 25).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2013). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. (cit. on pp. 56, 67, 88).

Hsu, Wu ron and Allan H. Murphy (1986). “The attributes diagram A geometrical framework for assessing the quality of probability forecasts”. In: *International Journal of Forecasting* 2.3, pp. 285 –293. DOI: `https://doi.org/10.1016/0169-2070(86)90048-8` (cit. on p. 58).

Huang, B., U. Cubasch, and C. Kadow (2018). “Seasonal prediction skill of East Asian summer monsoon in CMIP5 models”. In: *Earth System Dynamics* 9.3, pp. 985–997. DOI: `10.5194/esd-9-985-2018` (cit. on p. 21).

ICPO (2011). “Decadal and bias correction for decadal climate predictions”. In: *CLIVAR Publication Series* No. 150, 6 pp (cit. on pp. 33, 59).

Illing, S., C. Kadow, O. Kunst, and U. Cubasch (2014). “MurCSS: A Tool for Standardized Evaluation of Decadal Hindcast Systems”. In: *Journal of Open Research Software (JORS)* 2(1):e24. DOI: `http://dx.doi.org/10.5334/jors.bf` (cit. on pp. 21, 31, 45, 57).

Illing, S., C. Kadow, H. Pohlmann, and C. Timmreck (2018). “Assessing the impact of a future volcanic eruption on decadal predictions”. In: *Earth System Dynamics* 9.2, pp. 701–715. DOI: `10.5194/esd-9-701-2018` (cit. on p. 21).

ImageMagick Studio LLC (2015). “ImageMagick - a software suite to create, edit, compose, or convert bitmap images”. In: 6 (cit. on p. 25).

Jones, P. D., Lister D. H., Osborn T. J., et al. “Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010”. In: *Journal of Geophysical Research: Atmospheres* 117.D5. DOI: `10.1029/2011JD017139`. eprint: `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JD017139` (cit. on p. 59).

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”. English. In: *Journal Of Advances In Modeling Earth Systems* 5.2, 422–446. DOI: `{10.1002/jame.20023}` (cit. on pp. 32, 54).

Kadow, C., S. Illing, O. Kunst, et al. (2016). “Evaluation of Forecasts by Accuracy and Spread in the MiKlip Decadal Climate Prediction System”. In: *Meteorologische Zeitschrift*. DOI: `10.1127/metz/2015/0639` (cit. on pp. ii, 21, 54, 55, 57–59, 64).

Kadow, Christopher, Sebastian Illing, Igor Kröner, Uwe Ulbrich, and Ulrich Cubasch (2017). “Decadal climate predictions improved by ocean ensemble dispersion filtering”. In: *Journal of Advances in Modeling Earth Systems* 9.2, pp. 1138–1149. DOI: `10.1002/2016MS000787`. eprint: `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016MS000787` (cit. on p. 21).

Kalnay, Eugenia, Brian Hunt, Edward Ott, and Istvan Szunyogh (2006). “Ensemble forecasting and data assimilation: Two problems with the same solution?” In: (cit. on pp. 54, 55).

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner (2008). “Advancing decadal-scale climate prediction in the North Atlantic sector”. In: *Nature* 453.7191, pp. 84–88 (cit. on pp. 3, 53, 68, 76).

Keller, Jan D., Luis Kornblueh, Andreas Hense, and Andreas Rhodin (2008). “Towards a GME ensemble forecasting system: Ensemble initialization using the breeding technique”. In: *Meteorologische Zeitschrift* 17.6, pp. 707–718. DOI: {10.1127/0941-2948/2008/0333} (cit. on pp. 36, 58).

Kennedy, J. J., Rayner N. A., Smith R. O., Parker D. E., and Saunby M. “Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization”. In: *Journal of Geophysical Research: Atmospheres* 116.D14. DOI: 10.1029/2010JD015220. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010JD015220 (cit. on pp. 59, 60).

Kharin, V. V., Boer G. J., Merryfield W. J., Scinocca J. F., and Lee W.-S. “Statistical adjustment of decadal predictions in a changing climate”. In: *Geophysical Research Letters* 39.19. DOI: 10.1029/2012GL052647. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL052647 (cit. on p. 53).

Kosaka, Yu and Shang-Ping Xie (2013). “Recent global-warming hiatus tied to equatorial Pacific surface cooling”. English. In: *Nature* 501.7467, 403+. DOI: {10.1038/nature12534} (cit. on pp. 44, 53, 86).

Kröger, Jürgen, Holger Pohlmann, Frank Sienz, et al. (2017). “Full-field initialized decadal predictions with the MPI earth system model: an initial shock in the North Atlantic”. In: *Climate Dynamics*. DOI: 10.1007/s00382-017-4030-1 (cit. on p. 21).

Kruschke, Tim, Henning W. Rust, Christopher Kadow, Gregor C. Leckebusch, and Uwe Ulbrich (2014). “Evaluating Decadal Predictions of Northern Hemispheric Cyclone Frequencies”. In: *Tellus A* 66. DOI: {10.3402/tellusa.v66.22830} (cit. on pp. 43, 45, 65).

Kruschke, Tim, Henning W. Rust, Christopher Kadow, et al. (2015). “Probabilistic evaluation of Northern Hemisphere winter storm frequencies in the MiKlip decadal prediction system”. In: submitted to the same special issue of MetZ (cit. on pp. 4, 45, 54).

Kumar, Arun and Martin P. Hoerling (2000). “Analysis of a Conceptual Model of Seasonal Climate Variability and Implications for Seasonal Prediction”. In: *Bulletin of the American Meteorological Society* 81.2, pp. 255–264. DOI: 10.1175/1520-0477(2000)081<0255:AOACMO>2.3.CO;2. eprint: https://doi.org/10.1175/1520-0477(2000)081<0255:AOACMO>2.3.CO;2 (cit. on p. 54).

Kumar, Sanjiv, Venkatesh Merwade, James L. Kinter III, and Dev Niyogi (2013). “Evaluation of Temperature and Precipitation Trends and Long-Term Persistence in CMIP5 Twentieth-Century Climate Simulations”. In: *J. Climate* 26, 4168–4185. DOI: {10.1175/JCLI-D-12-00259.1} (cit. on p. 41).

Köhl, A. (2014). “Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic”. In: *Quarterly Journal Of The Royal Meteorological Society*. doi: 10.1002/qj.2347. DOI: doi:10.1002/qj.2347 (cit. on p. 45).

Lorenz, Edward N. (1963). “Deterministic Nonperiodic Flow”. In: *Journal of the Atmospheric Sciences* 20.2, pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. eprint: https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2 (cit. on p. 55).

Mao-Lin, Shen, Keenlyside Noel, Selten Frank, Wiegerinck Wim, and Duane Gregory S. “Dynamically combining climate models to “supermodel” the tropical Pacific”. In: *Geophysical Research Letters* 43.1, pp. 359–366. DOI: 10.1002/2015GL066562. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015GL066562 (cit. on p. 67).

Marotzke, Jochem, Wolfgang A. Müller, Freja S. E. Vamborg, et al. (2016). “MiKlip: A National Research Project on Decadal Climate Prediction”. In: *Bulletin of the American Meteorological Society* 97.12, pp. 2379–2394. eprint: https://doi.org/10.1175/BAMS-D-15-00184.1 (cit. on pp. 6, 10, 21, 31, 53–55, 57, 62, 76).

Mason, SJ and GM Mimmack (1992). “The use of bootstrap confidence-intervals for the correlation-coefficient in climatology”. English. In: *Theoretical And Applied Climatology* 45.4, 229–233. DOI: {10.1007/BF00865512} (cit. on p. 33).

Matei, Daniela, Holger Pohlmann, Johann Jungclaus, et al. (2012). “Two Tales of Initializing Decadal Climate Prediction Experiments with the ECHAM5/MPI-OM Model”. English. In: *Journal Of Climate* 25.24, 8502–8523. DOI: {10.1175/JCLI-D-11-00633.1} (cit. on pp. 3, 32, 39, 53).

Matheson, James E. and Robert L. Winkler (1976). “Scoring Rules for Continuous Probability Distributions”. In: *Management Science* 22.10, pp. 1087–1096. DOI: {10.1287/mnsc.22.10.1087}. eprint: http://dx.doi.org/10.1287/mnsc.22.10.1087 (cit. on p. 34).

Max-Planck-Institute for Meteorology (2015). “Climate Data Operators”. In: 1.6. License: GPL v2 (cit. on p. 25).

Meehl, Gerald A., Lisa Goddard, James Murphy, et al. (2009). “Decadal Prediction Can It Be Skillful?” English. In: *Bulletin Of The American Meteorological Society* 90.10, 1467—1485. DOI: {10.1175/2009BAMS2778.1} (cit. on pp. 1–3, 31, 53).

Meehl, Gerald A., Julie M. Arblaster, John T. Fasullo, Aixue Hu, and Kevin E. Trenberth (2011). “Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods”. English. In: *Nature Climate Change* 1.7, 360–364. DOI: {10.1038/NCLIMATE1229} (cit. on p. 44).

Meehl, Gerald A., Aixue Hu, Julie M. Arblaster, John Fasullo, and Kevin E. Trenberth (2013). “Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation”. In: *Journal of Climate* 26.18, pp. 7298–7310. DOI: 10.1175/JCLI-D-12-00548.1. eprint: https://doi.org/10.1175/JCLI-D-12-00548.1 (cit. on p. 53).

Meehl, Gerald A., Richard Moss, Karl E. Taylor, et al. (2014). “Climate Model Intercomparisons: Preparing for the Next Phase”. In: *Eos, Transactions American Geophysical Union* 95.9, pp. 77–78. DOI: 10.1002/2014EO090001 (cit. on pp. 3, 53).

Mehta, Vikram, Gerald Meehl, Lisa Goddard, et al. (2011). “Decadal Climate Predictability And Prediction Where Are We?” English. In: *Bulletin Of The American Meteorological Society* 92.5, 637–640. DOI: {10.1175/2010BAMS3025.1} (cit. on pp. 1, 53).

Menary, Matthew B., Daniel L. R. Hodson, Jon I. Robson, Rowan T. Sutton, and Richard A. Wood (2015). “A Mechanism of Internal Decadal Atlantic Ocean Variability in a High-Resolution Coupled Climate Model”. In: *Journal of Climate* 28.19, pp. 7764–7785. DOI: 10.1175/JCLI-D-15-0106.1. eprint: https://doi.org/10.1175/JCLI-D-15-0106.1 (cit. on p. 4).

Morice, Colin P., Kennedy John J., Rayner Nick A., and Jones Phil D. "Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set". In: *Journal of Geophysical Research: Atmospheres* 117.D8. DOI: 10.1029/2011JD017187. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JD017187 (cit. on p. 59).

Murphy, AH (1988). "Skill Scores Based On The Mean-square Error And Their Relationships To The Correlation-coefficient". English. In: *Monthly Weather Review* 116.12, 2417–2425. DOI: {10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2} (cit. on pp. 33, 57).

Murphy, AH and ES Epstein (1989). "Skill Scores And Correlation-coefficients In Model Verification". English. In: *Monthly Weather Review* 117.3, 572–581. DOI: {10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2} (cit. on pp. 33, 58).

Murray, R.J. and I Simmonds (1991). "A numerical scheme for tracking cyclone centres from digital data. Part II: application to January and July general circulation model simulations". In: 39, pp. 167–180 (cit. on p. 59).

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". English. In: *Geophysical Research Letters* 39, L22707. DOI: {10.1029/2012GL053326} (cit. on pp. 31, 37, 39, 45).

OpenSSH project (2015). "OpenSSH SSH client". In: OpenSSH is a derivative of the original free ssh 1.2.12 release from Tatu Ylönen. (cit. on p. 25).

Oracle Corporation (2015a). "Java language". In: Designed by James Gosling and Sun Microsystems, License: GNU General Public License (cit. on p. 25).

– (2015b). "MySQL - relational database management system". In: Original author(s) MySQL AB, License: GPL (version 2) (cit. on p. 25).

Palmer, T., R. Buizza, R. Hagedorn, et al. (2006). "Ensemble prediction: A pedagogical perspective." In: *ECMWF Newsletter* 106, pp. 10–17 (cit. on pp. 36, 58).

Pasternack, A., J. Bhend, M. A. Liniger, et al. (2018). "Parametric decadal climate forecast recalibration (DeFoReSt 1.0)". In: *Geoscientific Model Development* 11.1, pp. 351–368. DOI: 10.5194/gmd-11-351-2018 (cit. on p. 21).

Pattantyús-Ábrahám, Margit, Christopher Kadow, Sebastian Illing, et al. (2016). "Bias and Drift of the Medium-Range Decadal Climate Prediction System (MiKlip) validated by European Radiosonde Data". In: *Meteorologische Zeitschrift* 25.6, pp. 709–720. DOI: 10.1127/metz/2016/0803 (cit. on p. 21).

Pinto, Joaquim G., Thomas Spangehl, Uwe Ulbrich, and Peter Speth (2005). "Sensitivities of a cyclone detection and tracking algorithm: individual tracks and climatology". In: *Meteorologische Zeitschrift* 14.6, pp. 823–838. DOI: doi:10.1127/0941-2948/2005/0068 (cit. on p. 59).

Pohlmann, H., W. A. Müller, K. Kulkarni, et al. (2013). "Improved forecast skill in the tropics in the new MiKlip decadal climate predictions". In: *Geophysical Research Letters* 40.21, pp. 5798–5802. DOI: {10.1002/2013GL058051} (cit. on pp. 4, 21, 31, 32, 43–45, 54, 55).

Pohlmann, Holger, Johann H. Jungclaus, Armin Köhl, Detlef Stammer, and Jochem Marotzke (2009). “Initializing Decadal Climate Predictions with the GECCO Oceanic Synthesis: Effects on the North Atlantic”. In: *Journal of Climate* 22.14, pp. 3926–3938. DOI: 10.1175/2009JCLI2535.1. eprint: https://doi.org/10.1175/2009JCLI2535.1 (cit. on pp. 3, 53, 76).

Pohlmann, Holger, Jürgen Kröger, Richard J. Greatbatch, and Wolfgang A. Müller (2017). “Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific”. In: *Climate Dynamics* 49.7, pp. 2685–2693. DOI: 10.1007/s00382-016-3486-8 (cit. on p. 4).

Python Software Foundation (2015). “Python Language Reference”. In: 2.7.10. "G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995." (cit. on p. 25).

Ranilla, J., E. M. Garzón, and J. Vigo-Aguiar (2014). “High performance computing: an essential tool for science and engineering breakthroughs”. In: *The Journal of Supercomputing* 70.2, pp. 511–513. DOI: 10.1007/s11227-014-1279-6 (cit. on p. 77).

Rayner, Parker D. E., Horton E. B., et al. “Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century”. In: *Journal of Geophysical Research: Atmospheres* 108.D14. DOI: 10.1029/2002JD002670. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JD002670 (cit. on p. 59).

Schneider, Udo, Andreas Becker, Peter Finger, et al. (2011). “GPCC Full Data Reanalysis Version 6.0 at 2.5: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data”. In: *Global Precipitation Climatology Centre*. DOI: {10.5676/DWD_GPCC/FD_M_V6_250} (cit. on p. 32).

Shaffrey, L. C., D. Hodson, J. Robson, et al. (2017). “Decadal predictions with the HiGEM high resolution global coupled climate model: description and basic evaluation”. In: *Climate Dynamics* 48.1, pp. 297–311. DOI: 10.1007/s00382-016-3075-x (cit. on p. 4).

Sienz, Frank, Wolfgang A. Müller, and Holger Pohlmann (2016). “Ensemble size impact on the decadal predictive skill assessment”. In: *Meteorologische Zeitschrift* 25.6, pp. 645–655. DOI: 10.1127/metz/2016/0670 (cit. on pp. 4, 43, 44, 54).

Slurm Commercial Support and Development (2015). “Simple Linux Utility for Resource Management Workload Management”. In: "Slurm: Simple Linux Utility for Resource Management, A. Yoo, M. Jette, and M. Grondona, Job Scheduling Strategies for Parallel Processing, volume 2862 of Lecture Notes in Computer Science, pages 44-60, Springer-Verlag, 2003." License GNU General Public License 2 (cit. on p. 25).

Smith, Doug M., Stephen Cusack, Andrew W. Colman, et al. (2007). “Improved Surface Temperature Prediction for the Coming Decade from a Global Climate Model”. In: *Science* 317.5839, pp. 796–799. DOI: 10.1126/science.1139540. eprint: http://science.sciencemag.org/content/317/5839/796.full.pdf (cit. on pp. 3, 53, 76).

Smith, Doug M., Adam A. Scaife, George J. Boer, et al. (2013). “Real-time multi-model decadal climate predictions”. English. In: *Climate Dynamics* 41.11-12, 2875–2888. DOI: {10.1007/s00382-012-1600-0} (cit. on pp. 4, 39, 41, 43, 53).

Smith, Doug M., Rosie Eade, and Holger Pohlmann (2013b). “A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction”. In: *Climate Dynamics* 41.11, pp. 3325–3338. DOI: 10.1007/s00382-013-1683-2 (cit. on pp. 3, 4).

Spangehl, T., M. Schröder, S. Stolzenberger, et al. (2015). “Evaluation of the MiKlip decadal prediction system using satellite based cloud products”. In: *Meteorologische Zeitschrift*. submitted to the same special issue of MetZ (cit. on p. 45).

Stevens, Bjorn, Marco Giorgetta, Monika Esch, et al. (2013). “Atmospheric component of the MPI-M Earth System Model: ECHAM6”. English. In: *Journal Of Advances In Modeling Earth Systems* 5.2, 146–172. DOI: {10.1002/jame.20015} (cit. on pp. 32, 54).

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, et al. (2015). “Revealing skill of the MiKliP decadal prediction systems by three dimensional probabilistic evaluation”. In: *Meteorologische Zeitschrift*. submitted to the same special issue of MetZ (cit. on pp. 45, 54, 58).

Szeredi, M. (2015). “FUSE - Filesystem in Userspace”. In: License:GPL for kernel part, LGPL for Libfuse (cit. on pp. 25, 26).

Taylor, Karl E., Ronald J. Stouffer, and Gerald A. Meehl (2012). “An Overview Of CMIP5 And The Experiment Design”. English. In: *Bulletin of the American Meteorological Society* 93.4, 485–498. DOI: {10.1175/BAMS-D-11-00094.1} (cit. on pp. 4, 31, 32, 42, 53–55).

The NCO project (2015). “NCO - netCDF Operators”. In: 4. Zender, C. S. (2008): Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). Environmental Modelling and Software, 23 (10-11), pp. 1338–1342 (cit. on p. 25).

Thoma, M., R. Greatbatch, C. Kadow, and R. Gerdes (2015). “Decadal hindcasts initialised using observed surface wind stress: Evaluation and Prediction out to 2024”. In: *Geophysical Research Letters* (cit. on p. 21).

Timmreck, C., H. Pohlmann, S. Illing, and C. Kadow (2015). “The impact of stratospheric volcanic aerosol on decadal scale predictability”. In: *Geophysical Research Letters* (cit. on p. 21).

University Corporation for Atmospheric Research (2015). “NetCDF - Network Common Data Form”. In: 4 (cit. on p. 25).

Uppala, SM, PW Kallberg, AJ Simmons, et al. (2005). “The ERA-40 re-analysis”. English. In: *Quarterly Journal Of The Royal Meteorological Society* 131.612, B, 2961–3012. DOI: {10.1256/qj.04.176} (cit. on pp. 32, 55).

Vuuren, Detlef P. van, Jae Edmonds, Mikiko Kainuma, et al. (2011). “The representative concentration pathways: an overview”. In: *Climatic Change* 109.1, p. 5. DOI: 10.1007/s10584-011-0148-z (cit. on p. 2).

Weisheimer, A. and T. N. Palmer (2014). “On the reliability of seasonal climate forecasts”. In: *Journal of The Royal Society Interface* 11.96. DOI: 10.1098/rsif.2013.1162. eprint: http://rsif.royalsocietypublishing.org/content/11/96/20131162.full.pdf (cit. on pp. 54, 58).

Wilks, D.S. (2005). *Statistical Methods in the Atmospheric Sciences*. 2nd. Vol. 100. International Geophysics. Academic Press, Cornell University, 627pp (cit. on p. 88).

Wilks, S. D (2011). *Statistical Methods In The Atmospheric Sciences* (cit. on pp. 33, 58, 59).

Yeager, Stephen, Alicia Karspeck, Gokhan Danabasoglu, Joe Tribbia, and Haiyan Teng (2012). “A Decadal Prediction Case Study: Late Twentieth-Century North Atlantic Ocean Heat Content”. In: *Journal of Climate* 25.15, pp. 5173–5189. DOI: 10.1175/JCLI-D-11-00595.1. eprint: https://doi.org/10.1175/JCLI-D-11-00595.1 (cit. on p. 53).

Zappa, Giuseppe, Len C. Shaffrey, Kevin I. Hodges, Phil G. Sansom, and David B. Stephenson (2013). “A Multimodel Assessment of Future Projections of North Atlantic and European Extratropical Cyclones in the CMIP5 Climate Models”. In: *Journal of Climate* 26.16, pp. 5846–5862. DOI: `10.1175/JCLI-D-12-00573.1`. eprint: `https://doi.org/10.1175/JCLI-D-12-00573.1` (cit. on p. 65).

Zender, Charles (2008). “Analysis of Self-Describing Gridded Geoscience Data with NetCDF Operators (NCO)”. In: 23, pp. 1338–1342 (cit. on p. 56).

Acronyms

Term	Explanation
API	Application Programming Interface
BMBF	Bundesministerium für Bildung und Forschung
CMIP	Coupled Model Intercomparison Project
CMOR	Climate Model Output Rewriter
DCPP	Decadal Climate Prediction Project - A MIP within CMIP6
DRS	Dara Reference Syntax
DWD	Deutscher Wetterdienst - German weather service
ECHAM	atmospheric general circulation model, developed at the Max Planck Institute for Meteorology
ECMWF	European Center for Medium-Term Weather Forecasts
EDF	Ensemble Dispersion Filter
ENSO	El Nino Southern Oscillation
EMS	Ensemble Meta Selection
EMV	Ensemble Majority Vote
EnKF	Ensemble Kalman Filter
ERA40	Atmosphere Reanalyis of the ECMWF
ERA-Interim	Atmosphere Reanalyis of the ECMWF
ESGF	Earth System Grid Federation
GPCP	Global Precipitation Climatology Project
GPCC	Global Precipitation Climatology Centre
HPC	High Performance Computer
HadCRUT	Near-Surface Temperature dataset of the Hadley Centre and Climatic Research Unit
HadSST3	Sea-surface Temperature dataset version 3 of the Hadley Centre
IPCC	Intergovernmental Panel on Climate Project
LESS	Logarithmic Ensemble Spread Score
LESSS	Logarithmic Ensemble Spread Skill Score
LM	Lead Month - the months of the forecast independent of actual months
LY	Lead Year - the years of the forecast independent of actual years

MiKlip	Mittelfristige Klimaprognosen - Major project on decadal climate prediction in Germany
MiKlip-EDF	The EDF applied with the MiKlip baseline system
MiKlip-REF	The baseline system of MiKlip and the reference for the EDF experiment
MiKlip-REF-FF	The full-field initialization experiment of MiKlip called Prototype
MiKlip-REF-10	The 10 member experiment of MiKlip called Baseline1
MiKlip-REF-MR	The mixed-resolution experiment of MiKlip called Baseline1-MR
MiKlip-REF-UN	The uninitialized experiment as mixture of historical and rcp45
MIP	Model Intercomparison Project
MPI-ESM-LR	Max-Planck-Institute Earth System Model in the Low-Resolution version
MPI-OM	Max-Planck-Institute Ocean Model
MSESS	mean squared error skill score
NA	North Atlantic
ORAS4	Ocean Reanalyis System 4 of the ECMWF
PDO	Pacific Decadal Oscillation
PR	Precipitation
Prototype	The full-field initialization experiment of MiKlip - in this stidy called MiKlip-REF-FF
RCP	Representave Concentration Pathway - scenario experiment for climate projections within CMIP
RMSE	root mean squared error - in units of the researched variable
SST	Sea Surface Temperature
TAS	Near-Surface Air Temperature
WCRP	World Climate Research Programme

Tab. 5.2: Overview table of used acronyms.

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2$_{\varepsilon}$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

I hereby declare that this dissertation is a product of my own research endeavours and has been prepared without illegitimate assistance. The work is original except where indicated by reference in the text and no part of the dissertation has been submitted for any other degree. This dissertation has not been presented to any other university for examination.

Berlin, October 22th, 2018

Christopher Kadow

Acknowledgment

I would like to thank Ulrich Cubasch and Uwe Ulbrich for the opportunity to write my PhD thesis at the Institute of Meteorology of the Freie Universität Berlin. Fully aware of my progressive research style, I really appreciate the valuable advice and support from both of them - to keep the balance between producing new ideas and steady state research.

I'm thankful that my research is not only a theoretical thesis, but all three papers already turned into applied science. Presenting and communicating an actual forecast next to its systematic evaluation is the base for the official MiKlip forecast. Freva is installed and used at MiKlip, for CMIP6, at the DWD, at NCAR, and of course at FU Berlin. The EDF is part of the selection of MiKlip initialization methods, was reproducing its positive results in a different study, and is highlighting the climate modeling efforts of the FU Berlin at several international conferences.

The MiKlip project was an incredible scientific ride. Many discussions, conclusions, eyes opener, and obstacles were an adventure with fantastic colleagues having the same goal. Representative for all MiKlipers, I want to thank Jochem Marotzke and Freja Vamborg.

The Climate Modeling (Klimod) working group at FUB was always a big basket of themes which kept research variety ongoing. I'm thankful for being part of a group who well understood the blending of technical modeling and scientific research.

Being able to group a certain amount of smart people into a Freva team, was maybe the most satisfying effort at FUB. Many thanks goes to Sebastian Illing, who is an outstanding colleague to the research groups around Freva, Klimod, IfM, and MiKlip.

My parents always supported me, I'm thankful that I can do today what I'm doing right now. My whole family -with brothers, sisters, nieces, nephews, cousins, and friends around the globe- is like a soothing sea breeze for the soul.

The biggest *Thank You* goes to Mareike, who was a tremendous support and help during writing, writing, re-writing, writing, re-writing, etc.

> *Das schönste Glück des denkenden Menschen ist, das Erforschliche erforscht zu haben und das Unerforschliche ruhig zu verehren.*
>
> — **Johann Wolfgang von Goethe**
> (Poet and Naturalist)