Meteorologische Zeitschrift, PrePub DOI 10.1127/metz/2018/0871
© 2018 The authors

Temperature and sea ice hindcast skill of the MiKlip decadal prediction system in the Arctic

Daniel Senftleben[1*], Veronika Eyring[1,2], Axel Lauer[1] and Mattia Righi[1]

[1]Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR), Institut für Physik der Atmosphäre, 82234 Wessling, Germany
[2]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

(Manuscript received July 28, 2017; in revised form November 3, 2017; accepted November 17, 2017)

Abstract

In this study, hindcast skill for near-surface air temperature (TAS), sea surface temperature (SST), sea ice concentration, and sea ice area is assessed for the Arctic region using decadal simulations with the MiKlip decadal prototype prediction system. The prototype MiKlip system is based on the low-resolution version of the MPI-ESM model. In the simulations, a full field initialization of atmospheric and oceanic variables was used, but sea ice was not initialized. The hypothesis is that the increase in hindcast skill due to initialization found for TAS and SST in the North Atlantic in the prototype system compared to the historical simulations leads to enhanced skill also in the Arctic. However, the skill enhancement compared to the uninitialized experiments in the Arctic is generally weak. The hindcast skill only increases for SST and sea ice concentration along the east coast of Greenland and in the Fram Strait in lead years 2–5. Initialization additionally improves the skill in regionally integrated sea ice area (detrended) in the Greenland Sea, but only in lead year 1 and only in winter, and not in other Arctic regions. In order to assess whether additional initialization of sea ice concentration improves skill, we also analyse hindcasts and historical simulations performed with the MiKlip preoperational system that is based on the high-resolution version of the MPI-ESM. These simulations have nonetheless a negative bias in sea ice area in late summer of 1 to 3 million km^2. Noting that this is a much smaller ensemble than for the prototype system, the hindcast skill in North Atlantic TAS and SSTs is significantly reduced and not present when evaluated against ERA-Interim instead of HadCRUT4 data. Accordingly, in the Arctic, no additional skill compared to the prototype hindcasts is found. Our results underline the importance to assess the robustness of skill with different observational datasets and metrics. For future MiKlip simulations, we recommend to additionally initialize sea ice thickness or age, and to initialize the simulations in a different month to potentially enhance sea ice skill in the Arctic.

Keywords: Arctic, sea ice, decadal simulations, near term climate prediction, hindcast skill, climate change, ESMValTool, MiKlip

1 Introduction

Arctic sea ice is an important component of the climate system as it modulates the exchange of heat, moisture and momentum between ocean and atmosphere (Steele and Dickinson, 2016). Melting and freezing of Arctic sea ice also influences freshwater and salt fluxes into the ocean, thereby affecting the ocean circulation and thus climate (Komuro and Hasumi, 2003). In the last few decades, Arctic temperatures have increased at a rate that is twice as high as the global average (Bellucci et al., 2015), and a strong decline in Arctic summer sea ice extent has been observed over the same time period (Stroeve et al., 2012b). Climate models suggest a further retreat of the sea ice throughout the 21st century, with a potential for abrupt loss of Arctic sea ice (Holland, 2010). The loss of Arctic sea ice has several direct adverse impacts on the Arctic indigenous people and wildlife, and may also affect stakeholders on a larger scale, for example by an opening of Arctic shipping routes due to the sea ice decline (Paxian et al., 2010; Eyring et al., 2010; Melia et al., 2016) and by changing the atmospheric circulation and weather patterns (e.g., Cohen et al., 2014; Handorf et al., 2015).

Current state-of-the-art global climate models, such as those participating in the Fifth Phase of the Coupled Model Intercomparison Project (CMIP5, Taylor et al., 2012), include detailed representations of sea ice and their performance show substantial improvements with respect to previous generations of models (Flato et al., 2013; Notz et al., 2013; Stroeve et al., 2012a). Nevertheless, the reliability of climate predictions on time scales relevant to society and policymakers (years to a decade) still needs to be further assessed and improved.

On these time scales, the natural internal variability of the climate system plays an important role. In the field of decadal climate predictions, which typically focus on a time range of up to 10 years, the goal is to better capture the effects of internal variability of the climate system by initializing the models with obser-

[*]Corresponding author: Daniel Senftleben, Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR), Institut für Physik der Atmosphäre, 82234 Weßling, Germany, email: daniel.senftleben@dlr.de

vations of slowly-varying components of the Earth system such as ocean variables (Meehl et al., 2009; Smith et al., 2010). Numerous studies found that the initialization of climate models can improve the predictability of near-surface air temperature in certain regions (e.g., Pohlmann et al., 2009; Doblas-Reyes et al., 2013; Goddard et al., 2013; Jia and Delsole, 2013; Meehl and Goddard, 2013; Müller et al., 2004). Such improved prediction skill stems from a more realistic simulation of the natural variability, since the atmospheric and oceanic initialization allows the model experiments to start from the correct phase of relevant modes of natural variability (Müller et al., 2012).

Previous studies found that, on seasonal timescales, the prediction skill of sea ice improves by initializing the model with sea ice observations. For example, Day et al. (2014a) found that the inclusion of sea ice thickness in the model initialization process can significantly increase the predictability of Arctic sea ice concentration and extent and reduces forecast errors in Arctic near-surface air temperatures. The sea ice prediction skill hereby depends on the season. Germe et al. (2014) showed that winter Arctic sea ice in initialized simulations is potentially predictable for up to several years, but for summer the potential predictability does not exceed 2 years. A potential predictability of Arctic sea ice volume and extent of up to 3 years was also reported by Tietsche et al. (2014) in a set of four global climate models. Yeager et al. (2015) analysed the prediction skill of decadal trends in Arctic sea ice extent and showed that initialization improves Arctic winter skill in predicting the trend in sea ice extent, especially in the Atlantic.

The present study examines the skill of retrospective decadal climate predictions ("hindcasts") of Arctic sea ice. We investigate whether initializing decadal hindcasts leads to more skilful predictions in comparison to uninitialized ("historical") simulations. Here, we quantitatively assess the sea ice hindcast skill from decadal and historical climate simulations performed with the Max Planck Institute Earth System Model (MPI-ESM, Stevens et al., 2013; Giorgetta et al., 2013) as part of the German Federal Ministry of Education (BMBF) project MiKlip ("Mittelfristige Klimaprognosen", midterm climate predictions) (Marotzke et al., 2016). We apply a verification system for decadal hindcasts that we implemented into the Earth System Model Evaluation Tool (ESMValTool) (Eyring et al., 2016). Additionally, we assess the skill of the model in reproducing observations of two additional variables that are closely connected with sea ice: sea surface temperature, as a proxy for the oceanic influence, and near-surface air temperature for the atmospheric component.

The paper is structured as follows: Section 2 describes the observations, model experiments and methods used for the assessment of the hindcast skill. In Section 3, the results for hindcast skill in TAS, SST and sea ice are presented. The study concludes with a summary and discussion in Section 4.

2 Model experiments and methods

2.1 Model simulations

MPI-ESM is a coupled Earth system model with an atmospheric, an oceanic, a land biogeochemistry and a marine biogeochemistry component, contributing to CMIP5. We analyse simulations from MPI-ESM in its low-resolution configuration (MPI-ESM-LR) with the atmospheric component ECHAM6 (Stevens et al., 2013) resolved horizontally at 1.9°×1.9° and vertically with 47 levels up to 0.01 hPa (T63L47). The ocean component is the Max Planck Institute Ocean Model (MPIOM) (Jungclaus et al., 2013) configured with a bipolar orthogonal curvilinear C-grid (Marsland et al., 2003) with one pole over South Greenland and another one over Antarctica, a nominal resolution of 1.5°, and 40 vertical levels. Coupled to the MPIOM is a dynamic and thermodynamic sea-ice model based on Hibler (1979).

The MPI-ESM-LR forms the basis of the MiKlip decadal prototype prediction system (hereafter MPI-ESM-LR *prot*) (Marotzke et al., 2016; Kröger et al., in review). The MPI-ESM *prot* simulations consist of 30 decadal ensemble members: 15 are initialized with observationally-based data from the ocean reanalysis system 4 (ORAS4) (Balmaseda et al., 2013) and 15 with reanalysis data from the German contribution to Estimating the Circulation and Climate of the Ocean (GECCO2) (Köhl, 2015). In both cases, atmospheric and oceanic components are initialized applying the full-field initialization technique. First, the reanalyses are nudged into the coupled model to perform a so-called assimilation run. For the atmosphere, vorticity, divergence, temperature and sea level pressure from the ERA40 (Uppala et al., 2005) and ERA-Interim reanalyses (Dee et al., 2011), and for the ocean, 3-dimensional temperature and salinity fields are nudged in the assimilation run. No sea ice variables are nudged. Then, the retrospective decadal forecasts (the so-called "hindcasts") are started from the assimilation runs and freely integrated ten years into the future. The model is initialized with data from the assimilation run on January 1st each year between 1960 and 2013. This results in a set of 54 hindcast experiments, each simulating a different ten-year period and consisting of 30 ensemble members. Each ensemble member is initialized with slightly different initial conditions applying a 1-day lagged initialization.

In addition, an ensemble of ten uninitialized, long-term "historical" climate simulations have been performed with MPI-ESM-LR following the CMIP5 experiment protocol (Taylor et al., 2012). In our study, the output of the historical simulations is compared to the results from the initialized decadal hindcasts to assess whether the initialization improves hindcast skill. The historical simulations cover the period 1850–2005 driven with prescribed natural and anthropogenic forcings (Giorgetta et al., 2013). For this study, each of the ten ensemble members of the historical simulations

Table 1: Overview of MPI-ESM simulations analysed in this study (EM: ensemble members).

Model run	Time period	#EM	Comment
MPI-ESM-LR prot decadal hindcasts	10 years after initialization (1961–2023)	15 + 15	Decadal hindcasts initialized with GECCO2/ORAS4 reanalyses data with 15 ensemble members each; sea ice not initialized
MPI-ESM-LR historical simulations	1850–2005	10	Uninitialized historical simulations with MPI-ESM-LR (CMIP5)
MPI-ESM-LR RCP4.5 simulations	2006–2100	10	RCP 4.5 simulations with MPI-ESM-LR (CMIP5), used to extend MPI-ESM-LR historical runs until 2013
MPI-ESM-HR preop decadal hindcasts	10 years after initialization (1961–2023)	5	Decadal hindcasts in which SIC was additionally initialized
MPI-ESM-HR historical simulations	1850–2005	5	Uninitialized historical simulations with high-resolution configuration MPI-ESM-HR
MPI-ESM-HR RCP4.5 simulations	2006–2100	5	RCP 4.5 simulations with high-resolution configuration MPI-ESM-HR, used to extend MPI-ESM-HR historical runs until 2013

has been extended until 2013 with the corresponding ensemble member of the projection under the Representative Concentration Pathway (RCP) 4.5 (Thomson et al., 2011).

In addition, we analyse historical simulations and decadal hindcasts that were recently performed within MiKlip Phase 2 with the preoperational system based on the high-resolution model configuration (MPI-ESM-HR *preop*). In this setup, sea ice concentrations were also initialized in addition to atmospheric and oceanic variables in the decadal hindcasts, while sea ice thickness was not initialized. Compared to MPI-ESM-LR, the spatial and vertical resolution of the atmosphere is doubled (T127L95), and the nominal resolution of the ocean grid is improved from 1.9° to 0.4° (TP04/L40). All decadal hindcasts as well as the historical simulations with MPI-ESM-HR consist of 5 ensemble members each. All MPI-ESM-HR historical simulations and hindcasts cover the same time period as the ones with MPI-ESM-LR and were also extended to 2013 with their corresponding RCP 4.5 simulations. All model simulations are summarized in Table 1.

2.2 Observations and reanalysis data

As reference datasets in the calculations of the model prediction skill, we use different observations or reanalyses that are described in this section for each assessed variable.

For the evaluation of TAS, we use the European Centre for Medium-Range Weather Forecast Re-Analysis Interim (ERA-Interim) data (Dee et al., 2011) from 1979 to 2013. We also use HadCRUT4 (Morice et al., 2012) data which cover the time period 1850–2012 and are derived from over 4800 stations (for recent years) for land regions all over the globe (Jones et al., 2012), and for the ocean from merchant and naval vessels as well as fixed and drifting buoys (Kennedy et al., 2011).

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

Available observations of sea ice concentration (SIC), i.e. the area fraction of sea ice cover within each grid cell, are not very reliable before 1978 when satellite products became available. We therefore base the SIC analysis on satellite retrievals from the National Snow and Ice Data Center (NSIDC) (Walsh et al., 2015). These observations cover the period from 1978 to present. The satellite data are processed by two different retrieval algorithms: the Bootstrap (NSIDC-BT) (Comiso, 2000) and the NASA-Team (NSIDC-NT) (Cavalieri et al., 1996) algorithm. The main difference lies in the treatment of melt ponds that are nearly indistinguishable from open water in the satellite data. Whereas in NSIDC-BT sea ice concentrations are synthetically increased in summer to account for undetected melt ponds, NSIDC-NT does not contain such a correction. Since NSIDC-BT could potentially introduce a positive bias in melt-pond-free areas, we use both products as a reference in our sea ice evaluation (see also Notz, 2014).

2.3 Methods

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

We apply it here to assess the prediction skill of TAS, SST, SIC and sea ice area (SIA). The latter is not readily available and has been derived as the area integral of

SIC over a certain region. Due to the strong seasonality of sea ice, we analyse September and March means separately. The former typically represents the Arctic minimum ice conditions (summer), whereas the latter serves as a proxy for the winter situation with maximum sea ice area.

To assess the prediction skill of the four variables we calculate the anomaly correlation coefficient (ACC, also known as Pearson product-moment correlation coefficient) and the root-mean-square error (RMSE), using monthly mean values of the model and a reference observational dataset. Wherever possible, we apply both metrics to all variables, since the assessment of skill can be dependent on the choice of the metric (Hawkins et al., 2015).

In order to assess a possible reduction in RMSE by initialization, we define the RMSE skill score as a function of the ratio between the RMSE of decadal hindcasts and the RMSE of historical simulations:

$$RMSE_{\text{skill}} = 1 - \frac{RMSE_{\text{decadal}}}{RMSE_{\text{historical}}} \qquad (2.1)$$

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

The data processing also includes the computation of an ensemble average for each hindcast experiment (30 members for *prot*; 5 for *preop*) and the historical simulations (10 members for *prot*; 5 for *preop*). In the following, ACC and RMSE are always calculated for ensemble means. To apply the metrics to each grid cell, modelled sea ice and SST are first interpolated from their native irregular ocean grids to a regular 1° × 1° grid using the distance-weighted regridding method from the Climate Data Operators package (CDO, https://code.mpimet.mpg.de/projects/cdo). Higher resolutions of the target grid were tested but were found to have no impact on the results. As recommended by CLIVAR (2011), a cross-validated bias correction is applied to the decadal ensemble means in order to remove the mean bias from the data (e.g., Gangstø et al., 2013). Anomalies are calculated from the historical simulations and observations by subtracting their respective means over the analysed time periods.

In order to estimate the dependence of hindcast skill on the forecast time, we select different time samples from each hindcast, the so-called lead years. Following previous studies (e.g., Goddard et al., 2013; Müller et al., 2012; Kim et al., 2012), we analyse a set of three lead years: year 1, years 2–5 and years 6–9. For this we construct one climatology for each set of lead years from the decadal hindcasts. Lead-year-1 climatology consists of the first year of each hindcast experiment, whereas the climatology for lead years 2–5 consists of the average over the years 2 to 5 from each hindcast experiment. For the historical simulations and the observationally-based reference datasets, the lead-year climatologies are constructed by taking the same years as the ones used for calculating the hindcast climatologies from their respective time series. In most of the aforementioned studies, the lead-year climatologies were constructed by sampling over slightly different time periods depending on the selected lead time. This could cause a bias in the assessed prediction skill in particular for variables such as, for instance, sea ice extent, which have a large year-to-year variability: one lead-year climatology could include an exceptional year with a large peak or valley that is well predicted, but which a different climatology does not include. This would artificially increase the ACC skill of the former climatology. Thus, it is particularly important to sample the same period for all lead years. We therefore calculate the lead-year climatologies in a way that they all cover the same time period (i.e., 1979–2013). A drawback of this approach is that the climatologies are shorter and only include a subset of the decadal hindcasts for certain lead years.

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

3 Results

In the following, we first analyse the TAS hindcast skill for the North Atlantic, and discuss the role of different reference datasets (Section 3.1). We then compare time series of September SIA from decadal hindcasts, historical simulations and observations for different lead-year climatologies in the Arctic to identify possible biases and drifts (Section 3.2). Finally, we assess additional hindcast skill of temperature and sea ice in the Arctic compared to the uninitialized experiments with different metrics in different regions in Section 3.3.

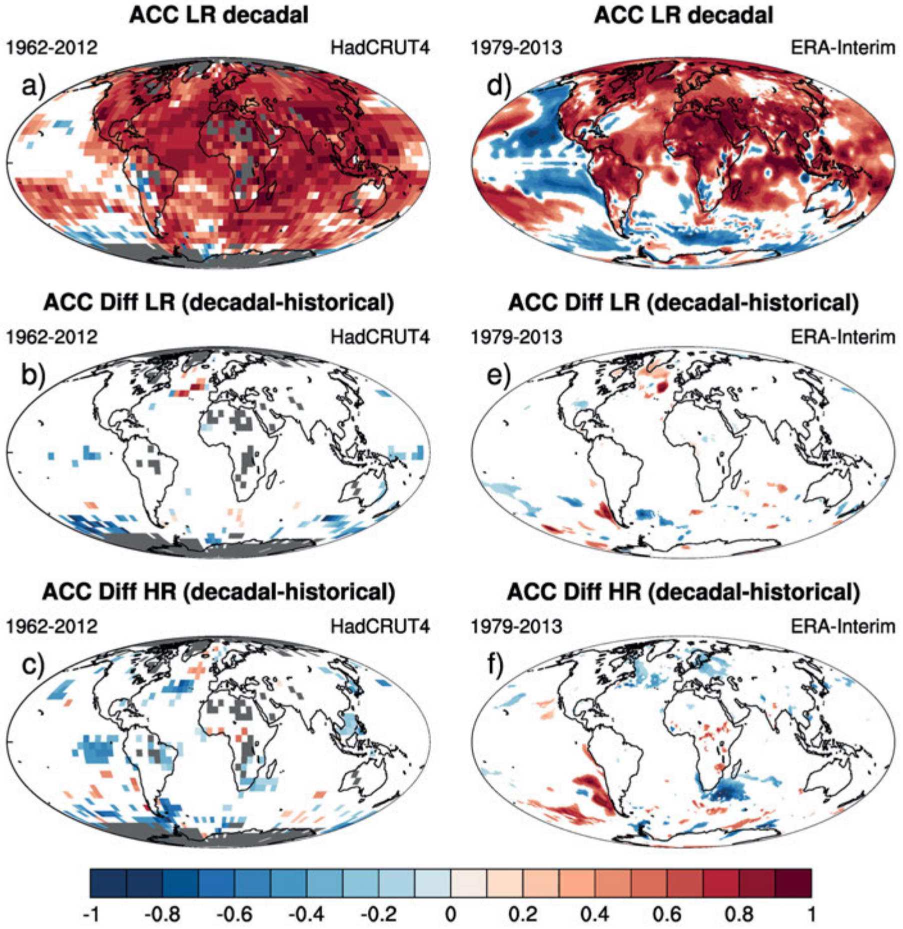

Figure 1: Ensemble-mean hindcast skill (ACC) of annual mean near-surface air temperature (TAS) averaged over the lead years 2–5 for MPI-ESM-LR *prot* decadal hindcasts (a, d) and with the anomaly correlation of MPI-ESM-LR historical simulations subtracted (b, e). Panels c and f are the same as b and e, but for MPI-ESM-HR *preop* simulations. The ACC is calculated with HadCRUT4 observations (1962–2012; left column, a–c) and ERA-Interim reanalyses (1979–2013; right column, d–f). All values shown are statistically significant at the 95 % confidence level according to a t-test for correlation coefficients. Grid cells with a dark grey colour denote missing values in the observations.

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

The ACC of TAS from decadal hindcasts and historical simulations for the lead years 2–5 is shown in Fig. 1. Similarly to Marotzke et al. (2016), we calculate the ACC for the MPI-ESM-LR *prot* system against HadCRUT4 data and confirm the significant skill improvement due to initialization in the North Atlantic Ocean (Fig. 1b). This forms our hypothesis for possibly improved skill in the Arctic (Section 3.3). In addition, we also show the results for the MPI-ESM-HR *preop* system. Since HadCRUT4 data have gaps in the Arctic we compare both model systems additionally with ERA-Interim data (right column of Fig. 1).

The ACC skill in the North Atlantic can be improved via initialization (Fig. 1b, e). Here, the ocean initialization improves prediction skill by increasing local SST skill due to the deep mixed layer in this region (Marotzke et al., 2016). In a predecessor version of the MPI-ESM-LR decadal prediction system, Müller et al. (2016) found significant skill improvements in the representation of the Atlantic Meridional Overturning Circulation (AMOC) in historical simulations of up to lead year 5, which may be the reason for enhanced TAS skill in the North Atlantic. This improvement is robust against the choice of reference dataset (compare panels b and e in Fig. 1).

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

In the MPI-ESM-HR *preop* hindcasts (Fig. 1c, f), on the contrary, we find no significant improvement in the hindcast skill in the North Atlantic due to initialization,

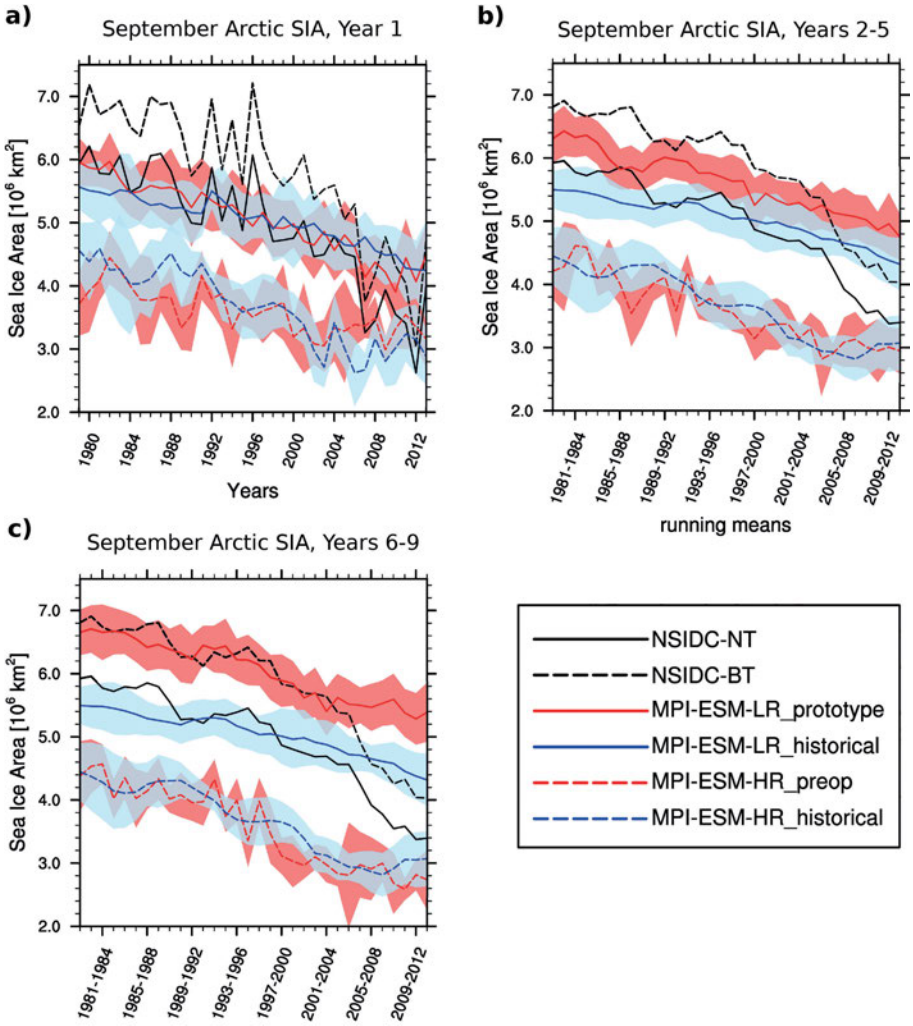

Figure 2: Time series of September mean Arctic sea ice area (SIA) from 1979 to 2013 from NSIDC-NT and NSIDC-BT observations, and the respective ensemble means of the following model simulations: MPI-ESM-LR historical simulations and *prot* decadal hindcasts, and MPI-ESM-HR historical simulations and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

independent of the reference dataset used (HadCRUT4, ERA-Interim). A possible reason for that could be the small ensemble size (5 ensemble members instead of 30 members in MPI-ESM-LR *prot*) not covering a large enough spread of different initial conditions. For example, Sienz et al. (2015) recommended at least 10 ensemble members (and even more in regions with a low signal-to-noise ratio), since initial conditions are never known exactly.

3.2 Time series of pan-Arctic sea ice area

Before assessing hindcast skill in the decadal simulations, we first look at the models' climatologies and trends to identify possible biases and drifts. Fig. 2 shows the evolution of pan-Arctic (60°–90° N) SIA between 1979 and 2013 calculated from the two observational NSIDC datasets and simulated by different MPI-ESM runs for different lead years. There is a relatively constant offset between the two NSIDC observations of roughly 1 million km^2 that results from differences in the retrieval algorithm (see Section 2.2). The historical simulations show a smaller bias compared to NSIDC-NT than to NSIDC-BT. This is consistent with results from earlier model versions (Notz et al., 2013). In agreement with both observational datasets, the ensemble means of all MPI-ESM-LR and MPI-ESM-HR simulations show a decline in Arctic summer SIA over this time period, but they underestimate sea ice decline after 2006.

The MPI-ESM-LR *prot* hindcasts show a strong drift in modelled SIA during the ten-year hindcast period: the simulated SIA is closest to NSIDC-NT observations in lead year 1 (Fig. 2a) and increases during the hindcast time (Fig. 2b) reaching a state that is closest to NSIDC-BT in lead years 6–9 (Fig. 2c). It hereby increasingly deviates from the historical simulations (blue solid line in Fig. 2), with a maximum offset between the two of around 1 million km^2 in lead years 6–9.

This drift could be caused by the so-called initialization shock: the model that has been nudged to observations returns to its biased equilibrium state immediately after the initialization process when the simulation starts to run freely. Since the assimilation runs are not performed with a coupled model, and since sea ice is not initialized, the initialization shock could be partly due to inconsistencies in the initial conditions. It can cause the model simulation to overshoot its preferred climatology, resulting in a larger error than that of the biased equilibrium (Meehl and Goddard, 2013). A similar drift has been found in the North Atlantic with the full-field initialization perturbing the overturning circulation, heat transport and associated SST and sea surface salinity in the region of the sub-polar gyre (Kröger et al., 2017). For the analysis presented here, the drift over the hindcast period has been corrected by applying a lead-time dependent cross-validated bias correction (CLIVAR, 2011) in the ESMValTool, as mentioned in Section 2.3.

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

The analysis shown in Fig. 2 was repeated using sea ice extent instead of SIA (not shown), but the main findings discussed in this section are not sensitive to whether SIE or SIA is used for the analysis.

3.3 Hindcast skill in the Arctic

The hypothesis for this study is that the enhanced hindcast skill in the North Atlantic in the MPI-ESM-LR *prot* system due to initialization may provide enhanced hindcast skill also in the Arctic for TAS, SST and SIC, which is the focus in the remainder of this study. Since MPI-ESM-HR *preop* does not show this improved skill (see Fig. 1), from here on we show results from the MPI-ESM-LR *prot* system only, although the entire analysis has also been done for the MPI-ESM-HR *preop* system to confirm the conclusions.

To investigate whether the skill improvement in North Atlantic TAS predictions could result in an improved skill in the Arctic, we repeat the analysis of Section 3.1 for TAS and SST in this region (Fig. 3), focussing on the September and March means (Fig. 3a–d). ACC for TAS evaluated against ERA-Interim data is generally high (above 0.6) with the exception of the Greenland Sea in March. In terms of improvement due to initialization, however, there is little to no skill gain (Fig. 3e–h). Solely along the east coast of Greenland and in the Fram Strait there is a statistically significant improvement of ACC skill in March for SST (Fig. 3d). Note that the skill in SST of the uninitialized MPI-ESM simulations is already high, i.e. any further improvement

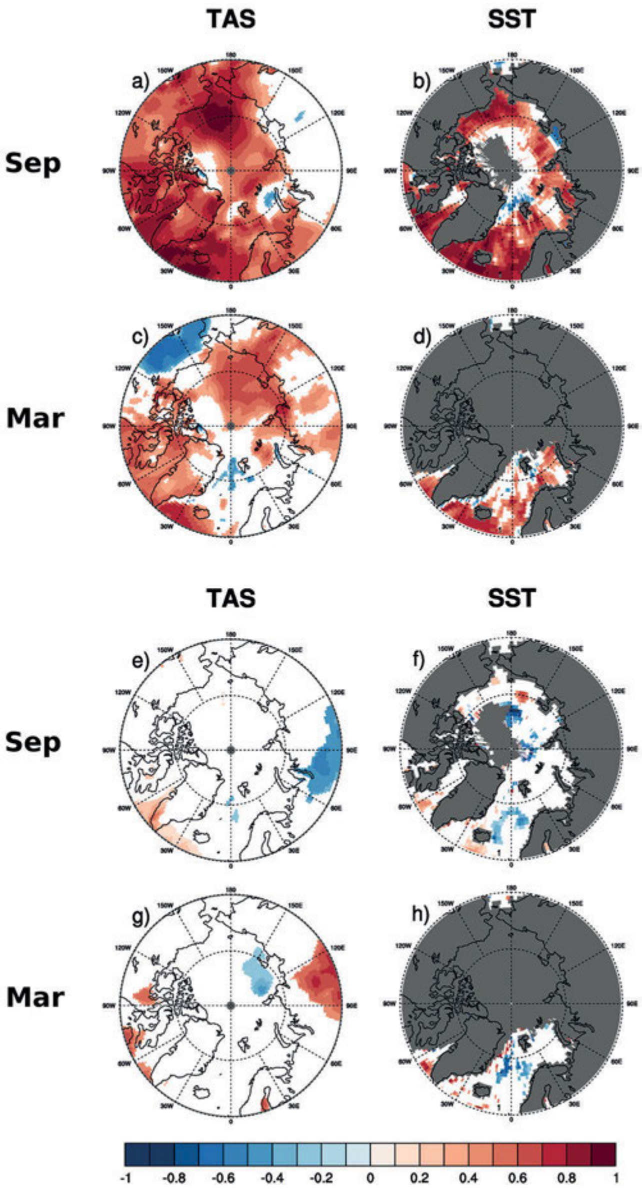

Figure 3: a–d: as Fig. 1, but for ACC of TAS and SST displayed in polar-stereographic projections in the Arctic and calculated against ERA-Interim and HadISST data, respectively. e–h: ACC difference between MPI-ESM-LR decadal and historical simulations. White grid cells denote values that are not statistically significant at the 95 % confidence level. Cells with dark grey colour represent missing values that either stem from gaps in the observational data or from constant SSTs due to ice coverage: in grid cells with sea ice, the SSTs in the model are set to a constant value of −1.9 °C.

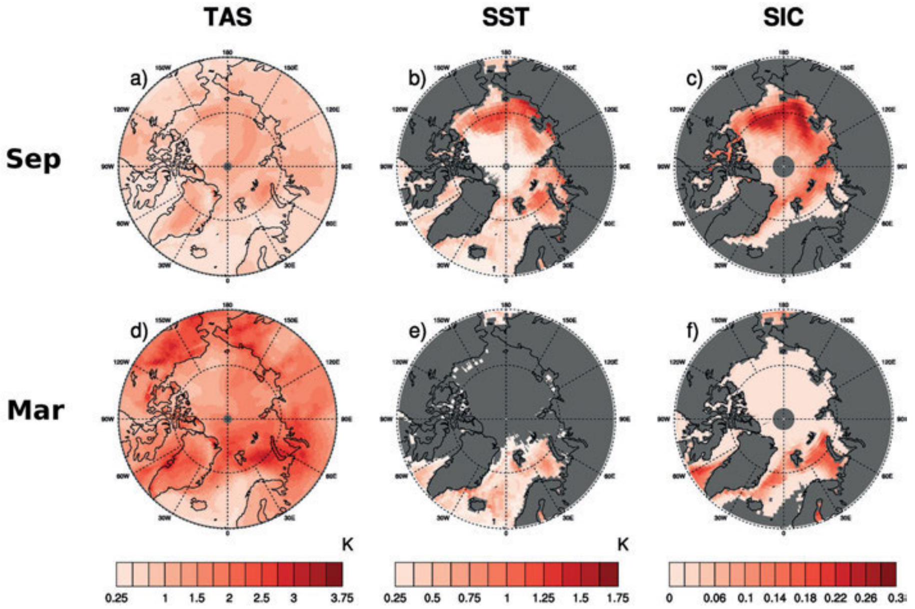

Figure 4: Arctic polar-stereographic contour maps of root mean square errors (RMSEs) between anomalies of the 30-member ensemble mean of the MPI-ESM-LR *prot* and respective reference datasets over the time period 1979–2013: ERA-Interim for TAS (a, d), HadISST for SST (b, e), NSIDC-NT for SIC (c, f).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

due to initialization can be expected to be rather small. We indeed find the largest improvement in a region where the ACC skill of the decadal hindcasts is not as high as in other regions (compare Fig. 3d to 3h), meaning that largest improvements occur in places where the historical simulations have particularly low skill.

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

We therefore use RMSE as a metric and calculate it for TAS, SST, and SIC. This is depicted in Fig. 4 for the MPI-ESM-LR *prot* hindcasts compared to the corresponding reference datasets. Note that all data represent anomalies to their respective climatological mean, with the hindcasts being lead month-wise bias corrected (see Section 2.3). This way, the RMSE is not affected by a constant offset between simulation and reference or a lead-time dependent drift (as seen in Fig. 2). For TAS, RMSEs are generally higher in March than in September (compare Fig. 4a to d). Especially in the northern part of the Greenland Sea and in the Kara and Barents Seas in the decadal hindcasts, we find RMSE values in March of up to 3 K compared to values smaller than 1.5 K in September. For both, SST and SIC, highest RMSE values are calculated for the East-Siberian and Beaufort Seas in September of 1.5 K (SST) and up to 0.3 (SIC), and for the Atlantic in March.

The change in model skill via initialization is shown in Fig. 5 for the three variables TAS, SST, and SIC, following Eq. (2.1). In September, there is a positive RMSE skill score of up to 0.5 (meaning the RMSE was halved) in the North Atlantic in TAS and SST (Figs. 5a and b), which is the same region for which an increase in TAS ACC skill is found (Fig. 1). A propagation of this RMSE improvement further northward into the Arctic Ocean is, however, not visible and in general not much improvement is found.

Hindcasts of the Arctic winter SSTs (Fig. 5e) show a statistically significant positive RMSE skill along the east coast of Greenland and in Fram Strait, due to the initialization of ocean temperature and salinity fields. This signal propagates into the decadal winter sea ice predictions decreasing RMSEs of SIC in the same area by about 30 % (Fig. 5f). Note that in the same area, the RMSE of SIC decadal hindcasts (Fig. 4f) is relatively high (around 0.16), which means that the absolute improvement via initialization in that region is also high. The March SST and SIC improvement in RMSE skill along the east coast of Greenland and in Fram Strait is also seen in the ACC (Fig. 3h). This is the only region in the Arctic in which we find statistically significant im-

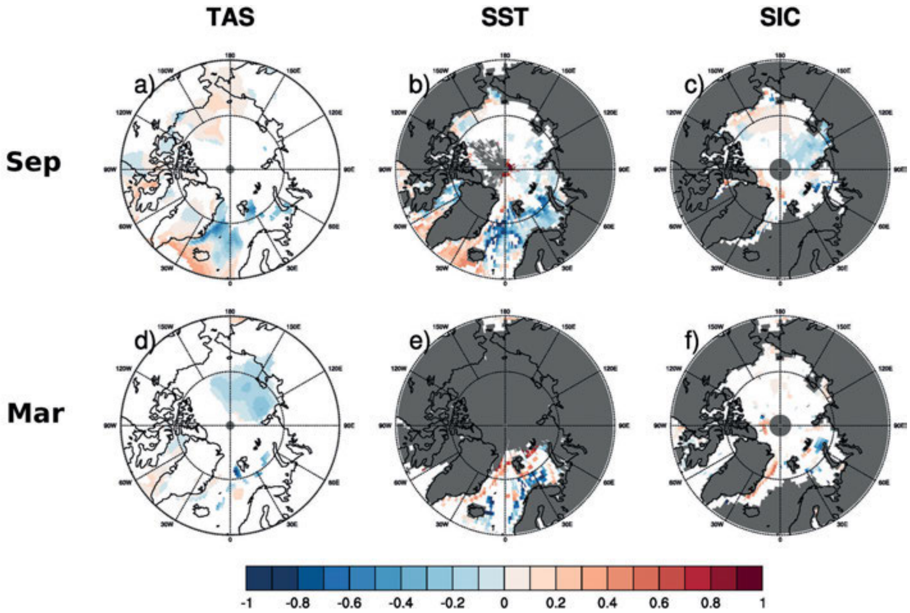

Figure 5: Same as Fig. 4, but for RMSE skill (2.1). Red (blue) colours indicate an improvement (degradation) of the model skill via initialization. All values shown are statistically significant at the 95 % confidence level following a significance test including the bootstrap algorithm (see Section 2.3).

provements by initialization that are robust against different metrics.

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

We further analyse the skill for SST and SIA for different Arctic regions. We only show those regions where improvements of hindcast skill could be detected: the Greenland and the Beaufort Seas. In order to investigate the effect of trends on the skill scores, we assess the ACC of detrended SST anomalies and detrended SIA anomalies for the two regions and for individual lead years (Fig. 6). In the Greenland Sea, we find an improvement in skill by initialization in lead year 1 in March for SIA. This corresponds to the RMSE reduction in SIC hindcasts seen in Fig. 5f and may be related to the ACC skill improvement in March SSTs in Fram Strait (Fig. 3 h). We do not find any other improvements for the Greenland Sea in September, neither for SST, nor for SIA. In the Beaufort Sea, September SST predictions are found to be improved via initialization only in lead year 5. This suggests that the high skill seen in Fig. 3b mainly stems from the model correctly reproducing the trend. The fact that we here analyse single lead years instead of a 4-year average as in Fig. 3, is another reason for the lower prediction skill shown in Fig. 6, since multi-year averages are easier to predict than single years (e.g., Goddard et al., 2013). Similar to SST, September SIA predictions are also improved by initialization in lead year 5 (and 6), emphasizing again the strong relationship in prediction skill between the two variables. However, we find no improvement in ACC for the Beaufort Sea in March.

In summary, a robust improvement across different variables and using different metrics could only be found in the Greenland Sea along the east coast of Greenland and in Fram Strait, and only for March. Thus, we conclude that the skill enhancement in the MiKlip prototype and preoperational systems compared to the uninitialized historical experiments in the Arctic is generally weak.

4 Summary and discussion

We assessed the skill in reproducing observed Arctic sea ice, sea surface temperature and near-surface air temperature in the MiKlip prototype decadal prediction system performed with the MPI-ESM-LR by developing and applying the ESMValTool evaluation system for decadal climate predictions. We focussed on the question whether the initialization of this model with observations improves the hindcast skill compared to uninitialized historical simulations in the Arctic. The evaluation was done by comparing model and observational data applying the metrics ACC and RMSE.

We find a statistically significant improvement in ACC skill for TAS by initialization in the North Atlantic,

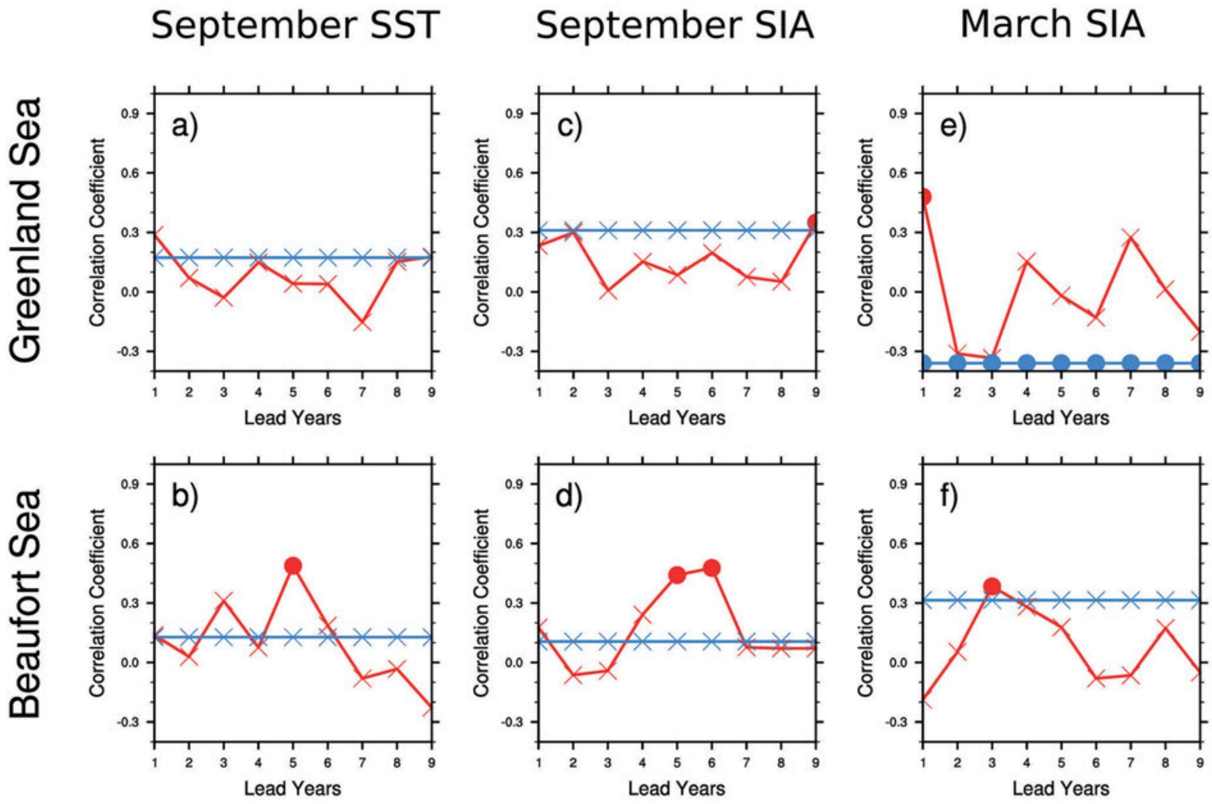

Figure 6: ACC against lead years for (a, b) regionally averaged September SST anomalies and (c–f) regionally integrated September and March SIA from MPI-ESM-LR *prot* decadal hindcasts (red) and MPI-ESM-LR historical simulations (blue). All data are detrended by subtracting the least squares linear trend. The two regions are the Greenland Sea (top row) and the Beaufort Sea (bottom row) and follow the definition of NSIDC with modifications. Reference for the correlation calculations are HadISST observations for SSTs and NSIDC-NT observations for SIA. The time period is 1982–2013. Filled circles indicate correlation coefficients that are statistically significant at the 95 % significance level according to a t-test for correlation coefficients.

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

ACC skill scores of Arctic TAS and SST reveal that this improvement in skill in the North Atlantic due to initialization does not propagate into the Arctic, and in some parts of it there is even a degradation in skill caused by the so-called initialization shock, i.e. a quick return of the initialized model to its biased equilibrium state after initialization (Meehl and Goddard, 2013). The model sometimes overshoots its intrinsic climatology, resulting in a larger bias than without initialization. This can be seen from time series of September pan-Arctic SIA showing a strong model drift in the MiKlip prototype system of about 1 million km^2 during the ten-year simulation periods. Pohlmann et al. (2016) found that the initialization shock in the MPI-ESM-LR decadal hindcasts stems from an overestimated trend in wind stress in the reanalyses used for initialization, leading to a displaced thermocline and large SST anomalies.

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

Why only certain regions are improved by initialization is still an open topic in MiKlip. In the Greenland Sea, the historical simulations show deficits in realistically simulating the year-to-year variability of sea ice. This is not expected from these long-term climate simulations, that aim at predicting the multi-year average climate conditions. The initialization, however, improves the year-to-year variability in that region because the simulations start from the correct phase of the observed natural variability. This directly increases the ACC skill. Further investigation of the underlying reasons for the regionality of improvement would require a) a sophisticated analysis of the parameterization of relevant processes in the model and b) a cross-variable assessment of multiple oceanic and atmospheric variables. Such an investigation is beyond the scope of this study.

Several studies suggest that the potential predictability limit for Arctic sea ice is higher than the actual prediction skill we were able to find (e.g., Guemas et al., 2014; Germe et al., 2014). Since atmospheric and oceanic variables included in the initialization process can improve regional temperature predictions, the inclusion of sea ice variables in the initialization process could potentially improve sea ice hindcasts. This has been done in seasonal forecasting with considerable success (e.g., Tietsche et al., 2014; Bunzel et al., 2016; Guemas et al., 2016). Therefore, we also analysed recent decadal hindcasts performed with the MiKlip preoperational decadal prediction system based on the high-resolution model configuration MPI-ESM-HR, where SIC was additionally initialized. However, improvement in TAS hindcast skill over the North Atlantic is not robust against different reference datasets in these simulations. Furthermore, timeseries of September pan-Arctic SIA show a strong negative bias of 1–3 million km^2 (dependent on reference observations and time) in all MPI-ESM-HR model runs. This bias is not present in March and indicates an unrealistic seasonal cycle of sea ice with too much melting at the end of summer in these simulations. Overall, the additional initialization of SIC brings no improvement in hindcast skill in the Arctic in the preoperational compared to the prototype system. These findings are consistent with several studies in the field of seasonal climate predictions suggesting that only initializing SIC but not any other sea ice quantities is not enough to improve sea ice forecasts. Only when also sea ice thickness or sea ice age is initialized together with SIC, seasonal forecasts of sea ice extent can be improved (e.g., Day et al., 2014a; Massonnet et al., 2015; Dirkson et al., 2017; Bushuk et al., 2017). This may be due to the fact that sea ice area has higher predictability in regions with thicker ice, with sea ice thickness being generally better predictable than sea ice area (Holland et al., 2010). We note that the MPI-ESM-HR decadal hindcasts available to date only consist of five ensemble members for each hindcast experiment as opposed to 30 members from MPI-ESM-LR. An increase in ensemble size was shown to improve the quality of decadal predictions (Sienz et al., 2015) by covering a higher variability range and could therefore change the specific conclusions for the MiKlip preoperational system.

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

5 Acknowledgements

This work was performed as part of the BMBF MiKlip Climate Model Validation by confronting globally Essential Climate Variables from models with observations (ClimVal project). We thank Bettina Gier, Robert Sausen (DLR), Felix Bunzel, Wolfgang Müller, and Dirk Notz (MPI-M Hamburg) for helpful comments on the manuscript and Melanie Bräu for her contributions to a subset of the sea-ice routines. The authors also thank the two anonymous reviewers for their very useful comments that helped to improve the manuscript.

References

Ahn, J., S. Hong, J. Cho, Y.-W. Lee, H. Lee, 2014: Statistical Modeling of Sea Ice Concentration Using Satellite Imagery and Climate Reanalysis Data in the Barents and Kara Seas, 1979–2012. – Remote Sens. **6**, 5520–5540. DOI: 10.3390/rs6065520.

Balmaseda, M.A., K.E. Trenberth, E. Källén, 2013: Distinctive climate signals in reanalysis of global ocean heat content. – Geophys. Res. Lett. **40**, 1754–1759. DOI: 10.1002/grl.50382.

Bellucci, A., R. Haarsma, N. Bellouin, B. Booth, C. Cagnazzo, B. van den Hurk, N. Keenlyside, T. Koenigk, F. Massonnet, S. Materia, M. Weiss, 2015: Advancements in decadal climate predictability: The role of nonoceanic drivers. – Rev. Geophys. **53**, 165–202. DOI: 10.1002/2014RG000473.

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

Bunzel, F., D. Notz, J. Baehr, W.A. Müller, K. Fröhlich, 2016: Observational uncertainty of Arctic sea-ice concentration significantly affects seasonal climate forecasts. – Geophys. Res. Lett., 1–8. DOI:10.1002/2015GL066928.

Bushuk, M., R. Msadek, M. Winton, G.A. Vecchi, R. Gudgel, A. Rosati, X. Yang, 2017: Skillful regional prediction of Arctic sea ice on seasonal timescales. – Geophys. Res. Lett., published online. DOI:10.1002/2017gl073155.

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

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

Cohen, J., J.A. Screen, J.C. Furtado, M. Barlow, D. Whittleston, D. Coumou, J. Francis, K. Dethloff, D. Entekhabi, J. Overland, J. Jones, 2014: Recent Arctic amplification and extreme mid-latitude weather. – Nature Geosci. **7**, 627–637. DOI:10.1038/ngeo2234.

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

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

Day, J.J., E. Hawkins, S. Tietsche, 2014a: Will Arctic sea ice thickness initialization improve seasonal forecast skill? – Geophys. Res. Lett. **41**, 7566–7575. DOI:10.1002/2014GL061694.

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

Dee, D.P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. – Quart. J. Roy. Meteor. Soc. **137**, 553–597. DOI:10.1002/qj.828.

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

Doblas-Reyes, F.J., I. Andreu-Burillo, Y. Chikamoto, J. García-Serrano, V. Guemas, M. Kimoto, T. Mochizuki, L.R.L. Rodrigues, G.J. van Oldenborgh, 2013: Initialized near-term regional climate change prediction. – Nature communications **4**, 1715. DOI:10.1038/ncomms2704.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? – Geophys. Res. Lett. **41**, 5620–5628. DOI: 10.1002/2014GL061146.

Eyring, V., I.S.A. Isaksen, T. Berntsen, W.J. Collins, J.J. Corbett, O. Endresen, R.G. Grainger, J. Moldanova, H. Schlager, D.S. Stevenson, 2010: Transport impacts on atmosphere and climate: Shipping. – Atmos. Env. **44**, 4735–4771. DOI:10.1016/j.atmosenv.2009.04.059.

Eyring, V., M. Righi, M. Evaldsson, A. Lauer, S. Wenzel, C. Jones, A. Anav, O. Andrews, I. Cionni, E.L. Davin, C. Deser, C. Ehbrecht, P. Friedlingstein, P. Gleckler, K.-D. Gottschaldt, S. Hagemann, M. Juckes, S. Kindermann, J. Krasting, D. Kunert, R. Levine, A. Loew, J. Mäkelä, G. Martin, E. Mason, A. Phillips, S. Read, C. Rio, R. Roehrig, D. Senftleben, A. Sterl, L.H. van Ulft, J. Walton, S. Wang, K.D. Williams, 2016: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP. – Geosci. Model Develop. Discuss. **8**, 7541–7661. DOI:10.5194/gmdd-8-7541-2015.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, M. Rummukainen, 2013: Evaluation of Climate Models. – Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 741–866. DOI:10.1017/CBO9781107415324.

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

Germe, A., M. Chevallier, D. Salas y Mélia, E. Sanchez-Gomez, C. Cassou, 2014: Interannual predictability of Arctic sea ice in a global climate model: regional contrasts and temporal evolution. – Climate Dyn. **43**, 2519–2538. DOI: 10.1007/s00382-014-2071-2.

Giorgetta, M.A., J.H. Jungclaus, C.H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, T. Crueger, M. Esch, K. Fieg, K. Glushak, V. Gayler, H. Haak, H.-D. Hollweg, T. Ilyina, S. Kinne, L. Kornblueh, D. Matei, T. Mauritsen, U. Mikolajewicz, W. Mueller, D. Notz, F. Pithan, T. Raddatz, S. Rast, R. Redler, E. Roeckner, H. Schmidt, R. Schnur, J. Segschneider, K.D. Six, M. Stockhause, C. Timmreck, J. Wegner, H. Widmann, K.-H. Wieners, M. Claussen, J. Marotzke, B. Stevens, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5. – J. Adv. Model. Earth Sys. **5**, 572–597. DOI: 10.1002/jame.20038.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S.J. Mason, B.P. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C.A.T. Ferro, D.B. Stephenson, G.A. Meehl, T. Stockdale, R. Burgman, A.M. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, T. Delworth, 2013: A verification framework for interannual-to-decadal predictions experiments. – Climate Dyn. **40**, 245–272. DOI:10.1007/s00382-012-1481-2.

Goessling, H.F., S. Tietsche, J.J. Day, E. Hawkins, T. Jung, 2016: Predictability of the Arctic sea ice edge. – Geophys. Res.Letters **43**, 1–9. DOI:10.1002/2015GL067232.

Guemas, V., E. Blanchard-Wrigglesworth, M. Chevallier, J.J. Day, M. Déqué, F.J. Doblas-Reyes, N.S. Fuèkar, A. Germe, E. Hawkins, S. Keeley, T. Koenigk, D. Salas y Mélia, S. Tietsche, 2014: A review on Arctic sea-ice predictability and prediction on seasonal to decadal timescales. – Quart. J. Roy. Meteor. Soc. **142**, 546–561. DOI: 10.1002/qj.2401.

Guemas, V., M. Chevallier, M. Déqué, O. Bellprat, F. Doblas-Reyes, 2016: Impact of sea ice initialization on sea ice and atmosphere prediction skill on seasonal timescales. – Geophys. Res. Lett. **43**, 3889–3896. DOI: 10.1002/2015gl066626.

Handorf, D., R. Jaiser, K. Dethloff, A. Rinke, J. Cohen, 2015: Impacts of Arctic sea ice and continental snow cover changes on atmospheric winter teleconnections. – Geophys. Res. Lett. **42**, 2367–2377. DOI:10.1002/2015gl063203.

Hawkins, E., S. Tietsche, J.J. Day, N. Melia, K. Haines, S. Keeley, 2015: Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems. – Quart. J. Roy. Meteor. Soc. **142**, 672–683. DOI:10.1002/qj.2643.

He, Y., B. Wang, M. Liu, L. Liu, Y. Yu, J. Liu, R. Li, C. Zhang, S. Xu, W. Huang, Q. Liu, Y. Wang, F. Li, 2017: Reduction of initial shock in decadal predictions using a new initialization strategy. – Geophys. Res. Lett. **44**, 8538–8547. DOI: 10.1002/2017gl074028.

Hibler, W.D., 1979: A Dynamic Thermodynamic Sea Ice Model. – J. Phys. Ocean. **9**, 815–846. DOI:10.1175/1520-0485(1979)009<0815:ADTSIM>2.0.CO;2.

Holland, M.M., 2010: Arctic sea ice and the potential for abrupt loss. – Geophysical Monograph Series, published online, DOI:10.1029/2008gm000787.

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

Jia, L., T. Delsole, 2013: Multi-year predictability of temperature and precipitation in multiple climate models. – Geophys. Res. Lett. **39**, 1–6. DOI:10.1029/2012GL052778.

Jones, P.D., D.H. Lister, T.J. Osborn, C. Harpham, M. Salmon, C.P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. – J. Geophys. Res. **117**, 5127.

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

Kennedy, J.J., N.A. Rayner, R.O. Smith, D.E. Parker, M. Saunby, 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. – J. Geophys. Res.**116**, D14104. DOI:10.1029/2010JD015220.

Kim, H.-M., P.J. Webster, J.A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. – Geophys. Res. Lett. **39**, published online. DOI:10.1029/2012gl051644.

Koenigk, T., C.K. Beatty, M. Caian, R. Döscher, K. Wyser, 2012: Potential decadal predictability and its sensitivity to sea ice albedo parameterization in a global coupled model. – Climate Dyn. **38**, 2389–2408. DOI:10.1007/s00382-011-1132-z.

Köhl, A., 2015: Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic. – Quart. J. Roy. Meteor. Soc. **141**, 166–181. DOI: 10.1002/qj.2347.

Komuro, Y., H. Hasumi, 2003: Effects of surface freshwater flux induced by sea ice transport on the global thermohaline circulation. – J. Geophys. Res. Oceans **108**, published online. DOI:10.1029/2002jc001476.

Kowalski, C.J., 1972: On the Effects of Non-normality on the Distribution of the Sample Product-moment Correlation Coefficient. – J. Roy. Statist. Soc. **21**, 1–12.

Kröger, J., H. Pohlmann, F. Sienz, J. Marotzke, J. Baehr, A. Köhl, M. Kameswarrao, I. Polkova, D. Stammer, F.S.E. Vamborg, W.A. Müller, 2017: Full-field initialized decadal predictions with the MPI Earth System Model: An initial shock in the North Atlantic. – Climate Dyn. DOI: 10.1007/s00382-017-4030-1.

Liu, C., K. Haines, A. Iwi, D. Smith, 2012: Comparing the UK Met Office Climate Prediction System DePreSys with idealized predictability in the HadCM3 model. – Quart. J. Roy. Meteor. Soc. **138**, 81–90. DOI:10.1002/qj.904.

Marotzke, J., W.A. Müller, F.S.E. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, K. Prömmel, H.W. Rust, D. Stammer, U. Ulbrich, C. Kadow, A. Köhl, J. Kröger, T. Kruschke, J. Pinto, G. , H. Pohlmann, M. Reyers, M. Schröder, F. Sienz, C. Timmreck, M. Ziese, 2016: MiKlip – a National Research Project on Decadal Climate Prediction. – Bull. Amer. Meteor. Soc., published online. DOI: 10.1175/BAMS-D-15-00184.1.

Marsland, S.J., H. Haak, J.H. Jungclaus, M. Latif, F. Röske, 2003: The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. – Ocean Model. **5**, 91–127. DOI:10.1016/s1463-5003(02)00015-x.

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

Meehl, G.A., L. Goddard, 2013: Decadal Climate Prediction: An Update from the Trenches. – Bull. Amer. Meteor. Soc. **3**, 1–82. DOI:10.1175/BAMS-D-12-00241.1.

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

Melia, N., K. Haines, E. Hawkins, 2016: Sea ice decline and 21^{st} century trans-Arctic shipping routes. – Geophys. Res. Lett. **43**, 9720–9728. DOI:10.1002/2016GL069315.

Morice, C.P., J.J. Kennedy, N.A. Rayner, P.D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. – J. Geophys. Res. Atmos. **117**, published online. DOI:10.1029/2011jd017187.

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

Müller, W.A., J. Baehr, H. Haak, J.H. Jungclaus, J. Krger, D. Matei, D. Notz, H. Pohlmann, J.S. Von Storch, J. Marotzke, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. – Geophys. Res. Lett. **39**, 1–7. DOI: 10.1029/2012GL053326.

Müller, V., H. Pohlmann, A. Düsterhus, D. Matei, J. Marotzke, W.A. Müller, M. Zeller, J. Baehr, 2016: Hindcast skill for the Atlantic meridional overturning circulation at 26.5° N within two MPI-ESM decadal climate prediction systems. – Climate Dyn. **49**, 2975–2990. DOI: 10.1007/s00382-016-3482-z.

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

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

Paxian, A., V. Eyring, W. Beer, R. Sausen, C. Wright, 2010: Present-day and future global bottom-up ship emission inventories including polar routes. – Env. Sci.Technol. **44**, 1333–1339. DOI:10.1021/es9022859.

Pohlmann, H., J.H. Jungclaus, A. Köhl, D. Stammer, J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. – J. Climate **22**, 3926–3938. DOI: 10.1175/2009JCLI2535.1.

Pohlmann, H., J. Kröger, R.J. Greatbatch, W.A. Müller, 2016: Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific. – Climate Dyn. **49**, 2685–2693. DOI:10.1007/s00382-016-3486-8.

Rayner, N.A., D.E. Parker, E.B. Horton, C.K. Folland, L.V. Alexander, D.P. Rowell, E.C. Kent, A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. – J. Geophys. Res. **108**, 4407. DOI:10.1029/2002JD002670.

Sienz, F., W.A. Müller, H. Pohlmann, 2015: Ensemble size impact on the decadal predictive skill assessment. – Meteorol. Z. **25**, 645–655. DOI:10.1127/metz/2016/0670.

Smith, D.M., R. Eade, N.J. Dunstone, D. Fereday, J.M. Murphy, H. Pohlmann, A. a. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. – Nature Geosci. **3**, 846–849. DOI:10.1038/ngeo1004.

Steele, M., S. Dickinson, 2016: The phenology of Arctic Ocean surface warming. – J. Geophys. Res. Oceans **121**, 6847–6861. DOI:10.1002/2016jc012089.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler, E. Roeckner, 2013: Atmospheric component of the MPI-M Earth System Model: ECHAM6. – J. Adv. Model. Earth Sys. **5**, 146–172. DOI: 10.1002/jame.20015.

Stroeve, J.C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, W.N. Meier, 2012a: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. – Geophys. Res. Lett. **39**, 1–7. DOI:10.1029/2012GL052676.

Stroeve, J.C., M.C. Serreze, M.M. Holland, J.E. Kay, J. Malanik, A.P. Barrett, 2012b: The Arctic's rapidly shrinking sea ice cover: A research synthesis. – Climatic Change **110**, 1005–1027. DOI:10.1007/s10584-011-0101-1.

Taylor, K.E., R.J. Stouffer, G.A. Meehl, 2012: An overview of CMIP5 and the experiment design. – Bull Amer. Meteor. Soci. **93**, 485–498. DOI:10.1175/BAMS-D-11-00094.1.

Thomson, A.M., K.V. Calvin, S.J. Smith, G.P. Kyle, A. Volke, P. Patel, S. Delgado-Arias, B. Bond-Lamberty, M.A. Wise, L.E. Clarke, J.A. Edmonds, 2011: RCP4.5: a pathway for stabilization of radiative forcing by 2100. – Climatic Change **109**, 77–94. DOI:10.1007/s10584-011-0151-4.

Tietsche, S., J.J. Day, V. Guemas, W.J. Hurlin, S.P.E. Keeley, D. Matei, R. Msadek, M. Collins, E. Hawkins, 2014: Seasonal to interannual Arctic sea ice predictability in current global climate models. – Geophys. Res. Lett. **41**, 1035–1043. DOI:10.1002/2013GL058755.

Uppala, S.M., P.W. Kållberg, A.J. Simmons, U. Andrae, V.D.C. Bechtold, M. Fiorino, J.K. Gibson, J. Haseler, A. Hernandez, G.A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R.P. Allan, E. Andersson, K. Arpe, M.A. Balmaseda, A.C.M. Beljaars, L.V.D. Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B.J. Hoskins, L. Isaksen, P.A.E.M. Janssen, R. Jenne, A.P. McNally, J.F. Mahfouf, J.J. Morcrette, N.A. Rayner, R.W. Saunders, P. Simon, A. Sterl, K.E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, J. Woollen, 2005: The ERA-40 re-analysis. – Quart. J. Roy. Meteor. Soc. **131**, 2961–3012. DOI:10.1256/qj.04.176.

Walsh, J., Chapman, W., Fetterer, F., 2015: Gridded monthly sea ice extent and concentration, 1850 onwards. Version 1, Tech. rep., NSIDC: National Snow and Ice Data Center, Boulder, Colorado USA, DOI:10.7265/N5833PZ5.

Wilks, D.S., 2011: Statistical Methods in the Atmospheric Sciences. – Int. Geophy. Ser. 102, published online. DOI: 10.1198/jasa.2007.s163.

Yang, Q., S.N. Losa, M. Losch, J. Liu, Z. Zhang, L. Nerger, H. Yang, 2017: Assimilating summer sea-ice concentration into a coupled ice–ocean model using a LSEIK filter. – Ann. Glaciol. **56**, 38–44. DOI:10.3189/2015AoG69A740.

Yeager, S.G., A.R. Karspeck, G. Danabasoglu, 2015: Predicted slowdown in the rate of Atlantic sea ice loss. – Geophys. Res. Lett. **42**, 10,704–10,713. DOI:10.1002/2015GL065364.