AGU PUBLICATIONS

Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE

10.1002/2016MS000787

Decadal climate predictions improved by ocean ensemble dispersion filtering

C. Kadow[1], S. Illing[1], I. Kröner[1], U. Ulbrich[1], and U. Cubasch[1]

[1]Institute of Meteorology, Freie Universität Berlin, Berlin, Germany

Key Points:
- Newly developed forecast method uses ocean ensemble dispersion filtering toward its ensemble mean
- Global and regional temperature as well as regional precipitation and cyclone predictions improve up to 5 years ahead
- It outperforms state-of-the-art predictions even with higher resolution and larger ensembles for typical setups with ensemble members <=10

Supporting Information:
- Supporting Information S1
- Figure S1
- Figure S2
- Figure S3
- Figure S4

Correspondence to:
C. Kadow,
Christopher.Kadow@met.fu-berlin.de

Citation:
Kadow, C., S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch (2017), Decadal climate predictions improved by ocean ensemble dispersion filtering, *J. Adv. Model. Earth Syst.*, *9*, 1138–1149, doi:10.1002/2016MS000787.

Received 29 AUG 2016
Accepted 5 APR 2017
Accepted article online 20 APR 2017
Published online 14 MAY 2017

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 time scales. 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 toward 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 toward the ensemble mean.

Plain Language Summary Decadal predictions aim to predict the climate several years in advance. Atmosphere-ocean interaction plays an important role for such climate forecasts. The ocean memory due to its heat capacity holds big potential skill. In recent years, more precise initialization techniques of coupled Earth system models (incl. atmosphere and ocean) have improved decadal predictions. Ensembles are another important aspect. Applying slightly perturbed predictions to trigger the famous butterfly effect results in an ensemble. Instead of evaluating one prediction, but the whole ensemble with its ensemble average, improves a prediction system. However, climate models in general start losing the initialized signal and its predictive skill from one forecast year to the next. Our study shows that the climate prediction skill of an Earth system model can be improved by a shift of the ocean state toward the ensemble mean of its individual members at seasonal intervals. We found that this procedure applying the average during the model run, called ensemble dispersion filter, results in more accurate results than the standard 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.

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 subseasonal scale, the ocean processes dominate the interannual to decadal time scale. As sea surface temperatures of ocean basins are key factors determining the atmospheric global mean temperature [*Meehl et al*., 2013, 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 multiannual 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*., 2011]. 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*., 2012; *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 nonlinear 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 setup of a new decadal forecast procedure, its evaluation methods, and the observational data used to validate the hindcast sets (see section 2). We show results of the global mean and regional temperature, precipitation, and winter cyclone predictions with the new method and its reference (see section 3), before we discuss and conclude this study (see section 4).

2. Modeling, Methods, and Data

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 three ensemble members or just initialization every fifth year.

The MiKlip setup [*Pohlmann et al*., 2013; *Kadow et al*., 2015] 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 five ensemble members is initialized every year on 1 January. The individual ensemble member of the full model system is started on different start days following 1 January to spread the ensemble—called lag-day initialization. The setup 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) 1–5 in the same time frame from 1979 to 2013 by shifting the

experiments (e.g., LY1 uses experiments initialized in 1978–2012, LY2 in 1977–2011, and so on) as suggested in the DCPP protocol. An "assimilation run" was set up to guide the MPI-ESM-LR model system toward an observational state. This reanalysis-like model run was used to start the prediction from. Therefore, the following reanalyses data were 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 subsection 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*., 2015]. 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*., 2015], 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.

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 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*., 2009]. 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 five 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 supporting information Text S3). 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 four of the five ensemble members when calculating the ensemble mean. Therefore, we have five combinations of four 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 (Figure 1a).

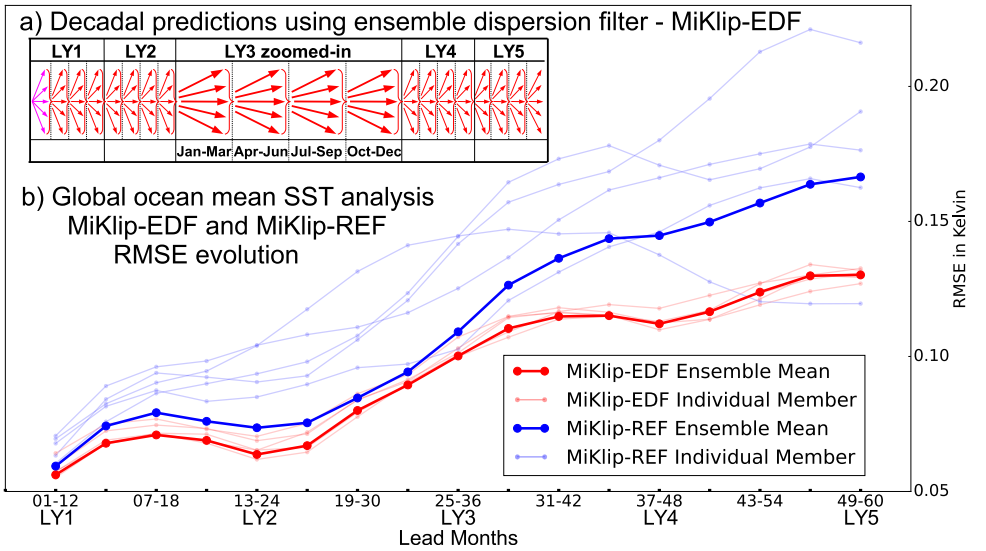

Figure 1. (a) (top) Schematic decadal climate prediction hindcast experiment setup of MiKlip-EDF in red. MiKlip-EDF consists of five ensemble members and 5 year integrations. The first 3 months of every experiment and ensemble member from MiKlip-EDF and MiKlip-REF are identical (magenta). Time frame 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.

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 1979–2013 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.* [2015] evaluating the MiKlip system, here the reference system MiKlip-REF.

The mean squared error skill score ($-\infty \leq \mathrm{MSESS} \leq 1$) 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 versus the observational climatology $\bar{O}$ compared to the observation O can be written as

$$MSESS(H,\bar{O},O) = 1 - \frac{MSE_H}{MSE_{\bar{O}}} = r_{HO}^2 - \left[r_{HO} - \frac{s_H}{s_O}\right]^2 \rightarrow 1 = \text{perfect skill score}$$

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

$$MSESS(H,R,O) = \frac{MSESS_H - MSESS_R}{1 - MSESS_R} \rightarrow 1 = \text{perfect skill score}$$

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*., 2015] which is symmetric around zero. To quantify the ensemble spread against the standard error, we use the average ensemble spread $\overline{\sigma_H^2}$ and the reference MSE_H

$$LESS = \ln\left(\frac{\overline{\sigma_H^2}}{MSE_H}\right) \rightarrow 0 = \text{perfect score}$$

If the MSE corresponds to the ensemble variance, the latter is a good estimate of the forecast uncertainty. If the ensemble variance is smaller than the MSE, the ensemble is said to be underdispersive (overconfident). An ensemble variance larger than the MSE indicates an overdispersive (underconfident) ensemble.

For the forecast reliability, we parameterize 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:

$$\textit{Reliability Classification} := \text{perfect} \mid \text{still useful} \mid \text{marginally useful} \mid \text{not useful}$$

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 smoothened field of reliability.

Observational and model data were spatially interpolated into a common $5^\circ \times 5^\circ$ 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., 2015]. Significance of the verification scores was estimated using a nonparametric bootstrap (1000-fold) approach [*Wilks*, 2011] taking autocorrelation 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.

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 [*Morice et al*., 2012]. It is a collaborative product with the ocean component HadSST3 [*Kennedy et al*., 2011] of the Met Office Hadley Centre and the land component CRUTEM4 of the Climatic Research Unit at the University of East Anglia [*Jones et al*., 2012]. 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 postprocessing 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 [*Rayner et al*., 2003] data set as a reference in the supporting information section to

evaluate the sea surface temperature biases of MiKlip-EDF and MiKlip-REF in comparison to the observations.

3. Results

The temporal development of the global ocean mean sea surface temperature as the RMSE in Kelvin compared to observations [*Kennedy et al.*, 2011] already indicates the benefits of the novel technique (Figure 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 (Figure 2). Here the MSESS compares MiKlip-EDF and MiKlip-REF against forecasting the climatological mean (Figure 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 three 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.

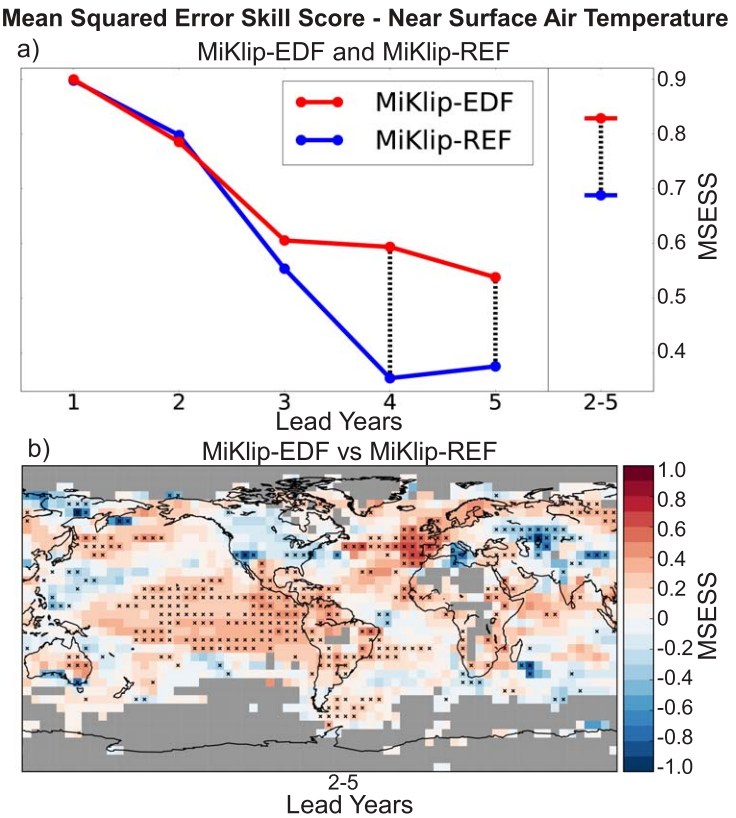

Figure 2. (a) Mean squared error skill score (MSESS) of the global mean temperature ensemble mean of MiKlip-REF (blue) and MiKlip-EDF (red) for LY1 to LY5 and LY2-5 compared to HadCrut4 with climatology as a reference prediction on the top. Significant differences of MiKlip-EDF to its reference prediction 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.

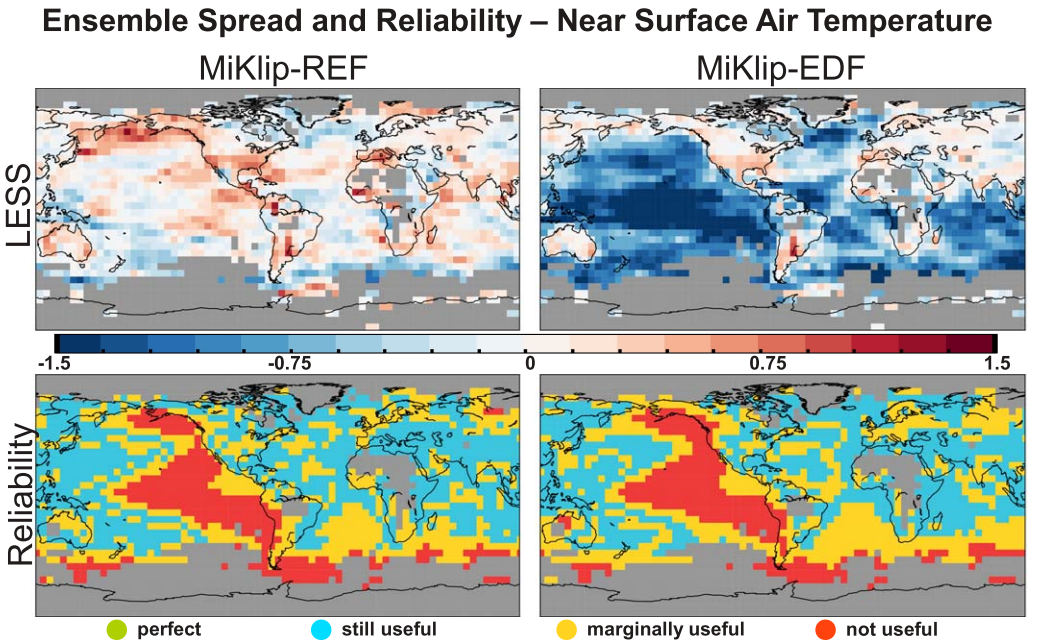

Figure 3. (top) Logarithmic ensemble spread score and (bottom) forecast reliability for near-surface air temperature in (left) MiKlip-REF and (right) MiKlip-EDF. Analyses are done for LY2-5 compared to HadCrut4. Gray areas indicate missing values with less than 90% data consistency in the observation. The analyses cover the time period from 1979 to 2013.

In addition to the improvement in the global mean temperature predictions, an enhancement of the skill on the regional scale can be seen (Figure 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 Figures 3a and 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

Table 1. Overview of Hindcast Systems Used in This Study[a]

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	AN	**3 months**
				T63L49	1.5°/L40	
MiKlip-REF	5	1979–2013	1974–2012	FF	AN	
				T63L49	1.5°/L40	
MiKlip-REF-10	**10**	1979–2013	1974–2012	FF	AN	
				T63L49	1.5°/L40	
MiKlip-REF-MR	5	1979–2013	1974–2012	FF	AN	
				T63**L95**	**0.4°/L40**	
MiKlip-REF-FF	5	1979–2013	1974–2012	FF	**FF**	
				T63L49	1.5°/L40	
MiKlip-REF-UN	5	1979–2013		T63L49	1.5°/L40	

[a]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 difference to the basic reference system MiKlip-REF.

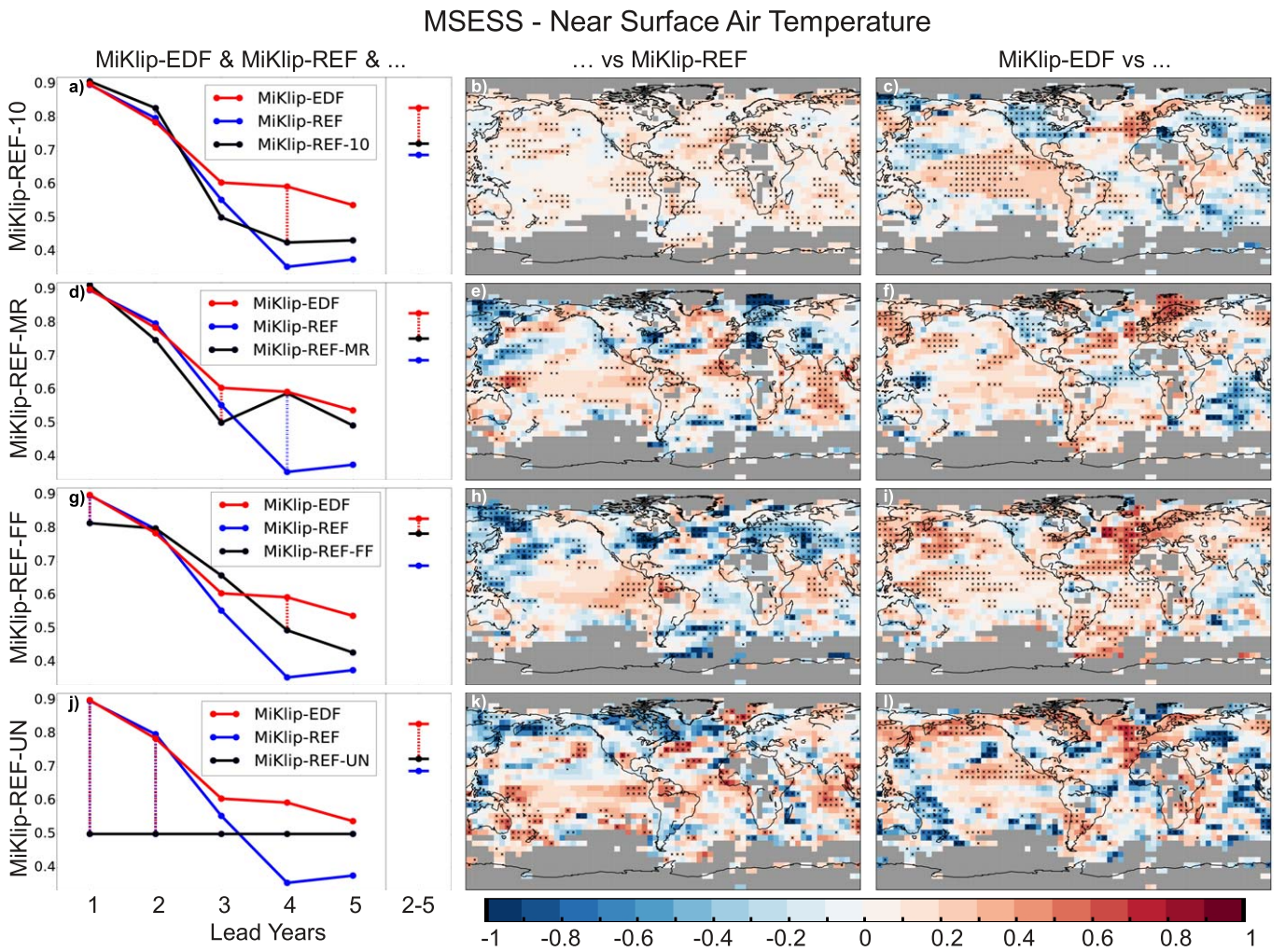

Figure 4. Mean squared error skill score (MSESS) of the global mean temperature with climatology as a reference prediction for (left column) LY1 to LY5 and LY2-5 of MiKlip-REF (blue), MiKlip-EDF (red), and other MiKlip reference data sets (black): (first row) MiKlip-REF-10 includes 10 instead of 5 ensemble 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 2.

forecast reliability analysis next to the spread evaluation shows no significant difference (Figures 3c and 3d). 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 to the free reference run. An additional and future approach could be bundling several independent five 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 (supporting information Figure S1).

For a more comprehensive temperature assessment, we also compared the new MiKlip-EDF data set to more recently developed sets of MiKlip experiments [*Marotzke et al.*, 2016] representing different decadal prediction strategies with the same model system (see Table 1). Figure 4 shows the comparison of MiKlip-EDF and MiKlip-REF to be directly compared with Figure 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

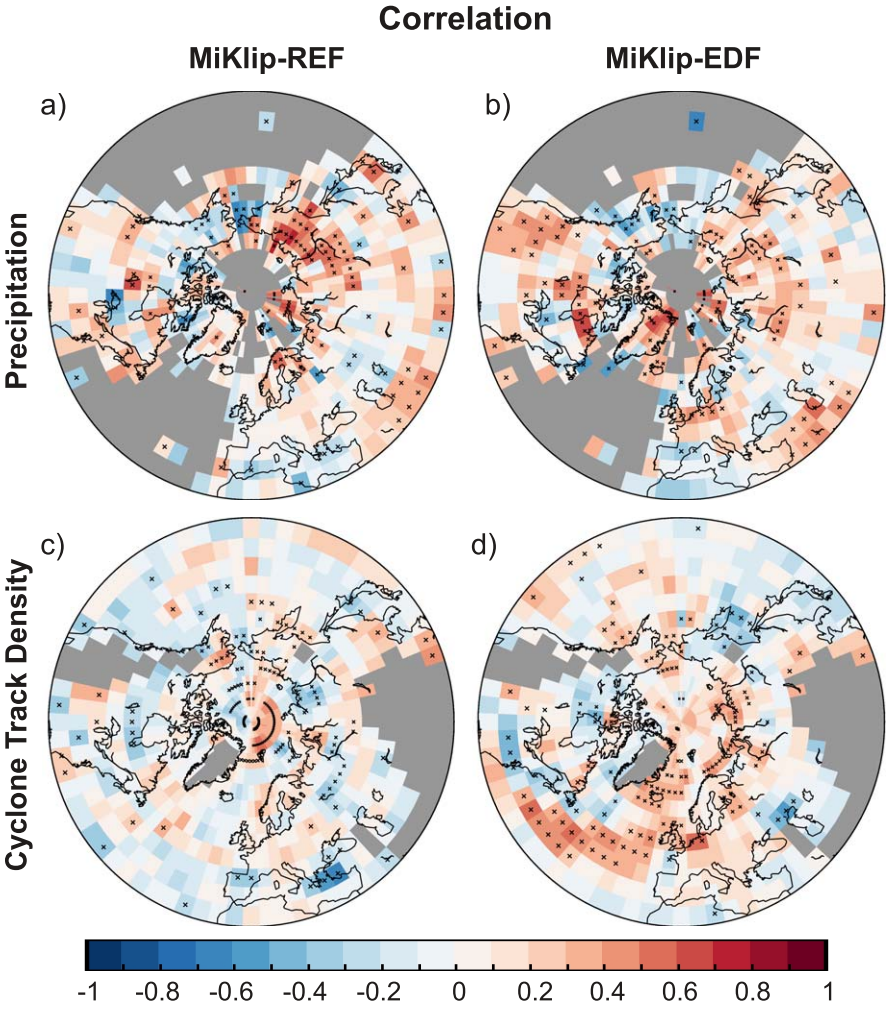

Figure 5. Correlations of (left) MiKlip-REF and (right) MiKlip-EDF (top) for precipitation compared to GPCC observations and (bottom) for DJF cyclone track density compared to ERA-Interim for the LY2-5 hindcasts sets over the period from 1979 to 2013. Significance is marked by black crosses. Gray shading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

other more sophisticated systems. However, these results especially in the global mean LY2-5 are not significant. The comparison against the uninitialized runs (Figure 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.

Forecasting long-term precipitation changes and multiannual 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 (supporting information Figure S2) in predicting large-scale anomalies on the global scale, which is in line with results of other studies [*Kadow et al.*, 2015]. We focus on the Northern Hemisphere and LY2-5 for a more detailed analysis of regional precipitation skill (Figures 5a and 5b). With respect to observations (Figure 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 (Figure 5b). Regions influenced by the ocean state and by atmospheric wind systems like cyclone tracks (see next paragraph and Figures 5c and 5d), 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 (supporting information Figure 3). 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 supporting information Figure S2).

Prediction of extratropical cyclones is another benchmark for decadal climate forecast systems [*Kruschke et al*., 2014]. We show the correlation of the lead winter (DJF) two to five cyclone track densities of MiKlip-REF and MiKlip-EDF compared to the ERA-Interim reanalysis in the Northern Hemisphere (Figures 5c and 5d). MiKlip-REF reveals no significant predictability (Figure 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 (Figure 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 (supporting information Figure 4).

4. Conclusions

The novel forecast technique presented here improves the multiannual 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 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 recentering 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. 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 subpolar 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*., 2009]. 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 *Shen et al*. [2016], 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.

Acknowledgments
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.

References

Balmaseda, M. A., K. Mogensen, and A. T. Weaver (2013), Evaluation of the ECMWF ocean reanalysis system ORAS4, *Q. J. R. Meteorol. Soc.*, *139*, 1132–1161, doi:10.1002/qj.2063.

Becker, A., P. Finger, A. Meyer-Christoffer, B. Rudolf, K. Schamm, U. Schneider, and M. Ziese (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, *Earth Syst. Sci. Data*, *5*, 71–99, doi:10.5194/essd-5-71-2013.

Boer, G. J., et al. (2016) The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, *Geosci. Model Dev.*, *9*, 3751–3777, doi:10.5194/gmd-9-3751-2016.

Breiman, L. (1996), Bagging predictors, *Mach. Learn.*, *24*, 123–140, doi:10.1007/BF00058655.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*, 553–597, doi:10.1002/qj.828.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and 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.

Evensen, G. (2003), The Ensemble Kalman Filter: Theoretical formulation and practical implementation, *Ocean Dyn.*, *53*, 343–367, doi:10.1007/s10236-003-0036-9.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016), Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev.*, *9*, 1937–1958, doi:10.5194/gmd-9-1937-2016.

Fyfe, J. C., W. J. Merryfield, V. Kharin, G. J. Boer, W.-S. Lee, and K. von Salzen (2011) Skillful predictions of decadal trends in global mean surface temperature, *Geophys. Res. Lett.*, *38*, L22801, doi:10.1029/2011GL049508.

Goddard, L., et al. (2013), A verification framework for interannual-to-decadal predictions experiments, *Clim. Dyn.*, *40*(1–2), 245–272, doi:10.1007/s00382-012-1481-2.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, 2nd ed., 763 pp., Springer, New York.

Hsu, W.-R., and A. H. Murphy (1986), The attributes diagram: A geometrical framework for assessing the quality of probability forecasts, *Int. J. Forecasting*, *2*, 285–293.

ICPO (2011), Decadal and bias correction for decadal climate predictions, *CLIVAR Publ. Ser. 150*, 6 pp., Southampton, U. K.

Illing, S., C. Kadow, O. Kunst, and U. Cubasch (2014), A tool for standardized evaluation of decadal hindcast systems, *J. Open Res. Software*, *2*(1), e24, doi:10.5334/jors.bf.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and 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*, D05127, doi:10.1029/2011JD017139.

Jungclaus, J. H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, and J. S. vonStorch (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 Syst.*, *5*, 422–446, doi:10.1002/jame.20023.

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

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*, doi:10.1594/PANGAEA.874231.

Kalnay, E., B. Hunt, E. Ott, and I. Szunyogh (2006), Ensemble forecasting and data assimilation: Two problems with the same solution?, in *Predictability of Weather and Climate*, pp. 157–180, Cambridge Univ. Press, Cambridge, U. K.

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner (2008) Advancing decadal-scale climate prediction in the North Atlantic sector, *Nature*, *453*, 84–88, doi:10.1038/nature06921.

Keller, J. D., L. Kornblueh, A. Hense, and A. Rhodin (2008), Towards a GME ensemble forecasting system: Ensemble initialization using the breeding technique, *Meteorol. Z.*, *17*, 707–718, doi:10.1127/0941-2948/2008/0333.

Kennedy, J. J., N. A. Rayner, R. O. Smith, M. Saunby, and D. E. Parker (2011), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850: 1. Measurement and sampling errors, *J. Geophys. Res.*, *116*, D14103, doi:10.1029/2010JD015218.

Kharin, V. V., G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W. S. Lee (2012), Statistical adjustment of decadal predictions in a changing climate, *Geophys. Res. Lett.*, *39*, L19705, doi:10.1029/2012GL05264.

Kosaka, Y., and S. P. Xie (2013), Recent global-warming hiatus tied to equatorial Pacific surface cooling, *Nature*, *501*, 403+, doi:10.1038/nature12534.

Kruschke, T., H. W. Rust, C. Kadow, G. C. Leckebusch, and U. Ulbrich (2014), Evaluating decadal predictions of northern hemispheric cyclone frequencies, *Tellus, Ser. A*, *66*, 22830, doi:10.3402/tellusa.v66.22830.

Kruschke, T., H. W. Rust, C. Kadow, W. A. Müller, H. Pohlmann, G. C. Leckebusch, and U. Ulbrich (2015), Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, *Meteorol. Z.*, *25*, 721–738, doi:10.1127/metz/2015/0641.

Kumar, A., and M. P. Hoerling (2000), Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction, *Bull. Am. Meteorol. Soc.*, *81*, 255–264.

Lorenz, E. (1963), Deterministic nonperiodic flow, *J. Atmos. Sci.*, *20*, 130–141, doi:10.1175/1520-0469(1963)020 < 0130:DNF>2.0.CO;2.

Marotzke, J., et al. (2016), MiKlip—A National Research Project on decadal climate prediction, *Bull. Am. Meteorol. Soc.*, *97*, 2379–2394, doi:10.1175/BAMS-D-15-00184.1.

Matei, D., J. Baehr, J. H. Jungclaus, H. Haak, W. A. Müller, and J. Marotzke (2012), Multiyear prediction of monthly mean Atlantic meridional overturning circulation at 26.5°N, *Science*, *335*, 76–79, doi:10.1126/science.1210299.

Meehl, G. A., et al. (2009), Decadal prediction, *Bull. Am. Meteorol. Soc.*, *90*, 1467–1485, doi:10.1175/2009BAMS2778.1.

Meehl, G. A., A. Hu, J. M. Arblaster, J. T. Fasullo, and K. E. Trenberth (2013), Externally forced and internally generated decadal climate variability associated with the interdecadal pacific oscillation, *J. Clim.*, *26*, 7298–7310, doi:10.1175/JCLI-D-12-00548.1.

Meehl, G. A., et al. (2014), Decadal climate prediction: An update from the trenches, *Bull. Am. Meteorol. Soc.*, *95*, 243–267, doi:10.1175/BAMS-D-12-00241.1.

Mehta, V., G. Meehl, L. Goddard, J. Knight, A. Kumar, M. Latif, T. Lee, A. Rosati, and D. Stammer (2011), Decadal climate predictability and prediction where are we?, *Bull. Am. Meteorol. Soc.*, *92*(5), 637–640.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, *J. Geophys. Res.*, *117*, D08101, doi:10.1029/2011JD017187.

Murphy, A. (1988), Skill scores based on the mean-square error and their relationships to the correlation-coefficient, *Mon. Weather Rev.*, *116*, 2417–2425, doi:10.1175/1520-0493(1988)116 < 2417:SSBOTM>2.0.CO;2.

Murphy, A., and E. Epstein (1989), Skill scores and correlation-coefficients in model verification, *Mon. Weather Rev.*, *117*, 572–581, doi: 10.1175/1520-0493(1989)117 < 0572:SSACCI>2.0.CO;2.

Murray, R. J., and I. A. Simmonds (1991), A numerical scheme for tracking cyclone centres from digital data. Part I: Development and operation of the scheme, *Aust. Meteorol. Mag.*, *39*, 155–166.

Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith (2006), Ensemble prediction: A pedagogical perspective, *ECMWF Newslett.*, *106*, 10–17.

Pinto, J. G., T. Spangehl, U. Ulbrich, and P. Speth (2005), Sensitivities of a cyclone detection and tracking algorithm: Individual tracks and climatology, *Meteorol. Z.*, *14*, 823–838.

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

Pohlmann, H., W. A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F. S. E. Vamborg, C. Kadow, S. Illing, and J. Marotzke (2013), Improved forecast skill in the tropics in the new MiKlip decadal climate predictions, *Geophys. Res. Lett.*, *40*, 5798–5802, doi:10.1002/2013GL058051.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and 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*(D14), 4407, doi: 10.1029/2002JD002670.

Shen, M.-L., N. Keenlyside, F. Selten, W. Wiegerinck, and G. S. Duane (2016), Dynamically combining climate models to "supermodel" the tropical Pacific, *Geophys. Res. Lett.*, *43*, 359–366, doi:10.1002/2015GL066562.

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

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy (2007), Improved surface temperature prediction for the coming decade from a global climate model, *Science*, *317*, 796–799, doi:10.1126/science.1139540.

Smith, D. M., R. Eade, and H. Pohlmann (2013), A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction, *Clim. Dyn.*, *41*, 3325–3338, doi:10.1007/s00382-013-1683-2.

Stevens, B., et al. (2013), Atmospheric component of the MPI-M Earth System Model: ECHAM6, *J. Adv. Model. Earth Syst.*, *5*, 146–172, doi: 10.1002/jame.20015.

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, M. Schröder, A. Mazurkiewicz, and A. Hense (2015), Revealing skill of the MiKliP decadal prediction systems by three-dimensional probabilistic evaluation, *Meteorol. Z.*, *25*, 657–671, doi:10.1127/metz/2015/0606.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Uppala, S. M., et al. (2005), The ERA-40 reanalysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012, doi:10.1256/qj.04.176.

Weisheimer, A., and T. N. Palmer (2014), On the reliability of seasonal climate forecasts, *J. R. Soc. Interface*, *11*, 20131162, doi:10.1098/rsif.2013.1162.

Wilks, D. (2011), *Statistical Methods in the Atmospheric Sciences*, 627 pp., Academic, Oxford, U. K.

Yeager, S., A. Karspeck, G. Danabasoglu, J. Tribbia, and H. Teng (2012) A decadal prediction case study: Late twentieth-century North Atlantic Ocean heat content, *J. Clim.*, *25*, 5173–5189, doi:10.1175/JCLI-D-11-00595.1.

Zappa, G., L. C. Shaffrey, and K. I. Hodges (2013), The ability of CMIP5 models to simulate North Atlantic extratropical cyclones, *J. Clim.*, *26*, 5379–5396, doi:10.1175/JCLI-D-12-00501.1.

Zender, C. S. (2008), Analysis of self-describing gridded geoscience data with netCDF Operators (NCO), *Environ. Modell. Software*, *23*(10), 1338–1342, doi:10.1016/j.envsoft.2008.03.004.