# Initialization and ensemble generation for decadal climate predictions: A comparison of different methods

Iuliia Polkova[1], Sebastian Brune[1], Christopher Kadow[2], Vanya Romanova[3], Gereon Gollan[4], Johanna Baehr[1], Rita Glowienka-Hense[3], Richard J. Greatbatch[4], Andreas Hense[3], Sebastian Illing[2], Armin Köhl[1], Jürgen Kröger[5], Wolfgang A. Müller[5], Klaus Pankatz[6] and Detlef Stammer[1]

Iuliia Polkova, iuliia.polkova@uni-hamburg.de

[1]Institute of Oceanography, Universität Hamburg, CEN, Hamburg, Germany.

[2]Institute for Meteorology, Freie Universität Berlin, Berlin, Germany.

[3]Meteorological Institute, University Bonn, Bonn, Germany.

[4]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2018MS001439

**Abstract.** Five initialization and ensemble generation methods are investigated with respect to their impact on the prediction skill of the German decadal prediction system "Mittelfristige Klimaprognose" (MiKlip). Among the tested methods, three tackle aspects of model-consistent initialization using the ensemble Kalman filter (EnKF), the filtered anomaly initialization (FAI) and the initialization method by partially coupled spin-up (MODINI). The remaining two methods alter the ensemble generation: the ensemble dispersion filter (EDF) corrects each ensemble member with the ensemble mean during model integration. And the bred vectors (BV) perturb the climate state using the fastest growing modes. The new methods are compared against the latest MiKlip system in the low-resolution configuration (Preop-LR), which uses lagging the climate state by a few days for ensemble generation and nudging toward ocean and atmosphere reanalyses for initialization. Results show that the tested methods provide an added value for the prediction skill as compared to Preop-LR in that they improve prediction skill over the eastern and central Pacific and different regions in the North Atlantic Ocean. In this respect, the EnKF and FAI show the most distinct improvements over

[5]Max Planck Institute for Meteorology,
Hamburg, Germany.

[6]Deutscher Wetterdienst, Offenbach am
Main, Germany.

Preop-LR for surface temperatures and upper ocean heat content, followed by the BV, the EDF and MODINI. However, no single method exists that is superior to the others with respect to all metrics considered. In particular, all methods affect the Atlantic Meridional Overturning Circulation in different ways, both with respect to the basin-wide long-term mean and variability, and with respect to the temporal evolution at the 26°N latitude.

**Keypoints:**

- Five initialization and ensemble generation methods are tested with respect to their impact on the skill of a decadal prediction system.
- Results show that the tested methods provide an added value for the prediction skill as compared to the reference prediction system.
- The study deals with dynamical consistency during initialization and ocean initial state uncertainty.

## 1. Introduction

Over the last years, decadal climate prediction has matured substantially to the point that it is now being developed into semi-operational applications [see, for example, *Smith et al.*, 2007; *Keenlyside et al.*, 2008; *Pohlmann et al.*, 2009; *Doblas-Reyes et al.*, 2011; *Kirtman et al.*, 2013; *Meehl et al.*, 2014; *Boer et al.*, 2016]. A recent example is the decadal climate prediction system MiKlip [developed during the German project "Mittelfristige Klimaprognose", *Marotzke et al.*, 2016] that within the near future will be taken over by the German Meteorological Office (DWD). Despite this enormous success, however, many important aspects need to be further improved to boost up the current level of prediction skill of such a system to what can be expected from theoretical considerations [e.g., *Griffies and Bryan*, 1997; *Branstator and Teng*, 2010, 2012; *Boer et al.*, 2013].

It can be expected that among the candidates leading to further improvement are the initialization as well as ensemble generation methods. With respect to initialization, we know that present practices that imply using different models for generating initial states and making predictions and/or using uncoupled data assimilation systems for initialization might be suboptimal due to presence of dynamic inconsistencies between the initial state and the prediction system leading to initialization shocks [e.g., *Mulholland et al.*, 2016; *Pohlmann et al.*, 2017; *Kröger et al.*, 2018]. With respect to ensemble generation, due to insufficient span of ensemble spread, some sources of prediction uncertainty are underrepresented [*Palmer*, 2000; *Germe et al.*, 2017]. In order to foster progress on both fronts, the aim of this paper is to analyze novel initialization and ensemble generation methods with respect to their potential effect on the skill of a decadal prediction sys-

tem. The new methods are tested in the framework of the MiKlip prediction system, but findings reported below should hold also for other systems.

Previously, the MiKlip prediction system, which is based on the Max Planck Institute for Meteorology Earth System Model (MPI-ESM), has undergone several development stages and evolved from the baseline system with ocean-only initialization to the pre-operational system with initialized atmosphere, ocean and sea ice components. Overall, the prediction system shows a robust skill for annual surface temperatures over large areas of the world ocean and for seasonal temperatures over Europe [*Müller et al.*, 2012; *Marotzke et al.*, 2016]. However, regions like the North Atlantic subpolar gyre, the central and eastern Pacific Ocean are especially sensitive to initialization; in these regions, the prediction system is swiftly losing prediction skill [*Pohlmann et al.*, 2017; *Kröger et al.*, 2018].

The current study is performed in the framework of the pre-operational MiKlip prediction system in low resolution configuration (Preop-LR). During the course of the study we investigate whether the MiKlip system can further benefit from initialization and ensemble generation methods such as an oceanic ensemble Kalman filter (EnKF), a filtered anomaly initialization (FAI), an initialization method by partially coupled spin-up (MODINI), an ensemble dispersion filter (EDF) and an oceanic bred vectors method (BV). Using these methods, which are described in detail in Section 2, we will investigate whether dynamically consistent initialization and improved representation of ocean initial state uncertainty may further improve the prediction skill of predictable components. The skill assessment is carried out on ten-member ensembles and covers the verification period 1962-2016 for all the initialization and ensemble generation methods.

The remainder of the paper is organized as follows: In Section 2, we briefly describe the MiKlip prediction system and give more details on the proposed initialization and ensemble generation methods. In Section 3, we analyze prediction skill for surface air temperature (SAT). Also, analyses of ocean heat content in the upper 700 meters (HC700) and the Atlantic Meridional Overturning Circulation (AMOC) are provided to better understand the performance of the different methods. Finally, Section 4 summarizes the results and provides the conclusions.

## 2. Methodology

All the retrospective initialized decadal predictions (hereafter initialized hindcasts) and the un-initialized historical simulations are based on the MPI-ESM version 1.2 [*Müller et al.*, 2018]. All simulations are performed in the low-resolution (LR) configuration MPI-ESM-LR. The atmospheric component of MPI-ESM is ECHAM6 [*Stevens et al.*, 2013] configured with a T63L47 resolution. The oceanic component MPIOM [*Jungclaus et al.*, 2013] is implemented with 1.5° horizontal resolution and 40 vertical levels.

### 2.1. Preoperational MiKlip Prediction System (Preop-LR)

Preop-LR is used as a reference, against which the methods described in the following Section 2.2 are compared. The Preop-LR assimilation is based on nudging the ocean toward ORAS4 temperature and salinity (T&S) anomalies [*Balmaseda et al.*, 2013] and the atmosphere toward ERA-40/ERA-Interim temperature, vorticity, divergence and surface pressure full-field values [*Uppala et al.*, 2005; *Dee et al.*, 2011]. ORAS4 anomalies are calculated with respect to the 1961-2005 climatology. The full-field initialization based on ORAS4 and GECCO2 reanalyses was tested earlier in the project [prototype systems

described by *Marotzke et al.*, 2016] and was dismissed at one of the development stages of the prediction system after *Kröger et al.* [2018] has demonstrated that MPI-ESM shows better skill for the North Atlantic when initialized from ORAS4 anomalies. Their study showed that nudging toward temperature and salinity full fields from the reanalysis induced ocean heat and mass transport changes, which triggered model adjustments through artificial heat sources and sinks in the forecast mode.

The Preop-LR nudging run is started from the historical simulation and is carried out over 1960-2016. The relaxation time in the ocean is 11 days, and in the atmosphere: 6 hours for vorticity, 24 hours for temperature and surface pressure, and 48 hours for divergence. The sea-ice concentration is nudged to the NSIDC data [*Fetterer et al.*, 2016] with the relaxation time of 11 days. The initialized hindcasts are started from the 1st of November over 1960-2016. Hindcast ensembles consist of 10 ensemble members and are generated by lagging the climate state by 1-9 days after the initialization date. The external forcing that is applied to initialized hindcasts is the same as for the uninitialized historical simulations: the CMIP5 solar irradiance data, aerosol and greenhouse gas concentrations over 1850–2005 [*Taylor et al.*, 2012], and the RCP4.5 pathway over 2006–2025 [*Giorgetta et al.*, 2013]. Further details concerning the Preop-LR experimental setup and differences to it from the test-suite experiments are given in Table 1.

### 2.2. Details of the Test-Suite Setup

The MiKlip project is organized in four modules focusing on initialization, evaluation, processes and regionalization and is concentrated around a single prediction system [www.fona-miklip.de, *Marotzke et al.*, 2016]. These modules provide tailored research for the central prediction system to enable its further advancement. Hence, when performing

the test-suite experiments for this study, initialization module followed the Preop-LR setup as close as possible, i.e., used the same model setup, model version, forcing and nudging fields and post-processing procedure to enable clean comparison. Finally, work-packages which provided minimum requirement of 10 ensemble members, yearly initialization over 1960-2015 were qualified for the test-suite. The minimum requirement for ensemble size and initialization period followed recommendations of the study by *Sienz et al.* [2016]. Thus, it is expected that differences between the performance of the proposed methods and Preop-LR are attributable to a particular initialization and ensemble generation method.

#### 2.2.1. Ensemble Kalman Filter (EnKF)

The oceanic EnKF analyzed in this study is the first attempt to directly insert ocean data into the MiKlip prediction system for initialization of decadal predictions. This method assimilates full values (i.e. not corrected for any model bias as for instance in *Counillon et al.* [2016]) of the monthly subsurface temperature and salinity profiles from the EN4 data [*Good et al.*, 2013]. Thus, in contrast to using nudging of reanalysis as in Preop-LR, another ocean model's biases do not enter the prediction system. The EnKF represents a weakly coupled data assimilation system. While the oceanic component of each ensemble member is updated by the oceanic EnKF once a month, the atmospheric component of each member is nudged towards ERA-40/ERA-Interim at every time step (10 min). The fluxes between the atmosphere and ocean are exchanged during coupling every hour. The EnKF in its local variant as used in this study is a follow up of the global variant approach described by *Brune et al.* [2015, 2018]. The method is based on the Parallel Data Assimilation Framework PDAF [*Nerger and Hiller*, 2012] with the localized

singular evolutive interpolated ensemble Kalman filter LSEIK [*Pham*, 2001].

The EnKF assimilation involves two spin-up phases to allow MPI-ESM adapting its climate and oceanic overturning circulation to the observed climate and to mitigate initialization shocks. Theses spin-ups **1)** and **2)** are followed by the actual coupled assimilation **3)**:

**1)** The first spin-up phase accounts for 300 years of atmospheric nudging in MPI-ESM toward a monthly climatology from ERA-40 calculated over the period 1958-1967 (1 ensemble member). No explicit restrictions are applied to the oceanic component. The atmospheric nudging methodology is similar to that in Preop-LR (see Section 2.1).

**2)** The second spin-up phase represents a 50-year coupled assimilation with atmospheric nudging and the oceanic EnKF. 16 ensemble members are generated from the last 16 years of the spin-up phase 1. The oceanic EnKF assimilates once per month full-value EN4 monthly climatological temperature and salinity profiles calculated by aggregating all available profiles over the period 1950-1959 in monthly bins. The technical details of the EnKF assimilation are described in step **3)**. The atmospheric nudging uses the same monthly climatology from ERA-40 as in spin-up phase 1.

**3)** The coupled assimilation, which serves as a pool of initial conditions for decadal predictions, is started at the end of spin-up phase 2 and covers the period 1958-2016, over which the full-value EN4 temperature and salinity profiles are assimilated into MPI-ESM through the oceanic EnKF, and ERA-40/ERA-Interim through nudging. In this set-up, no atmospheric nudging is applied near the air-sea boundary in the lower 5 atmospheric layers. Also no satellite data are assimilated in the EnKF. The main differences between

the EnKF assimilation with the localized variant as in this study and the assimilation with the global variant as in *Brune et al.* [2015, 2018] are horizontal localization (here, 5 degree radius), vertical localization within a single layer, parameter localization (parameter-wise) and an artificial inflation of the ensemble spread (between 1 and 1.01). The observation error is 1 K for temperature and 1 psu for salinity; it is uniformly applied at all grid cells similar to *Brune et al.* [2015, 2018]. Overall, the current EnKF setup leads to a stronger impact of oceanic observations as compared to the former EnKF setup with the global variant; with that we aim for a stronger impact of the oceanic component on the prediction skill. In contrast to Preop-LR, sea-ice nudging to NSIDC is not used because in the EnKF it leads to a degeneration of the oceanic state estimate near the ocean surface in regions close to the ice edge.

Hindcasts are initialized every year from the 1st of November over 1959-2016 and run for 10 years and 2 months (Table 1). Every hindcast member is a direct continuation of the corresponding assimilation member without any further assimilation applied. For the intercomparison in this study, we use the first 10 EnKF ensemble members.

#### 2.2.2. Filtered Anomaly Initialization (FAI)

Filtering variability from the non-native reanalysis that cannot be predicted by the MiKlip prediction system is implemented in FAI by projecting ocean reanalysis anomalies onto the modes of variability inherent to the prediction system. Similar methodologies have been implemented to eliminate the effect of higher frequency components (noise) on the numerical weather forecast skill, to obtain a correctly balanced initial state for data assimilation procedures, and to initialize long-lived stable modes for seasonal predic-

tions [e.g., *Williamson*, 1976; *Ballish*, 1981; *Branstator et al.*, 1993]. To this end, FAI is implemented as follows:

**1)** First, we derive modes of variability using the bivariate empirical orthogonal function (EOF) analysis applied to 15 ensemble members of the un-initialized historical simulations [from *Giorgetta et al.*, 2013]. The multivariate 3D-EOF methodology is similar to that in *Hawkins and Sutton* [2007]. Potential temperature and salinity October-anomalies used for the EOF analysis are calculated with respect to the period 1958-2005, for which the historical simulations and the ORAS4 ocean reanalysis overlap. Before the EOF analysis, anomalies are weighted by their contribution to density (i.e., thermal and haline expansion coefficients) and the grid-box area.

**2)** Next, we truncate the set of EOF-modes and project ORAS4 potential temperature and salinity anomalies onto the truncated set of EOFs. The truncation threshold is picked at an arbitrary point, at which the reconstruction loses a small amount ($\sim$3 %) of variance explained, retaining half of the EOF modes (360 out of 720). After the truncation, the reconstruction retains 40 % of variance explained as compared to the original ORAS4 reanalysis.

**3)** Assimilation and initialization of hindcasts: To produce coupled initial conditions for decadal hindcasts, the assimilation runs are performed for Octobers over the period 1960-2015. Here, potential temperature and salinity fields are nudged toward the reconstructed ORAS4 anomalies added to the climatology from a historical simulation. The ERA-40/ERA-Interim October states are nudged into the atmospheric component of MPI-ESM using the methodology described for Preop-LR (see Section 2.1). Different to Preop-LR, the FAI assimilation run requires restarts from a historical simulation on

the 30th of September of each year over the whole initialization period. The feasibility of 1-month assimilation as compared to conventional 12-months assimilation has previously been tested in that non-filtered Preop-LR-like experiments (nudging run and ensembles of initialized hindcasts) were carried out with 1-month nudging similar to FAI setup. Ensembles of the FAI hindcasts are started every year over 1960-2015, and are 10 years and 2 months long with 10 ensemble members (Table 1). As with Preop-LR, FAI uses lagged initialization to generate an ensemble of predictions. However, there are some differences in detail: Because the FAI assimilation runs are one-month long, initial conditions are sampled from 9-days long free runs following each assimilation.

#### 2.2.3. Model Initialization by Partially Coupled Spin-up (MODINI)

MODINI is proposed considering the importance of dynamically balanced ocean-atmosphere initial conditions in the equatorial oceans. In contrast to a 3-D initialization of the ocean and atmosphere, this method only uses surface wind-stress anomalies from the reanalysis to drive the ocean and the sea ice. All other feedbacks are maintained as in the fully coupled model. The merits of MODINI for the equatorial Pacific skill have earlier been shown by *Thoma et al.* [2015] and used as a benchmark to understand the initialization shock in the first generation of the MiKlip prediction system by *Pohlmann et al.* [2017]. These studies emphasize the importance of using high quality wind products for initialization of decadal predictions. The MODINI experiments are carried out in three phases:

**1)** The pre-initialization phase before 1958 consists of three historical simulations [*Müller et al.*, 2018].

**2)** Three assimilation runs are initialized from the different historical simulations. During assimilation phase, the ocean and sea ice components of MPI-ESM are forced by the wind stress anomalies from the atmospheric reanalyses. For the period 1958-1989, the wind-stress anomalies are estimated from ERA-40 [*Uppala et al.*, 2005] and for 1990-2016 from ERA-Interim [*Dee et al.*, 2011]. Using the bulk formulae of *Large and Yeager* [2009], 10 m wind velocities from the reanalysis are converted into wind stress. In particular, the wind stress $\boldsymbol{\tau}$ seen by the ocean model during the initialization is

$$\boldsymbol{\tau} = \boldsymbol{\tau}(\mathbf{u}_{rean}) - \boldsymbol{\tau}(\mathbf{u}_{rean})_{clim} + \boldsymbol{\tau}(\mathbf{u}_{model})_{clim}, \tag{1}$$

where $\mathbf{u} = (u, v)$ indicates the horizontal wind velocity from the model, $\mathbf{u}_{model}$, and from the reanalysis, $\mathbf{u}_{rean}$. Index $\boldsymbol{\tau}()_{clim}$ stands for wind stress climatology, which is for the combined reanalyses computed over the period 1958-2016 and for the model over 1958-2005. The model climatology is based on the three historical simulations used to initialize the assimilation runs. All climatologies are seasonally varying monthly means.

**3)** Initialization of hindcasts: 4 daily lagged ensemble members are generated for each of the 3 assimilation runs to construct 12 ensemble members in total. Initialized hindcasts are started yearly from 1960-2015, each hindcast covers 5 years and 2 months (Table 1). The ensemble members r1(i1–i4)p2, r2(i1–i4)p2 and r3(i1–i2)p2 are used for the comparison with Preop-LR and the rest of the test-suite.

### 2.2.4. Ensemble Dispersion Filter (EDF)

The EDF builds on the fact that an ensemble-mean prediction usually has a better skill than individual realizations. Considering that ensemble members contain predictable signals and evolved noise from initial perturbations, the ensemble mean will average out this noise, retaining the predictable component which leads to increased prediction skill

[*Kalnay et al.*, 2006]. Thus, during model integration, the EDF performs periodic ensemble mean resampling in order to control the ensemble spread. The ensemble resampling approach in EDF is close to that of jackknife resampling [*Quenouille*, 1956]. The EDF was shown to lead to more accurate predictions than those from the MiKlip prediction system in terms of global mean and regional temperature, precipitation and winter cyclones [*Kadow et al.*, 2017].

In more detail, the EDF initialized hindcasts are started from the Preop-LR assimilation run and are re-initialized every three months. Therefore, the first three months of Preop-LR and the EDF are identical. Before the fourth month gets started in the EDF, ten new ensemble members are calculated by ensemble-averaging the full-depth ocean temperature and the surface air temperature (SAT) from different combinations of 9 ensemble members chosen from the total set of 10 members. Thereby the temperature spread is largely reduced, while a certain spread is maintained. The procedure is repeated every three months over the necessary prediction range (in this study, 5 years; Table 1). Since the EDF is an add-on procedure, it depends on the underlying initialization strategy, for which it shifts climate predictions towards the in-run ensemble mean. Within this study, the EDF is carried out for the above outlined Preop-LR system (see Section 2.1). The decadal experiments are started on the 1st of November every year over the period 1960-2015 (Table 1).

#### 2.2.5. Bred Vectors (BV)

Several advanced ensemble generation methods based on oceanic singular vectors and anomaly transform have earlier been tested for the MiKlip prediction system in a low resolution configuration by *Marini et al.* [2016] and *Romanova and Hense* [2017], respec-

tively. These studies show that, in the first few lead years, perturbations induced by the atmospheric noise grow faster than oceanic perturbations, eventually dominating the contribution to the total ensemble spread in the later lead years. It is also demonstrated that large-scale ocean perturbations have lower growth rates. Among those, temperature perturbations have a larger effect on the error growth than salinity perturbations. The BV method discussed in this study is a follow-up of the method proposed by *Toth and Kalnay* [1993] and *Keller et al.* [2008]. It is designed to represent the dynamical modes in the variability-active regions and the regions of deep water formation. The advantages of the BV method are (i) reasonable instability patterns on the climatic timescales in perturbation sensitive regions and (ii) energy conservation as no additional energy sources or sinks are introduced to the perturbation fields. Thus, 10 oceanic bred-vector perturbations for each initialization date are calculated in parallel. For each perturbation, the following algorithm is implemented:

**1)** The BV routine starts with a random-noise perturbation applied to the unperturbed ocean state sampled from the Preop-LR nudging run (see Section 2.1) at a particular initialization date. Fastest growing errors resulting from the initial perturbation are bred over 5 iteration steps.

**2)** At each iteration a one year simulation is carried out, at the end of which the metric based on a total energy norm is applied to evaluate perturbation growth rates and to re-scale the perturbation to be used in the following iteration. The total energy norm contains the zonal and meridional contributions of the oceanic flow to the kinetic energy (the sum in the first parentheses) and the available potential energy (the term in the second parentheses):

$$\|x\|_E = (\frac{w_u}{2}\int u'^2\,dV + \frac{w_v}{2}\int v'^2\,dV) + (\frac{w_\rho}{2\rho_0}\int \frac{\rho'^2}{\rho_z}\,dV), \quad (2)$$

where, $u' = u_i - u_n, v' = v_i - v_n$, and $\rho' = \rho_i - \rho_n$ are the velocity and density anomalies, and indices $i, n$ indicate the two different oceanic state vectors which are compared by the norm. The weighting coefficients $w_u$, $w_v$ and $w_\rho$ are calculated such that zonal and meridional kinetic and potential energy components have equal contributions to the total energy [*Keller et al.*, 2008]. The perturbation is rescaled based on the ratio between the total energy growth at the initial and the final state of each iteration (a breeding cycle). The norm constrains initial perturbation to the geographical locations of the total energy growth. By definition, the norm conserves the energy and does not allow sinks or sources on global scale, i.e., the total energy is not changed, when the perturbation is added to the initial state. At the end of the fifth iteration, a further re-scaling coefficient is applied, which keeps perturbation amplitudes in the range of anomalies from the unperturbed state.

**3)** Initialization of hindcasts: Perturbations are added to the ocean potential temperature, salinity, zonal and meridional velocity fields (u&v) from the Preop-LR nudging run, from which the initialized hindcasts are started every year on the 1st of November over 1960-2016 and are 10 years and 2 months long (Table 1).

### 2.3. Verification Metrics and Data

The prediction skill of the test-suite is estimated for surface air temperature (SAT), sea surface temperature (SST) and ocean heat content in the upper 700 meters (HC700). To get some insight into the test-suite performance for the Atlantic Meridional Overturning Circulation (AMOC), we analyze the basin-wide long-term mean, standard deviation

and the time series at 1000 m depth, 26.5°N latitude. As verification data sets, we use HadCRUT4 for SAT [*Morice et al.*, 2012], HadISST1.1 for SST [*Rayner et al.*, 2003], the NOAA/NODC product for HC700 [*Levitus et al.*, 2012] and RAPID for AMOC at 26°N [*Smeed et al.*, 2016].

Within the MiKlip project, a verification tool (www-miklip.dkrz.de/plugins/) was developed for an ad-hoc evaluation of the MiKlip experiments and their comparison to the central MiKlip prediction system [plugins from *Illing et al.*, 2014; *Kadow et al.*, 2015; *Stolzenberger et al.*, 2015, and others]. Prediction skill in terms of correlation and mean squared error and an assessment of significance level follows the verification framework proposed by *Goddard et al.* [2013]. The EnKF is based on full-value initialization and requires a lead-time dependent bias correction. For a consistent analysis, the same bias correction procedure is applied to all test-suite experiments. The metrics are applied to calculate skill at lead years 1 and 2-5, covering the verification period 1962–2016. Thus, initialized hindcasts started in 1961-2015 are used for lead year 1 analysis and hindcasts started in 1960-2011 for lead years 2-5. The first two lead months (November and December) of all the initialized hindcasts are not part of the comparison. The skill assessment is carried out on 10-member ensembles for all the initialization and ensemble generation methods.

Thus, the following prediction skill metrics are analyzed:

**1)** The correlation skill, $r_{HO}$ is used, where index $O$ stands for the observational data, $H$ for the initialized hindcasts at a particular lead time.

**2)** The mean squared error skill score (MSESS) compares the mean squared error (MSE) from the test-suite experiment to either the observed climatology or Preop-LR. MSESS

is derived as $1 - (MSE_H/MSE_R)$, where index $R$ stands for the reference hindcasts. The skill score has a range from -$\infty$ to +1 and indicates improvements for the test-suite hindcasts (or the reference) in case of positive MSESS (or negative MSESS) values.

**3)** Evaluating ensemble prediction systems, it is common to compare the mean squared error of the ensemble mean prediction, $MSE_H$, with the average ensemble variance, $\sigma_H^2$ [*Fortin et al.*, 2014, and references therein]. If the ensemble variance is smaller (larger) than the mean squared error, the ensemble is considered to be underdispersive (overdispersive). As predictions are known to struggle accounting for all sources of uncertainties (due to initial state, model formulations and external forcing), they might show underdispersive ensemble spread. If an ensemble prediction has a perfect spread-to-error ratio (i.e., it equals one), this means that the ensemble spread is as large as the typical error between a single ensemble member and observations ($\sqrt{\sigma_H^2} = \sqrt{MSE_H}$). In this case, the ensemble spread is considered to be representative of uncertainties in the prediction. At the same time, the spread-to-error ratio does not indicate that the prediction error is small (for accurate prediction). For instance, the spread-to-error ratio can also equal one if a prediction system is characterized by large errors and large spread. Because of the latter, additionally to the common spread-to-error ratio, it is suggested to consider the metrics that describe characteristics of the predictions systems based on actual and potential prediction skill. The spread-to-error ratio in some studies is also known as the ensemble spread score [ESS, *Keller et al.*, 2011]. The ESS for the standardized variables can be defined as a function of the correlation coefficient (actual prediction skill) and the common variance of ensemble predictions, $p$ (a measure of sharpness or potential

prediction skill):

$$\mathrm{ESS} = \frac{1-p}{p+1-2\cdot r_{HO}\cdot\sqrt{p}}. \tag{3}$$

Details to the ESS follow the procedure described by *Glowienka-Hense et al.* [2018] and are briefly summarized in the Appendix A Ensemble Spread Score Derivation. An ESS of 1 indicates an optimal ensemble spread (with a flat analysis rank histogram) with the correlation being equal to $\sqrt{p}$. Values less than 1 suggest that the ensemble is too sharp (ensemble members evolve close to each other), which might be due to the ensemble generation procedure or due to model deficiencies. It is expected that a reduction in sharpness must lead to a reduction of correlation. Then reliability of an ensemble spread is inferred from a balance between the two terms. Also for this metric, there can be the case of ESS = 1 when an ensemble shows no potential predictability ($p = 0$).

## 3. Results

### 3.1. Prediction Skill for Surface Air Temperature (SAT) and the upper-ocean heat content (HC700)

The SAT prediction skill from the reference system Preop-LR in terms of correlation coefficients and MSESS with respect to HadCRUT4 climatology for lead years 2-5 is shown in Figs. 1-2. A comparison of correlation and MSE skill from the test-suite versus Preop-LR is shown in terms of correlation skill differences to Preop-LR (Fig. 1) and MSESS (Fig. 2), respectively. Apart from the eastern and central Pacific Ocean and the frontal area of the western-boundary currents, Preop-LR correlates well with SAT from the HadCRUT4 verification data set (Fig. 1). In the Pacific Ocean, there is an area of reduced skill in lead years 2-5 resembling a characteristic oscillation pattern, i.e., the Pacific Decadal Oscillation. Though all the test-suite methods show significant correlation

differences to Preop-LR in this area of the Pacific Ocean as estimated with the bootstrap method (Fig. 1), the correlation skill itself becomes significant only in the EnKF and FAI over the northern North Pacific and the central tropical Pacific and in the EnKF over the North Atlantic (Fig. S1).

A reduction of MSE as compared to Preop-LR accompanies correlation skill improvements in the eastern tropical Pacific (Fig. 2). Here, FAI, the EnKF and MODINI have the highest impact on the MSE as compared to Preop-LR. Further improvements are shown by FAI in the Gulf of Alaska and in the subtropical North Atlantic. Over the continents, the EnKF shows overall better MSE skill than Preop-LR. However, there are also large areas in the extratropics where the EnKF hindcasts are significantly outperformed by Preop-LR, which might be associated with an overestimation of variability strength in the EnKF ensemble as discussed by *Brune et al.* [2018]. In comparison with Preop-LR, MODINI shows large areas with reduced MSE skill, in particular in the tropics and the North Atlantic. The BV and EDF show some modest improvements in the eastern tropical Pacific and the subpolar North Atlantic as well.

Due to slow dynamics of the ocean and its large thermal capacity, the ocean is widely recognized as the memory of the climate system on decadal timescales [*Meehl et al.*, 2009; *Yeager and Robson*, 2017; *Yeager et al.*, 2018]. In the following, we would like to evaluate whether the test-suite shows an impact on the predictability of ocean heat content in the upper 700 meters (HC700; Fig. 3). Overall, in comparison to Preop-LR, the EnKF and FAI reveal better agreement with the NOAA/NODC heat content in the eastern and central Pacific Ocean. However, the EnKF shows negative correlation along the Canary Current and the Indian Ocean. In the Arctic Ocean, Preop-LR shows negative correlation

already in the first lead year (not shown). Also small areas of reduced skill in the first lead year grow in size in the following lead years in the central Pacific, the North Atlantic between 40°N and 50°N latitudes and the tropical Atlantic. The EDF and the BV, which build on the Preop-LR assimilation, inherit low skill areas of Preop-LR.

From the comparison of the prediction skill for SAT and HC700, two regions stand out where several test-suite methods bring the highest improvements: the North Atlantic Ocean and the central Pacific Ocean. In the following sections we take a closer look at these regions.

### 3.2. Skill for the Nino3.4 Region

Even though there are indications of temperature skill improvements in the tropical Pacific at lead years 2-5 (Fig. 1), the skill itself passes the significance test only in few places from the EnKF and FAI (Fig. S1). As the equatorial Pacific Ocean is characterized by strong interannual variability, we consider whether the test-suite also shows improvement for this timescale. The El Niño-Southern Oscillation (ENSO) is the dominant source of predictability at seasonal-to-interannual timescales. Various studies report that surface temperature anomalies in the tropical Pacific associated with ENSO have prediction skill up to one lead year [*Kumar et al.*, 2017, and the references therein].

We show sea surface temperature (SST) time series of the Nino3.4 region for the first lead year in Fig. 4. Though visually Preop-LR seems to follow closely HadISST1.1, the correlation coefficient amounts only to 0.56. Initialized hindcasts seem sometimes to lag the verification data set by one year. MODINI, which specifically targets to improve skill in the equatorial Pacific, shows larger amplitude of the ocean surface temperature response as compared to other procedures. Here and in the next section, we prefer to show the

accuracy skill in terms of the root-mean-square error (RMSE) rather than MSE, as the former has same units as the variable shown in the time series. Also to get an impression about the tolerable range of RMSE, it is compared with the standard deviation in the time series of the verification data set (STDobs). Overall, the range of correlation coefficients varies from 0.42 to 0.65. The correlation coefficients are higher and RMSE are smaller than in Preop-LR for FAI, the EnKF and the EDF, whereas the reverse is true, i.e. the skill is worse than in Preop-LR for BV and MODINI. This result shows a potential for the former methods at improving initialization of seasonal-to-interannual forecasts.

### 3.3. Skill for the North Atlantic Ocean

The North Atlantic subpolar gyre (SPG) is one of the key regions where initialization brings the most of improvements for the prediction skill at decadal timescales [*Hermanson et al.*, 2014; *Kröger et al.*, 2018; *Yeager and Robson*, 2017]. We analyze the time series of SAT and HC700 for the North Atlantic SPG (50°-60°N and 65°W-10°E). Here we use the same region as in *Kröger et al.* [2018] to enable comparison with previous MiKlip systems. At lead year 1, Preop-LR shows rather high correlation for the SPG SAT, with the smallest RMSE that cannot be beaten by the other test-suite experiments (Fig. 5).

For lead years 2-5, correlation coefficients for the SPG SAT from the test-suite experiments range from 0.80 to 0.90 (Fig. 6). Preop-LR shows a correlation value of 0.85. The FAI hindcasts follow closely the trend from historical simulations (not shown), also the multi-decadal variability in FAI is smaller than in the other initialized experiments. The variability from MODINI, evolving on top of the upward trend, provides relatively high correlation skill. The EnKF overestimates the SPG SAT changes in the 1970s and the

2000s. Overall, EnKF, BV and EDF slightly increase correlation for the North Atlantic SPG SAT as compared to Preop-LR.

Previous MiKlip systems showed some disagreements in the prediction skill over the North Atlantic SPG [based on full field versus anomaly initialization methods, *Marotzke et al.*, 2016]. Later *Kröger et al.* [2018] demonstrated that this region is sensitive to the initialization strategy and initial conditions which can cause an initialization shock of different extent. The authors also suggest that, in the case of a severe initialization shock, the problem can be detected in the assimilation run alone by analyzing time series of the regional ocean heat budget. In the first lead year, all the test-suite hindcasts demonstrate a cooling trend before the 1990s and a warming trend thereafter suggesting realistic HC700 changes (not shown). For lead years 2-5 (Fig. 7), all but the EnKF underrepresent the 1970s cooling. The EnKF simulates stronger variability than what is shown in the verification data set. FAI shows variations evolving on top of the upward trend, which is similar to the trend of historical simulations (not shown). MODINI shows large RMSE and no correlation for the SPG HC700, but, with a lag of about 3 years, has an evolution similar to the NOAA/NODC product with cooling in the 1970s and warming after 2000. Preop-LR and the EDF and to a smaller extent BV and FAI show a "spike" around 2005. A similar "spike" is also present in the study by *Marotzke et al.* [2016] for sea surface temperature and by *Kröger et al.* [2018] for HC700 for the previous MiKlip systems. It comes from hindcasts started in 2000 and 2001. For these initialization cases, the BV and FAI hindcasts show cooler SPG HC700 than in the verification data set, and Preop-LR and the EDF show a decrease of the SPG HC700 (Fig. S3). The EnKF hindcasts show an increase in HC700 for these start cases. Since Preop-LR, the EDF,

the BV and FAI are built on the ORAS4 anomaly nudging, in contrast to the EnKF and MODINI, this outlier event might be associated with the ORAS4 ocean initialization.

The AMOC variability is one of the important mechanisms for climate variability over the Atlantic Ocean and Europe [e.g., *Delworth et al.*, 2007; *Griffies and Bryan*, 1997]. *Smith et al.* [2013] show that some initialization approaches used for decadal predictions can distort the AMOC evolution. This could in turn reduce the skill of other climate variables such as sea surface temperature. We analyze the time series of the AMOC at 26.5°N latitude (Fig. 8). For all but the EnKF and MODINI experiments, ORAS4 is the source of the initial states. The ORAS4 AMOC is 3 SV weaker than that of the test-suite experiments and has a very strong downward trend. *Balmaseda et al.* [2013] also suggest that the AMOC in ORAS4 is substantially lower than in the observational data set RAPID. The EnKF shows a somewhat weaker AMOC than the other experiments but comparable magnitude of variability and a slight downward trend. MODINI shows a comparable magnitude of interannual variability and, in contrast to other experiments, a decrease of the AMOC before the 1990s and an increase thereafter. The AMOC from Preop-LR, the BV, and the EDF evolves closely in lead year 1. They adopted the major decadal variability and the downward trend from ORAS4. These are the experiments that share the same assimilation run. FAI and the historical simulations show less temporal variability in the ensemble mean than the other experiments. The historical simulations by design are not constrained with the observational estimates of ocean and atmosphere state parameters. And it seems that FAI in the North Atlantic does not sufficiently constrain the ocean state either. This variety of results shows once again how sensitive hindcasts are in this region to initialization strategies.

With the increased lead time (lead year 5 in lower panel in Fig. 8), all the test-suite experiments show less trend and less variability for the ensemble mean AMOC (also in the basin-wide AMOC variability in Fig. S5). The EDF and Preop-LR show some intensification of variability after 1990. In terms of the basin-wide long-term mean AMOC, Preop-LR, the BV, FAI and MODINI show similar AMOC structure including deep-water and bottom-water cells (Fig. S4). The EnKF and the EDF somewhat deviate from the expected structure: the EnKF shows a North Atlantic overturning cell that extends to the ocean bottom. For EDF, the deep-water cell is fractured into isolated cells, which extend to the bottom (Fig. S4).

We do not estimate any prediction skill for AMOC as observational record appears to be too short/sparse for a robust skill assessment. Also the ORAS4 ocean reanalysis used in this study is not an original source of initial states for all of the test-suite experiments, in order we could use it as the verification data set. Rather, the current AMOC analysis is carried out to make sure that new initialization methods do not disturb the overturning circulation into an unusual state. In fact, it shows that (i) the basin-wide AMOC mean differs in EDF and EnKF from that in other test-suite experiments, and (ii) the test-suite experiments seem to result in a variety of variability patterns with different strength of variance, which at 26.5°N latitude tends to decrease to the level of historical simulation by lead year 5.

### 3.4. Ensemble Spread Performance

The spread-to-error diagnostic, which is commonly used to test ensemble spread in predictions, suggests that Preop-LR and the test-suite largely underestimate the spread of surface temperature at the beginning of the forecast (lead year 1; Fig. S6). The advanced

ensemble generation methods such as BV and EnKF increase the first-year spread along the western-boundary currents and the extratropics. MODINI increases the spread in the North Atlantic as compared to Preop-LR. The EnKF seems to generate too large spread, which remains excessive for the later lead years, especially in the extratropics. FAI does not have impact on the first-year spread but has smaller RMSE than Preop-LR. The EDF by design reduces spread, but the RMSE is close to that of Preop-LR. For later lead times, the spread approaches the level of the errors; this feature is common to all the initialized hindcasts in the test-suite (lead year 5, Fig. S7) except in the EDF. As the spread grows and catches up with the errors, the spread-to-error ratio approaches optimal ($\sim 1$) values. At the same time, we know that the prediction skill diminishes with lead time. In the following, we attempt to interpret spread skill and prediction skill in parallel.

The ensemble spread score (ESS) for SAT for lead years 2-5 is shown as a function of the correlation skill and the ensemble sharpness in Fig. 9. The ESS and its components are calculated for standardized SAT (subtracting the mean and dividing by the standard deviation) as only then the ESS can be decomposed into these two terms. The Preop-LR hindcasts at lead years 2-5 are largely underdispersive (ESS $< 1$) over the ocean and overdispersive over some parts of the continents. Over large areas of the ocean, the system is characterized by relatively high sharpness and high correlation skill. If accurately estimating the first moment rather than the second moment is higher priority, then an underdispersive ensemble might not be a drawback as long as the system is characterized by high prediction skill in the region of interest. In Preop-LR, the regions in the central Pacific and along the western-boundary currents show relatively high sharpness but low prediction skill. This means that the hindcasts have a relatively high confidence,

although they do not predict the events seen in observations. In the northeast of the North Pacific, Preop-LR has rather low sharpness (high ensemble variance and low potential predictability) and low prediction skill.

Comparing the ESS and its components for the test-suite and for Preop-LR shows that (i) except for the North Atlantic, the ESS suggests an underdispersive ensemble over the ocean for all test-suite experiments, (ii) different test-suite experiments modify somewhat sharpness patterns, and (iii) as already described in Section 3.1 significant skill improvements are obtained from several test-suite methods in the North Atlantic and central and eastern Pacific Ocean. In detail, the spread scores of the test-suite experiments are closer to optimal values than for Preop-LR in the North Atlantic. In this region, the correlation skill from the test-suite is slightly improved compared to Preop-LR and the sharpness slightly reduced (ensemble variance increased). The EDF globally and MODINI in the tropics show increased sharpness. The EDF in the current setup reduces ensemble spread by design, which might make the method unsuitable for probabilistic forecasts. However, atmospheric variables such as precipitation and extra-tropical cyclones, which are not directly modified by EDF, show ensemble spread comparable to Preop-LR [*Kadow et al.*, 2017]. It is expected that post-processing or in-run methods using more than one independent bundle of members can improve temperature spread as compared to the current EDF setup.

Overall, with ESS $< 1$ the prediction system has larger potential than actual prediction skill [*Glowienka-Hense et al.*, 2018]. This suggests that the initial ensemble spread from the current ensemble generation methods is too small or that the overconfidence of the prediction system comes from the underrepresentation of some real-world processes. The

result of ESS $> 1$ as over some parts of the continents might suggest noisy system and lack of potential predictability. In addition, as was shown by *Marini et al.* [2016], the spread diagnostic can erroneously indicate overdispersiveness for ensemble predictions when spread skill assessment is carried out with respect to a verification data set with underrepresented variability resulting for instance from smoothed observations.

The reliability analysis is sensitive to the sample size. We boost the ensemble size for the ESS assessment through the "multi-initialization" ensemble of the test-suite accounting for 50 members (Fig. 9). Expectedly, the "multi-initialization" ensemble reduces sharpness and increases the ensemble variance. Interestingly, the ESS still suggests that over large areas of the ocean, the underdispersiveness remains. The exception is the North Atlantic, where the ensemble becomes notably overdispersive. The deterministic skill which is based on the ensemble mean further benefits from the large ensemble: as compared to a 10-member ensemble of Preop-LR, the "multi-initialization" ensemble mean shows globally increased correlation skill with the biggest improvement over the central and eastern Pacific Ocean and reduced MSE with the biggest improvement over the North Atlantic Ocean (Fig. S8).

## 4. Discussion and Summary

In this paper, through an intercomparison of several initialization and ensemble generation methods, we intrinsically address challenges that are fundamental for any decadal prediction system:

**(i.)** Identifying the importance of dynamical consistency for assimilation/ initialization methods.

**(ii.)** Enhancing forecast skill of a predictable component.

**(iii.)** Quantifying the impact of uncertainties in the ocean initial state on the skill of the prediction system.

To this end, five initialization and ensemble generation methods are investigated with respect to their impact on the skill of decadal prediction in the framework of the German decadal prediction system "Mittelfristige Klimaprognose" (MiKlip) in form of its low-resolution configuration (Preop-LR). Three tested methods, the ocean ensemble Kalman filter (EnKF), the filtered anomaly initialization (FAI) and the initialization using a partially coupled spin-up (MODINI), aim to improve prediction skill by a dynamically consistent assimilation/initialization. Whereas the EnKF and FAI address dynamical consistency between the model and the 3D ocean initial state, MODINI addresses the role of dynamically balanced initialization at the air-sea interface. The remaining two methods alter the ensemble generation approach: the ensemble dispersion filter (EDF) shifts the ocean state toward the ensemble mean during model integration. In contrast, the oceanic bred vectors (BV) perturb the climate state using the fastest growing modes.

**(i.)** Results presented above suggest that the EnKF shows the highest correlation skill improvements for surface air temperature (SAT) predictions. The EnKF skill for the ocean heat content in the upper 700 meters (HC700) is ambivalent: high skill in the North Pacific but low skill in the eastern tropical Atlantic and the Indian Ocean. For the North Atlantic subpolar gyre (SPG), where previous MiKlip initialization attempts showed that the prediction system is particularly sensitive to initialization [*Marotzke et al.*, 2016; *Kröger et al.*, 2018], the EnKF hindcasts are able to capture cooling and warming periods better than the other methods. However, the EnKF shows too strong variability and overshoots cooling and warming periods, which results in higher mean squared error

(MSE) than the other test-suite experiments. Thus, further work would be needed on tuning the EnKF variability. Increasing the horizontal and vertical localization in the EnKF assimilation scheme increases spatial distribution of information from ocean observations. In addition, decreasing observational error increases the impact of observations in comparison to the background model state. Both these sensitivities can be used to steer the impact of oceanic observations, at the same time ensuring model stability. However, when increasing localization, the ensemble size has to be increased as well. Another important aspect is treatment of biases (i.e., difference between model and observations). In this study, no bias-correction has been applied prior to the EnKF analysis. It is thus expected that the EnKF method could further benefit from the assimilation of unbiased observations and larger localization.

The FAI method improves SAT and HC700 correlation and reduces SAT MSE in the Pacific basin, but apparently filters too much of the variability in the North Atlantic Ocean. In the current experiment, we consider the explained variance in the filtered initial conditions to be rather low (40%). This may suggest that modes of variability of the reanalysis are not exactly compatible with the modes from the prediction system or that they are not yet sufficiently sampled by the available data used to construct the EOFs. In this respect, FAI might benefit from attempting to capture better the variability modes in the North Atlantic using a larger EOF-basis, using a different weighting method to determine better the structure of the modes in this region, or using regional (per-basin) EOFs, which have been shown to perform significantly better in e.g. reconstructing Atlantic sea level variability [*Meyssignac et al.*, 2012] in comparison to using global EOFs [*Carson et al.*, 2017]. Apart from this, the FAI-hindcasts show the highest skill for the ENSO region at

interannual timescales which is associated with the improved zonal momentum balance in the equatorial Pacific as compared to Preop-LR (not shown) and appear to have a substantial potential for further improvements in the future.

Given that MODINI only uses the wind stress data for initialization, it compares reasonably well with the other initialization methods that are tested here. This holds especially near the equator where the Coriolis force goes to zero and the balance between the pressure gradient and the wind stress is dominant. In terms of SAT, there is some hint of an improvement in the eastern Pacific. However, MODINI is outperformed in terms of the MSE skill, in particular over the tropics and the North Atlantic. This contrasts with *Thoma et al.* [2015], who show that MODINI has considerable potential as an initialization scheme, especially when hindcasting the eastern Pacific, the Pacific Decadal Oscillation (PDO) and global SAT. The relatively poor performance here is found even when considering the same hindcast period analyzed by *Thoma et al.* [2015], that is 1990-2006. The only differences to the previous study are (i) the wind stress product used for the initialization [they used the NCEP product from *Saha et al.*, 2010] and (ii) the time period over which MODINI initialization is carried out [here 1961-2011 and in *Thoma et al.*, 2015, 1980-2006]. The former points to the sensitivity of decadal hindcasts to the wind stress product used for initialization [see also *Pohlmann et al.*, 2017]. Although beyond the scope of the present study, it would be interesting to test the improved forcing data set based on ERA-40 described in *Brodeau et al.* [2010] for the hindcasts started between 1961 and 2000, where the wind speed is rescaled, especially in the tropics, using satellite measurements.

**(ii.)** Enhancing the forecast skill of a predictable component is generally a goal of all initialization methods. The ensemble dispersion filter (EDF) tested here specifically attempts to improve an ensemble mean prediction by reducing ensemble spread during integration. This method is an add-on procedure and thus its performance depends on the underlying initialization strategy (which is, in this study, same as in Preop-LR). For SAT and HC700, the EDF is largely mimicking Preop-LR, with some improvements in terms of correlation and MSE skill in the North Atlantic and the central Pacific Oceans. To a certain degree (in particular, in the North Atlantic), the EDF experiment is able to reproduce the results of the predecessor study by *Kadow et al.* [2017], who implemented the EDF with 5 ensemble members and 39 start dates. The EDF uses an assumption that an ensemble-mean prediction usually has a better skill than individual realizations, thus by averaging out the noise from the ensemble, the predictable component becomes more visible. However, the ensemble mean might not be a valid prediction in the sense that it cannot be compared to the single realization that is represented by observations. In addition, the applicability of the method for predicting extreme events is to be demonstrated, as rare extreme events could be smoothed out by the ensemble averaging. On the other hand, spread adjustment implemented during integration could be useful for a prediction system that suffers from an overdispersive spread. The EDF spread could be rescaled in the post-processing by the calibration techniques. Also using more than one independent bundle of members could further be tested to improve the current EDF spread.

**(iii.)** Two procedures tested here use advanced ocean perturbation methods to account for ocean initial state uncertainty. The ensemble generation method based on bred vectors (BV) perturbs the ocean state with the fastest growing modes. The EnKF uses an

assimilation ensemble, from which an ensemble of initialized predictions is started. In the MiKlip system, throughout all the development stages, the ensembles of decadal predictions were generated by perturbing the initial state using lagged initialization. Though, several advanced ensemble generation methods based on oceanic singular vectors and anomaly transform have earlier been tested for the MiKlip prediction system by *Marini et al.* [2016] and *Romanova et al.* [2017], respectively. A recent study by *Germe et al.* [2017] reports that lagging the ocean state by a few days might not be sufficient to properly represent the ocean initial state uncertainties, especially in the deep ocean. There has been a discussion that underdispersive ensembles (with an underestimated spread) lead to overconfident climate predictions indicating a necessity to increase ensemble spread [*Palmer*, 2000]. Other studies, on the other hand, demonstrate that decadal climate predictions might only be underdispersive in the first lead years, while thereafter ensembles actually become overdispersive [*Ho et al.*, 2013; *Marini et al.*, 2016]. In this respect, the MiKlip prediction system Preop-LR shows narrow spread at the beginning of surface temperature forecasts, but the ensemble spread grows relatively fast as compared to the root-mean-square error. Both ensemble generation methods, the EnKF and the BV, increase the spread. However, the EnKF tends to generate too large spread for surface temperature, especially in the extratropics. The BV in addition shows improvements over Preop-LR regarding the correlation skill and MSESS for lead years 2-5. In terms of the computational costs, the BV method requires additional resources for the iterative block, which depends on the number of iterations. A large number of iterations is required to obtain a good agreement between the bred vector growth rate and the forecast error, as shown in the statistical analysis of *Yang et al.* [2008] for the 1997-98 El-Nino event fore-

casts. By contrast, we aim at reduction of the computational costs. One way to reduce the costs is to use orthogonal BV rotation [*Wei et al.*, 2008]. In this case, the iterative block could be performed only once for one bred vector and the other perturbation patterns could follow the orthogonalization procedure. The excessive spread in the EnKF can possibly be adjusted through tuning the spread inflation coefficient.

As large portions of the world ocean including the abyss and ice covered regions remain coarsely sampled by observations and sampling prior to the early 2000s was coarser globally, ensemble generation for decadal predictions are expected to reflect this uncertainty in the ocean initial state. In this respect, the test-suite experiments at the beginning of surface temperature forecasts are largely underdispersive. While the ensemble-spread growth in decadal predictions might not be as serious an issue as in numerical weather predictions and seasonal predictions. For instance, the ensemble spread in the test-suite temperature predictions for the North Atlantic grows relatively fast. Another point is that decadal prediction studies, attempting to show usefulness to societal needs, often analyze "user-relevant" variables like surface temperature and precipitation. For these variables, the spread is largely dominated by the atmospheric noise [*Kleeman et al.*, 2003; *Marini et al.*, 2016]. The deep-ocean perturbations at the timescales that are relevant for decadal predictions are likely to have more effect on the ocean subsurface and the AMOC [*Zanna et al.*, 2011, 2012]. But these are the variables for which the decadal prediction skill is still treated with a special caution as the observational records for them are too sparse or too short to be used for verification [*Karspeck et al.*, 2015].

Also for ensemble reliability diagnostics, the importance of ensemble size should not be forgotten. In numerical weather predictions, the ensemble consists of about fifty realiza-

tions [*Bauer et al.*, 2015], while decadal predictions are often carried out with ten or even fewer ensemble members [*Taylor et al.*, 2012]. For decadal predictions, the experimental details are often a compromise between the ensemble size, the number of initialization dates, the length of the prediction and recently model resolution. *Ferro et al.* [2012]; *Sienz et al.* [2016]; *Benestad et al.* [2017]; *Yeager et al.* [2018] address this trade-off issue to find solutions for a more effective use of climate information given limited resources or/and allocating future resources. These studies suggest that with respect to each of the parameters a saturation level for prediction skill exists beyond which we do not necessarily gain new information from, for instance, increasing ensemble size or model resolution. A recent study by *Yeager et al.* [2018] suggests that about 20 and more members might be needed to detect robust differences between initialized and un-initialized hindcasts. *Sienz et al.* [2016] and *Yeager et al.* [2018] point out on the relationship between the optimal ensemble size and signal-to-noise ratio, meaning that some regions and variables might require larger ensembles. For the EnKF system, a thorough analysis of the unstable-neutral subspace, e.g. following *Carrassi et al.* [2018], would help to determine the minimum number of ensemble members which has to be used for decadal predictions. Alternatively to a large ensemble from a single model, *DelSole et al.* [2014] explain why multi-model ensembles might provide a complementary skill to that obtained from a single model. In this study we follow recommendations of *Sienz et al.* [2016] who advises to use at least 10 ensemble members and a verification period long enough to capture complete cycle of the decadal variation for decadal prediction experiments.

Overall, the comparison of the test-suite experiments leads to the following conclusions and future work:

• With respect to goals, all test-suite methods address different issues of the prediction system; with respect to prediction skill, several test-suite methods show added value for skill as compared to the reference method. As none of the tested methods is superior to others with respect to all points, it appears natural therefore to attempt to combine several methods in a future attempt to further improve the skill of decadal predictions.

• Considered advancements in initialization and ensemble generation show potential to further improve the prediction skill of the reference Preop-LR system for SAT and HC700 at lead years 1 and 2-5 in the central and eastern Pacific Ocean and the North Atlantic Ocean.

• It is expected that coupled data assimilation is a potential solution that combines improving dynamical consistency between model and initial state as well as balanced initialization between model components [*Penny et al.*, 2017]. Although assimilation in Preop-LR and the test-suite is carried out within the coupled framework, dynamical consistency in the initialized experiments is still not guaranteed. The EnKF is the first attempt to initialize the predictions within the MiKlip system with the native data assimilation system. We recall that the EnKF is the only method here that assimilates ocean temperature and salinity profiles directly into the prediction system, while still nudging the atmospheric reanalysis. While the EnKF allows for the transfer of atmospheric information into the ocean, it also may introduce biases from the atmospheric reanalysis. All the other initialization methods utilize ocean and atmosphere reanalyses obtained with a different model than that used for predictions. MODINI assimilates atmospheric variability from the reanalysis into the ocean model by partial coupling. The EnKF and FAI show the largest differences in the performance to the rest of the methods, both show benefits

and pitfalls, and are attractive with respect to tackling model consistency. Linking statistical improvements from these methods to concrete physical improvements is still needed to better understand their performance. Currently, the EnKF is being further tested for a potential implementation in the operational MiKlip system.

## Appendix A: Ensemble Spread Score Derivation

Ensemble Spread Score (ESS) dependence on Pearson's anomaly correlation and sharpness in terms of analysis of variance (ANOVA) is described in detail by *Glowienka-Hense et al.* [2018]. Thus, the ESS can be rationalized as an analytic function of the correlation coefficient and the common variance of the ensemble members. The total variance in the ensemble predictions over the whole verification period for a particular lead time can be expressed as:

$$\sigma_t^2 = \frac{1}{M \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{M} (H_{it} - H_{00})^2 = \sigma_e^2 + \sigma_a^2, \tag{A1}$$

where index $t = 1...T$ stands for the time-point in the verification interval, and $i = 1...M$ represents the ensemble members of the test-suite experiment. Index 0 indicates that the variable represents the average value, either over ensemble members or over the verification interval, or over both (index 00).

The mean of the ensemble variance over the verification interval is calculated as:

$$\sigma_e^2 = \frac{1}{M \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{M} (H_{it} - H_{0t})^2, \tag{A2}$$

and the variance of the ensemble mean predictions is:

$$\sigma_a^2 = \frac{1}{T} \sum_{t=1}^{T} (H_{0t} - H_{00})^2. \tag{A3}$$

As model sharpness is associated with potential predictability, we use notation $p$ for the former term. *Glowienka-Hense et al.* [2018] use the notation $ANOVA$ for $p$. Further derivations based on the terms presented above allow finding an expression for the common variance of the ensemble members, $p$, and mean squared error, $MSE$:

$$\begin{gathered} MSE = \frac{1}{T}\sum_{t=1}^{T}(H_{0t} - O_t)^2 = \sigma_a^2 + \sigma_o^2 - 2 \cdot r_{HO} \cdot \sigma_o \cdot \sigma_a + (H_{00} + O_0)^2, \\ p = \frac{\sigma_a^2}{\sigma_t^2} = \frac{\sigma_t^2 - \sigma_e^2}{\sigma_t^2} = 1 - \frac{\sigma_e^2}{\sigma_t^2}, \end{gathered} \tag{A4}$$

where $\sigma_o^2$ represents variability in the time-series of the verification (observational) data set and, respectively, $O_0$ represents the long-term mean value.

Standardizing variables with the total long-term mean $H_{00}$ and $O_0$ and variances $\sigma_t^2$ and $\sigma_o^2$ results in $\hat{H}$ and $\hat{O}$ with:

$$\begin{gathered} \hat{H}_{00} = \hat{O}_0 = 0, \\ \sigma_t^2(\hat{H}) = \sigma_O^2(\hat{O}) = 1, \\ \sigma_e^2(\hat{H}) = 1 - p, \\ MSE(\hat{H}, \hat{O}) = p + 1 - 2 \cdot r_{HO} \cdot \sqrt{p}. \end{gathered} \tag{A5}$$

Finally, ESS for the normalized variable can be found as:

$$\mathrm{ESS} = \frac{\sigma_e^2}{MSE} = \frac{1 - p}{p + 1 - 2 \cdot r_{HO} \cdot \sqrt{p}}. \tag{A6}$$

In this relationship, the lower values of sharpness, $p$, correspond to a broader band of acceptable balance between model and observations. Whereas, to achieve same statistical balance with higher values of sharpness, closer correspondence between $r_{HO}$ and $p$ is required. If ESS $< 1$, the prediction system has higher potential than actual prediction skill. If ESS $> 1$, the prediction system is overdispersive. The approximation based on

second moment is sensitive to sample size and is valid for large values of $T$ and $M$. Otherwise, correction factors for sampling should be used as e.g. in studies by *Ho et al.* [2013]; *Fortin et al.* [2014]; *Sospedra-Alfonso et al.* [2016]. More detail on the ESS dependence on correlation and sharpness are given by *Glowienka-Hense et al.* [2018].

**Acknowledgments.** The initialization and ensemble generation efforts presented here are carried out within the MiKlip project, which is funded by the German Ministry for Education and Research (BMBF): grant numbers 01LP1516A, 01LP1516B, 01LP1517D, 01LP1520G and 01LP1519B. We thank Thomas Spangehl, Tobias Stacke, Sebastian Hettrich and Kameswarrao Modali for discussions on the results of the test-suite experiments and Tina Dippe for her help with the MODINI experiments. The MPI-ESM is made available to the scientific community under a version of the MPI-M Software License Agreement (http://www.mpimet.mpg.de/en/science/models/license/). Model simulations were carried out at the German Climate Computing Centre (DKRZ) and are available from World Data Center for Climate (WDCC) via http://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=DKRZ_LTA_122_ds00001. Analysis for Figs. 1-3 is produced with the MURCSS tool [*Illing et al.*, 2014]. The configuration, restart files and model outputs of the Preop-LR system were made available by the German Meteorological Office (Dr. Klaus Pankatz) upon request. The uninitialized historical simulations were provided by the Max Planck Institute for Meteorology (Dr. Holger Pohlmann) upon request. The HadCRUT4 [*Morice et al.*, 2012] and HadISST1.1 [*Rayner et al.*, 2003] data sets for verification are available through the Integrated Climate Data Center (ICDC, http://icdc.cen.uni-hamburg.de). The NOAA/NODC ocean heat content

is provided via https://www.nodc.noaa.gov/OC5. Data from the RAPID-WATCH MOC monitoring project are funded by the Natural Environment Research Council and are freely available from www.rapid.ac.uk/rapidmoc.



## References

Ballish, B. A. (1981), A simple test of the initialization of gravity modes, *Monthly Weather Review*, *109*(6), 1318–1321.

Balmaseda, M. A., K. Mogensen, and A. T. Weaver (2013), Evaluation of the ECMWF ocean reanalysis system ORAS4, *Quarterly Journal of the Royal Meteorological Society*, *139*(674), 1132–1161.

Bauer, P., A. Thorpe, and G. Brunet (2015), The quiet revolution of numerical weather prediction, *Nature*, *525*(7567), 47.

Benestad, R., Sillmann, J., Thorarinsdottir, T. L., Guttorp, P., Mesquita, M. D. S., Tye, M. R., and others (2017), New vigour involving statisticians to overcome ensemble fatigue, *Nature Climate Change*, *7*(10), 697–703.

Boer, G., V. Kharin, and W. Merryfield (2013), Decadal predictability and forecast skill, *Climate Dynamics*, *41*, 1817–1833, doi:10.1007/s00382-013-1705-0

Boer, G. J., D. M. Smith, C. Cassou, F. Doblas-Reyes, G. Danabasoglu, B. Kirtman, Y. Kushnir, M. Kimoto, G. A. Meehl, R. Msadek, et al. (2016), The decadal climate prediction project (DCPP) contribution to CMIP6, *Geoscientific Model Development*, *9*(10), 3751.

Branstator, G., and H. Teng (2010), Two limits of initial-value decadal predictability in a CGCM, *Journal of Climate*, *23*(23), 6292–6311.

Branstator, G., and H. Teng (2012), Potential impact of initialization on decadal predictions as assessed for CMIP5 models, *Geophysical Research Letters*, *39*(12).

Branstator, G., A. Mai, and D. Baumhefner (1993), Identification of highly predictable flow elements for spatial filtering of medium-and extended-range numerical forecasts, *Monthly Weather Review*, *121*(6), 1786–1802.

Brodeau, L., B. Barnier, A.-M. Treguier, T. Penduff, S. Gulev (2010), An ERA40-based atmospheric forcing for global ocean circulation models, *Ocean Modelling*, *31*(3-4), 88–104.

Brune, S., L. Nerger, and J. Baehr (2015), Assimilation of oceanic observations in a global coupled Earth System Model with the SEIK filter, *Ocean Modelling*, *96*, 254–264.

Brune, S., A. Düsterhus, H. Pohlmann, W. Müller, and J. Baehr (2018), Time dependency of the prediction skill for the North Atlantic subpolar gyre in initialized decadal hindcasts, *Climate Dynamics*, 51 *(*5-6), 1947–1970.

Carrassi, A., Bocquet, M., Bertino, L. and Evensen, G. (2018), Data assimilation in the geosciences: An overview of methods, issues, and perspectives, *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535.

Carson, M., Köhl, A., Stammer, D., Meyssignac, B., Church, J., Schröder, J., Wenzel, M. and Hamlington, B. (2017), Regional sea level variability and trends, 19602007: A comparison of sea level reconstructions and ocean syntheses, *Journal of Geophysical Research: Oceans*, *122*(11), 9068–9091.

Counillon, F., Keenlyside, N., Bethke, I., Wang, Y., Billeau, S., Shen, M. and Bentsen, M. (2016), Flow-dependent assimilation of sea surface temperature in isopycnal coordinates with the Norwegian Climate Prediction Model, *Tellus A*, *68*, 32437.

Dee, D. P., S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597.

DelSole, T., Nattala, J., and Tippett, M. K. (2014), Skill improvement from increased ensemble size and model diversity, *Geophysical Research Letters*, *41*(20), 7331-7342.

Delworth, T. L., R. Zhang, and M. E. Mann (2007), Decadal to centennial variability of the Atlantic from observations and models, *GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION*, *173*, 131–148.

Doblas-Reyes, F., M. Balmaseda, A. Weisheimer, and T. Palmer (2011), Decadal climate prediction with the European Centre for Medium-Range Weather Forecasts coupled forecast system: Impact of ocean observations, *Journal of Geophysical Research: Atmospheres*, *116*(D19).

Fetterer, F., K. Knowles, W. Meier, and M. Savoie (2016), Sea ice index, version 2, *Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center*, doi: http://doi.org/10.7265/N5736NV7.

Ferro, C. A., Jupp, T. E., Lambert, F. H., Huntingford, C., and Cox, P. M (2012), Model complexity versus ensemble size: allocating resources for climate prediction, *Phil. Trans. R. Soc. A.*, *370*(1962), 1087–1099.

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte (2014), Why should ensemble spread match the RMSE of the ensemble mean?, *Journal of Hydrometeorology*, *15*(4), 1708–1713.

Germe, S. F., J. Mignot, A. Fedorov, S. Nguyen, and D. Swingedouw (2017), The impacts of oceanic deep temperature perturbations in the North Atlantic on decadal climate variability and predictability, *Climate Dynamics*, doi:10.1007/s00382-017-4016-z

Giorgetta, M. A., J. Jungclaus, C. H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, T. Crueger, M. Esch, K. Fieg, et al. (2013), Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project Phase 5, *Journal of Advances in Modeling Earth Systems*, *5*(3), 572–597.

Glowienka-Hense, R., A. Hense, T. Spangehl, and M. Schröder (2018), Common metrics of calibration for continuous gaussian and exceedance data, *Geophysical Model Development*, doi:https://doi.org/10.5194/gmd-2018-141.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. J. Mason, et al. (2013), A verification framework for interannual-to-decadal predictions experiments, *Climate Dynamics*, *40*(1-2), 245–272.

Good, S. A., M. J. Martin, and N. A. Rayner (2013), EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, *Journal of Geophysical Research: Oceans*, *118*(12), 6704–6716.

Griffies, S. M., and K. Bryan (1997), Predictability of North Atlantic multidecadal climate variability, *Science*, *275*(5297), 181–184.

Hawkins, E., and R. Sutton (2007), Variability of the Atlantic thermohaline circulation described by three-dimensional empirical orthogonal functions, *Climate Dynamics*, *29*(7-8), 745–762.

Hermanson, L., R. Eade, N. H. Robinson, N. J. Dunstone, M. B. Andrews, J. R. Knight, A. A. Scaife, and D. M. Smith (2014), Forecast cooling of the Atlantic subpolar gyre

and associated impacts, *Geophysical research letters*, *41*(14), 5167–5174.

Ho, C. K., E. Hawkins, L. Shaffrey, J. Broecker, L. Hermanson, J. M. Murphy, D. M. Smith, and R. Eade (2013), Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion, *Geophysical Research Letters*, *40*(21), 5770–5775.

Illing, S., C. Kadow, K. Oliver, and U. Cubasch (2014), Murcss: A tool for standardized evaluation of decadal hindcast systems, *Journal of Open Research Software*, *2*(1), doi: http://doi.org/10.5334/jors.bf

Jungclaus, J., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, and J. Storch (2013), Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth System Model, *Journal of Advances in Modeling Earth Systems*, *5*(2), 422–446.

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, *Meteorologische Zeitschrift*, pp. 1–13.

Kadow, C., S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch (2017), Decadal climate predictions improved by ocean ensemble dispersion filtering, *Journal of Advances in Modeling Earth Systems*, *9*(2), 1138–1149.

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, Cambridge Univ. Press, Cambridge, U.K.*, pp. 157–180.

Karspeck, A., D. Stammer, A. Köhl, G. Danabasoglu, M. Balmaseda, D. Smith, Y. Fujii, S. Zhang, B. Giese, H. Tsujino, et al. (2015), Comparison of the Atlantic meridional

overturning circulation between 1960 and 2007 in six ocean reanalysis products, *Climate Dynamics*, *49*(3), 957–982.

Keenlyside, N., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner (2008), Advancing decadal-scale climate prediction in the North Atlantic sector, *Nature*, *453*(7191), 84.

Keller, J. D., L. Kornblueh, A. Hense, and A. Rhodin (2008), Towards a GME ensemble forecasting system: Ensemble initialization using the breeding technique, *Meteorologische Zeitschrift*, *17*(6), 707–718.

Keller, J. D., and A. Hense (2011), A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms, *Meteorologische Zeitschrift*, *20*(2), 107–117.

Kirtman, B., S. Power, A. Adedoyin, G. Boer, R. Bojariu, I. Camilloni, F. Doblas-Reyes, A. Fiore, M. Kimoto, G. Meehl, et al. (2013), Near-term climate change: projections and predictability. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Kleeman, R., Y. Tang, and A. M. Moore (2003), The calculation of climatically relevant singular vectors in the presence of weather noise as applied to the ENSO problem, *Journal of the Atmospheric Sciences*, *60*(23), 2856–2868.

Kröger, J., H. Pohlmann, F. Sienz, J. Marotzke, J. Baehr, A. Köhl, K. Modali, I. Polkova, D. Stammer, F. Vamborg, and W.A. Müller (2018), Full-field initialized decadal predictions with the MPI Earth System Model: An initial shock in the North Atlantic, *Climate Dynamics*, *51*, 2593–2608, doi:10.1007/s00382-017-4030-1.

Kumar, A., Z.-Z. Hu, B. Jha, and P. Peng (2017), Estimating ENSO predictability based on multi-model hindcasts, *Climate Dynamics*, *48*(1-2), 39–51.

Large, W. G., and S. Yeager (2009), The global climatology of an interannually varying air–sea flux data set, *Climate Dynamics*, *33*(2-3), 341–364.

Levitus, S., J. I. Antonov, T. P. Boyer, O. K. Baranova, H. E. Garcia, R. A. Locarnini, A. V. Mishonov, J. Reagan, D. Seidov, E. S. Yarosh, et al. (2012), World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010, *Geophysical Research Letters*, *39*(10).

Marini, C., I. Polkova, A. Köhl, and D. Stammer (2016), A comparison of two ensemble generation methods using oceanic singular vectors and atmospheric lagged initialization for decadal climate prediction, *Monthly Weather Review*, *144*(7), 2719–2738.

Marotzke, J., W. A. Müller, F. S. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, et al. (2016), MiKlip-a national research project on decadal climate prediction, *Bulletin of the American Meteorological Society*, *97*(12), 2379–2394.

Meyssignac, B., M. Becker, W. Llovel, and A. Cazenave (2012), An assessment of two-dimensional past sea level reconstructions over 1950–2009 based on tide-gauge data and different input sea level grids, *Surv. Geophys.*, *33*, 945–972, doi:10.1007/ s10712-011-9171-x

Meehl, G., L. Goddard, J. Murphy, R. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. Giorgetta, A. Greene, E. Hawkins, et al. (2009), Decadal prediction: Can it be skillful?, *Bulletin of the American Meteorological Society*, *90*(10), 1467–1485, doi: 10.1175/2009BAMS2778.1

Meehl, G. A., L. Goddard, G. Boer, R. Burgman, G. Branstator, C. Cassou, S. Corti, G. Danabasoglu, F. Doblas-Reyes, E. Hawkins, et al. (2014), Decadal climate prediction: an update from the trenches, *Bulletin of the American Meteorological Society*, *95*(2), 243–267.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *Journal of Geophysical Research: Atmospheres*, *117*(D8).

Müller, W.A., J. Baehr, H. Haak, J. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J. Storch, and J. Marotzke (2012), Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology, *Geophysical Research Letters*, *39*(22).

Müller, W.A., J. H. Jungclaus, T. Mauritzen, J. Baehr, M. Bittner, R. Budich, F. Bunzel, M. Esch, R. Ghosh, H. Haak, T. Ilyina, T. Kleine, L. Kornblueh, H. Li, K. Modali, H., D. Notz, H. Pohlmann, E. Roeckner, I. Stemmler, F. Tian, J. Marotzke (2018), A higher-resolved version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), *Journal of Advances in Modeling Earth Systems*, *10*(7), 1383–1413.

Mulholland, D. P., Laloyaux, P., Haines, K., and Balmaseda, M. A.(2016), Origin and impact of initialization shocks in coupled atmosphere-ocean forecasts, *Monthly Weather Review*, *143*(11), 4631–4644.

Nerger, L., and W. Hiller (2012), Software for ensemble-based data assimilation systems, *Computers and Geosciences*, *55*, 110–118.

Palmer, T. N. (2000), Predicting uncertainty in forecasts of weather and climate, *Reports on Progress in Physics*, *63*(2), 71.

Penny, S. G., and Hamill, T. M. (2017), Coupled data assimilation for integrated earth system analysis and prediction, *Bulletin of the American Meteorological Society*, *97*(7), ES169–ES172.

Pham, D. T. (2001), Stochastic methods for sequential data assimilation in strongly nonlinear systems, *Monthly Weather Review*, *129*(5), 1194–1207.

Pohlmann, H., J. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke (2009), Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic, *Journal of Climate*, *22*(14), 3926–3938.

Pohlmann, H., J. Kröger, R. J. Greatbatch, and W. A. Müller (2017), Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific, *Climate Dynamics*, *49*(7-8), 2685–2693.

Quenouille, M. (1956), Notes on bias in estimation, *Biometrika*, *43*(3-4), 353–360.

Rayner, N., D. E. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *Journal of Geophysical Research: Atmospheres*, *108*(D14).

Romanova, V., and A. Hense (2017), Anomaly transform methods based on total energy and ocean heat content norms for generating ocean dynamic disturbances for ensemble climate forecasts, *Climate Dynamics*, *49*, 731–751, doi:10.1007/s00382-015-2567-4

Romanova, V., A. Hense, S. Wahl, S. Brune, and J. Baehr (2017), Skill assessment of different ensemble generation schemes for retrospective predictions of surface fresh-

water fluxes on inter and multi-annual timescales, *Meteorologische Zeitschrift*, doi: 10.1127/metz/2017/0790

Saha, S., S. Moorthi, H.-L. Pan, X. Wu, J. Wang, S. Nadiga, P. Tripp, R. Kistler, J. Woollen, D. Behringer, et al. (2010), The NCEP climate forecast system reanalysis, *Bulletin of the American Meteorological Society*, *91*(8), 1015–1058.

Sienz, F., W. A. Müller, and H. Pohlmann (2016), Ensemble size impact on the decadal predictive skill assessment, *Meteorologische Zeitschrift*, *25*, 645–655.

Smeed, D., G. McCarthy, D. Rayner, B. Moat, W. Johns, M. Baringer, and C. Meinen (2016), Atlantic meridional overturning circulation observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturning Circulation and Heatflux Array-Western Boundary Time Series) array at 26N from 2004 to 2015, *British Oceanographic Data Centre/Natural Environment Research Council*, doi:10.5285/1a774e53-7383-2e9a-e053-6c86abc0d8c7

Smith, D., S. Cusack, A. Colman, C. Folland, G. Harris, and J. Murphy (2007), Improved surface temperature prediction for the coming decade from a global climate model, *Science*, *317*(5839), 796–799.

Smith, D. M., R. Eade, and H. Pohlmann (2013), A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction, *Climate Dynamics*, *41*(11-12), 3325–3338.

Sospedra-Alfonso, R., Merryfield, W. J., and Kharin, V. V. (2016), Representation of snow in the Canadian seasonal to interannual prediction system. Part II: Potential predictability and hindcast skill, *Journal of Hydrometeorology*, *17*(9), 2511–2535.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, et al. (2013), Atmospheric component of the MPI-M Earth System Model: ECHAM6, *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172.

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, M. Schröder, A. Mazurkiewicz, and A. Hense (2015), Revealing skill of the MiKlip decadal prediction system by three-dimensional probabilistic evaluation, *Meteorologische Zeitschrift*, doi: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, *Bulletin of the American Meteorological Society*, *93*(4), 485–498.

Thoma, M., R. J. Greatbatch, C. Kadow, and R. Gerdes (2015), Decadal hindcasts initialized using observed surface wind stress: Evaluation and prediction out to 2024, *Geophysical Research Letters*, *42*(15), 6454–6461.

Toth, Z., and E. Kalnay (1993), Ensemble forecasting at NMC: The generation of perturbations, *Bulletin of the American Meteorological Society*, *74*(12), 2317–2330.

Uppala, S. M., P. Kållberg, A. Simmons, U. Andrae, V. d. Bechtold, M. Fiorino, J. Gibson, J. Haseler, A. Hernandez, G. Kelly, et al. (2005), The ERA-40 re-analysis, *Quarterly Journal of the Royal Meteorological Society*, *131*(612), 2961–3012.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu (2008), Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system, *Tellus A*, *60*(1), 62–79.

Williamson, D. L. (1976), Normal mode initialization procedure applied to forecasts with the global shallow water equations, *Monthly Weather Review*, *104*(2), 195–206.

Yang, S.-C., Kalnay, E., Cai, M., and Rienecker, M. (2008), Bred Vectors and Tropical Pacific Forecast Errors in the NASA Coupled General Circulation Model, *Monthly Weather Review*, *136*(4), 1305–1326.

Yeager, S., and J. Robson (2017), Recent progress in understanding and predicting Atlantic decadal climate variability, *Current Climate Change Reports*, *3*(2), 112–127.

Yeager, S. G., Danabasoglu, G., Rosenbloom, N., Strand, W., Bates, S., Meehl, G., and others (2018), Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model, *Bulletin of the American Meteorological Society*, 1876–1886, doi: https://doi.org/10.1175/BAMS-D-17-0098.1.

Zanna, L., P. Heimbach, A. M. Moore, and E. Tziperman (2011), Optimal excitation of interannual Atlantic Meridional Overturning Circulation variability, *Journal of Climate*, *24*(2), 413–427.

Zanna, L., P. Heimbach, A. Moore, and E. Tziperman (2012), Upper-ocean singular vectors of the North Atlantic climate with implications for linear predictability and variability, *Quarterly Journal of the Royal Meteorological Society*, *138*(663), 500–513.

**Table 1.** Summary of experiments.

| Experiment | Assimilation | | | Initialized hindcasts* | Ensemble generation* |
|---|---|---|---|---|---|
| | Ocean | Atmosphere | Sea ice | | |
| Preop-LR | Nudging to ORAS4 T&S anomalies | Nudging to ERA40/ ERA-Interim full field | NSIDC anomalies | yearly started over 1960-2016, 10-years long | Lagging by 1-9 days after the start date, 10 members |
| Bred Vectors (BV) | | | | yearly started from the Preop-LR assimilation over 1960-2016, 10-years long | BV-based perturbations for T&S, and u&v, 10 members |
| Ensemble Dispersion Filter (EDF) | | | | yearly started from the Preop-LR assimilation over 1960-2015, 5-years long. Re-initialization every 3 months from the ensemble mean for T and SAT | EDF, 10 members |
| Ensemble Kalman Filter (EnKF) | EN4 T&S full value | | no | yearly started over 1960-2016, 10-years long | EnKF, 16 members |
| Filtered anomaly initialization (FAI) | 1-month nudging to filtered ORAS4 T&S anomalies | 1-month nudging to ERA40/ ERA-Interim full field | 1-month nudging to NSIDC anomalies | yearly started over 1960-2015, 10-years long | Lagging by 1-9 days after the start date 10 members |
| Model initialization by partially coupled spin-up (MODINI) | Reanalysis wind-stress anomalies seen by MPIOM | ECHAM6 response to MPIOM using the coupled model dynamics | as for the ocean | yearly started over 1960-2015, 5-years long | Lagging by 1-4 days from 3 assimilation runs 12 members |

*Across the methods, the skill assessment is carried out on 10-member ensembles and the verification period 1962-2016.

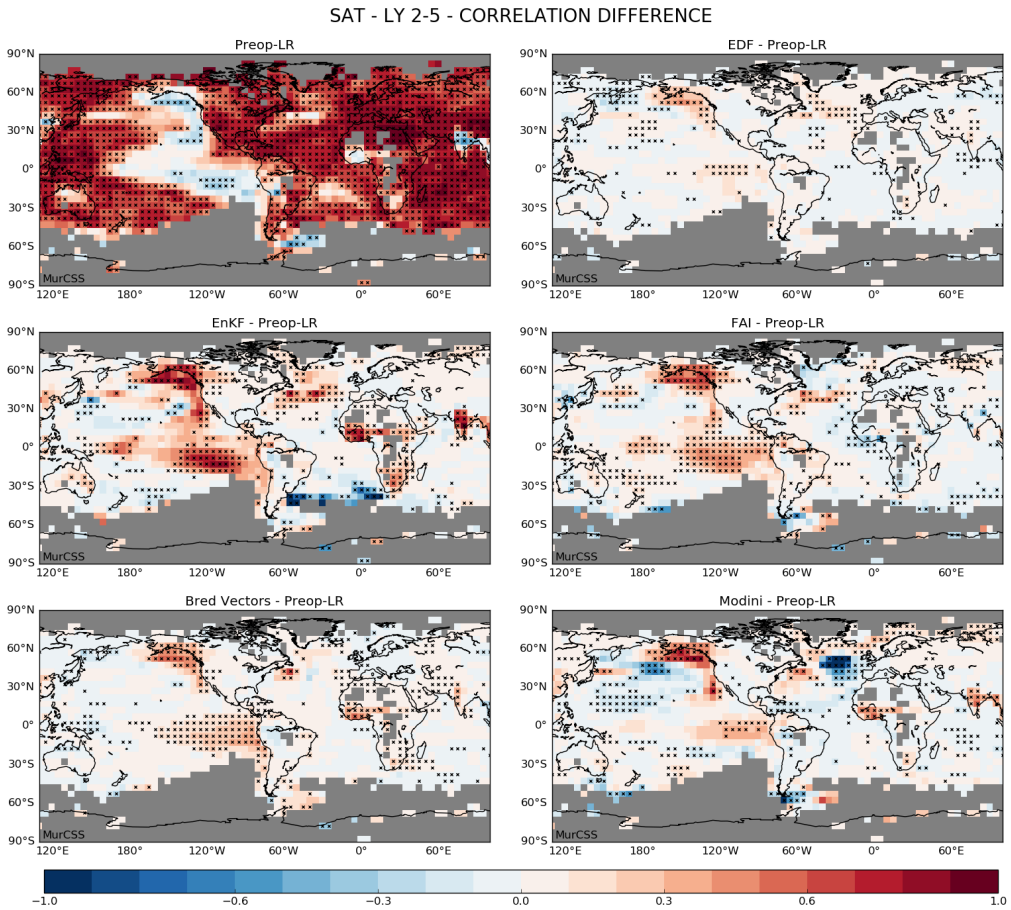


**Figure 1.** Correlation skill for SAT w.r.t. HadCRUT4 for lead years 2-5 from Preop-LR (top left panel) and the correlation skill difference between a particular test-suite experiment and Preop-LR (first, second and third rows). Hindcasts initialized from 1960 to 2011 are evaluated. Stippling indicates statistical significance estimated with the bootstrap method that the value is positive at the 95 % confidence level. Areas are masked where time series from the observational data set contain missing values.

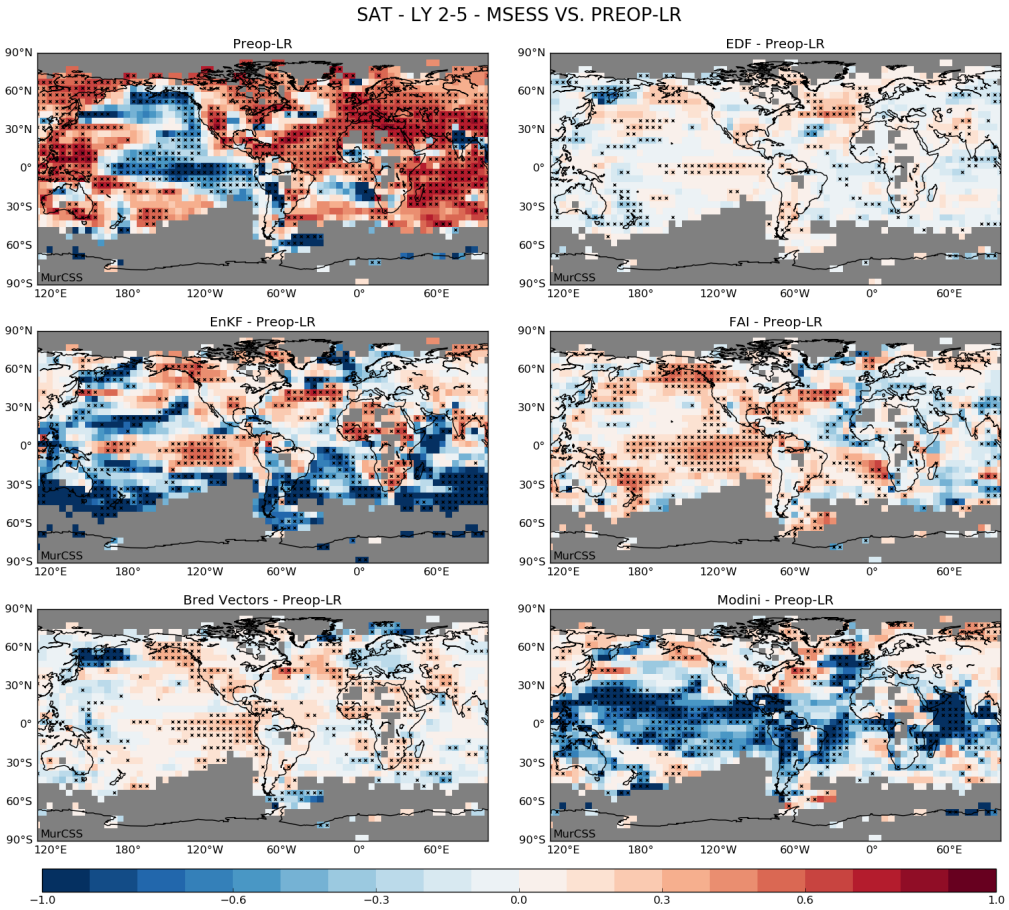


**Figure 2.** MSESS for SAT w.r.t. the HadCRUT4 climatology for lead years 2-5 from Preop-LR (top left panel). MSESS for SAT w.r.t. Preop-LR for the test-suite experiments (first, second and third rows). Hindcasts initialized from 1960 to 2011 are evaluated. The range of MSESS is from -$\infty$ to +1. Stippling indicates significant MSESS values as estimated with the bootstrap method at the 95 % confidence level.

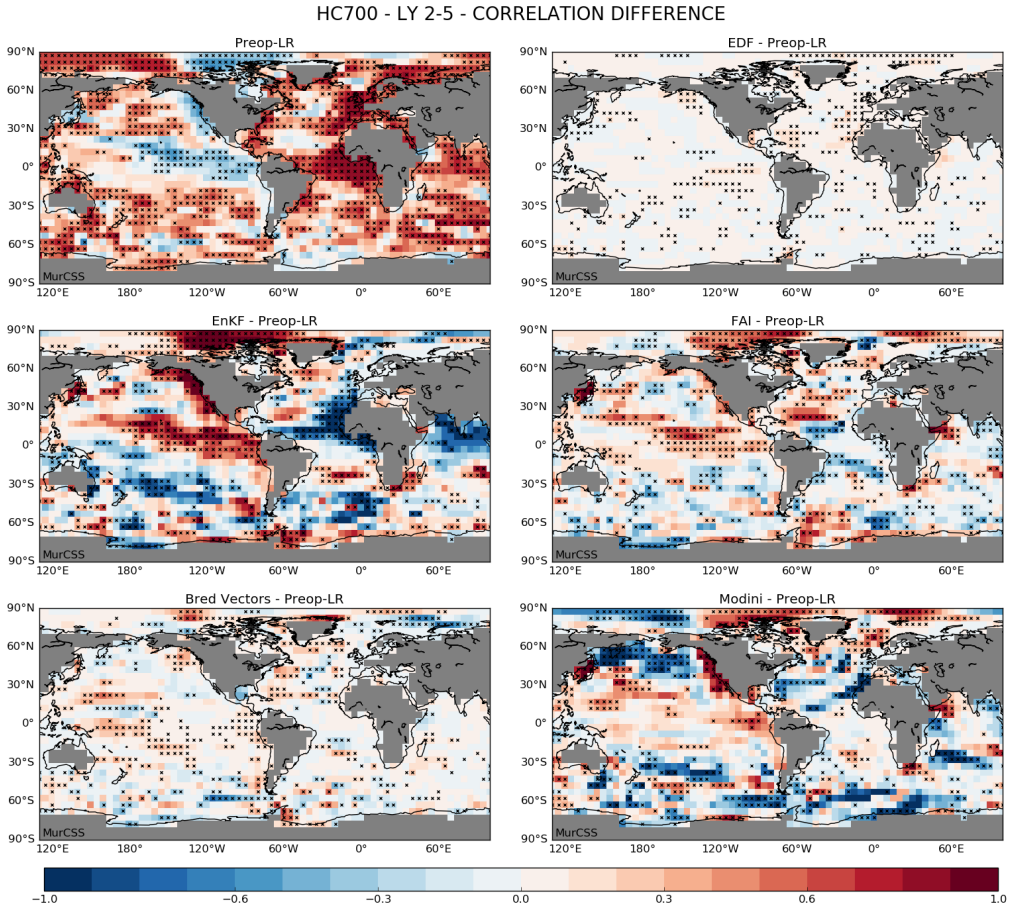


**Figure 3.** Correlation skill for HC700 w.r.t. the NOAA/NODC product for lead years 2-5 from Preop-LR (top left panel) and the correlation skill difference between a particular test-suite experiment and Preop-LR (first, second and third rows). Hindcasts initialized from 1960 to 2011 are evaluated. Stippling indicates statistical significance estimated with the bootstrap method at the 95 % confidence level.

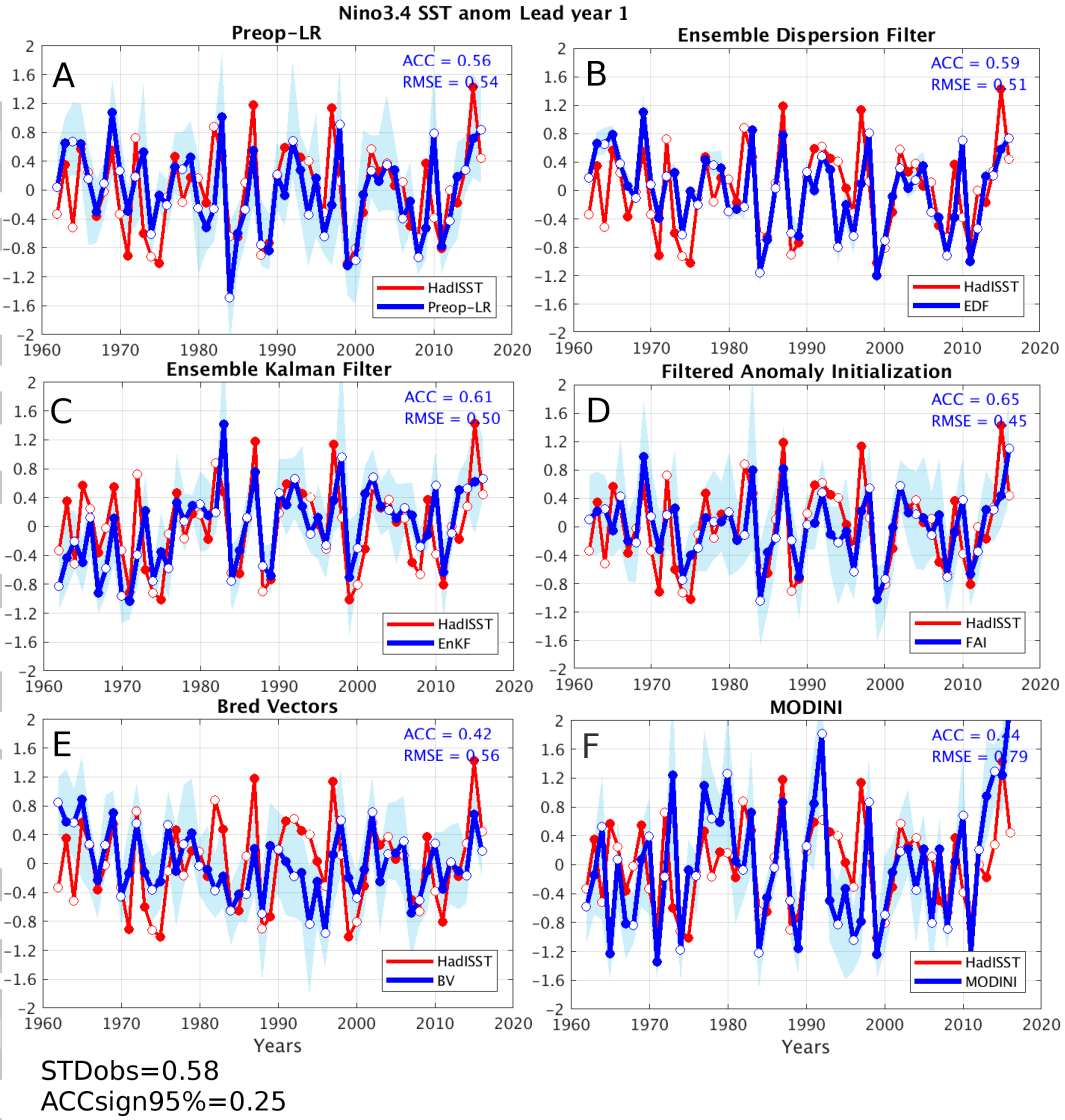


**Figure 4.** Time series of SST anomalies (°C) averaged over the Nino 3.4 region (5°S-5°N and 170W°-120°W) from HadISST1.1 (red) and the ensemble mean hindcasts at lead year one (blue). Shaded is the range between the minimum and maximum ensemble members. Empty (filled) circles indicate even (odd) years. ACC is the anomaly correlation coefficient and RMSE – root mean squared error for the ensemble mean hindcasts which are initialized over the period 1961-2015. STDobs represents variability in the observational data set (in °C). ACCsign95% gives a threshold for significant correlation coefficients. It is assessed using the $t$-test at 5% level, using the autocorrelation of the time series to estimate the degrees of freedom.

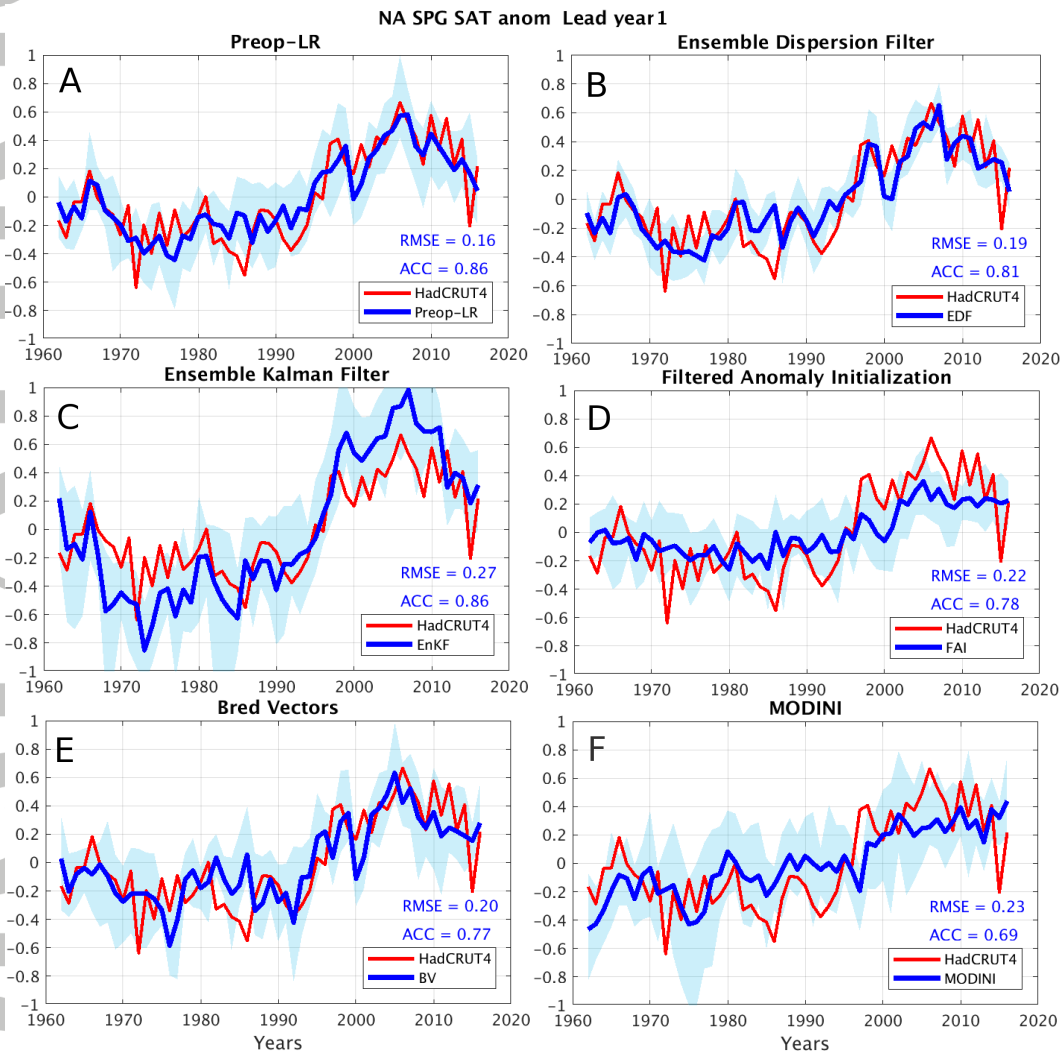


**Figure 5.** Same as in Figure 4 for the North Atlantic subpolar gyre (SPG) SAT in °C (50°-60°N and 65°W-10°E) from HadCRUT4 (red) and the ensemble mean hindcasts at lead year 1 (blue) initialized over the period 1961-2015.

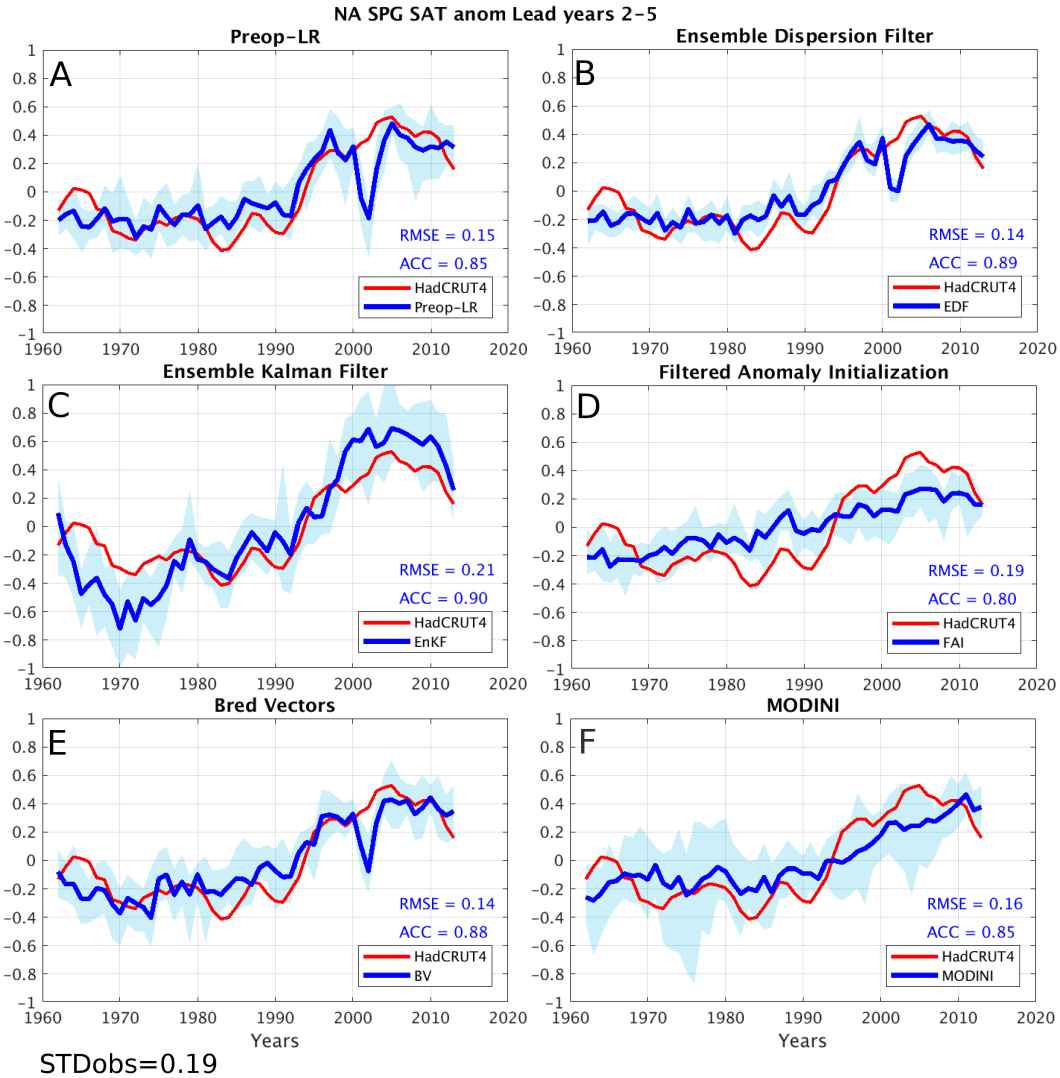


**Figure 6.** Same as in Figure 4 for the North Atlantic SPG SAT in °C from HadCRUT4 (red) and the ensemble mean hindcasts at lead years 2-5 (blue) initialized over the period 1960-2011. 4-year running mean is applied to the HadCRUT4 data.

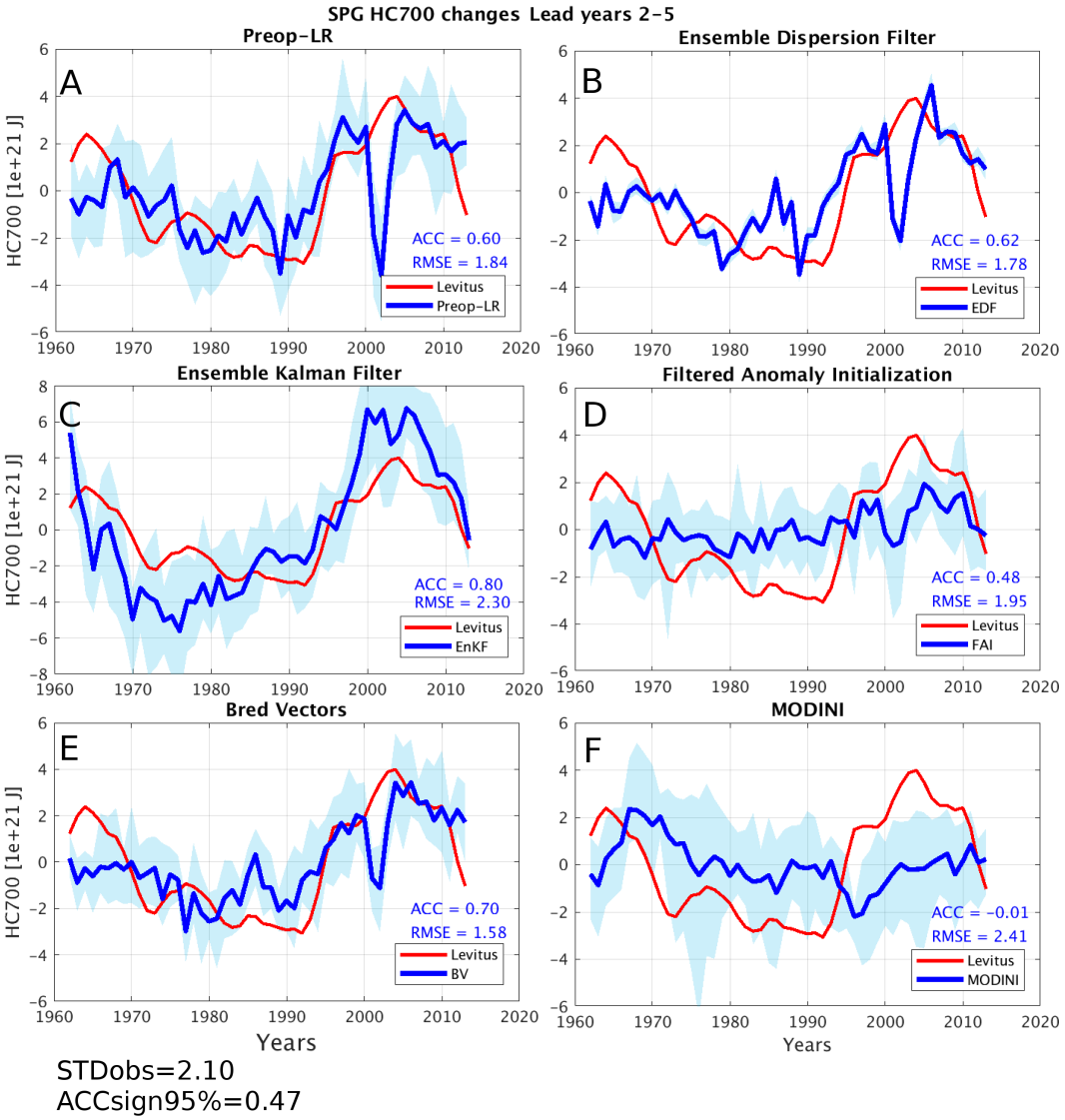


**Figure 7.** Time series of the North Atlantic SPG OHC (1e+21 J) from NOAA/NODC (red) and the initialized hindcasts at lead years 2-5 (blue). In bold solid is the ensemble mean and shading indicates the range of the ensemble members. ACC is the anomaly correlation coefficient and RMSE – root mean squared error for the ensemble mean initialized hindcasts started over the period 1960-2011. 4-year running mean is applied to the NOAA/NODC data. STDobs represents variability in the observational data set. ACCsign95% gives a threshold for significant correlation coefficients, it is estimated using the $t$-test at 5% level, using the autocorrelation of the time series to estimate the degrees of freedom.

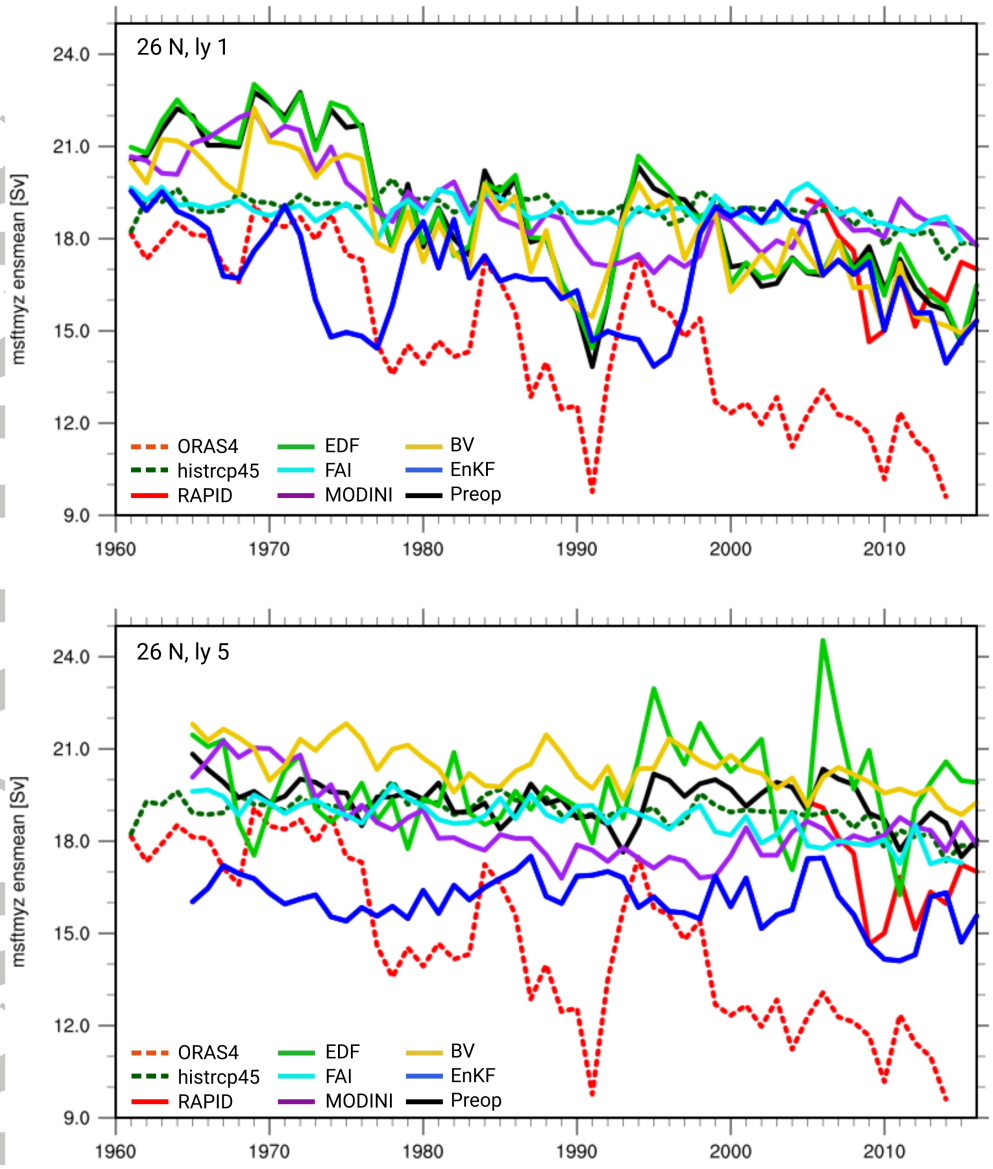


**Figure 8.** The time series of the Atlantic meridional overturning circulation (in SV) at 26.5°N latitude 1000 m depth from the initialized hindcasts (Preop-LR - black, the EnKF - blue, the BV - yellow, FAI - cyan, MODINI - magenta and the EDF - green) in the 1st (upper panel) and 5th (lower panel) lead years and the un-initialized historical simulation (green dashed), the ORAS4 reanalysis (red dashed) and the RAPID observations (red solid).

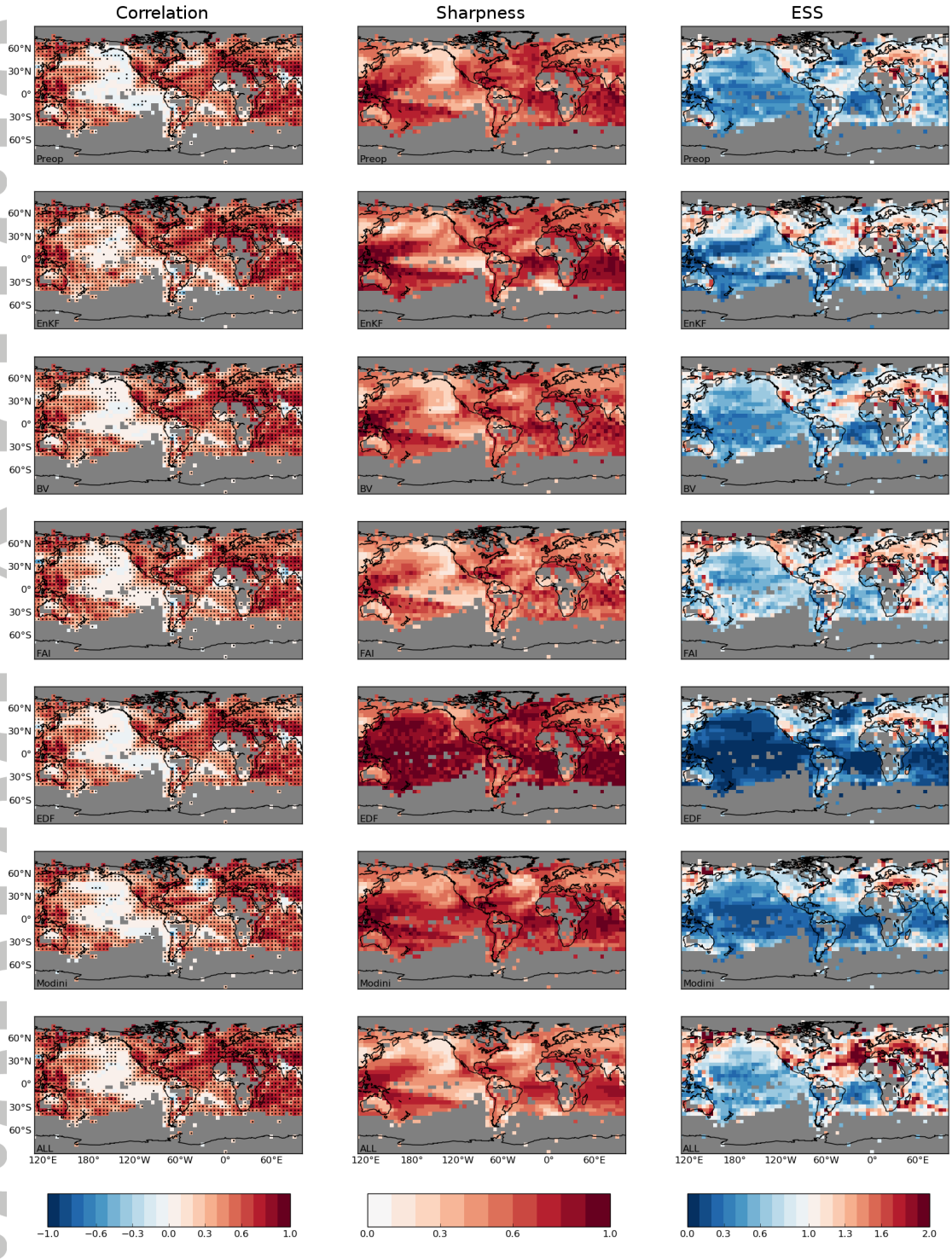


**Figure 9.** Correlation ($r_{HO} \times |r_{HO}|$; first column) and sharpness ($p$, second column) according to the decomposition of ESS (third column) for standardized SAT at lead years 2-5 from Preop-LR and the test-suite experiments. Skill is assessed w.r.t. HadCRUT4 for hindcasts initialized from 1960 to 2011. Stippling on correlation patterns indicates statistical significance at the 95 % confidence level.