Decadal climate prediction has progressed from being an avant-garde enterprise of only a few modeling groups to the scientific mainstream within less than a decade (Smith et al. 2007; Keenlyside et al. 2008; Pohlmann et al. 2009; Mochizuki et al. 2010; Kirtman et al. 2013; Meehl et al. 2014). Responding to both the new research opportunities and the enhanced societal requirements for information about near-term future climate change (e.g., WMO 2011; Kirtman et al. 2013), the German Federal Ministry for Education and Research has for the period 2011–19 funded a comprehensive national project on decadal climate prediction, Mittelfristige Klimaprognose (MiKlip; midterm climate forecast). This paper summarizes the scientific, strategic, and structural lessons learned from MiKlip so far.

A decadal prediction system simulates not only the climate response to future natural and anthropogenic forcing but also the future evolution of internal climate variability, caused by chaotic processes. Because chaos fundamentally limits climate predictability, a decadal prediction must be initialized from the observed state of those components of the climate system that provide a multiyear “memory,” usually but not exclusively the ocean (e.g., Bellucci et al. 2015a). Relevant ocean memory arises from the persistence of ocean heat content anomalies, especially where the atmosphere interacts with deep oceanic mixed layers, such as in the North Atlantic and North Pacific Subpolar Gyres (e.g., Mochizuki et al. 2010; Guemas et al. 2012; Matei et al. 2012b). Ocean memory possibly also arises from properly initialized ocean circulation and hence “slow” ocean dynamics [e.g., Matei et al. (2012b); a comprehensive review of the principles behind decadal prediction was recently provided by Kirtman et al. (2013)].

The quality of a decadal prediction system is assessed—in analogy to a seasonal prediction system—by performing a set of hindcasts (retrospective predictions) and by evaluating these hindcasts against the observed climate evolution. This evaluation step requires a sufficiently powerful observing system and is therefore usually limited to the period since around 1960. Assessing the gain in prediction skill that is obtained through the initialization is a core element of decadal prediction research, although for the users of such a prediction it matters little whether skill arises from the expected change in forcing or from the initialized internal variability.

The MiKlip project aims to establish and improve a decadal climate prediction system that by the end of the project can be transferred to the German
The abstract for this article can be found in this issue, following the table of contents. DOI:10.1175/BAMS-D-15-00184.1
period is subdivided into five development stages of usually 18 months in length. Each transition from one development stage to the next marks a well-defined and easy-to-communicate point in time for collecting, synthesizing, and implementing recommendations for changes in the global prediction system. Three generations of the prediction system are now available, termed baseline0, baseline1, and prototype (Table 1). Because of the relative timing of CMIP5 and the MiKlip start, we could use the CMIP5 initialized simulations (hindcasts) as our starting point, a set that we redubbed for MiKlip use as baseline0. Already during development stage 1, we defined and performed the next set of hindcasts (baseline1), using an initialization procedure and initialization data different from baseline0. Based on the research during development stage 1, we have defined and executed during development stage 2 the experiments with the prototype system. We have not defined a prediction generation for development stage 3 (see “Discussion and conclusions” section); at this writing, we are at the beginning of development stage 4.

From baseline0 to baseline1. Our design of baseline1 started from the recognition that baseline0 performed poorly in the tropics. Following Matei et al. (2012b), the initial conditions in baseline0 were constructed from a simulation with the Max Planck Institute Ocean Model (MPIOM; Jungclaus et al. 2013) forced by the National Centers for Environmental Prediction (NCEP)—National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996). The three-dimensional ocean temperature and salinity anomalies of the forced ocean run were added to the coupled model climatology; in a step with the coupled model called the assimilation run, the ocean hydrography was nudged to this sum of fields. The coupled model state resulting from the assimilation run was used as the initial condition for the 10-yr-long hindcast simulations. While this simple initialization gave excellent hindcast skill for North Atlantic sea surface temperature (SST) and even some skill in central European summer surface air temperature (Müller et al. 2012), the initialization led to degraded performance for SST in the tropics, compared to the uninitialized (historical) CMIP5 simulations (Figs. 1a,d; Müller et al. 2012; Bellucci et al. 2015b). This poor performance in the tropics may have arisen from the very simple initialization procedure, leading to a lack of balance between zonal wind stress and ocean surface pressure gradient in the coupled model (Thoma et al. 2015) or from the observations used in the procedure (e.g., McGregor et al. 2012; Lee et al. 2013; Pohlmann et al. 2016, manuscript submitted to Geophys. Res. Lett.).

A test suite of three-member hindcast ensembles with yearly start dates from 1961 onward explored various alternative initialization procedures. For each initialization, hindcast skill was evaluated for some

| Table 1. Experiments performed in MiKlip. In MPI-ESM-LR, LR stands for low resolution, T63 horizontally with 47 levels in the atmosphere and nominally 1.5° horizontal resolution and 40 levels in the ocean. In MPI-ESM-MR, MR stands for mixed resolution, T63 with 95 levels in the atmosphere and 0.4° horizontal resolution with 40 levels in the ocean. |
|-----------------|-----------------|-----------------|
| **Models**      | **Baseline0**   | **Baseline1**   |
| **Initialization** | **Prototype**   | **Prototype**   |
| **Ocean**       | **atmosphere**  | **MR**: 5       |
| **3D temperature (T)–salinity (S) anomalies from MPIOM forced with NCEP–NCAR reanalysis** | **Assimilation run** | **LR: 10**       |
| **3D T–S anomalies from ORAS4** | **ERA-40 and ERA-Interim; vorticity, divergence, log (p), T; full field** | **30 (15 each with initialization from GECCO2 and ORAS4)** |
| **Ensemble size** | **Start years**  | **Start years**  |
predefined measures such as global-mean surface temperature, North Atlantic SST index, and, for years 2004–10, the Atlantic meridional overturning circulation (AMOC) at 26.5°N. These evaluations suggested initializing the ocean with temperature and salinity anomalies from the Ocean Reanalysis System 4 (ORAS4; Balmaseda et al. 2013) reanalysis and the atmosphere from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) and ECMWF interim reanalysis (ERA-Interim; Dee et al. 2011; Table 1).

Baseline1 shows much improved correlation skill for tropical surface temperature, compared to baseline0, while maintaining positive skill in North Atlantic surface temperature (Fig. 1; see also Pohlmann et al. 2013). Almost all regions with negative correlation in baseline0 show positive correlation in baseline1 (tropical Atlantic, Africa, Indian Ocean, and western Pacific). Only the eastern Pacific continues to show negative skill, although less pronounced than in baseline0, in a pattern resembling the Pacific decadal oscillation (see also Mochizuki et al. 2010; Guemas et al. 2012). The improvement in tropical SST hindcast skill in baseline1 has led to a substantial improvement also in hindcast skill for global-mean surface temperature (Pohlmann et al. 2013).

Compared against the uninitialized (historical) simulations, initialization continues to provide additional skill primarily in the North Atlantic, owing to the deep mixed layers and associated long-lived heat

![Fig. 1. Evolution during the MiKlip project of the ensemble-mean hindcast skill [(left) anomaly correlation; (right) with anomaly correlation of historical simulations subtracted] of surface air temperature averaged over the lead years 2–5 in the low-resolution model version MPI-ESM-LR. Observations are from Hadley Centre/Climatic Research Unit, version 4 (HadCRUT4; Jones et al. 2012). The period is 1961–2012; (a),(d) baseline0, (b),(e) baseline1, and (c),(f) prototype. Hindcast ensemble size is 3 for baseline0, 10 for baseline1, and 30 for prototype; historical ensemble size is 3 for baseline0, 10 for baseline1, and 15 for prototype. Crosses denote skill different from zero exceeding the 95% confidence level; significant negative skill indicates where the initialization causes skill degradation. The figure was created with the evaluation tool provided by the central evaluation system (Illing et al. 2014).]
content anomalies there (Fig. 1e). Because the skill enhancement in the North Atlantic is supported by robust physical understanding (e.g., Matei et al. 2012b), we have confidence in this result, although the region covers only a small portion of the globe. Notice that northeastern North Atlantic SST skill relative to the historical simulations in baseline0 is inflated because of one particularly improbable historical realization within the small ensemble of three; the larger ensemble size in baseline1, both in initialized and historical simulations, means that skill assessment is more robust (see “Evaluation of prediction system generations” sections). The baseline1 hindcasts track the observed time series of North Atlantic Subpolar Gyre SST quite well and much better than the historical simulations, with the exception of a large and unexplained drop centered around year 2002 (Fig. 2). In particular, the hindcasts also show the downward trend beginning in 2005 [as was found earlier by Hermanson et al. (2014) with the Met Office decadal prediction system], and our predictions suggest that this downward trend is not reversed until the end of the current decade.

From baseline1 to prototype. The design of the prototype system was based on a far more comprehensive assessment compared to the design of baseline1. Suggestions for modifications were collected from each MiKlip subproject; a number of suggestions for modified initialization could readily be implemented and tested. The first suggestion is based on the recognition that the German contribution to Estimating the Circulation and Climate of the Ocean 2 (GECCO2) ocean reanalysis (Köhl 2015) provides an improved initial state compared to its predecessor GECCO [which was used earlier in Pohlmann et al. (2009), Matei et al. (2012b), and Kröger et al. (2012)]. The model comprises higher horizontal and vertical resolution, the domain is now fully global including the Arctic, and the simulation has been extended into the most recent years. Benefits of the new assimilation can be seen in several GECCO2 solution properties crucial for decadal prediction, such as ocean heat content, which, compared to the reference simulation (without assimilation), shows reduced and more realistic interdecadal variability. The AMOC at 26.5°N agrees excellently between the reanalysis and the observations (Fig. 3; Köhl 2015).

The workflow for producing initial conditions from GECCO2 has been modified so that the data needed for the initialization are available for quasi-operational use. Such availability, ideally with no more than a 1-month delay, cannot currently be obtained through the full-blown and computationally intensive four-dimensional variational data assimilation (4D-Var) method on which GECCO2 is based. This drawback is overcome here by performing shorter independent optimization runs toward the end of the assimilation window and further by appending a brief unconstrained run with unadjusted forcing for the final period. This modification in the workflow might make 4D-Var more broadly applicable not only for reanalyses but also for predictions.

The second suggestion for modified initialization concerns the use of full-field rather than anomaly initialization in the ocean, reflecting a more general tendency in the decadal prediction field (Smith et al. 2013a; Meehl et al. 2014; Polkova et al. 2014). A simulation closer to the observed mean state, instead of the coupled model’s, offers conceptual advantages because some important climate processes such as sea ice formation and melt and atmospheric tropical stability are sensitive to the background state. Moreover, full-field initialization obviates the need to compute anomalies separately.

A suite of three-member test hindcast ensembles, using each of ORAS4 and GECCO2 in both anomaly and full-field ocean initialization, suggested that all three initialization alternatives to the baseline
initialization (cf. Figs. 1b,e) led to improvements in the eastern tropical Pacific, the Indian Ocean, and the region in the northwestern North Atlantic where the three-member subensemble of baseline1 showed a relative minimum in skill (not shown). Although the skill was not improved everywhere, we concluded from the results of the initialization module (Polkova et al. 2014) and our additional test ensemble that the prototype system should use full-field initialization. The differences between ORAS4 and GECCO2 were only slight (not shown), so we used both initialization fields side by side.

Most baseline0 and baseline1 hindcasts were performed with the Max Planck Institute Earth System Model, low resolution (MPI-ESM-LR; T63 with 47 levels in the atmosphere and nominally 1.5° horizontal resolution and 40 levels in the ocean). The Max Planck Institute Earth System Model, mixed resolution (MPI-ESM-MR; T63 with 95 levels in the atmosphere; 0.4° horizontal resolution with 40 levels in the ocean), has yielded only modest benefit in the hindcasts (Pohlmann et al. 2013), just as in the CMIP5 historical simulations (Jungclaus et al. 2013). Clear exceptions exist where use of the higher vertical resolution is essential, such as for the QBO (Pohlmann et al. 2013; see “Evaluation of prediction system generations” section). But given the computational constraints, we decided against the use of MPI-ESM-MR in the prototype system.

Instead, the prototype system employs a much larger ensemble than before. With increasing ensemble size, the ensemble-mean correlation with observations is expected to increase, while the uncertainty of the skill estimate and the risk of finding spurious skill are expected to decrease (Murphy 1990; Kumar et al. 2001; Scaife et al. 2014a). These expectations are confirmed in baseline1 for the North Atlantic SST index and central European summer surface temperature (Fig. 4; Sienz et al. 2016). The prototype system thus comprises 30 ensemble members instead of 10, with 15 members each based on ORAS4 and GECCO2 (Table 1).

Hindcast ensembles are generated in baseline0 and baseline1 through lagged initialization, meaning that the model initial state at the nominal start day (1 January of any given start year) is taken from the state a few days earlier or later. The chaotic nature of the atmospheric model solution implies that the realizations soon drift away from each other and develop their own weather histories. But this procedure does not explore the possible ocean initial conditions that within uncertainty bounds are consistent with the available observations. Therefore, MiKlip aims at the development of alternative ensemble-generation procedures that explore the possible initial states more fully (see also Du et al. 2012).

Four procedures have been tested: empirical oceanic singular vectors (Molteni et al. 1996; Marini et al. 2016), the anomaly transform (Wei et al. 2006; Romanova and Hense 2015), a multiassimilation run approach in which the assimilation is based on several realizations of a historical run (Keenlyside et al. 2008), and the singular evolutive interpolated Kalman (SEIK) filter (Pham et al. 1998; Brune et al. 2015). Unfortunately, no robust improvement compared to the lagged initialization has been found; if there

Fig. 3. (a) Anomalies of the global-mean upper-ocean (0–700 m) heat content from the reference run (Ref, no assimilation) and GECCO2 in comparison to the estimates from Levitus et al. (2009) and Ishii and Kimoto (2009). (b) Comparison of the monthly mean meridional overturning at 26.5°N in Sv (1 Sv = 10⁶ m³ s⁻¹) from GECCO2 (red) and the reference run (black) with observations from Rapid Climate Change (RAPID; green; e.g., Cunningham et al. 2007; McCarthy et al. 2012). Updated from Köhl (2015), including a correction to the curve from the reference run in (a).
is improvement, this is compensated by additional problems such as an overestimation of the internal variability by the ensemble spread in some, though not all, variables (Marini et al. 2016). A speculative interpretation of this result suggests that on the time scales relevant here, variability even in the ocean interior might be dominated by the forcing from atmospheric internal variability. Because the more sophisticated ensemble-generation methods do not yet provide a clear path forward, we use the same lagged initialization procedure in the prototype system as in baseline0 and baseline1.

Given the large effort that went into designing and executing the prototype system, the comparison against baseline1 for surface temperature averaged over lead years 2–5 is a little sobering. We see incremental improvement in the correlation with observations, such as in the eastern tropical Pacific and the central North Atlantic (Figs. 1b,c), but the skill improvement by initialization has not increased against baseline1, except around Drake Passage and the Indian Ocean portion of the Southern Ocean (Figs. 1e,f). The anticipated improvements from the combination of enhanced ensemble size and full-field initialization have thus not materialized for all quantities.

EVALUATION OF PREDICTION SYSTEM GENERATIONS. The evaluation module pursues two related but distinct objectives; first, data-oriented evaluation of the prediction system and, second, process-oriented evaluation beyond the estimation of forecast skill for standard model output. Much of the data-oriented work stems from the recognition that observational datasets often provide insufficient spatiotemporal coverage or quality to enable a comprehensive evaluation of the prediction system. Therefore, considerable work is required on these observational datasets themselves. For example, global precipitation data over both land and ocean have been reprocessed for the period 1988–2008 to deliver daily maps with a grid resolution of $1^\circ \times 1^\circ$ and $2.5^\circ \times 2.5^\circ$, with a traceable estimate of the uncertainty (Schamm et al. 2014; Andersson et al. 2016a,b). As another example, variations in terrestrial water storage since 2002 have been inferred from GRACE satellite gravity measurements and used for the evaluation of the MiKlip hindcasts (Zhang et al. 2015).

The work on verification and process-oriented evaluation takes as its starting point the recommendations by Goddard et al. (2013). These include bias adjustment, typical spatial and temporal scales of aggregation, and verification of the hindcast ensemble proceeding along two lines. The first line of verification focuses on the mean square error skill score (MSESS), which tests whether the ensemble mean of a prediction outperforms a reference prediction, measured against a verification dataset. In the simple case of climatology as reference forecast, the MSESS combines the correlation between anomalies, the conditional bias (the prediction system systematically overestimates or underestimates the magnitude of anomalies), and the unconditional bias (difference

Fig. 4. Correlation with observations for (a) annual-mean North Atlantic (20°–60°N, 10°–80°W) SST and (b) central European (40°–45.5°N, 10°–30°E) summer (Jun–Aug) surface temperature for baseline1 hindcasts averaged over lead years 2–5 (red) and historical runs (blue). Shown is the dependence of the correlation on the ensemble size $k$; the vertical lines are 95% confidence intervals. The dots at $k = 1$ give the correlations for the single members. Numbers at the top are $p$ values for the correlation skill scores of baseline1, with historical simulations as the reference prediction. The approximate theoretical correlation–ensemble size relation is given by the red (baseline1) and blue (historical) solid lines, based on Murphy (1990). The observations used are HadISST1.1 (Rayner et al. 2003) for SST and Climate Research Unit Time Series version 3.10 (CRU TS3.10) (Harris et al. 2014) for surface temperature over land. From Sienz et al. (2016), reproduced with permission (www.schweizerbart.de).
between time averages; Murphy 1988). In some results shown here, the anomaly correlation is used because the conditional bias is assumed small and the unconditional bias has been subtracted. The second line of verification focuses on the full probabilistic hindcast derived from the ensemble. We use a variant of the rank probability skill score (RPSS), which assesses whether the ensemble spread of predictions accurately represents the forecast uncertainty (e.g., Kadow et al. 2015).

The central evaluation system is constantly expanded with contributions from the MiKlip evaluation module and, together with its reference data pool for verification, resides on the same data server as the entire MiKlip prediction output. The analyses are collected into a database ensuring reproducibility and transparency. Providing the central evaluation system to the entire MiKlip project is also an effective training tool, especially for those researchers who have only recently joined the rapidly expanding field of decadal prediction.

Applying the central evaluation system to the three MiKlip hindcast generations has identified a problem with the full-field initializations that to our knowledge has so far escaped attention. While the prototype hindcasts tend to provide the highest skill for North Atlantic Subpolar Gyre SST in later lead years, early lead years display a marked degradation in skill. This degradation is most pronounced in a drop in correlation skill in the initializations with ORAS4 and an increase in RMSE in the initializations with GECCO2 (Fig. 5). Presumably this skill degradation is related to model drift upon initialization with a state that builds on an incompatible climatology. Figure 5 furthermore illustrates the limitation of our testing procedure with small test ensembles—it is only the full prototype ensemble that identifies the consequences of the drift and forces us to readdress the question of full-field versus anomaly initialization.

As an example of evaluating probabilistic forecasts of discrete events with the RPSS, we analyze whether wind storms related to intense extratropical cyclones occur at a frequency that is either below normal, normal, or above normal for the Northern Hemisphere extended winter season (October through March; Fig. 6; Kruschke et al. 2015). The analysis combines the 29 realizations from all three MiKlip generations available at that time. Using climatology as the reference leads to RPSS-based skill over most of the Northern Hemisphere (not shown; Kruschke et al. 2015). Against the historical simulations as reference, however, additional skill arises in only a few regions, the most prominent of which are the entrance of the North Pacific storm track over eastern Asia and the northwestern Pacific. Similar but less pronounced and less coherent skill enhancement occurs at the entrance of the North Atlantic storm track along the North American east coast and the American sector of the Arctic Ocean (Fig. 6; Kruschke et al. 2015).

![Fig. 5](image_url)
For the analysis shown in Fig. 6, Kruschke et al. (2015) developed and used a bias correction that goes beyond the one recommended in Goddard et al. (2013). The standard correction method is effectively an adjustment of the mean that only depends on lead time. But in a changing climate, model drift following initialization depends also on start year (Kharin et al. 2012). Kruschke et al. (2015) therefore combined the bias correction by Gangstø et al. (2013), which is formulated as a third-order polynomial in lead time, with the drift correction proposed by Kharin et al. (2012) by making the coefficients of the third-order polynomial a linear function of the start year.

We mention here four further examples of evaluating hindcast skill for quantities other than the surface temperature. First, the baseline1-MR version shows prediction skill for the QBO for lead times of up to 4 years. Here, it is essential to use the atmospheric initialization as well as the high vertical resolution in the atmosphere for basic process representation (Pohlmann et al. 2013; see also Scaife et al. 2014b). Second, the MESS and ensemble reliability have been computed for zonal-mean geopotential height. The only weak dependence of the skill measures on lead time suggests that geopotential height, changes in external forcing are the main source of skill (Stolzenberger et al. 2015). Third, baseline1 displays significant prediction skill for the AMOC at 26.5°N (Müller et al. 2016, manuscript submitted to Climate Dyn.), confirming the earlier results obtained with a system predating the CMIP5 (Matei et al. 2012a), although the physical cause of the prediction skill appears to be different. And fourth, baseline1 shows multiyear potential prediction skill for carbon uptake by the North Atlantic Subpolar Gyre, arising from the improved representation of SST through the initialization (Li et al. 2016).

**Fig. 6.** Hindcast skill for whether wind storms related to intense extratropical cyclones occur at a frequency that is either below normal, normal, or above normal, for the Northern Hemisphere extended winter season (number of tracks within 1,000 km per period Oct–Mar), in a 29-member ensemble constructed from all three MiKlip generations. Skill score is the RPSS; the reference predictions are the historical simulations, and the verification dataset is the ERA reanalyses. (a) Hindcast of winters 2–5 and (b) hindcast of winters 2–9; significant skill scores are indicated ($\alpha < 5\%$) as black dots, and areas of strong inconsistencies between ERA-40 and ERA-Interim are masked out (gray). From Kruschke et al. (2015), reproduced with permission (www.schweizerbart.de).

**PROCESSES AND MODEL DEVELOPMENT.**

One MiKlip module aims to understand better the processes causing decadal variability, to improve existing model components, and to incorporate additional climate subsystems that are relevant for decadal climate predictions. Substantial effort is devoted to exploring the effects of model resolution. For example, a higher-resolution (T106) version of the CMIP3 atmospheric model ECHAM5 revealed that a significant fraction of the convective precipitation over and south of the Gulf Stream can be explained by the variability of the underlying SST, especially in summer (Hand et al. 2014; see also Minobe et al. 2008). Higher horizontal resolution in both atmosphere and ocean is expected to improve the teleconnections between the North Atlantic and Europe (e.g., Minobe et al. 2008; Hand et al. 2014), which are weaker at the T63 atmospheric horizontal resolution used in MiKlip than in reanalyses (e.g., Müller et al. 2012; Ghosh et al. 2016). Increasing the atmospheric horizontal resolution to T127 is therefore high on MiKlip’s list of priorities.

The subpolar North Atlantic and its interaction between gyre and overturning circulations are important for the northward oceanic heat transport and thus for Atlantic warming events such as in the 1990s (Robson et al. 2012a) and the 1920s (Müller et al. 2012).
et al. 2015), including their predictions [Robson et al. (2012b) and Müller et al. (2014), respectively]. These results underscore the importance of reducing the misplacement of the Gulf Stream and the North Atlantic Current that is ubiquitous in CMIP5 climate models (e.g., Flato et al. 2013), including the MPI-ESM (Jungclaus et al. 2013).

Hindcast skill is markedly degraded by not including the effects of volcanic eruptions (Fig. 7; Timmreck et al. 2016). MiKlip has therefore developed a volcano code package that enables the running of a new ensemble of predictions if a major volcanic eruption occurs in the future. The volcano code package is implemented in a two-step procedure. In the first step, the volcanic radiative forcing is calculated offline with a global aerosol–climate model; in the second step, this forcing is included in the MiKlip system. As a consequence of this two-step procedure, the underlying climate model for producing the predictions remains unchanged, obviating the need to retune the model (Mauritsen et al. 2012) and to create new control and historical simulations.

**Downscaling the Decadal Predictions.** Climate information is often required at a substantially higher spatial resolution than is available from the global climate models, particularly for regional-scale impact studies. The representation of processes such as orographic rain, mesoscale circulations, or wind gusts improves as resolution is refined. For this reason, MiKlip has developed a coordinated regional downscaling component for the decadal predictions. The two main research questions pursued in MiKlip are (i) whether predictive skill can be found also on the much smaller regional and local scales and (ii) whether the downscaling adds value to the global predictions. The geographical focus lies on Europe and Africa. Because the regional models rely on the global results, there is necessarily some time lag between constructing the global hindcast ensembles and their use in downscaling.

Downscaling implies additional uncertainty (e.g., Räisänen 2007; Flato et al. 2013); therefore, different approaches are employed in MiKlip to assess the robustness of the results. These approaches are coordinated with respect to model grids, initialization, and data processing [analogous to the Coordinated Regional Climate Downscaling Experiment (CORDEX) contribution to CMIP5; e.g., Kotlarski et al. 2014]. For Europe, the ensemble consists of the two regional climate models (RCMs) Consortium for Small-Scale Modelling in Climate Mode (COSMO-CLM or CCLM; Rockel et al. 2008) and Regional-Scale Model (REMO; Jacob 2001), and a statistical–dynamical method. For Africa, three RCMs are used: CCLM, REMO, and Weather Research and Forecasting (WRF) Model (Skamarock and Klemp 2008).

The regionalization for Europe maintains or slightly enhances the skill inherited from the baseline global hindcasts for annual-mean surface temperature (Fig. 8). Given the user orientation of downscaled predictions, we show here the combined skill from forcing changes and initialized internal variability; skill score is MSESS evaluated against the European daily high-

![Fig. 7. Detrended time series of hindcast and observed globally averaged surface temperature from 1962 to 2004; the figure shows anomalies relative to the mean over this period for HadCRUT3v observations (Brohan et al. 2006; black), baseline1-LR (blue), and baseline1-LR without volcanic eruptions (b1-NVA; red). The blue and red curves are each based on three realizations. The purple line indicates the 10-ensemble-member mean of baseline1-LR. The standard deviation is indicated by the hatched areas. (a) Lead year 1 and (b) lead years 2–5. The numbers indicate the anomaly correlation coefficient between the hindcasts and the observations over the whole period. The gray shaded region right above the x axis shows the time series of annually averaged stratospheric aerosol optical depth (Stenchikov et al. 1998; and updates). From Timmreck et al. (2016), reproduced with permission.](image_url)
Fig. 8. Ensemble-mean hindcast skill over Europe for near-surface air temperature for lead time 2–5 yr, starting yearly from 1961 to 2004. Skill score is MSESS evaluated against E-OBS (Haylock et al. 2008), with climatology as the reference forecast; positive values indicate better skill than climatology. (a) MPI-ESM-LR baseline; (b) RCM ensemble (combined from CCLM and REMO). Given the user orientation of downscaled predictions, we show here the combined skill from forcing changes and initialized internal variability.

DISCUSSION AND CONCLUSIONS. MiKlip is well poised to deliver its decadal prediction and evaluation systems to the DWD for operational use by 2019. Placing a single global prediction system in the focus of a major research effort has demonstrated benefits such as the rapid development of alternative initialization strategies, sophisticated evaluation methods for quantities beyond the surface temperature, and regional applications of the global predictions. Such rapid progress would have been impossible at any single institution in Germany, no matter how scientifically powerful or well-funded.

At least five major issues remain unsettled and must be tackled by MiKlip in the coming years:

1) We have not yet converged on a best initialization procedure of our prediction ensemble. Some hindcasts suffer from degraded skill right after initialization, in particular when full-field initialization is used. This effect presumably is related to using an assimilation model, either statistical or dynamical, that is different from the model used in the hindcasts (Kröger et al. 2012). Furthermore, it is unsatisfactory that our initial condition ensemble is unable to explore the full uncertainty range of the initial ocean state.

2) The teleconnections between SST and surface temperature over land are not robust enough in our model. While MiKlip has successfully reproduced the observed connection between the SST in the tropical Atlantic and the West African monsoon (Paeth et al. 2016, manuscript submitted to Meteorologische Zeitschrift), prediction...
skill for North Atlantic SST translates into only some, but not sufficient, skill over Europe (Müller et al. 2012). The required higher-resolution version of MPI-ESM has until recently not been available, owing to some unrealistic features in an earlier control run (J. Jungclaus 2014, personal communication). These problems have now been overcome, and we will perform the next set of production runs with an atmospheric model with resolution T127 (MPI-ESM-HR).

3) The availability of the MPI-ESM-HR brings into even sharper relief the computing resource issue that we already faced when applying the MR version of our system. Because higher resolution usually implies smaller possible ensemble size, we experience a palpable trade-off between more realistic representation of physical processes on the one hand and the translation of this representation into prediction skill on the other hand. With a new computer available to MiKlip since July 2015, the competition for resources between resolution and ensemble size has subsided somewhat, but in the foreseeable future hindcasts with MPI-ESM-HR will be limited to an ensemble size of 10.

4) When starting MiKlip, we underestimated the difficulty of implementing suggested model improvements. Any modification to the climate model itself requires a retuning (e.g., Mauritsen et al. 2012), a new control run with constant forcing to make sure the model simulates a stable climate, and a new ensemble of historical runs as a reference for assessing skill enhancement through initialization. Being tied to the general MPI-ESM development implies that the cycle of model versions rests outside of MiKlip’s immediate control and occurs in intervals longer than sometimes desired by MiKlip. On the other hand, MiKlip does not command the personnel resources needed to maintain an independent climate model, and even if it did, separating its model development from that of the MPI-ESM would not use resources efficiently—

![Fig. 9. MSESS for wind energy output for (top) years 1–4, (middle) years 2–5, and (bottom) years 6–9 of the (left column) baseline0 3-member ensemble mean, (second column) baseline1 10-member ensemble mean, (third column) prototype ORAS4 10-member subensemble mean, and (right column) prototype GECCO2 10-member subensemble mean. Reference prediction is the ensemble mean of the uninitialized historical runs, using 3 realizations for baseline0 and 10 members for the other generations; verification dataset is the wind energy output downscaled from ERA-Interim (Dee et al. 2011). Adapted from Moemken et al. (2016).](image)
MiKlip would maintain a full-blown climate model for decadal prediction alone.

For generational cycles of the prediction system that are defined not through different model versions but through different initialization procedures, a much faster turnover can be implemented. The 18-month turnover originally envisioned in MiKlip, however, proved to be overambitious for a sustained mode of operation. We therefore decided not to produce a set of hindcasts during development stage 3 and have instead focused our effort on a comprehensive evaluation of the prototype system. A sustained 18-month turnover would imply that we could never explore the full implication of a generation of hindcasts, including the effects on downscaling, before designing the generation after. We thus tentatively recommend for later operational use to allow for a more relaxed cycle of prediction system generations, with intervals of 2–3 years rather than 18 months.

5) We have so far focused almost exclusively on evaluating the hindcasts and not on constructing and issuing our own exploratory forecasts, although we do participate in the multimodel real-time decadal prediction exercise led by the Hadley Centre (Smith et al. 2013b). We have also started a dialogue with potential users of the MiKlip forecasts and have now added subprojects that develop such a dialogue systematically. Issuing our own forecasts requires further exploration of how to communicate the strengths and weaknesses of the forecast in a manner both accurate and easy to grasp. MiKlip plans to tackle this challenge over the coming years because without this communication component an operational system would remain incomplete.

ACKNOWLEDGMENTS. MiKlip is funded by the German Federal Ministry for Education and Research (BMBF) under grant agreements 01LP11nnx, where nn ranges from 04 to 70 and x ranges from A to F. All simulations were carried out at the German Climate Computing Centre (DKRZ), which also provided all major data services. We thank Bjorn Stevens, the anonymous reviewers, and Editor Michael Alexander for comments on an earlier version of the manuscript.

REFERENCES


Ghosh, R., W. A. Müller, J. Bader, and J. Baehr, 2016: Impact of observed North Atlantic multidecadal variations to European summer climate: A quasi-


