The DOI for this manuscript is doi: 10.1175/BAMS-D-15-00184.1

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

MiKlip – a National Research Project on Decadal Climate Prediction

Jochem Marotzke[1], Wolfgang A. Müller[1], Freja S. E. Vamborg[1], Paul Becker[2], Ulrich Cubasch[3], Hendrik Feldmann[4], Frank Kaspar[2], Christoph Kottmeier[4], Camille Marini[5], Iuliia Polkova[5], Kerstin Prömmel[3], Henning W. Rust[3], Detlef Stammer[5], Uwe Ulbrich[3], Christopher Kadow[3], Armin Köhl[5], Jürgen Kröger[1], Tim Kruschke[3,6], Joaquim G. Pinto[7,8], Holger Pohlmann[1], Mark Reyers[7], Marc Schröder[2], Frank Sienz[1], Claudia Timmreck[1], Markus Ziese[2]

1. Max Planck Institute for Meteorology, Hamburg, Germany
2. Deutscher Wetterdienst (DWD), Offenbach, Germany
3. Institute of Meteorology, Freie Universität Berlin, Berlin, Germany
4. Institute for Meteorology and Climate Research (IMK-TRO), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
5. Institute of Oceanography, Center for Earth System Research and Sustainability (CEN), University of Hamburg, Hamburg, Germany
6. GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
7. Institute for Geophysics and Meteorology, University of Cologne, Cologne, Germany
8. Department of Meteorology, University of Reading, Reading, UK

Jochem Marotzke, Max Planck Institute for Meteorology, Bundesstrasse 53, 20146 Hamburg, Germany; jochem.marotzke@mpimet.mpg.de; *phone: +49-40-41173-311; fax: +49-40-41173-366 (corresponding author)*

Capsule Summary

A German national project coordinates research on improving a global decadal climate prediction system for future operational use.

Abstract

MiKlip, an eight-year German national research project on decadal climate prediction, is organized around a global prediction system comprising the climate model MPI-ESM together with an initialization procedure and a model evaluation system. This paper summarizes the lessons learned from MiKlip so far; some are purely scientific, others concern strategies and structures of research that targets future operational use.

Three prediction-system generations have been constructed, characterized by alternative initialization strategies; the later generations show a marked improvement in hindcast skill for surface temperature. Hindcast skill is also identified for multi-year-mean European summer surface temperatures, extra-tropical cyclone tracks, the Quasi-Biennial Oscillation, and ocean carbon uptake, among others. Regionalization maintains or slightly enhances the skill in European surface temperature inherited from the global model and also displays hindcast skill for wind-energy output. A new volcano code package permits rapid modification of the predictions in response to a future eruption.

MiKlip has demonstrated the efficacy of subjecting a single global prediction system to a major research effort. The benefits of this strategy include the rapid cycling through the prediction-system generations, the development of a sophisticated evaluation package usable by all MiKlip researchers, and regional applications of the global predictions. Open research questions include the optimal balance between model resolution and ensemble size, the appropriate method for constructing a prediction ensemble, and the decision between full-field and anomaly initialization.

Operational use of the MiKlip system is targeted for the end of the current decade, with a recommended generational cycle of two to three years.

1. Background and Philosophy

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 2011–2019 funded a comprehensive national project on decadal climate prediction, MiKlip (from the German **Mi**ttelfristige **Kli**ma**p**rognose; mid-term 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 multi-year “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 meteorological service DWD for operational use. To serve this dual purpose – pre-operational predictions combined with research progress – MiKlip is organized around a hub consisting of a global climate prediction system, in turn comprising the climate model MPI-ESM (Giorgetta et al. 2013) together with an initialization procedure. Around this hub, the research is organized in four modules focusing on initialization, evaluation, processes and modelling, and regionalization.

The MiKlip hub furthermore provides a central evaluation system. The evaluation system, the necessary observational data, as well as the entire set of MiKlip prediction results conform to the CMIP5 data standards (Taylor et al. 2012) and reside on a dedicated data server. The MiKlip server makes the prediction results and evaluation system immediately accessible to the entire MiKlip community, thereby providing a crucial interface between production on the one hand and research and evaluation on the other hand.

The structure of MiKlip differs notably from other community efforts in decadal climate prediction, especially the decadal-prediction portion of the Coupled Model Intercomparison Project Phase 5 (CMIP5; see Kirtman et al. 2013; Meehl et al. 2014). CMIP5 comprises sixteen different decadal prediction systems and thus offers a much richer spectrum of modeling approaches than does MiKlip, which focuses on a single global prediction system. On the other hand, MiKlip can produce quick and tailored research responses that help modify its prediction system. MiKlip could hence cycle through a greater number of generations of its prediction system, compared to the cycle defined by the different phases of CMIP; this faster cycle enables faster learning from successive generations (see Section 2).

A project that conceptually rests in between MiKlip and CMIP is “Seasonal-to-decadal climate Prediction for the improvement of European Climate Services” (SPECS, http://www.specs-fp7.eu/), funded by the European Union Framework Program 7. SPECS comprises six European climate prediction systems and thus shares with CMIP the multi-model approach. SPECS shares with MiKlip the strategy to coordinate research within the project and to coordinate improvements of the prediction systems; however, SPECS is not designed to provide the same interactive cycle of prediction-system improvements as does MiKlip. Overall, the approaches by MiKlip, SPECS, and CMIP complement each other.

The remainder of this paper is dedicated to the following scientific and strategic topics. Section 2 documents how we explored a variety of initialization methods and developed a strategy for deciding among them. These decisions have resulted in the succession of three generations of the MiKlip global decadal prediction system. Section 3 demonstrates that the systematic effort in prediction evaluation and verification has led to identification of prediction skill in many new quantities, such as multi-year-mean seasonal surface temperature over Europe, Northern-Hemisphere mid-latitude storm tracks, the Quasi-Biennial Oscillation (QBO), and carbon uptake by the North Atlantic. Section 4 presents aspects of enhanced process understanding and, in particular, how the development of a volcano code package enables us to include in future predictions the occurrence of a major volcanic eruption. Section 5 discusses how the regionalization of the predictions has made possible the identification of regional forecast skill. Section 6 provides a synthesis of the lessons learned from MiKlip so far.

2. Three generations of the global prediction system

The MiKlip funding period is subdivided into five development stages of usually eighteen months 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 re-dubbed 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 section 6); 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 ocean model MPIOM (Jungclaus et al. 2013) forced by the NCEP/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 initial condition for the ten-year-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 (Figure 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).

A test suite of three-member hindcast ensembles with yearly start dates from 1961 onwards explored various alternative initialization procedures. For each initialization, hindcast skill was evaluated for some pre-defined measures such as global-mean surface temperature, North Atlantic SST index, and, for years 2004–2010, the Atlantic Meridional Overturning Circulation (AMOC) at 26.5°N. These evaluations suggested initializing the ocean with temperature and salinity anomalies from the ORAS4 (Balmaseda et al. 2013) re-analysis and the atmosphere from the ERA40 (Uppala et al. 2005) and ERA-Interim (Dee et al. 2011) re-analyses (Table 1).

Baseline1 shows much improved correlation skill for tropical surface temperature, compared to baseline0, while maintaining positive skill in North Atlantic surface temperature (Figure 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-content anomalies there (Figure 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 Section 3). The baseline1 hindcasts track the observed time series of North Atlantic subpolar-gyre SST quite well and much better than do the historical simulations, with the exception of a large and unexplained drop centered around year 2002 (Figure 2). In particular, the hindcasts also show the downward trend beginning in 2005 (as was found earlier by Hermanson et al. 2014 with the UK MetOffice 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 sub-project; a number of suggestions for modified initialization could readily be implemented and tested.

The first suggestion is based on the recognition that the GECCO2 ocean re-analysis (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 re-analysis and the observations (Figure 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 one-month delay, cannot currently be obtained through the full-blown and computationally intensive four-dimensional variational (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 baseline1 initialization (cf., Figure 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 sub-ensemble 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 low-resolution model version MPI-ESM-LR (T63 with 47 levels in the atmosphere and nominally 1.5° horizontal resolution and 40 levels in the ocean). The mixed-resolution version 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 Section 3). 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 (Figure 4; Sienz et al. 2016). The prototype system thus comprises thirty ensemble members instead of ten, with fifteen 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 multi-assimilation-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 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 timescales 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 (Figure 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 (Figure 1e,f). The anticipated improvements from the combination of enhanced ensemble size and full-field initialization have thus not materialized for all quantities.

3. 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 spatio-temporal 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 re-processed for the period 1988–2008 to deliver daily maps with a grid resolution of 1° by 1° and 2.5° by 2.5°, with a traceable estimate of the uncertainty (Schamm et al. 2014; Andersson et al. 2016a, 2016b). 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 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 (Figure 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 re-address 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 extra-tropical cyclones occur at a frequency that is either below normal, normal, or above normal, for the Northern-Hemisphere extended winter season (October through March; Figure 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 Northwest 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 (Figure 6, Kruschke et al. 2015).

For the analysis shown in Figure 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 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 four 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 MSESS and ensemble reliability have been computed for zonal-mean geopotential height. The only weak dependence of the skill measures on lead time suggests that for 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), confirming the earlier results obtained with a system pre-dating 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).

4. 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. 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 (Figure 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 re-tune the model (Mauritsen et al. 2012) and to create new control and historical simulations.

5. 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 CORDEX contribution to CMIP5, e.g., Kotlarski et al. 2014). For Europe, the ensemble consists of the two regional climate models (RCMs) COSMO-CLM (CCLM, Rockel et al. 2008) and REMO (Jacob 2001), and a statistical-dynamical method. For Africa, three RCMs are used, CCLM, REMO, and WRF (Skamarock and Klemp 2008).

The regionalization for Europe maintains or slightly enhances the skill inherited from the baseline1 global hindcasts for annual-mean surface temperature (Figure 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 E-OBS (Haylock et al. 2008), with climatology as the reference forecast. The RCM ensemble consists of simulations with CCLM as well as with REMO, and it maintains the skill in western and southern Europe and shows an increase in parts of central, eastern and northern Europe (Figure 8).

Added value of the downscaling has been found for strong precipitation events over Central Europe; the RCM CCLM clearly outperforms the baseline0 global model in the representation of the frequency of days with precipitation larger than about 20 mm/day (not shown; Mieruch et al. 2014). Furthermore, while the global-model ensemble is overconfident (ensemble spread smaller than the error, a feature that is ever more pronounced with increasing precipitation intensity), the regional-model ensemble is reliable out to very large intensities.

A statistical-dynamical downscaling approach comprising a combination of weather-typing and CCLM simulations has been used to explore the predictability of wind-energy output over central Europe (Reyers et al. 2015). The skill score used is the MSESS, the reference prediction is the downscaled historical simulation, and the verification data set is the downscaled wind-energy output of ERA-Interim for the period 1979–2010. While no skill is found for any lead time for baseline0, positive skill is obtained for short forecast periods of baseline1 and prototype, particularly over central Europe; prototype-GECCO2 outperforms all other systems over Poland for lead years 2–5 (Figure 9). Hindcast skill is highest for autumn and lowest for summer over central Europe (not shown), indicating a clear dependency of the predictive skill on season (Moemken et al. 2016).

6. Discussion and conclusions

MiKlip is well poised to deliver its decadal prediction and evaluation systems to the German meteorological service 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),

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 (Johann Jungclaus, 2014, personal communication). These problems have now been overcome, and 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 ten.

(4) When starting MiKlip we underestimated the difficulty of implementing suggested model improvements. Any modification to the climate model itself requires a re-tuning (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 – 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 multi-model 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 sub-projects that develop such a dialog 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.

7. Acknowledgements

MiKlip is funded by the German Federal Ministry for Education and Research (BMBF). 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.

8. References

Andersson, A., M. Ziese, F. Dietzsch, M. Schröder, A. Becker, and K. Schamm, 2016a: HOAPS/GPCC global daily precipitation data record with uncertainty estimates using satellite and gauge based observations at 1.0°, 10.5676/DWD_CDC/HOGP_100/V001.

——, 2016b: HOAPS/GPCC global daily precipitation data record with uncertainty estimates using satellite and gauge based observations at 2.5°, 10.5676/DWD_CDC/HOGP_250/V001.

Balmaseda, M. A., K. Mogensen, and A. T. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quart. J. Roy. Meteor. Soc.*, **139**, 1132-1161.

Bellucci, A., R. Haarsma, N. Bellouin, B. Booth, C. Cagnazzo, B. van den Hurk, N. Keenlyside, T. Koenigk, F. Massonnet, S. Materia, and M. Weiss, 2015a: Advancements in decadal climate predictability: The role of nonoceanic drivers. *Rev. Geophys.*, **53**, 165-202, 10.1002/2014rg000473.

Bellucci, A., R. Haarsma, S. Gualdi, P. J. Athanasiadis, M. Caian, C. Cassou, E. Fernandez, A. Germe, J. Jungclaus, J. Kröger, D. Matei, W. Müller, H. Pohlmann, D. Salas y Melia, E. Sanchez, D. Smith, L. Terray, K. Wyser, and S. Yang, 2015b: An assessment of a multi-model ensemble of decadal climate predictions. *Climate Dyn.*, **44**, 2787-2806, 10.1007/s00382-014-2164-y.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.-Atmos.*, **111**, D12106, 10.1029/2005JD006548

Brune, S., L. Nerger, and J. Baehr, 2015: Assimilation of oceanic observations in a global coupled Earth system model with the SEIK filter. *Ocean Modell.*, **96**, 254-264, 10.1016/j.ocemod.2015.09.011.

Cunningham, S. A., Torsten Kanzow, Darren Rayner, Molly O. Baringer, William E. Johns, Jochem Marotzke, Hannah R. Longworth, Elizabeth M. Grant, Joël J.-M. Hirschi, Lisa M. Beal, C. S. Meinen, and H. L. Bryden, 2007: Temporal variability of the Atlantic meridional overturning circulation at 26.5°N. *Science*, **317**, 935-938, DOI: 10.1126/science.1141304.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553-597.

Du, H., F. J. Doblas-Reyes, J. García-Serrano, V. Guemas, Y. Soufflet, and B. Wouters, 2012: Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. *Climate Dyn.*, **39**, 2013-2023, 10.1007/s00382-011-1285-9.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen, 2013: Evaluation of Climate Models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to* *the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, and Coauthors, Eds., Cambridge University Press, 741-866.

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

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

Giorgetta, M. A., and Coauthors, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Sys.*, **5**, 572-597.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245-272.

Guemas, V., F. J. Doblas-Reyes, F. Lienert, Y. Soufflet, and H. Du, 2012: Identifying the causes of the poor decadal climate prediction skill over the North Pacific. *J. Geophys. Res.-Atmos.*, **117**, D20111, 10.1029/2012jd018004.

Hand, R., N. Keenlyside, N.-E. Omrani, and M. Latif, 2014: Simulated response to inter-annual SST variations in the Gulf Stream region. *Climate Dyn.*, **42**, 715-731, doi: 10.1007/s00382-013-1715-y.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *Int. J. Climatol.*, **34**, 623-642, 10.1002/joc.3711.

Haylock, M. R., N. Hofstra, A. M. G. K. Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *J. Geophys. Res.-Atmos.*, **113**, D20119, 10.1029/2008jd010201.

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. *Geophys. Res. Lett.*, **41**, 5167-5174, 10.1002/2014gl060420.

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

Ishii, M., and M. Kimoto, 2009: Reevaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections. *J. Oceanogr.*, **65**, 287-299.

Jacob, D., 2001: A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin. *Met. Atmos. Phys.*, **77**, 61-73.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.-Atmos.*, **117**, D05127, 10.1029/2011JD017139.

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

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. *Meteor. Zeitschr.*, 10.1127/metz/2015/0639.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-471.

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

Kharin, V. V., G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W. S. Lee, 2012: Statistical adjustment of decadal predictions in a changing climate. *Geophys. Res. Lett.*, **39**, L19705, 10.1029/2012gl052647.

Kirtman, B., S. B. Power, J. A. Adedoyin, G. J. Boer, R. Bojariu, I. Camilloni, F. J. Doblas-Reyes, A. M. Fiore, M. Kimoto, G. A. Meehl, M. Prather, A. Sarr, C. Schär, R. Sutton, G. J. v. Oldenborgh, G. Vecchi, and H. J. Wang, 2013: Near-term climate change: projections and predictability. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, and Coauthors, Eds., Cambridge University Press 953-1028.

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

Kotlarski, S., K. Keuler, O. B. Christensen, A. Colette, M. Déqué, A. Gobiet, K. Goergen, D. Jacob, D. Lüthi, E. van Meijgaard, G. Nikulin, C. Schär, C. Teichmann, R. Vautard, K. Warrach-Sagi, and V. Wulfmeyer, 2014: Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model Dev.*, **7**, 1297-1333, 10.5194/gmd-7-1297-2014.

Kröger, J., W. A. Müller, and J. S. von Storch, 2012: Impact of different ocean reanalyses on decadal climate prediction. *Climate Dyn.*, **39**, 795-810.

Kruschke, T., H. W. Rust, C. Kadow, W. A. Müller, H. Pohlmann, G. C. Leckebusch, and U. Ulbrich, 2015: Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms. *Meteor. Zeitschr.*, 10.1127/metz/2015/0641.

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate*, **14**, 1671-1676.

Lee, T., D. E. Waliser, J.-L. F. Li, F. W. Landerer, and M. M. Gierach, 2013: Evaluation of CMIP3 and CMIP5 wind stress climatology using satellite measurements and atmospheric reanalysis products. *J. Climate*, **26**, 5810-5826, 10.1175/jcli-d-12-00591.1.

Levitus, S., J. I. Antonov, T. P. Boyer, R. A. Locarnini, H. E. Garcia, and A. V. Mishonov, 2009: Global ocean heat content 1955-2008 in light of recently revealed instrumentation problems. *Geophys. Res. Lett.*, **36**, L07608, 10.1029/2008GL037155.

Li, H., T. Ilyina, W. A. Müller, and F. Sienz, 2016: Decadal predictions of the North Atlantic CO_2 uptake. *Nature Comm.*, **7**, 10.1038/ncomms11076.

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. *Mon. Wea. Rev.*, in press, doi: 10.1175/MWR-D-1115-0350.1171.

Matei, D., J. Baehr, J. H. Jungclaus, H. Haak, W. A. Müller, and J. Marotzke, 2012a: Multiyear prediction of monthly mean Atlantic meridional overturning circulation at 26.5ºN. *Science*, **335**, 76-79.

Matei, D., H. Pohlmann, J. Jungclaus, W. Müller, H. Haak, and J. Marotzke, 2012b: Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. *J. Climate*, **25**, 8502-8523.

Mauritsen, T., B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, U. Mikolajewicz, D. Notz, R. Pincus, H. Schmidt, and L. Tomassini, 2012: Tuning the climate of a global model. *J. Adv. Model. Earth Sys.*, **4**, M00A01, doi:10.1029/2012MS000154.

McCarthy, G., E. Frajka-Williams, W. E. Johns, M. O. Baringer, C. S. Meinen, H. L. Bryden, D. Rayner, A. Duchez, C. Roberts, and S. A. Cunningham, 2012: Observed interannual variability of the Atlantic meridional overturning circulation at 26.5 degrees N. *Geophys. Res. Lett.*, **39**, L19609, 10.1029/2012GL052933.

McGregor, S., A. Sen Gupta, and M. H. England, 2012: Constraining wind stress products with sea surface height observations and implications for Pacific Ocean sea level trend attribution. *J. Climate*, **25**, 8164-8176, 10.1175/jcli-d-12-00105.1.

Meehl, G. A., and Coauthors, 2014: Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteor. Soc.*, **95**, 243-267.

Mieruch, S., H. Feldmann, G. Schädler, C. J. Lenz, S. Kothe, and C. Kottmeier, 2014: The regional MiKlip decadal forecast ensemble for Europe: the added value of downscaling. *Geosci. Model Dev.*, **7**, 2983-2999, 10.5194/gmd-7-2983-2014.

Minobe, S., A. Kuwano-Yoshida, N. Komori, S. P. Xie, and R. J. Small, 2008: Influence of the Gulf Stream on the troposphere. *Nature*, **452**, 206-209.

Mochizuki, T., M. Ishii, M. Kimoto, Y. Chikamoto, M. Watanabe, T. Nozawa, T. T. Sakamoto, H. Shiogama, T. Awaji, N. Sugiura, T. Toyoda, S. Yasunaka, H. Tatebe, and M. Mori, 2010: Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 1833-1837.

Moemken, J., M. Reyers, B. Buldmann, and J. G. Pinto, 2016: Decadal predictability of regional scale wind speed and wind energy potentials over Central Europe. *Tellus Ser. A - Dyn. Meteor. Oceanogr.*, **68**, 29199, 10.3402/tellusa.v68.29199.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Müller, V., H. Pohlmann, H. Haak, D. Matei, J. Marotzke, W. A. Müller, and J. Baehr, 2016: Hindcast skill for the Atlantic meridional overturning circulation at 26.5°N within two MPI-ESM decadal climate prediction systems. *Climate Dyn.*, submitted.

Müller, W. A., J. Baehr, H. Haak, J. H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J. S. von Storch, and J. Marotzke, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.*, **39**, L22707, 10.1029/2012GL053326.

Müller, W. A., D. Matei, M. Bersch, J. H. Jungclaus, H. Haak, K. Lohmann, G. P. Compo, P. D. Sardeshmukh, and J. Marotzke, 2015: A twentieth-century reanalysis forced ocean model to reconstruct the North Atlantic climate variation during the 1920s. *Climate Dyn.*, **44**, 1935-1955.

Müller, W. A., H. Pohlmann, F. Sienz, and D. Smith, 2014: Decadal climate predictions for the period 1901–2010 with a coupled climate model. *Geophys. Res. Lett.*, **41**, 2100-2107, 10.1002/2014gl059259.

Murphy, A. H., 1988: Skill scores based on the mean-square error and their relationships to the correlation-coefficient. *Mon. Wea. Rev.*, **116**, 2417-2425.

Murphy, J. M., 1990: Assessment of the practical utility of extended range ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **116**, 89-125.

Paeth, H., A. Paxian, D. Sein, D. Jacob, H.-J. Panitz, M. Warscher, A. Fink, H. Kunstmann, M. Breil, T. Engel, A. Krause, J. Toedter, and B. Ahrens, 2016: Decadal and multi-year predictability of the West African monsoon and the role of dynamical downscaling. *J. Climate*, submitted.

Pham, D. T., J. Verron, and M. C. Roubaud, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Sys.*, **16**, 323-340.

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

Pohlmann, H., J. Kröger, R. J. Greatbatch, and W. A. Müller, 2016: Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific. *Geophys. Res. Lett.*, submitted.

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

Polkova, I., A. Köhl, and D. Stammer, 2014: Impact of initialization procedures on the predictive skill of a coupled ocean-atmosphere model. *Climate Dyn.*, **42**, 3151-3169, 10.1007/s00382-013-1969-4.

Räisänen, J., 2007: How reliable are climate models? *Tellus Ser. A - Dyn. Meteor. Oceanogr.*, **59**, 2-29, 10.1111/j.1600-0870.2006.00211.x.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.-Atmos.*, **108**, 4407, 10.1029/2002jd002670.

Reyers, M., J. G. Pinto, and J. Moemken, 2015: Statistical-dynamical downscaling for wind energy potentials: evaluation and applications to decadal hindcasts and climate change projections. *Int. J. Climatol.*, **35**, 229-244, 10.1002/joc.3975.

Robson, J., R. Sutton, K. Lohmann, D. Smith, and M. D. Palmer, 2012a: Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. *J. Climate*, **25**, 4116-4134, 10.1175/jcli-d-11-00443.1.

Robson, J. I., R. T. Sutton, and D. M. Smith, 2012b: Initialized decadal predictions of the rapid warming of the North Atlantic Ocean in the mid 1990s. *Geophys. Res. Lett.*, **39**, L19713, 10.1029/2012gl053370.

Rockel, B., A. Will, and A. Hense, 2008: The regional climate model COSMO-CLM (CCLM). *Meteor. Zeitschr.*, **17**, 347-348, 10.1127/0941-2948/2008/0309.

Romanova, V., and A. Hense, 2015: Anomaly transform methods based on total energy and ocean heat content norms for generating ocean dynamic disturbances for ensemble climate forecasts. *Climate Dyn.*, 10.1007/s00382-015-2567-4.

Scaife, A. A., and Coauthors, 2014a: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514-2519, 10.1002/2014gl059637.

Scaife, A. A., M. Athanassiadou, M. Andrews, A. Arribas, M. Baldwin, N. Dunstone, J. Knight, C. MacLachlan, E. Manzini, W. A. Mueller, H. Pohlmann, D. Smith, T. Stockdale, and A. Williams, 2014b: Predictability of the quasi-biennial oscillation and its northern winter teleconnection on seasonal to decadal timescales. *Geophys. Res. Lett.*, **41**, 1752-1758, 10.1002/2013gl059160.

Schamm, K., M. Ziese, A. Becker, P. Finger, A. Meyer-Christoffer, U. Schneider, M. Schröder, and P. Stender, 2014: Global gridded precipitation over land: a description of the new GPCC first guess daily product. *Earth Syst. Sci. Data*, **6**, 49-60, doi:10.5194/essd-6-49-2014.

Sienz, F., H. Pohlmann, and W. A. Müller, 2016: Ensemble size impact on the decadal predictive skill assessement. *Meteor. Zeitschr.*, 10.1127/metz/2016/0670.

Skamarock, W. C., and J. B. Klemp, 2008: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J. Comput. Phys.*, **227**, 3465-3485, 10.1016/j.jcp.2007.01.037.

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

Smith, D. M., R. Eade, and H. Pohlmann, 2013a: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Climate Dyn.*, **41**, 3325-3338.

Smith, D. M., and Coauthors, 2013b: Real-time multi-model decadal climate predictions. *Climate Dyn.*, **41**, 2875-2888.

Stenchikov, G. L., I. Kirchner, A. Robock, H. F. Graf, J. C. Antuna, R. G. Grainger, A. Lambert, and L. Thomason, 1998: Radiative forcing from the 1991 Mount Pinatubo volcanic eruption. *J. Geophys. Res.-Atmos.*, **103**, 13837-13857, 10.1029/98jd00693.

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, M. Schröder, A. Mazurkiewicz, and A. Hense, 2015: Revealing skill of the MiKliP decadal prediction systems by three-dimensional probabilistic evaluation. *Meteor. Zeitschr.*, DOI: 10.1127/metz/2015/0606.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 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. *Geophys. Res. Lett.*, **42**, 6454-6461, 10.1002/2015GL064833.

Timmreck, C., H. Pohlmann, C. Kadow, and S. Illing, 2016: The impact of stratospheric volcanic aerosol on decadal scale predictability. *Geophys. Res. Lett.*, **43**, 834–842, 10.1002/2015GL067431.

Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961-3012.

Wei, M. Z., Z. Toth, R. Wobus, Y. J. Zhu, C. H. Bishop, and X. G. Wang, 2006: Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus Ser. A - Dyn. Meteor. Oceanogr.*, **58**, 28-44.

WMO, 2011: Climate Knowledge for Action - A Global Framework for Climate Services. *WMO-No. 1065*, 247 pp.

Zhang, L., H. Dobslaw, C. Dahle, I. Sasgen, and M. Thomas, 2015: Validation of MPI-ESM decadal hindcast experiments with terrestrial water storage variations as observed by the GRACE satellite mission. *Meteor. Zeitschr.*, DOI: 10.1127/metz/2015/0596.

9. Tables

	baseline0	baseline1	prototype
Models	MPI-ESM-LR MPI-ESM-MR	MPI-ESM-LR MPI-ESM-MR	MPI-ESM-LR
Initialization ocean	3-D T/S anomalies from MPIOM forced with NCEP/NCAR reanalysis	3-D T/S anomalies from ORAS4	3-D T/S (full field) from GECCO2 and from ORAS4
Initialization atmosphere	Assimilation run	ERA40 and ERA-Interim; vorticity, divergence, log p, T; full-field	ERA40 and ERA-Interim; vorticity, divergence, log p, T; full-field
Ensemble size	LR: 3 (10) MR: 3	LR: 10 MR: 5	30 (15 each with initialization from GECCO2 and ORAS4)
Start years	LR: 1961–2013; yearly for 3 realizations 1961–2000: five-yearly for 10 realizations 2001–2013: yearly for 10 realizations MR: 1961–2000: five-yearly 2001–2012: yearly	LR: 1961–2014: yearly MR: 1961–2013: yearly	1961–2014: yearly

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.

10. Figure captions

Figure 1. Evolution during the MiKlip project of the ensemble-mean hindcast skill (anomaly correlation, left column; with anomaly correlation of historical simulations subtracted, right column) of surface air temperature averaged over the lead years 2–5 in the low-resolution model version MPI-ESM-LR. Observations are from HadCRUT4 (Jones et al. 2012); the period is 1961–2012. (a) and (d), baseline0; (b) and (e), baseline1; (c) and (f), prototype. Hindcast ensemble size is three for baseline0, ten for baseline1, and thirty for prototype; historical ensemble size is three for baseline0, ten for baseline1, and fifteen 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).

Figure 2. SST index for the North Atlantic subpolar gyre (40–60°N, 0–60° W) from 1960–2020. Shown is the five-year running mean for the observations (HadISST1.1; Rayner et al. 2003; solid black) and the ensemble mean over ten realizations of MPI-ESM-LR historical simulations extended with the RCP4.5 scenario (dashed black). Further shown is the time mean over the first five prediction years for each start year, for the ensemble-mean of baseline1 MPI-ESM-LR hindcasts and predictions (solid red; last start year 2015). The whiskers show the range (minimum to maximum) of the ensemble.

Figure 3. (a) Anomalies of the global-mean upper-ocean (0–700m) 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 = 10^6 m^3s^{-1}) from GECCO2 (red) and the reference run (black) with observations from 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).

Figure 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 (JJA) 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 CRU TS3.10 (Harris et al. 2014) for surface temperature over land. From Sienz et al. (2016); reproduced with permission (www.schweizerbart.de).

Figure 5. Correlation (a) and RMSE (b) against HadISST1.1 (Rayner et al. 2003) of the SST index for the North Atlantic subpolar gyre (40–60°N, 0–60° W), against lead year and for all MiKlip generations and the historical simulations. Baseline0 and baseline1 outperform the historical simulations for almost all lead years. The prototype (pr) hindcasts sometimes provide the highest skill, as is the case for most lead years when using GECCO2 and correlation skill (a). But sometimes the prototype hindcasts provide the lowest skill, especially for early lead years, as is the case when using ORAS4 and correlation skill (a) as well as when using either ORAS4 or GECCO2 and the RMSE (b).

Figure 6. Hindcast skill for whether wind storms related to intense extra-tropical 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 1000 km per period October–March), 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 ($\alpha < 5$ %) as black dots, areas of strong inconsistencies between ERA40 and ERA-Interim are masked out (grey). From Kruschke et al. (2015); reproduced with permission (www.schweizerbart.de).

Figure 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; (b) lead years 2–5. The numbers indicate the anomaly correlation coefficient between the hindcasts and the observations over the whole period. The grey-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.

Figure 8. Ensemble-mean hindcast skill over Europe for near-surface air temperature for lead time 2–5 years, 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 baseline1; (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.

Figure 9. MSESS for wind-energy output for years 1–4 (upper row), years 2–5 (middle row), and years 6–9 (lower row) of the baseline0 3-member ensemble mean (left column), baseline1 10-member ensemble mean (second column), prototype ORA-S4 10-member sub-ensemble mean (third column), and prototype GECCO2 10-member sub-ensemble mean (last column). Reference prediction is the ensemble mean of the uninitialized historical runs, using 3 realizations for baseline 0 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).

11. Figures

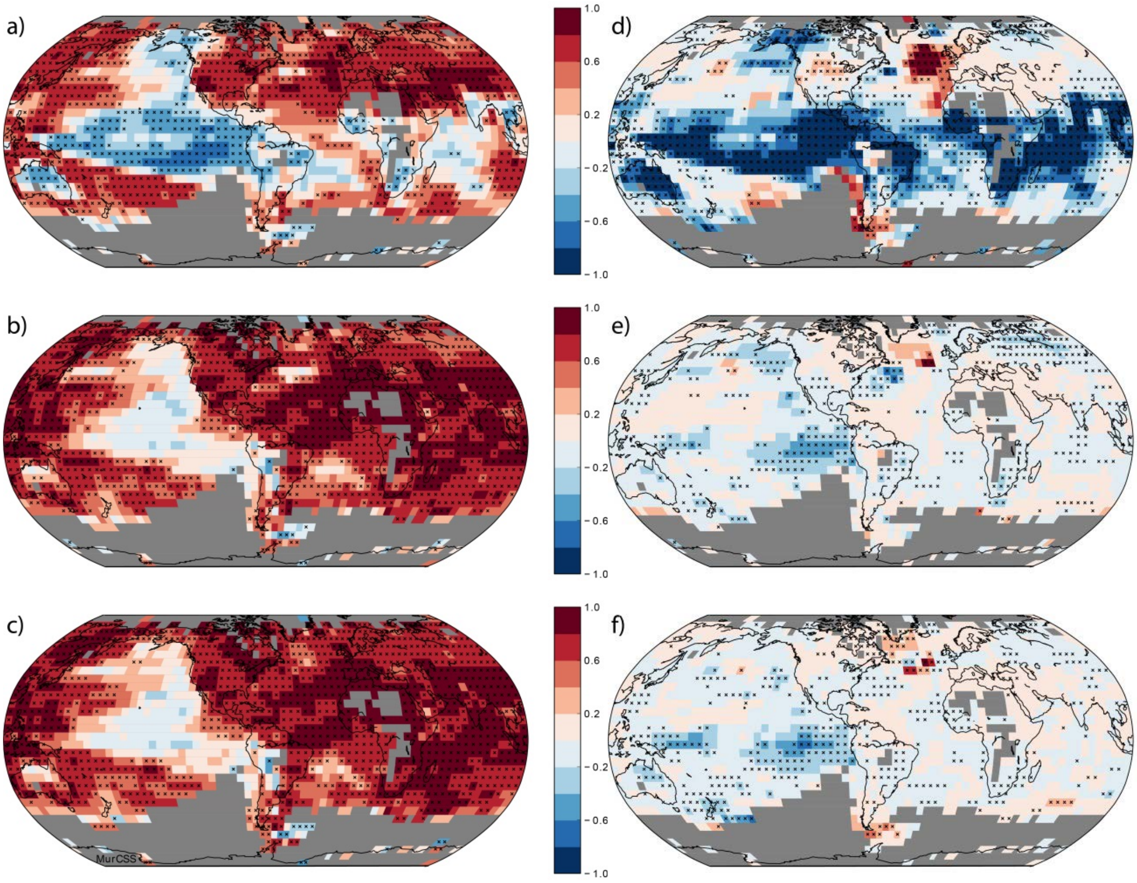

Figure 1. Evolution during the MiKlip project of the ensemble-mean hindcast skill (anomaly correlation, left column; with anomaly correlation of historical simulations subtracted, right column) of surface air temperature averaged over the lead years 2–5 in the low-resolution model version MPI-ESM-LR. Observations are from HadCRUT4 (Jones et al. 2012); the period is 1961–2012. (a) and (d), baseline0; (b) and (e), baseline1; (c) and (f), prototype. Hindcast ensemble size is three for baseline0, ten for baseline1, and thirty for prototype; historical ensemble size is three for baseline0, ten for baseline1, and fifteen 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).

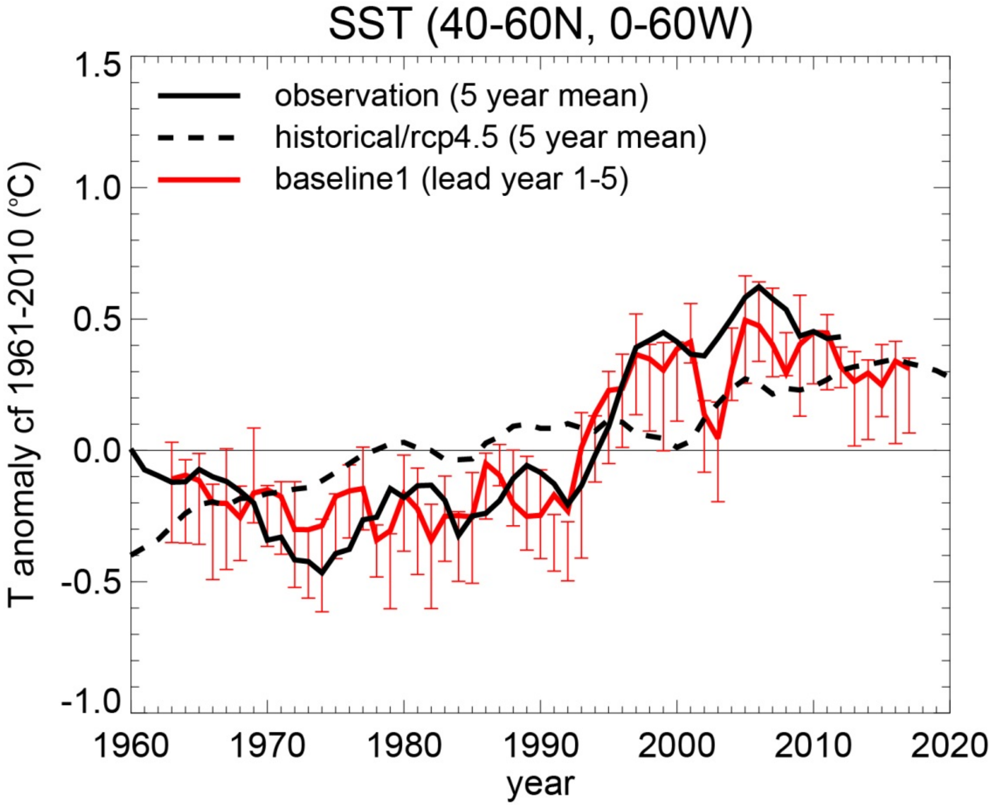

Figure 2. SST index for the North Atlantic subpolar gyre (40–60°N, 0–60° W) from 1960–2020. Shown is the five-year running mean for the observations (HadISST1.1; Rayner et al. 2003; solid black) and the ensemble mean over ten realizations of MPI-ESM-LR historical simulations extended with the RCP4.5 scenario (dashed black). Further shown is the time mean over the first five prediction years for each start year, for the ensemble-mean of baseline1 MPI-ESM-LR hindcasts and predictions (solid red; last start year 2015). The whiskers show the range (minimum to maximum) of the ensemble.

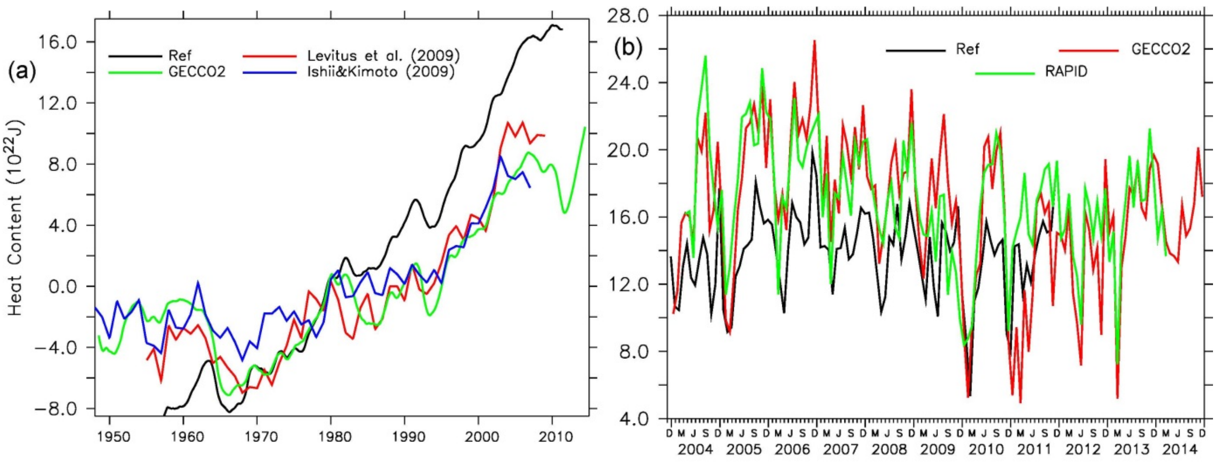

Figure 3. (a) Anomalies of the global-mean upper-ocean (0–700m) heat content from the reference run (Ref, no assimilation) and GECCO2 in comparison to the estimates from Levitus et al. (2009) and Ishii and Kimoto (2009). (b) Comparison of the monthly-mean meridional overturning at 26.5°N in Sv (1 Sv = 10^6 m^3s^{-1}) from GECCO2 (red) and the reference run (black) with observations from 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).

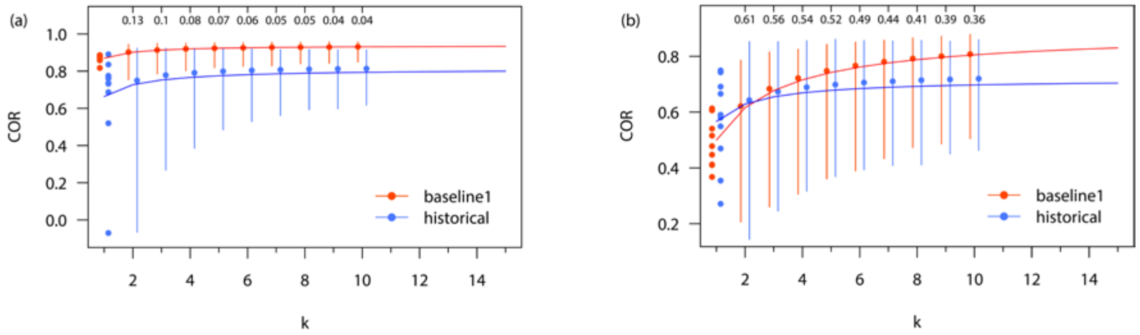

Figure 4. Correlation with observations for (a) annual mean North Atlantic (20–60°N, 10–80° W) SST and (b) central European (40–45.5°N, 10–30° E) summer (JJA) surface temperature, for baseline1 hindcasts averaged over lead years 2–5 (red) and historical runs (blue). Shown is the dependence of the correlation 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 CRU TS3.10 (Harris et al. 2014) for surface temperature over land. From Sienz et al. (2016); reproduced with permission (www.schweizerbart.de).

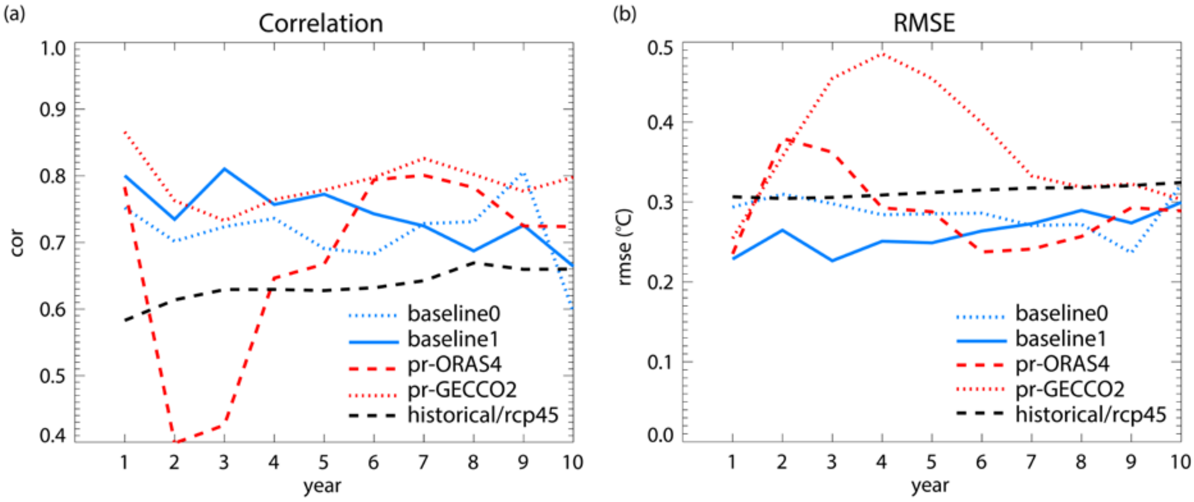

Figure 5. Correlation (a) and RMSE (b) against HadISST1.1 (Rayner et al. 2003) of the SST index for the North Atlantic subpolar gyre (40–60°N, 0–60° W), against lead year and for all MiKlip generations and the historical simulations. Baseline0 and baseline1 outperform the historical simulations for almost all lead years. The prototype (pr) hindcasts sometimes provide the highest skill, as is the case for most lead years when using GECCO2 and correlation skill (a). But sometimes the prototype hindcasts provide the lowest skill, especially for early lead years, as is the case when using ORAS4 and correlation skill (a) as well as when using either ORAS4 or GECCO2 and the RMSE (b).

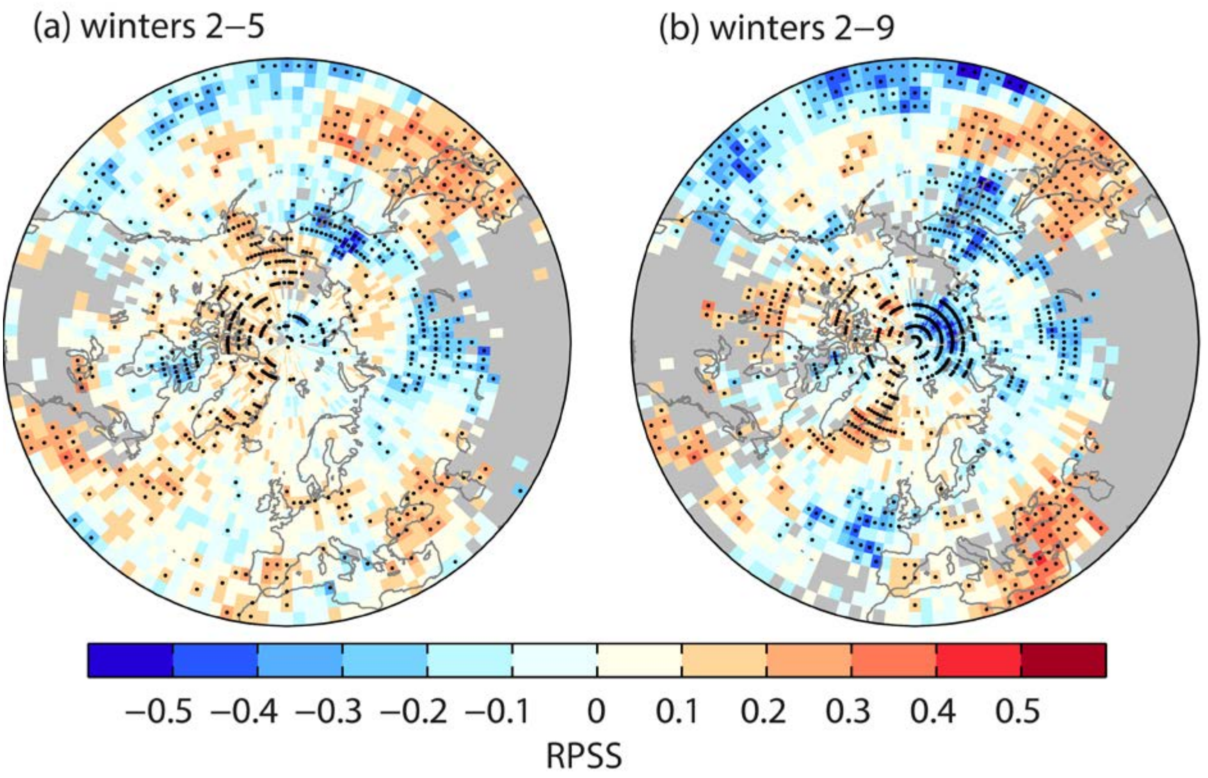

Figure 6. Hindcast skill for whether wind storms related to intense extra-tropical cyclones occur at a frequency that is either below normal, normal, or above normal, for the Northern-Hemisphere extended winter season (number of tracks within 1000 km per period October–March), in a 29-member ensemble 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 ($\alpha < 5$ %) as black dots, areas of strong inconsistencies between ERA40 and ERA-Interim are masked out (grey). From Kruschke et al. (2015); reproduced with permission (www.schweizerbart.de).

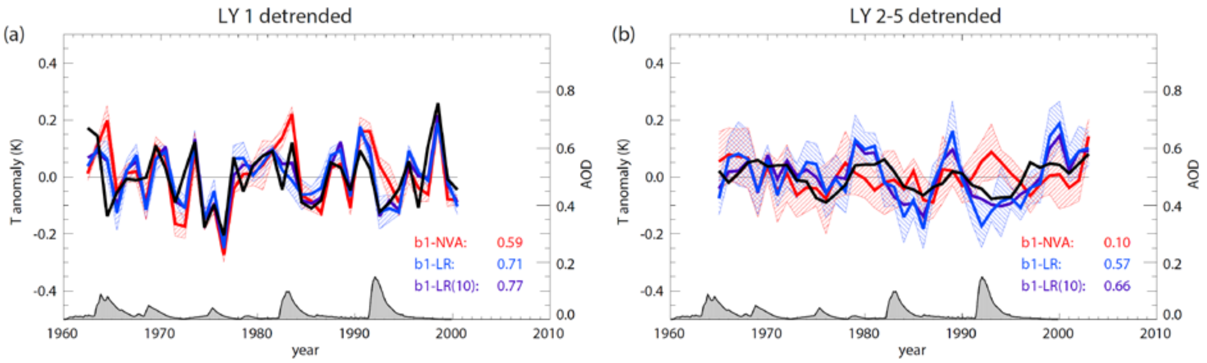

Figure 7. Detrended time series of hindcast and observed globally averaged surface temperature from 1962 to 2004; the figure shows anomalies relative to the mean over this period for HadCRUT3v observations (Brohan et al. 2006) (black), baseline1-LR (blue), and baseline1-LR without volcanic eruptions (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; (b) lead years 2–5. The numbers indicate the anomaly correlation coefficient between the hindcasts and the observations over the whole period. The grey-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.

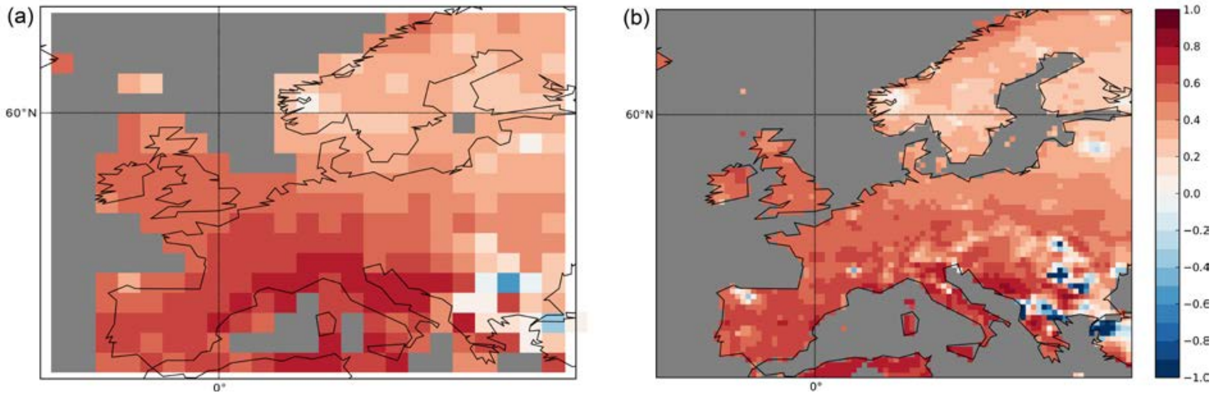

Figure 8. Ensemble-mean hindcast skill over Europe for near-surface air temperature for lead time 2–5 years, starting yearly from 1961 to 2004. Skill score is MSESS evaluated against E-OBS (Haylock et al. 2008), with climatology as the reference forecast; positive values indicate better skill than climatology. (a) MPI-ESM-LR baseline1; (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.

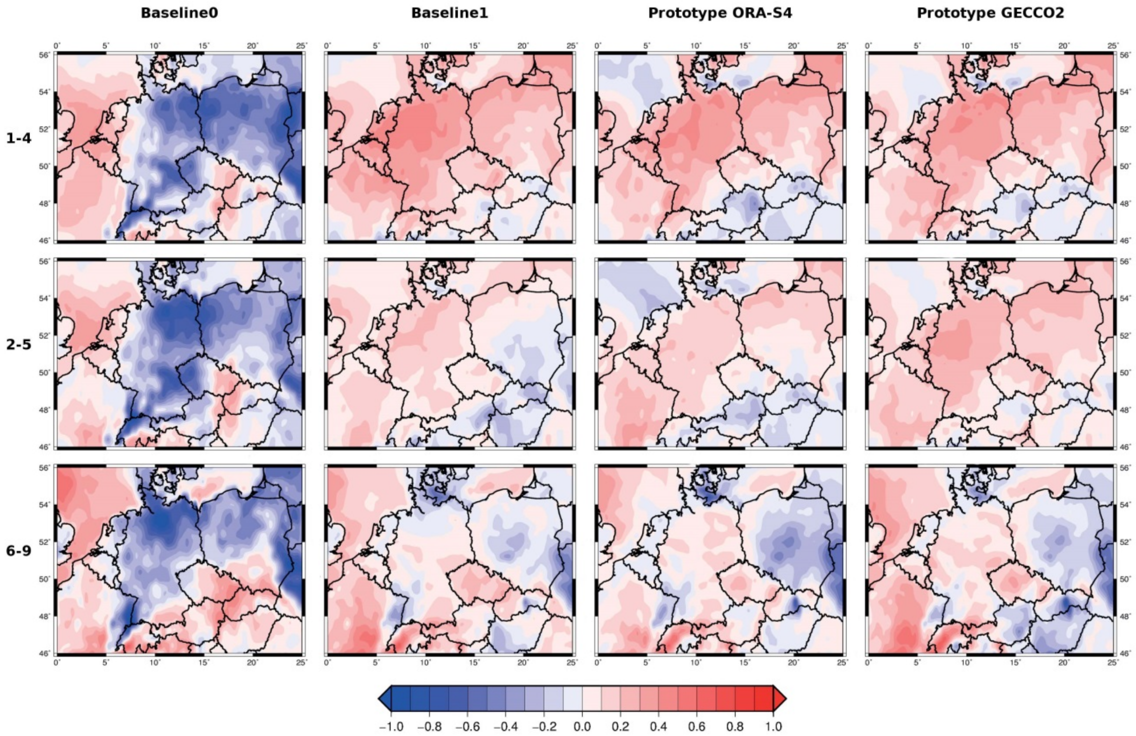

Figure 9. MSESS for wind-energy output for years 1–4 (upper row), years 2–5 (middle row), and years 6–9 (lower row) of the baseline0 3-member ensemble mean (left column), baseline1 10-member ensemble mean (second column), prototype ORA-S4 10-member sub-ensemble mean (third column), and prototype GECCO2 10-member sub-ensemble mean (last column). Reference prediction is the ensemble mean of the uninitialized historical runs, using 3 realizations for baseline 0 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).