Meteorologische Zeitschrift, Vol. **25**, No. 6, 709–720 (published online December 1, 2016)
© 2016 The authors

Bias and Drift of the Medium-Range Decadal Climate Prediction System (MiKlip) validated by European Radiosonde Data

Margit Pattantyús-Ábrahám[1*], Christopher Kadow[2], Sebastian Illing[2], Wolfgang A. Müller[3], Holger Pohlmann[3] and Wolfgang Steinbrecht[1]

[1]Deutscher Wetterdienst, Meteorological Observatory Hohenpeissenberg, Germany
[2]Institut für Meteorologie, Freie Universität Berlin, Berlin, Germany
[3]Max-Planck-Institute for Meteorology, Hamburg, Germany

(Manuscript received April 26, 2016; in revised form August 29, 2016; accepted October 3, 2016)

Abstract

Quality controlled and homogenized radiosonde observations have been used to validate decadal hindcasts of the MPI-Earth-System-Model for Europe (excl. some Eastern European countries). Simulated temperatures have a cold bias of 1 to 4 K, increasing with height throughout the free troposphere over Europe. This implies that the simulated troposphere is less stable than observed by the radiosondes over Europe. Simulated relative humidity is 10 to 40 % higher than observed. Part of the humidity bias, 10 to 25 % relative humidity, is due to the simulated lower temperature, but the remainder indicates that modelled water vapour pressure is too high in the free troposphere above Europe. After full-field initialization with oceanic state, the atmospheric temperature bias changes over the first couple of years, with a relaxation time of 5 years near the surface (850 hPa) and less than 1 year near the tropopause (200 hPa). Anomaly correlations, mean-square error and logarithmic ensemble spread score indicate small improvements in hindcasted tropospheric temperatures over Europe when going from ocean anomaly initialisation to ocean anomaly initialisation plus full field atmospheric initialisation, and then to full field ocean initialisation plus full field atmospheric initialisation. In the stratosphere, these changes have little effect. For humidity, correlations and skill scores are much poorer, and little can be said about changes over Europe due to different initializations.

Keywords: Decadal prediction, MiKlip, evaluation, radiosonde, bias

1 Introduction

Projections of future climate have been made by many Earth-system-models, e.g. those coordinated by the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al., 2012; Meehl et al., 2014). Typically, the effects of model boundary conditions, that is impacts of the expected atmospheric composition (aerosol and greenhouse gases), and natural forcings have been analysed for time scales of decades to centuries (Smith et al., 2013; Meehl et al., 2007). Over the last 10 to 15 years, scientific and economic interest has turned to shorter term, decadal, climate predictions (Pohlmann et al., 2009; Smith et al., 2007; Keenlyside et al., 2008; Marotzke et al., 2016). On decadal time-scales not only boundary conditions are important, but also the initial atmospheric and oceanic state. Decadal climate projections are of particular scientific interest. If skilful enough, they could be useful for decisions on administrative and political time scales, such as infrastructure planning and adaptation to expected climate variations. An appropriate model system with good enough resolution in atmosphere and ocean, with realistic internal variations, and careful model initialization with observational data are key requirements for skilful climate predictions on the decadal scale.

The German decadal climate prediction program (MiKlip, Mittelfristige Klimaprognose) has contributed to extensive testing and further development of the Max-Planck-Institute Earth System Model (MPI-ESM) (Jungclaus et al., 2013; Stevens et al., 2013; Marotzke et al., 2016). A major part of MiKlip was to investigate the impact of different initialization techniques (Pohlmann et al., 2013). This can be done by comparing the performance of hindcasts with the actual evolution of the observed atmosphere (Hawkins et al., 2014; Stolzenberger et al., 2015). A challenge in this context are systematic differences between the hindcasts and the actual observations, i.e. bias and drift (Hawkins et al., 2014; Eade et al., 2014).

The aim of this paper is to analyse MiKlip hindcasts in the free atmosphere over Europe, up to the middle stratosphere, by comparison to observational data from homogenized radiosonde measurements (Pattantyús-Ábraháam and Steinbrecht, 2015). This allows quantification of model biases and drifts, as well as evaluation of hindcast skills. The main focus is on the com-

[*]Corresponding author: Margit Pattantyús-Ábrahám, Deutscher Wetterdienst, Meteorological Observatory Hohenpeissenberg, Albin-Schweiger-Weg 10, 82383 Hohenpeissenberg, Germany, e-mail: Margit.Pattantyus-Abraham@dwd.de

DOI 10.1127/metz/2016/0803

Table 1: Overview of the MiKlip experiments.

Experiment name	Historical-LR	Baseline 0-LR	Baseline 1-LR	Baseline 1-MR	Prototype-LR
Resolution	ocean: 1.5°, 40L; atmosphere T63L47 up to 0.1 hPa	ocean: 1.5°, 40L; atmosphere T63L47 up to 0.1 hPa	ocean: 1.5°, 40L; atmosphere T63L47 up to 0.1 hPa	ocean: 0.4°, 40L; atmosphere T63L95 up to 0.1 hPa	ocean: 1.5°, 40L; atmosphere T63L47 up to 0.1 hPa
Initialisation	no initialisation	yearly, lagged 1-day of atmosphere and ocean	yearly, lagged 1-day of atmosphere and ocean	yearly, lagged 1-day of atmosphere and ocean	yearly, lagged 1-day of atmosphere and ocean
Ocean initialisation	no initialisation	3D temperatures and salinity anomalies from MPI ocean model forced by NCEP reanalysis	anomalies from ORA-S4	anomalies from ORA-S4	full field: temperature and salinity from ORA-S4m, GECCO2
Atmosphere initialisation	no initialisation	no nudging	full field: vorticity, divergence temperature and pressure (log) from ERA	full field: vorticity, divergence temperature and pressure (log) from ERA	full field: vorticity, divergence temperature and pressure (log) from ERA
Ensemble size	15	3 (10 for every 5th year)	10	5	2 × 15
Years	1850–2014	1961–2013	1961–2014	1961–2013	1961–2014
Abbreviation	H-LR	B0-LR	B1-LR	B1-MR	Pr-LR

parison of hindcasted and observed profiles of temperature and relative humidity over Europe. Among other things, these profiles provide key information on the vertical stability of the atmosphere and on severe weather indicators. For quantitative intercomparison of predictive skills of the MPI-ESM and its diverse initialisation techniques (Goddard et al., 2013), the Mean Square Error Skill Score (MSESS), its decomposition, and the Logarithmic Ensemble Spread Score (LESS) (Kadow et al., 2015) are used. These scores are implemented in the 'MurCSS' plug-in of the MiKlip central evaluation system (Illing et al., 2014).

The paper is organized as follows: The next section explains the different hindcast experiments. It also describes the European radiosonde data used for validation. Section 3 is devoted to comparisons and evaluation of model performance. Conclusions are presented in Section 4.

2 Data

2.1 Design of the decadal prediction experiments

Hindcasts in this study come from the Max-Planck-Institute's earth system model (MPI-ESM) (Stevens et al., 2013; Pohlmann et al., 2013; Müller et al., 2012). The ocean component of this coupled atmosphere-ocean system is provided by the Max-Planck-Institute ocean model (MPI-OM) (Jungclaus et al., 2013), and ECHAM6 is the atmospheric component (Giorgetta et al., 2013). Hindcasts of the coupled MPI-ESM model have contributed to the fifth Coupled Model Intercomparison Project (CMIP5, Doblas-Reyes et al. (2013)).

Within MiKlip, different initialization techniques were tested for both ocean and atmosphere. Table 1 provides an overview of these experiments, including the resolution used by different parts of the earth system model. All hindcast experiments (B0, B1, Pr) were initialized and re-started for every model-year. These restarts included several members, forming a new ensemble for each start-year. All ensemble members contain 10 model-years of data. Baseline 0 is the MPI-ESM coupled model as in CMIP5. In the baseline 0 low resolution (B0-LR) hindcast, the ocean was initialized by nudging to 3-dimensional temperature and salinity anomaly data from the MPI Ocean Model forced by NCEP reanalysis (Kalnay et al., 1996). No initialisation was applied to the model atmosphere (Marotzke et al., 2016). Baseline 1, low and mixed resolution runs (B1-LR and B1-MR, respectively) used ocean anomaly initialisation from the Ocean Reanalysis System 4 (ORA-S4) data set (Balmaseda et al., 2013). In addition, full-field initialisation was used for the atmosphere, based on vorticity, divergence, temperature and log pressure data from the European Center for Medium-range Weather Forecasts (ECWMF) reanalyses ERA-40 (Uppala et al., 2005) and ERA-Interim (Dee et al., 2011). The "prototype" low-resolution experiment (Pr-LR) applied the same reanalysis fields as B1-LR for atmospheric initialisation, but used full-field temperature and salinity initialisation for the ocean, based on ORA-S4

(Balmaseda et al., 2013) and GECCO2 data (Köhl, 2015). In this paper, however, only the Pr-LR hindcasts with ORA-S4 initialisation were used to focus solely on the effect of full-field initialisation. B0-LR had 3 ensemble members (10 members every fifth starting year), B1-LR had 10 members, B1-MR 5 members, and Pr-LR 15 members with ORA-S4 initialisation (plus 15 members with GECCO2 initialisation which were not analysed here).

To assess the added value of initialisation, uninitialised (historical boundary condition) experiments were used as well. These historical (H-LR) hindcasts use the same model configuration as the initialised runs, but randomly sampled initialized on the 1st of January 1850. In all runs, historical natural forcings (earth orbit, solar variability, natural tropospheric and volcanic stratospheric aerosol) and anthropogenic forcings (greenhouse gases, ozone, anthropogenic sulfate aerosol, land-use changes) are applied until 2015 (Giorgetta et al., 2013).

2.2 Homogenized radiosonde observations

Radiosonde (RS) data from operational upper-air stations serve as the reference for our validation. Since the main goal of the MiKlip project is to provide decadal climate prediction for Central Europe, and the radiosonde network is most dense and robust in Europe, our focus is on the European region. The RS data were collected from the archive maintained by Deutscher Wetterdienst and from the International Geophysical Radiosonde Archive (Durre et al., 2006).

Because inhomogeneities in radiosonde time series are a common problem, homogenization of the RS temperature data was done first, both for Gemany (Pattantyús-Ábraháam and Steinbrecht, 2015), and, following the same methodology, for other European stations. Since our homogenization has not yet been extended to radiosondes from the former Soviet Union and a few other Eastern European countries, these are not part of the current validation. See Fig. 1 for the European radiosonde stations used. About 50 RS stations were selected for the validation. Selection criteria were: 1) at least 30 years of data; 2) two or more soundings per day (day-time and night-time); 3) no multi-year gaps in the data; 4) uniform coverage for Europe. Temperature data were taken at 11 standard pressure levels: 850 hPa, 700 hPa, 500 hPa, 400 hPa, 300 hPa, 250 hPa, 200 hPa, 150 hPa, 100 hPa, 70 hPa, 50 hPa, and for the time period from 1960–2013. Accuracy of the homogenized radiosonde monthly mean data is better than 0.2 K in the troposphere and better than 0.4 K in the lower stratosphere.

Relative humidity data from RS soundings were used only in the lower troposphere (850, 700, 500 hPa), and only after the mid 1990s, when radiosondes with good humidity sensors (e.g. Vaisala RS80, RS92) became routine. Before about 1990, and at higher levels, RS humidity data have large errors and uncertainties, including instrument dependent variance and bias (Elliott and Gaffen, 1991; Miloshevich et al., 2004; Dai et al., 2011). These low quality RS humidity records were not used here.

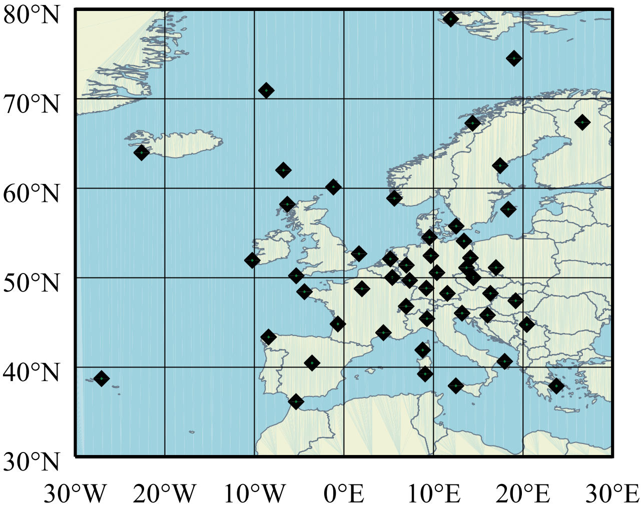

Figure 1: Radiosonde stations used for the validation.

For severe weather probability analyses, daily data were used. For bias and skill score comparisons, monthly means were used.

3 Results

3.1 Model bias

Bias is the systematic difference between model-simulated and radiosonde-observed quantities at the same pressure level (model minus observation). Figure 2a shows the vertical profile of the lead-year 2 annual mean temperature bias of the Pr-LR experiments for five European regions. The inset shows how the bias profile over Germany varies between the different model experiments. Generally, the model hindcasts have a negative temperature bias, i.e. hindcasted temperatures are up to 4 K lower than observed by the radiosondes. The only exception is the B1-MR experiment, which uses a higher vertical resolution, and gives higher temperatures than observed near the 100 hPa pressure level in the lower stratosphere. Generally, the seasonal variation of these biases in the troposphere is rather small and decreases with the height (and is not shown here). Most months of the year show similar bias, although the smallest negative biases (0.5 K warmer than in the annual mean) do occur in April and November. In July and August the negative model bias has a peak about 0.7 to 1 K colder than the annual mean bias below the tropopause (not shown).

The mean vertical bias profiles all show a similar pattern (see Figure 2a): The bias between hindcasts and observations increases from around 1 K near the surface to 2 to 4 K near the 200 hPa level, which is near the tropopause. The increase of the cold bias with altitude

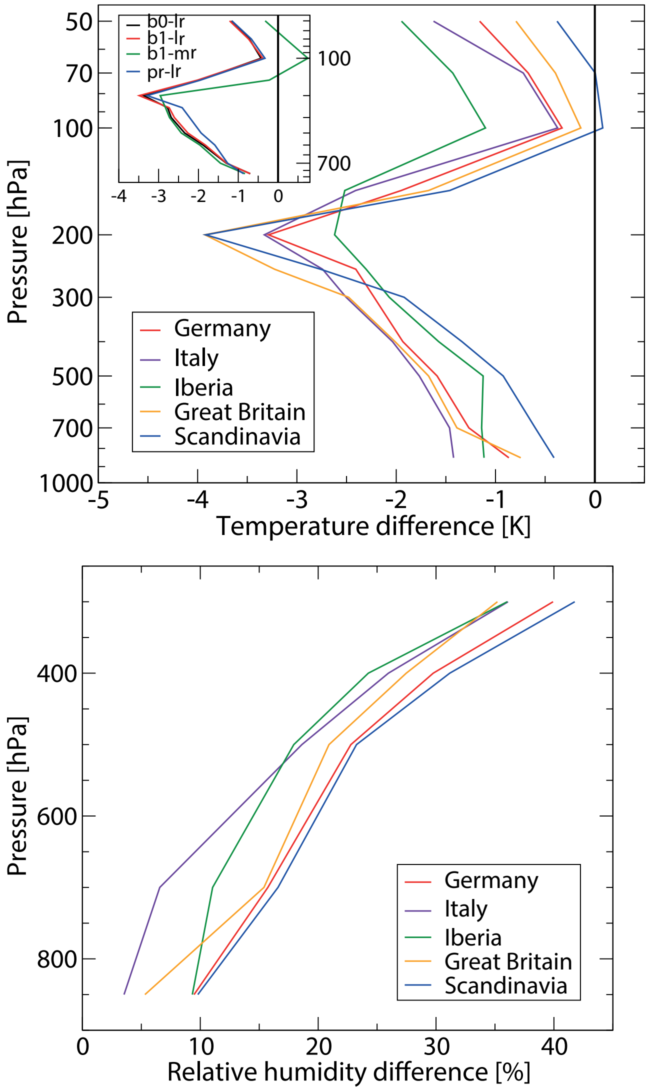

Figure 2: Annual mean bias in lead year 2 of the Prototype experiment (Pr-LR) over five European regions. Top: temperature, bottom: relative humidity. Inset in the top figure shows the mean vertical bias profile for the different experiments over Germany.

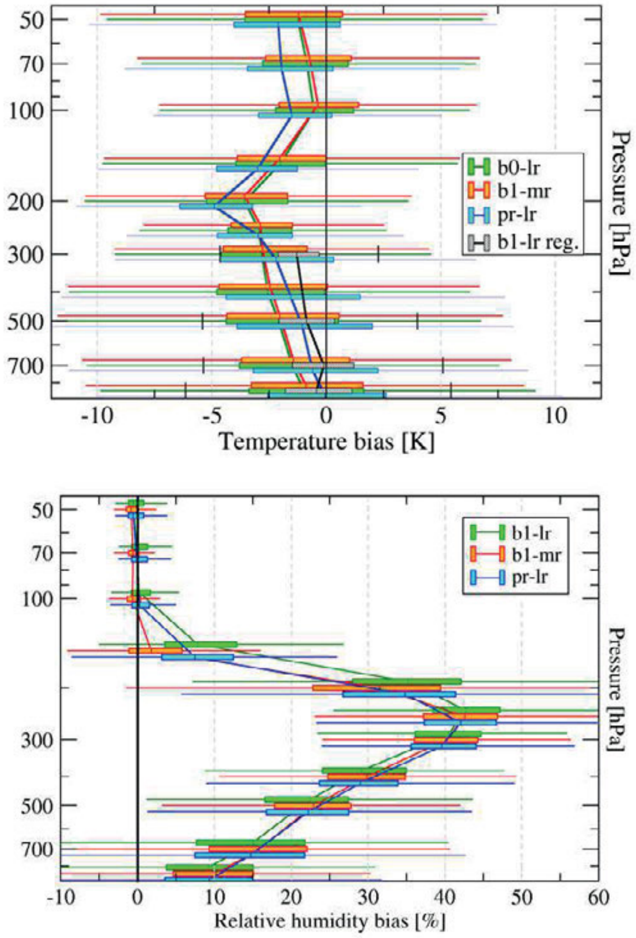

Figure 3: Box-Whisker (minimum / 1st quartile / median / 3rd quartile / maximum) plots of the bias profiles for lead year 2 of respective ensemble members. Top: temperature, bottom: relative humidity. red: B1-LR, green: B1-MR, blue: Pr-LR.

implies that the hindcasted troposphere has less vertical stability than observed by the radiosondes. In the lower stratosphere, above 200 hPa, the bias drops until 100 hPa, above which level it increases again. There are some regional differences in the temperature bias profile.

Figure 2b shows that the hindcasts overestimate relative humidity throughout the troposphere. This is the case over most of the annual cycle (not shown). Only at 850 and 700 hPa a small seasonal variation of the humidity bias is found, with an amplitude smaller than 5 % relative humidity, and the smallest model overestimation in June, the largest in winter. The general positive humidity bias increases with height, from 5 to 15 % relative humidity near the surface to 30 to 40 % relative humidity near 300 hPa. Part of the humidity bias, 10 % to 25 % relative humidity, can be explained by the 1 to 4 K lower temperatures of the hindcasts. According to the Clausius-Clapeyron relation, lower temperature implies lower water vapour saturation pressure, hence higher relative humidity (Etling, 2002). However, the remainder of the observed humidity bias implies that modelled water vapour pressure is generally too high above Europe.

While Figure 2 only shows the average profiles, Figure 3 gives additional information on the range of observed bias seen in lead year 2 for the different ensemble members. Clearly, the positive humidity bias is present in almost all cases. Negative temperature bias, however, is present in many, but not all cases. It occurs nearly always in the upper troposphere, but near the surface and in the stratosphere positive temperature bias occurs quite often as well.

The considerable negative temperature bias of the hindcasts, which also increases with height in the troposphere, indicates that the global model underestimates atmospheric stability. Together with higher relative humidities, this has substantial impact on the probability of convective situations and severe weather events, predicted e.g. by the K-Index (George, 1960). This severe weather index combines vertical temperature lapse rate

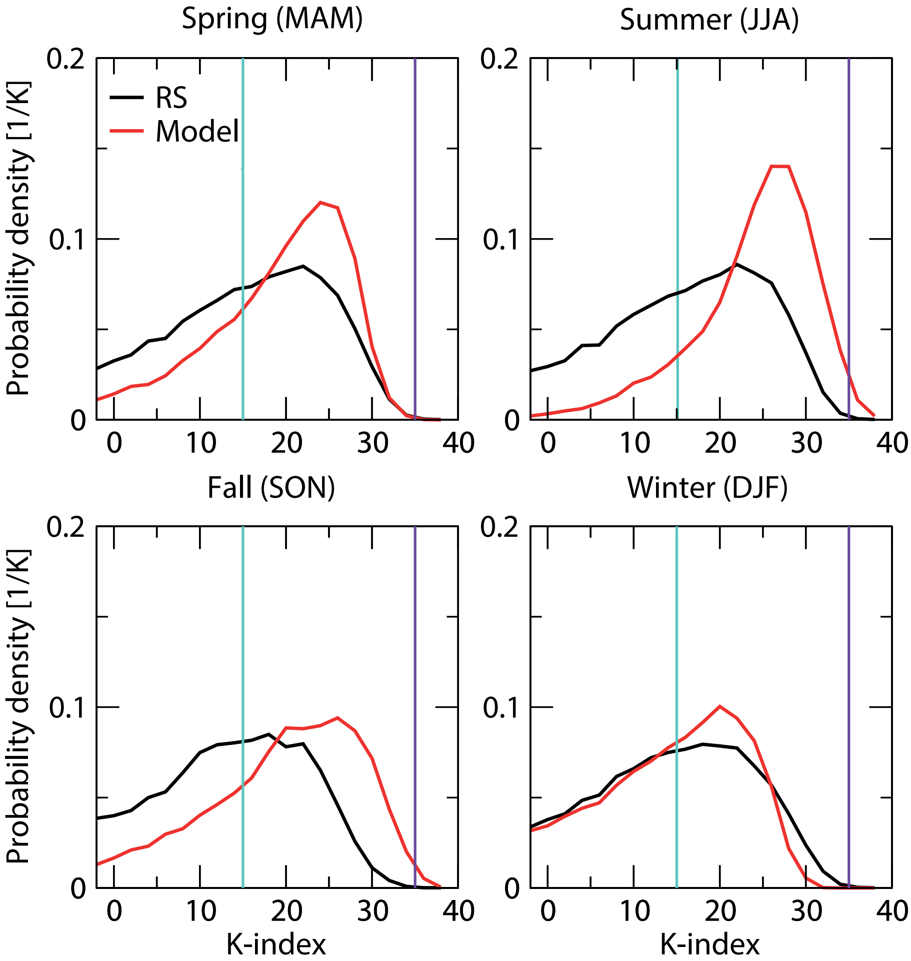

Figure 4: Observed and hindcasted probability distribution of the K-index for the four seasons over Germany for lead year 2. Red line: hindcast, black line: RS observations. Cyan vertical line: maximum threshold for no thunderstorm likely; purple line: minimum threshold for very unstable atmosphere, thunderstorms very likely.

and moisture content of the lower troposphere. It is calculated as $K = T_{850} - T_{500} + T_{d850} - T_{dd700}$, where T is temperature, T_d is dew point temperature, T_{dd} is dewpoint depression, and subscript values denote the pressure level in hPa. Figure 4 compares radiosonde-observed (black) and hindcasted (red) probability density distributions for the K-index and four seasons over Germany.

K-index values below 15 (cyan vertical line in Figure 4) indicate a stable atmosphere, where thunderstorms are not likely. K-index values above 35 (purple line in Figure 4) indicate a very unstable moist atmosphere with high potential for thunderstorms. Figure 4 shows clear differences between observed and hindcasted probability density distributions of the K-index. In all seasons, except for winter, the hindcasted probability density distributions are shifted towards higher K-Index values, and are also narrower. Thus, the model simulations give substantially higher probabilities for convective and severe weather events than the radiosondes. This is most pronounced in summer and fall, whereas simulated and observed probability densities agree much better in winter. Without careful, seasonally adapted bias and drift correction the global model simulations should, therefore, not be used directly for forecasting the likelihood of convective and severe weather events, e.g. thunderstorms.

3.2 Model drift

Since it is possible that model bias changes with time after initialisation (Smith et al., 2013), we also checked for systematic lead year dependence.

Resulting drifts of the model's temperature and humidity bias are plotted in Figure 5 for the 500 hPa level over Germany. Typical 1σ uncertainties are shown for the Pr-LR experiment only. In the B0-LR, B1-LR, and B1-MR hindcasts, the temperature bias remains constant with lead year, i.e. model drift is small or negligible. The Pr-LR hindcasts, however, show substantial lead-year dependence, i.e. significant drift. The bias between modelled and observed temperature becomes larger with hindcast lead year, typically by −0.5 to −0.1 K per lead year. Similar results are seen for other European regions.

The drift of the Pr-LR hindcasts is most likely a consequence of the full field initialization applied to the

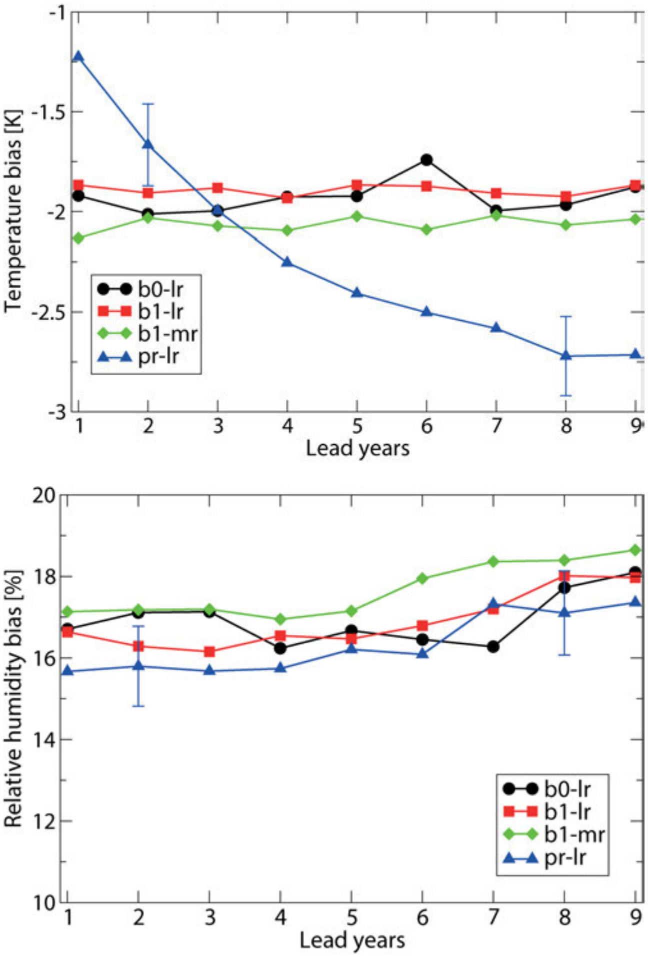

Figure 5: Lead time dependence of the bias over Germany. Top: temperature at 500 hPa, bottom: relative humidity at 700 hPa. black: B0-LR, red: B1-LR, green: B1-MR, blue: Pr-LR. Error bars show ±1 standard deviation of the mean over ensemble members for lead years 2 and 8 of Pr-LR. Similar standard deviations are seen for other experiments and lead years.

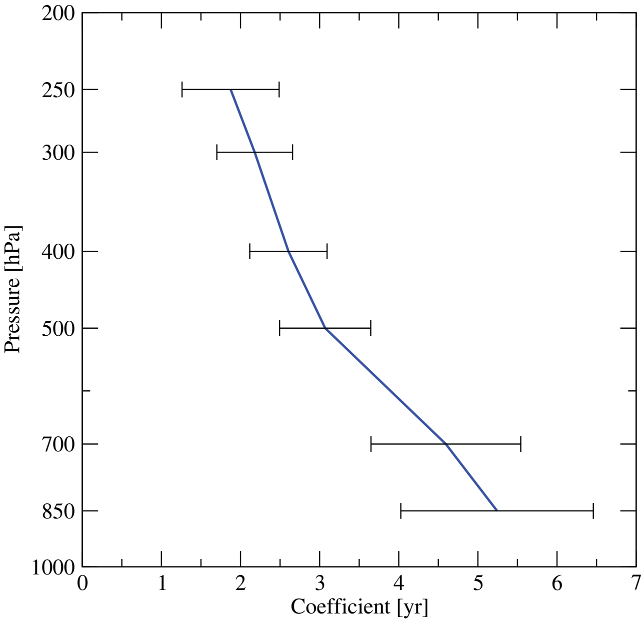

Figure 6: Height dependence of the temperature bias decay time τ, obtained by fitting the tropospheric Pr-LR bias time series in Fig 5 (see text). Error bars give the uncertainty of the fits. In the stratosphere, at levels 200 hPa and above, there is no clear decay, and meaningful decay time constants cannot be obtained.

ocean in this experiment (Marotzke et al. (2016), Jürgen Kröger, personal communication), whereas the anomaly field initialization applied to the ocean in the B0 and B1 experiments does not seem to result in drifting model bias.

The temperature drift of the Pr-LR experiment is present at all tropospheric levels. It seems to decay exponentially:

$$T(t) = [T_0 - T_e]\exp\left(-\frac{t}{\tau}\right) + T_e, \qquad (3.1)$$

where T_0 is the temperature bias at lead year 1, T_e is the equilibrium or asymptotic temperature bias of the model, and τ is the decay time coefficient of the bias. Figure 6 shows the height dependence of decay time τ obtained by fitting Eq. 3.1 at several pressure levels. In the lowermost troposphere, τ is largest. This is probably caused by the ocean's large thermal inertia. Decay time τ decreases with height until 200 hPa. In the stratosphere, 100 hPa and above, bias does not show exponential decay, and is usually constant after the third lead year.

For relative humidity (Figure 5 bottom), all model experiments show a slight drift towards higher humidity, by about 0.1 % relative humidity per lead year. Different from temperature, the model drifts for relative humidity appear very similar for all experiments.

3.3 Forecast accuracy and uncertainty

A major goal of verification is to quantify how well time series from the model hindcasts correspond to the observations. High correlation between modelled and observed anomaly time series, for example, is desirable (Pohlmann et al., 2013; Müller et al., 2014). Anomalies have the hindcasted or observed annual cycle removed, respectively. This eliminates most systematic biases. Figure 7 (top) shows good anomaly correlation between Pr-LR hindcasts and observations, both using unhomogenized (left) and homogenized (right) RS observations. Over most of Europe, the anomaly correlations exceed 0.5. Interestingly, higher anomaly correlations are seen over much of Europe, when the homogenized RS observations are used (right). This might be expected, because inhomogeneous RS data include artificial jumps. Since corresponding spurious jumps or trends in the model hindcasts are unlikely, the correlation between RS time series and hindcast improves for the homogenized RS time series, as seen in Figure 7.

Another important question is, whether initialization (or model changes) improve the hindcasts, e.g. by giving better Mean Square Error Skill Score (MSESS) (Goddard et al., 2013; Stolzenberger et al., 2015).

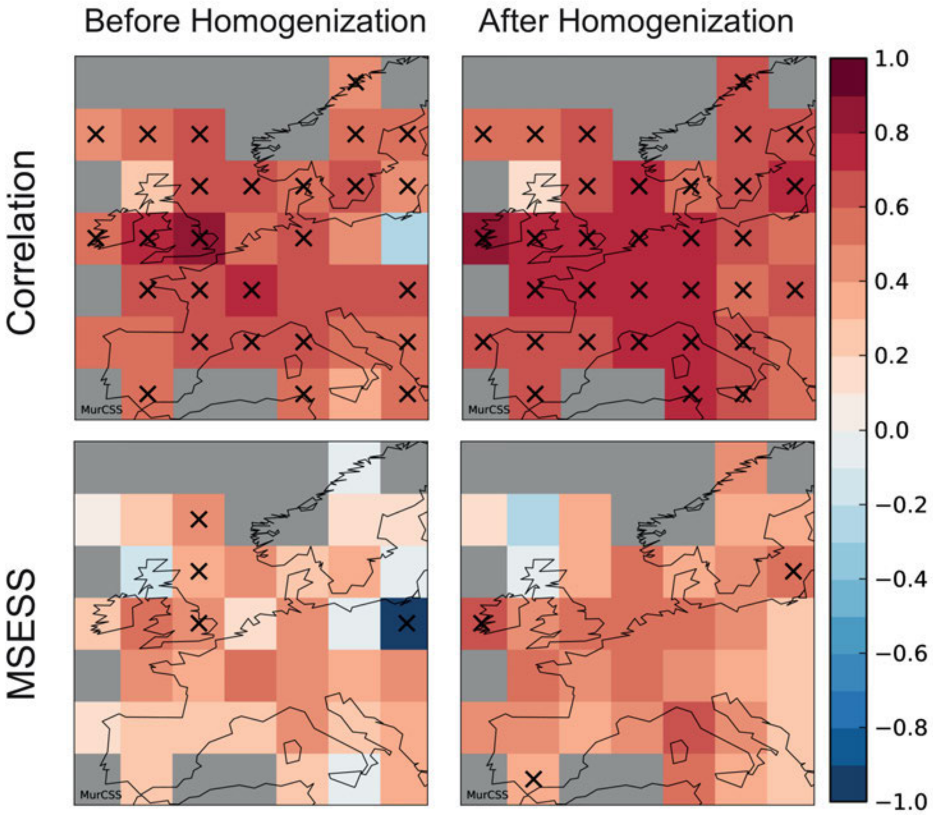

Figure 7: Top: Correlation between temperature anomaly times series from Pr-LR hindcasts and temperature anomaly time series from RS observations. Left: Using raw RS observations, before homogenization. Right: Using homogenized RS observations. Bottom: Same, but for the Mean Square Error Skill Score (MSESS, see Eq. 3.3), comparing mean square error between Pr-LR hindcasts and RS observations against the mean square error between climatology and RS observations. All results in the Figure are for 500 hPa temperature, the period 1970 to 2000, and lead years 2 to 5. Black '×'-s denote statistical significant results with 5 % error probability.

MSESS is based on the Mean Square Error (MSE), or difference, between hindcasted ($H(t,e)$) and observed anomalies ($O(t)$), where t is the temporal index, and e represents the ensemble member (Murphy, 1988; Goddard et al., 2013). MSE describes how closely the ensemble members reproduce the observed anomaly time series.

$$MSE_H = \frac{1}{TXE}\sum_{t=1}^{T}\sum_{e=1}^{E}(H(t,e) - O(t))^2, \qquad (3.2)$$

MSESS, then, compares MSE (accuracy) of a test prediction against MSE of a reference prediction (R). R could be the climatology, an uninitialized prediction, or another hindcast experiment:

$$MSESS(H,R,O) = 1 - \frac{MSE_H}{MSE_R}. \qquad (3.3)$$

A perfect hindcast H would have $MSE_H = 0$ ($MSE_R \neq 0$), and therefore in MSESS$(H,R,O) = 1.0$. Positive MSESS values mean that $MSE_H < MSE_R$, indicating that the hindcast is an improvement over the reference.

The bottom panels of Figure 7 show the MSESS comparing Pr-LR results against climatology for temperature at 500 hPa in the free troposphere. On the left, the hindcast results are relative to the raw RS data, on the right the homogenized RS data are used. In both cases, positive MSESS skill score indicate that Pr-LR hindcasts are an improvement over the climatology. Interestingly, this improvement again becomes clearer when the better, homogenized RS data are used. In this case, the MSESS scores are ≈ 0.1 higher than for the original inhomogeneous RS data. Both, for anomaly correlation and MSESS, Figure 7 indicates that using the successfully homogenized RS data improves hindcast verification metrics.

Ensemble spread (due to varying initial conditions) can be used as a measure of forecast uncertainty. Based on Keller et al. (2008), Kadow et al. (2015) proposed the Logarithmic Ensemble Spread Score (LESS) to quantify skill in forecast spread.

$$LESS = \ln\left(\frac{\overline{\sigma^2_{\hat{H}}}}{\overline{\sigma^2_R}}\right), \qquad (3.4)$$

where the average ensemble spread

$$\overline{\sigma^2_{\hat{H}}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{E-1}\sum_{e=1}^{E}\left(\hat{H}_{et} - \hat{H}_t\right)^2 \qquad (3.5)$$

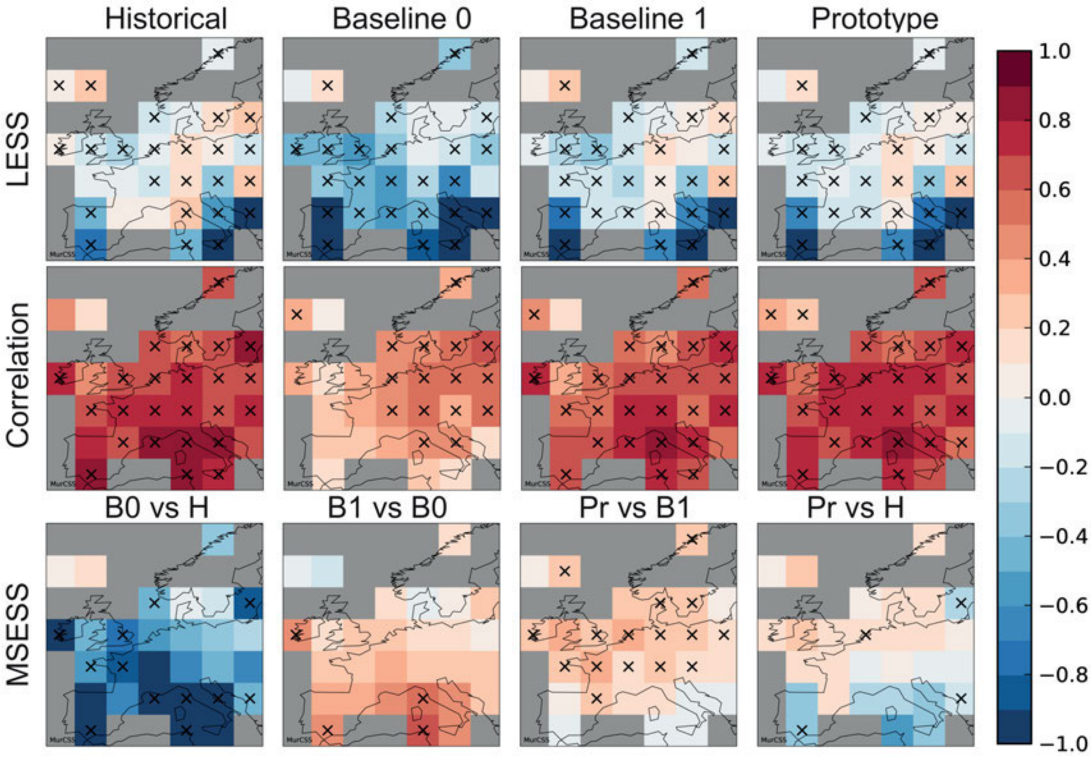

Figure 8: LESS (top), anomaly correlation (middle) and MSESS (bottom) of historical (H), B0-LR, B1-LR, and Pr-LR hindcasts using homogenized RS temperature observations. Period of experiment starts: 1968–2007, lead time: 2–5 years, 500 hPa pressure level. Black '×'-s denote statistical significant results with 5 % error probability.

is calculated using the bias (and conditional bias) corrected ensemble members $\hat{H}_{et}$ and ensemble mean $\hat{H}_t$. E and T are the number of ensemble members and time steps considered. The reference spread σ_R^2 is calculated as follows:

$$\sigma_R^2 = \frac{1}{T-2}\sum_{t=1}^{T}\left(\hat{H}_t - O_t\right)^2 . \qquad (3.6)$$

LESS is positive (negative) if the hindcasts overestimate (underestimate) the observed variance. The optimal value for LESS is 0.

Based on LESS, anomaly correlation, and MSESS, Figure 8 compares the skills of the H-LR, B0-LR, B1-LR and Pr-LR model experiments against each other (MSESS), or against the RS observations (LESS and anomaly correlation). Results are for tropospheric temperature at 500 hPa over Europe, and for lead-years 2 to 5, initialized during the period 1968–2007. Simulations and homogenized RS observations are interpolated to a 5° × 5° grid, as suggested by Goddard et al. (2013). The H-LR historical runs (which do not use data assimilation and yearly initialisation) give the highest anomaly correlations (see Figure 8). Anomaly correlations for B0-LR are lowest, but increase again for B1-LR and Pr-LR. This is reflected in the MSESS results (Figure 8 bottom). Despite the use of data assimilation and yearly initialisation, the B0-LR hindcasts loose skill compared to H-LR, at least over the European region. This is somewhat disappointing. B1-LR then show higher skill (higher MSESS values) than B0-LR, and Pr-LR also shows higher skill than B1-LR. Ultimately this indicates that the Pr-LR full-field initialisation of atmosphere and ocean provides better hindcasts for mid-tropospheric temperature over Europe than the anomaly initialisations in B0-LR and B1-LR. However, the MSESS results on the bottom right of Figure 8 show that even the Pr-LR simulations do not fully recover the skill lost when going from H-LR historical runs to the initialized runs.

For forecast spread, as measured by LESS, the top panels of Figure 8 indicate LESS values close to 0 over most of Europe, except Southern Europe. There is hardly any difference between the H-LR historical runs, and B1-LR to Pr-LR hindcasts, all of which provide realistic ensemble spread compared to the RS observations. Only B0-LR results indicate too narrow ensemble spread as well.

Figure 9 shows the corresponding hindcast skills for lower stratospheric temperatures at the 100 hPa pressure level. Again, high anomaly correlations, largely above 0.6, are seen for all experiments. Different from the troposphere, the MSESS values show substantial improvements from the H-LR to the B0-LR experiment, but no improvement or even deterioration from B0-LR to B1-LR and from B1-LR to Pr-LR, especially in South-Eastern Europe. Since anomaly correlations are fairly high, largely above 0.6, this deterioration means

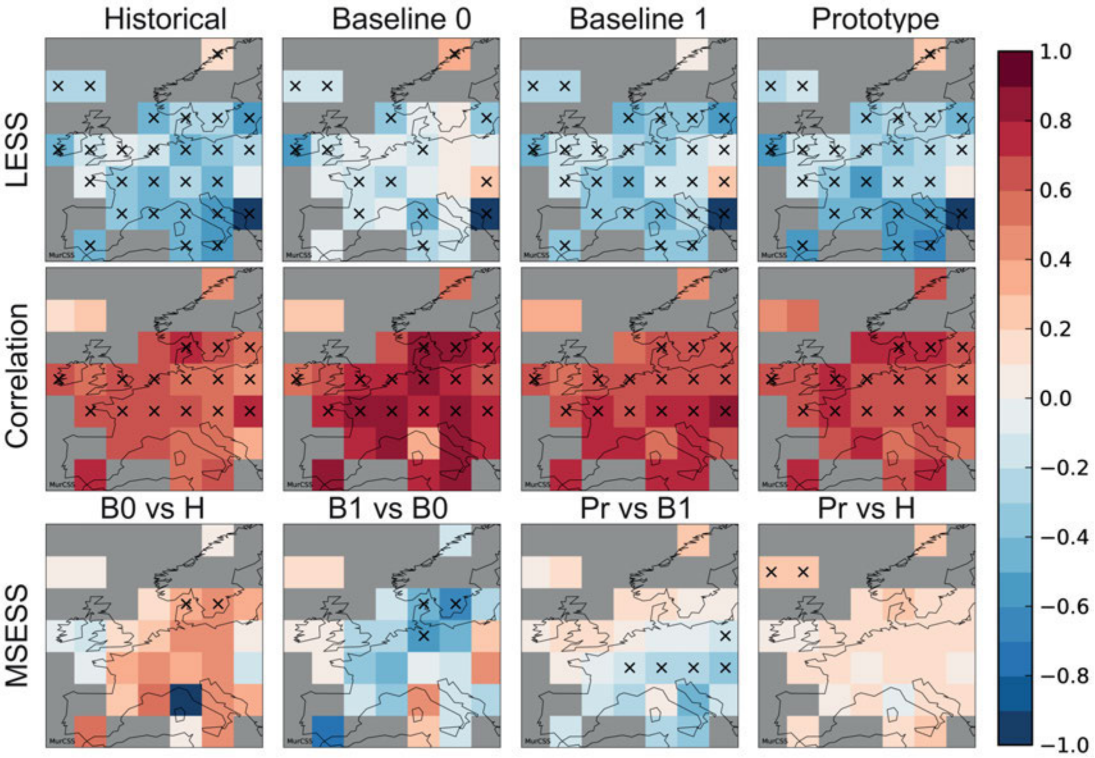

Figure 9: Same as Figure 8 but for temperatures at the 100 hPa pressure level.

only a slight drop in accuracy for the Pr-LR experiment, which still performs better than the uninitialized H-LR simulation.

LESS (Figure 9 top) is generally negative for most experiments, indicating that model spread for stratospheric temperature is generally smaller than observed by the radiosondes. Interestingly, LESS is best in the stratosphere for the B0-LR runs, where LESS is worst in the troposphere (compare Figure 8).

Finally, results for 700 hPa relative humidity are presented in Figure 10. Note that for humidity only data between 1993 and 2006 and for lead years 2–5 were used. Compared to temperature, agreement between relative humidity hindcasts and RS data is much poorer for all experiments. LESS shows that the observed spread is not well captured by the hindcasts, with little difference between experiments. If we look at model accuracy (anomaly correlation and MSESS), a large spatial variance is present, and results look a lot like noise. H-LR performed best in parts of Western and South Eastern Europe, poorly in N-Europe. B0-LR showed improvement compared to H-LR in Central-Europe and around the Baltic, but performed poorer in the West. B1-LR implies anti-correlation over a large part of Europe. Pr-LR gives overall smaller anti-correlation than B1-LR, but even then positive correlation in the Iberian-Peninsula only. Overall, simulated relative humidity anomalies do not correlate well with the RS time series, and little can be said about changes or improvements between the different hindcast experiments.

4 Conclusions

We used upper-air temperatures and humidities from radiosonde observations to evaluate the performance of decadal hindcasts of the MPI-Earth-System-Model above Europe. The model simulates a colder and moister free troposphere than observed, by 1 to 4 K, and by 10 to 40 % relative humidity, respectively. Part of the humidity bias, 10 to 25 % relative humidity, is due to the lower hindcasted temperature, but the remainder indicates that modelled relative humidity is too high in the free troposphere above Europe. Generally, the cold bias of the hindcasted temperature increases with altitude. This means less vertical stability for the simulated troposphere. Without appropriate corrections, lower vertical stability and higher humidity of the simulations have significant implications, e.g. for the forecast of severe weather events over Europe.

The model hindcasts are initialized at the beginning of each year using different combinations of observed full or anomaly fields of the oceanic or atmospheric state. Comparison of different initializations for the model indicates that the atmospheric temperature bias over Europe does not depend much on the atmospheric initialization. For ocean initialization, however, temperature bias remains constant with time for model initialization with observed ocean anomaly fields, but drifts over the first couple of years when the full ocean fields are used. In this case, the time scale on which the model atmospheric temperature bias relaxes towards

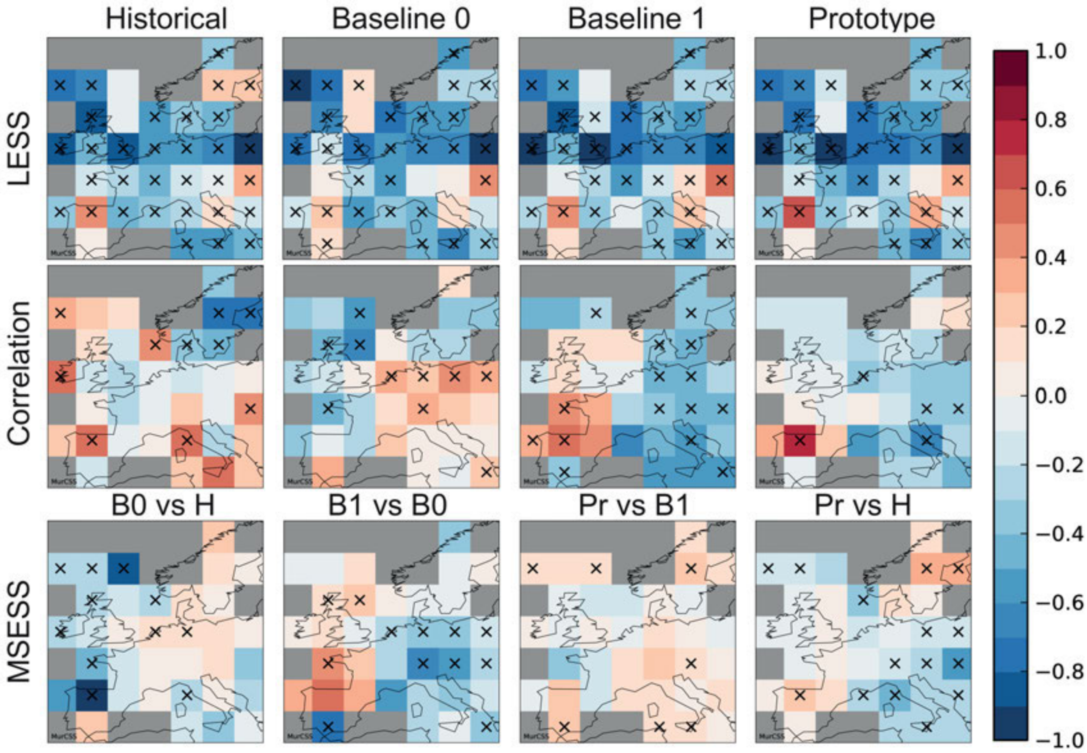

Figure 10: Same as Figure 8 but for relative humidity at the 700 hPa pressure level.

its final value changes from about 5 years near the surface (850 hPa) to less than 1 year near the tropopause (200 hPa).

Using the European radiosonde temperature data as a reference, the different model hindcasts were evaluated on the basis of correlation of time series, mean-square error skill score (MSESS) and logarithmic ensemble spread score (LESS), which measure forecast error and forecast spread. Both in the troposphere and lower stratosphere (100 hPa), the largest effects are seen for the change from unitialized runs (H) to ocean anomaly initalization (B0-LR). In the troposphere, this change leads to lower correlation and lower MSESS, whereas in the lower stratosphere it leads to slight improvements in MSESS and a fairly large improvement in forecast spread (LESS). The addition of atmospheric initialization (B0 to B1) and the change to full field ocean initialization (B1 to Pr) lead to small improvements in the troposphere over Europe, i.e. slightly larger correlations and better MSESS. In the stratosphere, these changes have little effect on correlation but deteriorate MSESS slightly. LESS remains similar in nearly all initialized experiments except B0-LR, both in the troposphere and stratosphere above Europe. For humidity, correlations and skill scores are much poorer, and little can be said about changes due to different initializations.

Temperature correlations above 0.5 and LESS values around 0 indicate that the decadal MPI-ESM forecasts do have predictive skills over Europe. However, without appropriate bias and drift corrections, the model simulations should not be used for direct forecasts. Severe weather indices calculated directly from the model, for example, would indicate substantially higher probability for convective and severe weather events than the radiosonde observations. Better characterization of the substantial temperature and humidity bias and drift found in this study are one of the tasks in the next phase of the German decadal climate prediction research programme (MiKlip II).

Acknowledgments

The authors thank Andreas Hense and an anonymous reviewer for their many helpful and constructive comments making our work more comprehensible. We thank Henning Rust for his fruitful discussions on many issues of validating decadal climate predictions, and a thorough checking of the manuscript. Support by the German Federal Ministry of Education and Research (BMBF) programme MiKlip, project MOSQUITO, Grant 01LP1106A, is greatly appreciated.

Abbreviations

RS	radiosonde
DWD	Deutscher Wetterdienst
MPI	Max-Planck Institute for Meteorology
MPI-OM	MPI Ocean Model
MPI-ESM	MPI Earth System Model

ECHAM6	ECmwf HAMburg: the atmospheric component of MPI-ESM
CMIP5	Coupled Model Intercomparison Project Phase 5
NCEP	National Center for Environmental Prediction
ECMWF	European Center for Medium-Range Weather Forecasts
ORA-S4	Ocean Reanalysis System 4
GECCO2	German Contribution to Estimating the Circulation and Climate of the Ocean 2
MiKlip	Mittelfristige Klimaprognose
H	historical runs of the MPI-ESM
B0	MPI-ESM baseline 0 hindcast experiment
B1	MPI-ESM baseline 1 hindcast experiment
Pr	MPI-ESM prototype hindcast experiment
LR	low resolution
MR	mixed resolution
MSE	mean square error
MSESS	mean square error skill score
LESS	logarithmic ensemble spread score

References

Balmaseda, M., K. Mogensen, A. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis ORAS4. – Quart. J. Roy. Meteor. Soc. **139**, 1132–1161, DOI: 10.1002/qj.2063.

Dai, A., J. Wang, P. Thorne, D. Parker, L. Haimberger, X. Wang, 2011: A new approach to homogenize daily radiosonde humidity data. – J. Climate **24**, 965–991, DOI: 10.1175/2010JCLI3816.1.

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

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

Durre, I., R. Vose, D. Wuertz, 2006: Overview of the integrated global radiosonde archive. – J. Climate **19**, 53–68.

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

Elliott, W., D. Gaffen, 1991: On the utility of radiosonde humidity archives for climate studies. – Bull. Amer. Meteor. Soc **72**, 1507–1520, DOI: 10.1175/1520-0477(1991)072<1507:OTUORH>2.0.CO;2.

Etling, D., 2002: Theoretische Meteorologie: Eine Einführung. – Springer-Verlag, Berlin, Heidelberg, New York, 356 pp.

George, J., 1960: Weather forecast for Aeronautics. – Academic Press.

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

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

Hawkins, E., B. Dong, J. Robson, R. Sitton, D. Smith, 2014: The interpretation and use of biases in decadal climate predictions. – J. Climate **27**, 2931–2947, DOI: 10.1175/JCLI-D-13-00473.1.

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

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

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

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. Mo, C. Ropelewski, R. Wang, R. Jenne, D. Joseph, 1996: The NCEP/NCAR 40-year reanalysis project. – Bull. Amer. Meteor. Soc **77**, 437–470, DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

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

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

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

Marotzke, J., W.A. Müller, F.S.E. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, K. Prömmel, H.W. Rust, D. Stammer, U. Ulbrich, C. Kadow, A. Köhl, J. Kröger, T. Kruschke, J.G. Pinto, H. Pohlmann, M. Reyers, M. Schröder, F. Sienz, C. Timmreck, M. Ziese, 2016: MiKlip – a national research project on decadal climate prediction. – Bull. Amer. Meteor. Soc., published online: 17 June 2016, DOI: 10.1175/BAMS-D-15-00184.1

Meehl, G., T. Stocker, W. Collins, P. Friedlingstein, A. Gaye, J. Gregory, A. Kitoh, R. Knutti, J. Murphy, A. Noda, S. Raper, I. Watterson, A. Weaver, Z. Zhao, 2007: Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, chapter 10: Global Climate Projections. – Cambridge University Press.

Meehl, G., L. Goddard, G. Boer, R. Burgman, G. Branstator, C. Cassou, S. Corti, G. Danabasoglu, F. Doblas-Reyes, E. Hawkins, A. Karspeck, M. Kimoto, A. Kumar, D. Matei, J. Mignot, R. Msadek, A. Navarra, H. Pohlmann, M. Rienecker, T. Rosati, E. Schneider, D. Smith, R. Sutton, H. Teng, G van. Oldenborgh, G. Vecchi, S. Yeager, 2014: Decadal climate prediction: An update from the trenches. – Bull. Amer. Meteor. Soc. **95**, 243–267, DOI: 10.1175/BAMS-D-12-00241.1.

Miloshevich, L., A. Paukkunen, H. Vömel, S. Oltmans, 2004: Development and validation of a time-lag correction for Vaisala radiosonde humidity measurements. – J. Atmos. Oceanic Technol. **21**, 1305–1327, DOI: 10.1175/1520-0426(2004)021<1305:DAVOAT>2.0.CO;2.

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

Müller, W.A., H. Pohlmann, F. Sienz, D. Smith, 2014: Decadal climate predictions for the period 1901–2010 with a coupled climate model. – Geophys. Res. Lett. **41**, 2100–2107, DOI: 10.1002/2014GL059259

Murphy, A., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. – Mon. Wea. Rev. **116**, 2417–2424, DOI: 10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Pattantyús-Ábrahám, M., W. Steinbrecht, 2015: Temperature trends over Germany from homogenized radiosonde data. – J. Climate **28**, 5699–5715, DOI: 10.1175/JCLI-D-14-00814.1.

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

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

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

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

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

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

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

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