Evaluation and Bias Correction of Regional Climate Model Results Using Model Evaluation Measures

ROBERT SCHOETTER AND PETER HOFFMANN
Meteorological Institute, KlimaCampus, University of Hamburg, Hamburg, Germany

DIANA RECHID
Max Planck Institute for Meteorology, KlimaCampus, Hamburg, Germany

K. HEINKE SCHLÜNZEN
Meteorological Institute, KlimaCampus, University of Hamburg, Hamburg, Germany

(Manuscript received 5 August 2011, in final form 3 April 2012)

ABSTRACT

For the assessment of regional climate change the reliability of the regional climate models needs to be known. The main goal of this paper is to evaluate the quality of climate model data that are used for impact research. Temperature, precipitation, total cloud cover, relative humidity, and wind speed simulated by the regional climate models Climate Local Model (CLM) and Regional Model (REMO) are evaluated for the metropolitan region of Hamburg in northern Germany for the period 1961–2000. The same evaluation is performed for the global climate model ECHAM5 that is used to force the regional climate models. The evaluation is based on comparison of the simulated and observed climatological annual cycles and probability density functions of daily averages. Several model evaluation measures are calculated to assure an objective model evaluation. As a very selective model evaluation measure, the hit rate of the percentiles is introduced for the evaluation of daily averages. The influence of interannual climate variability is considered by determining confidence intervals for the model evaluation measures by bootstrap resampling. Evaluation shows that, with some exceptions, temperature and wind speed are well simulated by the climate models; whereas considerable biases are found for relative humidity, total cloud cover, and precipitation, although not for all models in all seasons. It is shown that model evaluation measures can be used to decide for which meteorological parameters a bias correction is reasonable.

1. Introduction

Climate model output is increasingly used in climate change impact, vulnerability, and adaptation studies for statistical downscaling (e.g., Wilby 2003; Hoffmann et al. 2012) and statistical–dynamical downscaling (e.g., Fuentes and Heimann 2000) and for the provision of physically consistent forcing data for impact models like hydrological models (e.g., van Pelt et al. 2009), ecosystem models (e.g., Huntley et al. 2008), and biometeorological models (e.g., Muthers et al. 2010). Results from impact studies can be very sensitive to the meteorological data used. For this reason, evaluation of regional climate model results is crucial for the interpretation of results from impact studies (e.g., Dibike et al. 2008).

Biases in regional climate model simulation results can arise from shortcomings of the regional climate models themselves, but also from erroneous forcing data. Regional climate model (RCM) evaluations are mostly done for RCMS forced with reanalysis data over a given time period; the RCM results are then compared to observed data (e.g., Kotlarski et al. 2005; Christensen et al. 2008). This evaluation approach is not within the scope of the present study, because transient runs of regional climate models forced by global climate models are needed for studies that deal with regional climate change impacts. Transient regional climate model integrations have, for example, been evaluated by Jacob et al. (2007) with the focus on seasonal and annual
averages of temperature and precipitation. In the present study, we evaluate results of the RCMs Climate Local Model (CLM; Rockel et al. 2008) at 0.165° (≈18 km) and Regional Model (REMO; Jacob 2001) at 0.088° (≈10 km) horizontal resolution, both driven by the general circulation model (GCM) known as European Centre for Medium-Range Weather Forecasts–Hamburg fifth-generation GCM–Max Planck Institute Ocean Model (ECHAM5-MPIOM; Roeckner et al. 2003; Jungclaus et al. 2006) at T63 resolution (≈200 km at 0° latitude). The results of ECHAM5-MPIOM, CLM, and REMO are evaluated for the metropolitan region of Hamburg in northern Germany for the period 1961–2000. This area (Fig. 1) is close to the coast and is one of the areas selected by the German Research Ministry to develop adaptation measures for climate change within the project known as Klimawandel in Regionen Zukunftsfähig Gestalten—Nord (regional strategies concerning climate changes in the metropolitan area of Hamburg; KLIMZUG-NORD). Thus, the RCM results are used for various purposes in the framework of the development of climate change adaptation measures.

The evaluation of a climate model operating in climate mode and not in hindcast mode needs to be based on comparisons of the statistics of the simulated and observed climate. In the present study, the climatological annual cycles, the variance of time series of monthly averages, and the empirical probability density functions (PDF) of daily averages are evaluated. The degree of agreement between simulated and observed climate is quantified by model evaluation measures. Model evaluation measures introduced by Keuler (2006) are used for evaluation of monthly averages. The skill score (SSC) from Perkins et al. (2007) and an additional model evaluation measure, the hit rate of the percentiles (HRP) introduced in the present article, are used for evaluation of the PDFs of daily averages.

Two major issues have to be addressed in the evaluation: uncertainties of the observational datasets and climate variability. The issue of uncertainties in observational datasets is addressed by using more than one observational dataset for evaluation. The issue of inter-annual climate variability is addressed by calculation of confidence intervals for the model evaluation measures by bootstrap resampling. The influence of long-term climate variability is investigated by evaluation of different realizations of simulated climate from the same climate model against each other.

The quantitative climate model evaluation is carried out to determine if it is reasonable to apply a bias correction to the RCM data before they are used in impact studies. For the RCMs evaluated here, recommendations of which meteorological parameters should be bias corrected and which not will be given.

The structure of this paper is the following: the setup of the climate model simulations to be evaluated as well as the data processing is described in section 2. The observational datasets used for model evaluation are described in section 3. The model evaluation measures and the bootstrap-resampling technique are described in section 4. Evaluation results are presented and discussed in section 5. In section 6, an example of using model evaluation measures to determine whether bias correction is reasonable is presented. Conclusions are drawn in section 7.

2. Regional climate model simulations

In this section, the setup of the regional climate models evaluated as well as the data processing is described.

a. Setup of the regional climate models

The global GCM datasets are taken from the ECHAM5-MPIOM simulations for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) (Roeckner 2010; Roeckner et al. 2010a,b). The ECHAM5-MPIOM control climate simulations are driven by observed anthropogenic forcings [CO$_2$, CH$_4$, N$_2$O, chlorofluorocarbons (CFCs), O$_3$, and sulfate] for the period 1860–2000. These simulations neglect natural forcings from volcanic aerosols and variable solar activity. The three different members of the control period were initialized in different years from a 500-yr preindustrial control run to account for internal climate variability.

The global ECHAM5-MPIOM ensemble was used to drive the RCM REMO first for Europe at 0.44° horizontal resolution. In a second nesting step, the ensemble is further downscaled to 0.088° for Germany with 27
vertical levels. For the third realization a different version of the REMO model was used.

The RCM CLM was directly nested into the global ECHAM5-MPIOM fields for Europe at 0.165° horizontal resolution with 32 vertical levels.

b. Data processing

For REMO, hourly values of 2-m temperature, 2-m dewpoint temperature, total cloud cover, 10-m wind speed, and convective and stratiform precipitation are available. Hourly total precipitation sums are calculated by adding convective and stratiform precipitation sums. Hourly relative humidity is calculated from 2-m temperature and 2-m dewpoint temperature by using the Magnus formula (e.g., Hupfer and Kuttler 2006). From hourly values, daily and monthly averages (2-m temperature, 2-m relative humidity, 10-m wind speed, and total cloud cover) are calculated. For precipitation the integral values are calculated.

For CLM, 3-hourly values of 2-m temperature, 2-m dewpoint temperature, and total cloud cover; hourly values of 10-m wind speed in west–east and north–south directions; and convective and stratiform liquid and solid precipitation sums are available. Hourly values of 10-m wind speed and total precipitation are derived; 3-hourly values of relative humidity are calculated from 2-m temperature and 2-m dewpoint temperature. Again, daily and monthly averages are derived from hourly or 3-hourly values.

For ECHAMS5, 6-hourly values of 2-m temperature, 2-m dewpoint temperature, total cloud cover, 10-m wind speed in west–east and south–north directions, and convective and stratiform precipitation are available. Daily and monthly averages are derived with the same methods as for CLM and REMO.

3. Observation based data

a. Datasets

1) ERA-40 REANALYSIS

The 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) covers the period from September 1957 to August 2002. The spatial resolution is about 1.125°. The values for temperature, total cloud cover, relative humidity, and wind speed are taken from the 6-hourly instantaneous fields. Precipitation is taken as the integral value from the first 6 h after each assimilation step. It has to be kept in mind that a general underestimation of precipitation in the ERA-40 reanalysis data for the extratropical regions has been found by Uppala et al. (2005).

2) CRU TS 2.1

The Climatic Research Unit (CRU) TS 2.1 time series (Mitchell and Jones 2005) covers the period from 1901 to 2002. Data are provided on a 0.5° regular grid. From CRU TS 2.1, monthly values of temperature, precipitation, and total cloud cover are used in the present study. In the 1961–90 climatology for northern Europe, New et al. (1999) report a mean-square cross validation error of about 15% for precipitation, 0.5 K for summer temperatures, 1.0 K for winter temperatures, 0.3 oktas (3.75%) for total cloud cover, and 1 m s⁻¹ for wind speed.

3) GPCC FULL DATA REANALYSIS VERSION 4

Among the various Global Precipitation Climatology Centre (GPCC) products (Rudolf and Schneider 2005; Rudolf et al. 2010) the “Full Data Reanalysis Version 4” (GPCC-V4), which covers the period from 1901 to 2009, is used in the present study. The data are available on a 0.5° regular grid. The number of rain gauges used for interpolation is above 300 in the evaluation domain for all months of the evaluation period. This indicates that the interpolation error is below 10% (Rudolf and Schneider 2005).

The GPCC-V4 data are not corrected for systematic undercatch of precipitation. Estimates of the climatological correction factors derived by the method of Legates and Willmott (1990) are provided with the data. Correction factors are substantial for precipitation evaluation: during the winter season they reach a value of up to 1.6 in the evaluation domain. (The difference between corrected and uncorrected precipitation data is shown in Fig. 7.) Correction factors delivered with the GPCC data appear to range at the upper bound of published precipitation correction factors. For example, correction factors derived by Adam and Lettenmaier (2003) and by Yang et al. (2005) are about 1.2–1.3 in winter season, thus much below the GPCC-V4 correction factors.

4) EUROPEAN CLIMATE ASSESSMENT AND DATA 3.0

The European Climate Assessment and Data (ECA&D-3.0) data (Haylock et al. 2008) cover the period from 1950 to 2009. Daily average temperature values as well as daily precipitation sums are used in the present study. ECA&D-3.0 consists of station data interpolated on a 0.25° regular grid. Precipitation is not corrected for systematic undercatch.

5) DWD-REGNIE

The Regionalisierung von Niederschlagshöhen (regionalization of precipitation amounts; REGNIE) gridded
daily precipitation sums have been provided for the period 1961–2000 by the German Meteorological Service (DWD) (A. Köcher 2010, personal communication) for the purpose of the present evaluation study. The precipitation data are on a regular 30° × 60° grid. REGNIE data have been created by the DWD in three steps: first the climatological spatial distribution of the monthly precipitation sum for the period 1961–90 has been interpolated to the target grid by multiple linear regression using height above sea level, longitude, latitude, and the slope of the terrain in relation to the prevailing wind. The precipitation anomalies at the stations have then been interpolated onto the REGNIE grid. As a third step the anomalies have been multiplied with the background field values to give the final value for precipitation. REGNIE precipitation values used in the present study are not corrected for systematic undercatch.

6) DWD ROUTINE OBSERVATIONS (ROD)

The DWD routine observations (ROD) are station data. From these, daily averages of total cloud cover, relative humidity, and wind speed are used in the present study. The station distribution map is shown in Fig. 2. The average of all available station records is calculated for every day during the evaluation period.

b. Uncertainties of datasets

Datasets used in the present study are gridded station data, the ERA-40 reanalysis data, and the ROD, which are area-averaged station data. Gridded station data are created by spatial interpolation of point measurements. Uncertainties in these datasets arise mainly from measurement errors and the method used for spatial interpolation. The contribution of both types of uncertainties to the total uncertainty depends on the meteorological parameter. For example, station measurements of temperature are accurate to 0.1 K (Haylock et al. 2008), whereas observations of total cloud cover are quite uncertain.

Different methods used for spatial interpolation can lead to significantly different results if the interpolated meteorological field has large gradients (e.g., the station data have low spatial representativeness). The spatial representativeness of point measurements depends on the meteorological parameter. The gridded datasets used are created with (slightly) different interpolation methods but are based on the same measurement dataset. For this reason, usage of different gridded datasets helps to capture interpolation uncertainty but not necessarily the measurement uncertainty.

The ERA-40 data can be assumed to be more independent of the gridded station datasets because they use not only station data but also other sources such as radiosonde and satellite data. Additionally, ERA-40 data are subject to characteristics of the ECMWF model and the method used for assimilation. As precipitation is the result of 6 h of model integration it is expected to be very sensitive not only to initial data but also to model characteristics. Differences between ERA-40 data and gridded station data capture many different types of uncertainties. However, one has to keep in mind that the station data do influence the ERA-40 data, and shortcomings of the ECMWF model do introduce new uncertainties.

The described observational datasets are not suitable for trend detection because of the changing number of stations contributing to the interpolated fields and because of the inhomogeneities in records of single stations. Trends in regional climate models can therefore not be evaluated using these datasets.

c. Evaluation domain

As mentioned, the focus of the evaluation is on northern Germany. Since the CRU TS 2.1, ECA&D-3.0, GPCC-V4, and DWD-REGNIE data are only defined for land points, the evaluation domain is chosen to be a rectangular area situated at 52.5°–54.5°N and 8.5°–11.5°E (Fig. 1), with the water surfaces excluded. The ERA-40 data are selected in a way to best fit this evaluation domain. The climate model data are averaged over the whole domain; all model grid points with a land fraction below 0.5 are omitted. The gridded observation data as well as the station data are also averaged over the evaluation domain.
4. Evaluation method

a. Evaluation of monthly averages

Several measures for the evaluation of monthly averages (Table 1) have been suggested by Keuler (2006). These model evaluation measures are calculated for every pair of climate model and observational dataset. One value for BIAS is calculated for every month of the year. All realizations of simulated climate by one given climate model version (3 for CLM and ECHAM5; 2 for REMO) are merged as they are equally probable estimates of the average simulated climate. REMO-3 is here not merged with REMO-2 and REMO-1 because of the different model version.

The model evaluation measures of Keuler (2006) are applicable to all meteorological parameters of interest. However, some shortcomings have to be mentioned. All model evaluation measures deal with the mean or the standard deviation of monthly averages. It is well known (e.g., Wilks 1995) that the mean and standard deviation are not robust statistical measures and, therefore, are very sensitive to climate variability. Furthermore, the model evaluation measures “climatological temporal correlation” (CTCO) and “ratio of yearly amplitudes” (ROYA) have different effective ranges for different meteorological parameters. If a meteorological parameter has a dominant annual cycle (e.g., temperature), CTCO and ROYA tend to yield values near the optimum value of 1; for example, temperature biases are very unlikely to become dominant above the annual amplitude of temperatures in the given evaluation domain.

b. Evaluation of daily averages

The agreement between simulated and observed PDFs of daily averages is quantified with the model evaluation measures shown in Table 2. The SCC was introduced by Perkins et al. (2007). The model evaluation measures are calculated for every pair of climate model and observational dataset for each of the four seasons December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON). Again, all realizations of simulated climate are merged for every climate model. Bin widths used for construction of the empirical PDFs, which in turn are needed to calculate the SCC, are chosen to be 1 K for temperature, 10% for total cloud cover, 4% for relative humidity, and 1 m s\(^{-1}\) for wind speed. The bin widths have been chosen in the order of the uncertainty of the observational data as it does not make sense to discriminate the data with a higher accuracy than they can be observed.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Denomination</th>
<th>Formula</th>
<th>Optimum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>Bias</td>
<td>(\text{BIAS}_m = \overline{M}_m - \overline{O}_m)</td>
<td>0</td>
</tr>
<tr>
<td>MAMD</td>
<td>Mean absolute monthly difference</td>
<td>(\text{MAMD} = \frac{1}{12} \sum_{m=1}^{12}</td>
<td>\overline{M}_m - \overline{O}_m</td>
</tr>
<tr>
<td>ROYA</td>
<td>Ratio of yearly amplitudes</td>
<td>(\text{ROYA} = \frac{\max(\overline{M}_m) - \min(\overline{M}_m)}{\max(\overline{O}_m) - \min(\overline{O}_m)})</td>
<td>1</td>
</tr>
<tr>
<td>CTCO</td>
<td>Climatological temporal correlation</td>
<td>(\text{CTCO} = \frac{\frac{1}{12} \sum_{m=1}^{12} (\overline{M}_m - \overline{M})(\overline{O}<em>m - \overline{O})}{\sqrt{\left(\frac{1}{12} \sum</em>{m=1}^{12} (\overline{M}<em>m - \overline{M})^2\right)\left(\frac{1}{12} \sum</em>{m=1}^{12} (\overline{O}_m - \overline{O})^2\right)}})</td>
<td>1</td>
</tr>
<tr>
<td>RATV</td>
<td>Ratio of temporal variances</td>
<td>(\text{RATV} = \frac{\sigma^2_m}{\sigma^2_O})</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Denomination</th>
<th>Formula</th>
<th>Optimum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC</td>
<td>Skill score</td>
<td>(\text{SSC} = \sum_i \min(\text{PDF}<em>{M,i}, \text{PDF}</em>{O,i}))</td>
<td>1</td>
</tr>
<tr>
<td>HRP</td>
<td>Hit rate of the percentiles</td>
<td>(\text{HRP} = \sum_{i=1}^{99} \left{ \frac{1}{99} \text{ if }</td>
<td>P_{M,i} - P_{O,i}</td>
</tr>
</tbody>
</table>
The value of the SSC is 1 for a perfect overlap of simulated and observed PDFs. Because of the limited sample size, a perfect agreement of the empirical PDFs is very improbable. Therefore, the effective optimum value for the SSC is below 1. The value of the SSC is 0 if the probability density functions have no overlap at all. For most evaluation cases this is a very improbable situation: even if the model has large deficiencies some overlap of the PDFs will be found. The SSC is therefore easy to calculate but not very selective, as the optimum value of 1 as well as the weakest possible score of 0 are not in the effective range of this model evaluation measure.

The motivation for introduction of the HRP is two-fold: at first, the model evaluation measure has an effective range between 0 and 1 independent of the variable to be evaluated. At second, HRP allows to account for an "allowed deviation" between simulated and observed percentiles. The allowed deviation is used to express that a difference between simulated and observed percentiles does not necessarily mean that there is a model shortcoming as there might for example be an uncertainty of the observational data. For the calculation of the HRP, the first to 99th percentiles of simulated and observed daily averages are calculated using the percentile function of the MATLAB software. The HRP is then defined as the fraction of percentiles with a difference between observed and simulated percentiles below the allowed deviation. The choice of the allowed deviation depends on the purpose and the type of the evaluation study. If RCMs driven with reanalysis data are evaluated, the allowed deviation should be an estimate of the uncertainty of the observations. In the present study, the allowed deviation should account for both: uncertainty of observations and climate variability. In the following we describe how we determine the value for the allowed deviation.

In an first step, one has to decide if the allowed deviation shall depend on the value of the percentiles. In the present study, the allowed deviation for temperature, wind speed, and relative humidity is chosen to be independent on the actual percentile \( U(P_{Oi}) = U \). Thus, for example, it is assumed that temperatures of \(-5^\circ C\) have the same uncertainty as temperatures of \(+25^\circ C\). For total cloud cover we assume that the allowed deviation is twice as large for a total cloud cover of 50% as for a total cloud cover of 0% and 100%, with a linear scaling between the minimum and maximum value.

In a second step, the values of the allowed deviations are calculated by consideration of the interannual climate variability and the uncertainty of the observational datasets. The contribution of both uncertainties is estimated by a learning process based on the available observational datasets. The contribution of the interannual climate variability to the allowed deviation is determined by calculation of the HRP for different bootstrap resamples of the observational data as a function of the chosen allowed deviation. To account for temporal correlation, the daily data are resampled as blocks, one block containing all the daily averages during one given season (e.g., all daily data from JJA-1965). Because of statistical variability, the median of the HRP is 0 if the allowed deviation is 0 and increases to 1 for large allowed deviations. This behavior is shown in Fig. 3 for the example of MAM temperatures (black crosses). The allowed deviation due to interannual climate variability is defined as the allowed deviation where the median of the HRP is 0.95 (\( U_{VAR} \), black line in Fig. 3). The allowed deviation due to uncertainties of the observational data is obtained by calculating the HRP as a function of the allowed deviation for two observational datasets (gray crosses in Fig. 3). Again, the allowed deviation due to uncertainties of the observational data is defined as the allowed deviation, for which the HRP is 0.95 (\( U_{OBS} \), gray line in Fig. 3).

With the described method, two allowed deviations are obtained for each season: one considering the interannual climate variability and one considering the uncertainty of the observational data. The evaluation
could be performed with a different allowed deviation for each season, but in the present study we choose to use the maximum allowed deviation of all four seasons per type of uncertainty to avoid time-dependent changes in the allowed deviation. The total allowed deviation is then defined in a pragmatic way as the square root of the sum of the squares of allowed deviations from interannual climate variability and from observations. The allowed deviations obtained by this method are 1.4 K for temperature, 6% for total cloud cover, 4% for relative humidity, and 0.6 m s\(^{-1}\) for wind speed. As the allowed deviation for wind speed appears lower than published estimates it is set to 1 m s\(^{-1}\), consistent with values in New et al. (1999).

c. How to obtain confidence intervals for the model evaluation measures

The few realizations of simulated climate and the one realization of past climate provide only a small sample of the statistical population of climate model results and observations. For this reason, model evaluation measures are only representative for the samples of model results and observations used in the present study but not necessarily for the basic population. Confidence intervals for the model evaluation measures are calculated by bootstrap resampling (Efron 1979) to get an estimate of their statistical spread with respect to interannual climate variability. By resampling of blocks containing all daily data in one given season in simulated and observed data, \(10^5\) bootstrap resamples are drawn and then the model evaluation measures are calculated for each. In this way, \(10^5\) values are obtained for each model evaluation measure. From these values, the 5th, 50th, and 95th percentiles are calculated and graphically presented in section 5 as the median and the 90% confidence intervals of the model evaluation measures.

When dealing with bootstrap resampling, a major issue is the serial correlation in the data (e.g., Wilks 1995). The higher the serial correlation, the more the confidence intervals are underestimated. By the resampling of blocks containing all daily data in one season and not, for example, the days in the samples of simulated and observed data, the issue of serial correlation is reduced to serial correlations over a lag of one year or more. Even then the serial correlation cannot be ignored (e.g., Kendon et al. 2008). Therefore, to gain confidence in the results achieved, different realizations of simulated climate from the same climate model are evaluated against each other in section 5. If the confidence intervals created by resampling are too small, model evaluation measures obtained in comparing different realizations of simulated climate can be expected to yield low values.

In section 5, a BIAS is called significantly positive if the 5th percentile of the BIAS values obtained by resampling of the data is positive, and significantly negative if the 95th percentile of the BIAS values obtained by resampling is negative.

5. Results

a. Temperature

The climatological annual cycle of temperature is shown in Fig. 4. Differences between the observational datasets (black lines) are small. At first glance simulated annual cycles of temperature (gray lines) agree well with the observed, meaning they reflect the large annual cycle. However, a significant negative BIAS is found in ECHAM5 for SON and DJF, whereas in CLM a significant negative BIAS appears for JJA and SON. The fact that no significant BIAS is found in REMO is reflected by the values for the “mean absolute monthly difference” (MAMD) whose median values range between 0.70 and 0.75 K for ECHAM5 and 0.75 and 0.82 K for CLM, but only from 0.45 to 0.49 K for REMO (not shown).

Observed and simulated PDFs of temperature are shown for DJF in Fig. 5. The PDFs of the observational data are skewed to the left and therefore large negative daily temperature anomalies occur with a higher probability than large positive daily temperature anomalies. Climate models are able to capture this skewed PDF pattern. Regional climate models CLM and REMO do simulate the major part of the PDF better than ECHAM5, but both regional climate models produce a peak in the PDF at 0°C. This unrealistic peak can cause
problems when dealing with the data (e.g., calculation of the number of frost days). The problem of the peak at 0°C is present in results from older versions of CLM and REMO but it is eliminated in the new versions (e.g., the REMO version used to create the REMO-3 data). REMO-3 is only shown in Fig. 5. For DJF temperatures, the SSC and the HRP (Fig. 6) are of limited help in the evaluation. SSC is nearly the same for the three models because a slightly better simulation of the shape of the PDF by the RCMs when compared with ECHAM5 cancels out the negative effect of the peak at 0°C. HRP yields higher values for the regional climate models than for ECHAM5 because the peak at 0°C does not cause many percentiles of the PDF to be shifted by more than the allowed deviation. This example shows that singularities in climate data that may appear are not always detectable by model evaluation measures. The example also shows that a general improvement of simulated climate by the regional climate models compared with the global climate model is not always found.

b. Precipitation

Six observational datasets are available to evaluate monthly precipitation sums: four datasets based on gauge measurements without undercatch correction, one dataset with corrected precipitations, and the ERA-40 reanalysis data. Climatological annual cycles of all observations are shown in Fig. 7. All uncorrected datasets exhibit nearly the same annual cycle. The reasons are the large number of stations in the evaluation domain and the good spatial representativeness of monthly precipitation sums. Therefore, the interpolation error for gridded monthly precipitation sums is small. The corrected dataset (dash–dotted line) differs much more from the four uncorrected datasets than the four uncorrected datasets differ among themselves. Even if the correction factors provided by GPCC are large relative to correction factors derived by other authors, this example shows that using different observational datasets capturing only one type of uncertainty (in this case the interpolation uncertainty) can lead to bad interpretations. ERA-40 reanalysis data show an underestimation of precipitation that is especially large in the summer season, consistent with Uppala et al. (2005). As precipitation overcatch by rain gauges is very improbable, the ERA-40 precipitation...
data are assumed to be too low and are not used for model evaluation purpose.

Simulated precipitation lies between corrected and uncorrected observational data except from late spring to late summer. In JJA, ECHAM5 suffers from a significant underestimation of precipitation, whereas both regional climate models show a significant overestimation. The overestimation found for JJA is different from what Feldmann et al. 2008 report for a mountainous region in southwestern Germany: they found no overestimation of precipitation during JJA in the same regional climate model runs. Figure 7 also shows that the decision whether a climate model has a positive or negative precipitation BIAS in DJF mainly depends on the choice between corrected and uncorrected precipitation data. In impact studies problems might arise when hydrological models are calibrated with uncorrected precipitation data. An evaluation of daily precipitation is not shown in the present study as no corrected daily precipitation data are available. Furthermore, as pointed out by Bohnenstengel et al. (2011), the spatial representativeness of daily precipitation data is very bad and thus comparison of daily values would need to consider large uncertainties.

c. Total cloud cover

The climatological annual cycle of total cloud cover is shown in Fig. 8. In the observations (black lines), total cloud cover has a maximum in winter and a minimum in summer. This pattern is reproduced by the climate models, but in CLM (gray line with dots) the amplitude of the annual cycle is too low, resulting in a significant positive BIAS in cloud cover in JJA and SON. In ECHAM5 (gray line), a significant positive BIAS is found in MAM. No significant BIAS is found in REMO results (gray line with crosses).

The differences between the observational datasets are quite large. Therefore, the same model evaluated against different observational datasets yields quite different values for MAMD (Fig. 9). The large BIAS of CLM in JJA and SON causes higher MAMD values for CLM than for REMO. The too-high cloud covers simulated by CLM for the period May to September might be the physical reason for the underestimation of JJA and SON temperatures in this model.

PDFs of JJA daily average total cloud cover are shown in Fig. 10. In ECHAM5 and CLM, days with total cloud cover above 90% and more are too frequently simulated, while intermediate total cloud covers are underestimated. REMO is able to reproduce the observed PDF. The SSC and the HRP confirm this subjective impression (Fig. 11). However, the SSC is much less selective than the HRP. For example, the PDF simulated by CLM still obtains a median SSC of 0.64 when compared to the ERA-40 reanalysis data and an SSC of 0.67 when compared to the ROD data. In application studies, the simulated values of total cloud cover from ECHAM5 and CLM might lead to unrealistic results, especially in JJA and SON. REMO results are within the allowed deviation.

d. Relative humidity

The climatological annual cycle of relative humidity (Fig. 12) is characterized by a maximum in winter and a minimum in summer. This feature is covered by the climate models, but the simulated relative humidity (gray lines) is generally larger than the observed relative humidity (black lines). A significant positive BIAS is
found for all models and all seasons except for SON in ECHAM5 and in both transition seasons in REMO.

PDFs of relative humidity in JJA are presented in Fig. 13, with model evaluation measures in Fig. 14. The observed PDF is reproduced well by ECHAM5 whereas both regional climate models have deficiencies. This behavior is captured by both model evaluation measures, but the HRP is much more subject to interannual climate variability than the SSC, especially for REMO. The reason is that the whole PDF simulated by REMO is shifted by a constant offset toward higher values of relative humidity compared with observations. The offset has about the magnitude of the allowed deviation. In such cases, resampling of the data can lead to a wide range of values for the HRP, a behavior that is the trade-off for its higher selectiveness.

For the evaluated climate model results, the use of uncorrected values of relative humidity might lead to unrealistic results in impact studies. The fact that CLM simulates too high a frequency of high relative humidity is physically consistent with the too-high frequency of high total cloud covers. This type of physical consistency is not found for ECHAM5: the PDF of relative humidity is simulated well, while the PDF of total cloud cover is not.

e. Wind speed

For wind speed, the evaluation period is reduced to 1981–2000 because the number of station records in the ROD dataset available before 1980 is low. Average wind speed is higher in DJF than in JJA because of the higher number of passing low pressure systems in DJF. Climate models reproduce this annual cycle, with slightly smaller annual amplitude (Fig. 15). ECHAM5 and CLM exhibit a significant positive BIAS in JJA, whereas REMO has a significant negative BIAS in DJF. The variance of the time series of monthly averages is underestimated in both RCMs: the median value of the ratio of temporal variances (RATV) measure ranges from 0.47 to 0.61 for CLM and from 0.39 to 0.50 for REMO (not shown). This was not found for the other meteorological parameters. For ECHAM5 there is only a slight underestimation of wind speed variance with a RATV from 0.69 to 0.89. This example shows that even for monthly average values, RCMs can produce different characteristics of the time series than the forcing GCM.

In all four seasons, the PDFs of daily averages are well simulated by the models. As an example the PDF for
JJA is shown in Fig. 16 and the corresponding model evaluation measures in Fig. 17.

For wind speed, the results obtained from the evaluation of the climatological annual cycle and from the evaluation of the PDFs of daily averages are different. In the first case significant BIASes are found, while in the latter case model evaluation measures indicate that the PDFs are within the uncertainty due to interannual climate variability and observational data. The reason for this difference in results is that the BIAS does not account for the uncertainty of the observations, whereas the HRP does. Also, the described underestimation of monthly variability can be the reason why the comparatively small BIASes become significant. The conclusion of wind speed evaluation is that the amplitude of the annual cycle is slightly underestimated by the climate models, but because simulated PDFs do not differ that much from observed PDFs for any model or season, a bias correction of climate model results is not recommended.

f. Influence of long-term climate variability

The impact of long-term climate variability on the evaluation results is investigated by the calculation of the HRP for one realization against another realization of control climate simulations of the same climate model. For example, the PDF of JJA temperatures from CLM-1 is compared to the PDF of JJA temperatures from CLM-2 and from CLM-3.

The median HRP values range from 0.97 to 1.00 for total cloud cover, from 0.96 to 1.00 for relative humidity, from 0.99 to 1.00 for wind speed, and from 0.87 to 1.00 for temperature. For temperature, larger deviations of the HRP from the optimum value 1 are observed, the largest in DJF and MAM. This result is not surprising, as time series of temperature can be expected to be subject to higher serial correlation than time series of the other meteorological parameters evaluated in the present study. However, for temperature, the HRP also yields high values. Evaluation of model results from different realizations of simulated climate from the same climate model give a hint that the issue of long-term climate variability is not as important for model evaluation as for the detection of climate change signals. The main reason is that the allowed deviation used for HRP calculation considers not only the interannual climate variability but also the uncertainty of the observational data. Therefore, systematic too-low confidence intervals...
due to serial correlation are not as important as in the case where only the uncertainty due to climate variability is investigated. The last is relevant for detecting climate change signals.

It has to be mentioned that the number of realizations available for testing the influence of climate variability is very small and the three available realizations might not capture the whole spectrum of climate variability. A larger number of realizations would therefore be needed for detailed studies of simulated climate variability.

6. Usage of model evaluation measures to decide if bias correction is reasonable

In this section, an example of using model evaluation measures as a basis to decide if bias correction is reasonable is presented. Hoffmann et al. (2012) developed a statistical model for the urban heat island (UHI) of Hamburg in northern Germany. The statistical model is based on relationships between the nocturnal UHI and daily averages of total cloud cover, relative humidity, and wind speed. For 1971–2000, considerable differences between observed UHI and statistically calculated UHI appear when using the CLM results as input for the statistical model (Fig. 18). Evaluation results presented in section 5 show that significant biases are found in CLM results for total cloud cover as well as for relative humidity. Bias correction of CLM output seems a promising approach for obtaining better results from the statistical model for present day climate.

The UHI climatology was determined for every month of the year by Hoffmann et al. (2012). To use the same approach with bias-corrected data, a bias correction needs to be done for the monthly PDFs of relative humidity and total cloud cover. Here, a bias correction is recommended if the HRP values are below 0.95 for comparison of model results with both observation based datasets. The statistical bias correction method that we apply is described in Piani et al. (2010a). Instead of the gamma distribution used for precipitation by Piani et al. (2010a), a beta distribution is fitted to both observations and model output. The beta distribution is the theoretical PDF that is best suited to fit the empirical PDFs of cloud cover and relative humidity (e.g., Wilks 1995). The domain of definition of the beta distribution between 0 and 1 helps to assure that corrected cloud cover and relative humidity are in the range between 0% and 100%.

FIG. 16. PDFs of simulated (gray lines) and observed (black lines) daily average wind speed in JJA for the period 1981–2000.

FIG. 17. Model evaluation measures SSC and HRP for wind speed in JJA. Two observational datasets are used for evaluation.

FIG. 18. Climatological annual cycle of the UHI of Hamburg calculated with the statistical model described in Hoffmann et al. (2012). In black: statistical model driven with ROD (crosses) and ERA-40 reanalysis data (dots). In gray: statistical model driven with original (squares) and with bias-corrected (circles) CLM data.
After the bias correction, the model output does not agree exactly with observations because the theoretical PDFs fitted to the data do not exactly match the empirical PDFs. An important question is, therefore, if the bias correction really has improved the data. This is quantified by the HRP, the most selective measure for the evaluation of daily averages. The HRP values for relative humidity and for total cloud cover are displayed before and after the bias correction in Tables 3 and 4. After the bias correction, the HRP is above 0.95 for every month for both meteorological parameters. This shows that bias correction sufficiently improves the CLM output. As the correction employed is different for every month of the year, unrealistic jumps in the time series of bias-corrected data might appear (Piani et al. 2010b). As these effects are not important for the results obtained in Hoffmann et al. (2012), the authors chose not to apply special transition functions to remove month-to-month jumps in the bias-corrected time series.

The statistical model of the UHI driven with bias-corrected CLM output (Fig. 18) yields much better results for the UHI than those with uncorrected data. The ERA-40 reanalysis data have been used as a reference dataset for the derivation of the bias correction. It should be noted that the climate change signals for the UHI of Hamburg do not change considerably with or without bias-corrected data (not shown). However, the bias correction presented in this section does not conserve the physical consistency of meteorological data. For the statistical model of the UHI this is not a serious constraint, but in the case of a dynamical downscaling with a complex high-resolution numerical model, physical inconsistencies in the bias-corrected data could lead to problems.

7. Conclusions

In the present article, temperature, precipitation, total cloud cover, relative humidity, and wind speed from the GCM ECHAM5 and from the RCMs CLM and REMO have been evaluated for the metropolitan region of Hamburg. All models are able to reproduce general characteristics of the climatologies. The best model performance was found for wind speed, the weakest model performance for relative humidity. Summer precipitation is underestimated by the GCM and overestimated by both RCMs. Winter precipitation is simulated well; it lies in between corrected and uncorrected precipitation data. Total cloud cover is simulated correctly by REMO and shows strong biases in ECHAM5 and CLM results. Temperature PDFs are generally well simulated, with the exception of a peak at 0°C in the RCM data used in the present study, but removed in the more recent versions of the RCMs. The RCM results mostly agree better with observations than do the results from the forcing GCM. Thus they improve the GCM result and mostly provide additional value. However, the improvement is not found for all seasons and all meteorological variables as, for example, for relative humidity in JJA.

Model evaluation measures used for the evaluation of monthly and daily averages can help to objectify subjective impressions when doing climate model evaluation. The newly introduced hit rate of the percentiles is more selective than the skill score, but also more sensitive toward climate variability. HRP can be used to decide if bias correction can be recommended from the viewpoint of model evaluation. Whether or not bias correction is useful depends on the application. Furthermore, the evaluation measures can be used for evaluation of the correction itself. Evaluation of different realizations of simulated climate from the same climate model against themselves shows that at least for the small sample of available realizations, the issue of long-term climate variability has no large impact on evaluation results. The issue of long-term climate variability can be expected to be larger when focusing on the detection of climate change signals. A large ensemble of realizations of simulated climate would be needed to quantify the

<table>
<thead>
<tr>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.73</td>
<td>0.68</td>
<td>0.45</td>
<td>0.21</td>
<td>0.31</td>
<td>0.12</td>
<td>0.10</td>
<td>0.22</td>
<td>0.30</td>
<td>0.86</td>
<td>0.97</td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
</tr>
</tbody>
</table>

TABLE 4. Model evaluation measure HRP for total cloud cover simulated by CLM compared with the ERA-40 reanalysis data before and after bias correction.
influence of climate variability for both model evaluation and trend detection results with higher accuracy.

Last but not least it has to be mentioned that the climate model evaluation presented in the present article is based on the evaluation of the statistics of time series and is done from the viewpoint of a climate model data user. From the viewpoint of a climate model developer, a more detailed analysis of the representation of physical processes or of single weather types that occur might be a promising model evaluation strategy.

Acknowledgments. The present work was performed in the framework of the project KLIMZUG-NORD funded under Grant 01LR0805D by the German Federal Ministry of Education and Research.

The ENSEMBLES data used in this work were funded by the EU FP6 Integrated Project ENSEMBLES (Contract 505539), whose support is gratefully acknowledged.

Thanks also go to the Deutscher Wetterdienst, who kindly provided routine measurements and the REGNIE dataset. The present work is supported in parts through the Cluster of Excellence “CliSAP” (EXC177), University of Hamburg, funded through the German Science Foundation (DFG). Felix Ament is acknowledged for discussions on uncertainties of observational datasets. Climate model data evaluated in the present study were downloaded from the “World Data Center for Climate” database (http://ecera-www.dkrz.de) with the exception of the third realization of present climate simulated by REMO (REMO-3), which was directly provided to the project KLIMZUG-NORD.

The REMO-1 simulation (Jacob 2010; Jacob et al. 2008) was commissioned by the German Federal Environment Agency (UBA), and the REMO-2 simulation (Jacob et al. 2010) was commissioned by the German Federal Institute of Hydrology (BfG). REMO-3 was created within the project KLIMZUG-NORD. The CLM-1, CLM-2, and CLM-3 runs were created at the MPI-M in the so-called consortium runs (Hollweg et al. 2008; Lautenschlager et al. 2010a,b) commissioned by the Federal Ministry of Education and Research, BMBF. The ERA-40 data have been downloaded from http://www.dwd.de/webwerdis; last accessed on 6 May 2010.


The present work was performed in the framework of the project KLIMZUG-NORD funded under Grant 01LR0805D by the German Federal Ministry of Education and Research.

The ENSEMBLES data used in this work were funded by the EU FP6 Integrated Project ENSEMBLES (Contract 505539), whose support is gratefully acknowledged.

Thanks also go to the Deutscher Wetterdienst, who kindly provided routine measurements and the REGNIE dataset. The present work is supported in parts through the Cluster of Excellence “CliSAP” (EXC177), University of Hamburg, funded through the German Science Foundation (DFG). Felix Ament is acknowledged for discussions on uncertainties of observational datasets. Climate model data evaluated in the present study were downloaded from the “World Data Center for Climate” database (http://ecera-www.dkrz.de) with the exception of the third realization of present climate simulated by REMO (REMO-3), which was directly provided to the project KLIMZUG-NORD.

The REMO-1 simulation (Jacob 2010; Jacob et al. 2008) was commissioned by the German Federal Environment Agency (UBA), and the REMO-2 simulation (Jacob et al. 2010) was commissioned by the German Federal Institute of Hydrology (BfG). REMO-3 was created within the project KLIMZUG-NORD. The CLM-1, CLM-2, and CLM-3 runs were created at the MPI-M in the so-called consortium runs (Hollweg et al. 2008; Lautenschlager et al. 2010a,b) commissioned by the Federal Ministry of Education and Research, BMBF. The ERA-40 data have been downloaded from http://www.dwd.de/webwerdis; last accessed on 6 May 2010.
