Vol. 61: 41–56, 2014
doi: 10.3354/cr01242

CLIMATE RESEARCH
Clim Res

Published September 2

FREE ACCESS

A height dependent evaluation of wind and temperature over Europe in the CMIP5 Earth System Models

Annemarie Devis*, Nicole P. M. van Lipzig, Matthias Demuzere

Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E, 3001 Heverlee (Leuven), Belgium

ABSTRACT: To date, evaluation studies of global circulation models (GCMs) generally focus on large-scale circulation variables. However, since the resolution of GCMs is increasing and their interaction with the surface is improving, GCM representations of near-surface variables are becoming increasingly realistic. For downscaling practices, this implies that near-surface variables might become suitable as predictors in statistical models, and might increase the added value of dynamical models. This study focuses on the representation of wind and temperature in the lowest 1.5 km of the atmosphere over Europe as simulated by 6 of the earth system models (ESMs) from the Coupled Model Intercomparison Project Phase 5. The evaluation is based on the representation of the variables' probability density functions (PDFs) using ERA-Interim reanalysis data as the reference. The PDF biases are analyzed according to their scale and origin. Above coastal bays and capes, small-scale biases in the ESMs result in unskillful wind speed PDFs up to 400 m. High orography affects wind speeds throughout the lowest 1.5 km of the atmosphere, especially during summer and night. Apart from these small-scale biases, the surface wind speed PDFs north of 45° N are well represented by all the ESMs. Therefore, these PDFs can be considered skillful inputs for statistical downscaling practices. South of 45° N, winds are affected by a large-scale bias originating from errors in the representation of the large-scale circulation, especially during winter. For temperature, near-surface levels as well as upper-atmospheric levels are affected by small-scale and large-scale biases. Large-scale biases are adopted by the downscaling models, underlining the importance of model evaluation before downscaling.

KEY WORDS: Earth System Models · Evaluation · Probability density function · Wind speed · Temperature · Near-surface · CMIP5

1. INTRODUCTION

The latest realizations of general circulation model (GCM) experiments are available from the fifth phase of the Coupled Model Intercomparison Project (CMIP5). Some of these experiments are conducted using earth system models (ESMs), which include earth system components and processes (e.g. terrestrial and ocean carbon cycles) and thereby account for multiple (positive and negative) feedbacks in the overall system (Taylor et al. 2012).

The CMIP5 experiments are frequently used for future projections and downscaling purposes (Jones & Carvalho 2013, Knutson et al. 2013, Roehrig et al. 2013). In a downscaling approach, the GCM data are rescaled to a finer resolution. A dynamical downscaling uses the GCM data at the lateral boundaries to drive a higher-resolution regional climate model (RCM), while a statistical model is based on the relation between the large-scale GCM data and small-scale observations. Since GCM data are used as input variables in downscaling models, the accuracy of the downscaling result to a large extent depends on the quality of the GCM (Wilby et al. 1998). Small-scale biases in the GCMs related to the representation of the small-scale features can be overcome by

*Corresponding author: annemarie.devis@ees.kuleuven.be

the downscaling (Gómez-Navarro et al. 2011, Schmidli et al. 2006); however, biases at larger scale will affect the downscaling practice. In a dynamical downscaling, the large-scale bias will be carried over to the regional model (Flaounas et al. 2013), while in a statistical downscaling, the large-scale bias is often accounted for using bias-corrections (Ehret et al. 2012). Because there is no 'one best GCM' (Gleckler et al. 2008, Knutti et al. 2010), users of downscaling practices need to know which variables and which GCMs are suitable for which purposes.

During recent years, the methodology for model evaluation has tended to shift from the use of statistical performance metrics (Gleckler et al. 2008), measuring the ability of the GCM to simulate the mean and/or standard deviation, to a more probability density based approach, assessing the performance of the GCM to represent the entire distribution (Schoetter et al. 2012). Also, regime-oriented approaches have been developed by dividing data into categories that describe physically distinct regimes to identify processes that might be responsible for particular errors (Jakob 2010, Huth et al. 2008, Demuzere et al. 2009).

Recent published work on the performance of CMIP5 experiments evaluated the simulation of the free tropospheric circulation, temperature and humidity in ESMs (Brands et al. 2013), the mean and extreme temperatures over Europe in ocean–atmosphere coupled models (Cattiaux et al. 2013a) and the simulation of El Niño in the Tropical Pacific Ocean (Yang & Giese 2013). Collins et al. (2011) and Dufresne et al. (2013), respectively, focused on the evaluation of HadGEM2 and IPSL-CM5. Generally, the outcomes of CMIP5 evaluations are in line with the results based on CMIP3 simulations. They indicate overly strong westerlies in the Northern Hemisphere mid-latitudes during winter (Brands et al. 2013, van Ulden & van Oldenborgh 2005, Vial & Osborn 2012) and overly warm (cold) summer temperatures in Central and Eastern (Western) Europe (Cattiaux et al. 2013b).

Circulation-based evaluation practices generally focus on mean sea-level pressure and upper-atmospheric levels (500 to 850 hPa), because GCMs are developed to represent the large-scale circulations and because these levels are commonly used for downscaling practices. However, in recent years, GCMs have become more sophisticated by incorporating more components and feedbacks, and the resolution (horizontal and vertical) of the underlying atmospheric models has increased. Consequently, for some practical purposes, the question arises of whether downscaling is still beneficial and if statistical downscaling models would not profit from predictors selected from lower level variables, instead of the commonly used 850 to 500 hPa atmospheric large-scale circulation variables. Devis et al. (2013) showed that depending on atmospheric conditions, lower level variables can be suitable in statistical downscaling models. By making use of variables describing the lower atmosphere, instead of the large-scale circulation variables, less uncertainty will be added through the statistical downscaling implementation when modeling near-surface conditions.

This study provides an evaluation of 6 of the newly available CMIP5 ESM datasets for temperature and wind speed over Europe. The evaluation is performed on all model levels from the lowest model level up to approximately 1500 m (~850 hPa). The levels above 1500 m have been studied in previous evaluations (Brands et al. 2013). The evaluation focuses on the representation of the probability density function (PDF) using the skill score developed by Perkins et al. (2007).

Because of its multi-level approach, this paper provides insight into the vertical dependency of the performance of the model, and defines the vertical level down to which ESMs are skillful in representing the reanalysis conditions concerning their resolution. In addition to the vertical dependency of the model bias, attention is also given to the spatial and seasonal dependency of the performance of the ESMs.

The area under study is presented in Fig. 1. The domain covers flat and orographic land, open and closed sea and coastal areas within 27 to 72° N, 32° W to 35° E. The climate of the study area is mostly controlled by the unstable nature of the North-Atlantic dynamics (Collins et al. 2011): westerlies carry moist air from the Atlantic, while easterlies bring cold (warm) continental air in winter (summer) (Brands et al. 2011a).

2. DATA

2.1. Earth System Models

The evaluation is performed on 6 ESMs of the CMIP5 project, the model realisations used for the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (AR5). The considered ESMs are listed in Table 1 together with their resolution specifications. These ESMs have already been considered by Brands et al. (2012), with the exception of MPI-ESM-LR, which provides data only at pressure

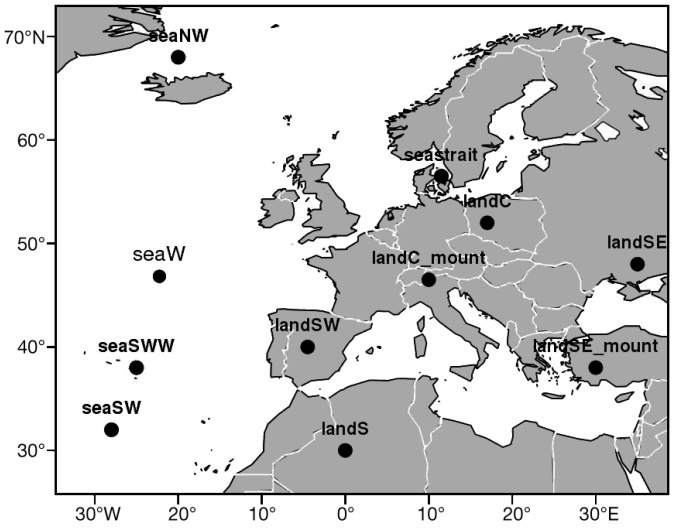

Fig. 1. Study area with specific locations indicated

levels and not at model level resolution. ESM data are obtained from the Earth System Grid Federation (ESGF) platform gateways of the BADC (British Atmospheric Data Center) node and the DKRZ (German Climate Computing Center) node. This study uses data from the historical CMIP5 project experiment, which imposes changing conditions (consistent with the observations) of atmospheric composition (including CO_2) due to the following factors: land-use, anthropogenic and volcanic influences, solar forcing, emissions or concentrations of short-lived species, and natural and anthropogenic aerosols or their precursors (Taylor et al. 2012). The historical ensemble experiment offers long-term data from 1850 to 2005. Instantaneous 12:00 and 00:00 h UTC temperature and east- and westward wind records are obtained from 1979 to 2005 (the reference data are only available since 1979), at model levels up to 1.5 km.

2.2. ERA-Interim

The ERA-Interim data produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Dee et al. 2011) are used as a reference. Instantaneous 12:00 and 00:00 h UTC temperature and east- and westward winds are obtained for the model levels up to 1.5 km. The data have a horizontal resolution of approximately 0.75° × 0.75°, which for the study area corresponds to a resolution of ~50 km in meridional direction and ~80 km in zonal direction.

Brands et al. (2013) analyzed the uncertainty of reanalysis data by comparing ERA-Interim with JRA-25 data (Onogi et al. 2007) for Europe and Africa. The reanalysis uncertainty of the 850 hPa level temperature and east- and westward wind components is very small over Europe. The largest uncertainty is found in the 2 m temperature in winter in the south of Europe, where JRA-25 is systematically warmer than ERA-Interim. For land areas north of 45° N during winter and spring, the differences are negligible.

2.3. Data treatment

Common GCM evaluation methods generally aggregate all GCMs to one standard grid as it enables the direct comparison of the results of the various models. However, bringing all GCMs to one standard resolution hinders the evaluation of higher-

Table 1. CMIP5 ESM models included in the presented evaluation study. The model level heights shown in this table refer to the average of all model level heights over the entire domain and time period

Model (abbreviation)	Modeling center	Horizontal resolution	Model level heights (m; below 1.5 km)	Land sea mask types	Reference
MIROC-ESM (MIROC)	MIROC	2.8 × 2.8°	44; 166; 420; 858; 1503	Land, sea	Watanabe et al. (2011)
CanESM2 (CanESM)	CCCma	2.8 × 2.8°	42; 126; 232; 351; 471; 592; 715; 846; 999; 1185; 1411; 1683	Land, sea	Chylek et al. (2011)
NorESM1-M (NorESM)	NCC	1.5 × 1.9°	66; 247; 600; 1163; 1926	Land, coast, sea	Kirkevag et al. (2008), Seland et al. (2008)
IPSL-CM5-MR (IPSL)	IPSL	1.5 × 1.27°	37; 113; 219; 369; 579; 867; 1247; 1736	Land, coast, sea	Dufresne et al. (2013)
HadGEM2-ES (HADGEM)	MOHC (additional realizations by INPE)	1.875 × 1.25°	20; 80; 179; 320; 500; 720; 980; 1279; 1619	Land, coast, sea	Collins et al. (2011)
CNRM-CM5 (CNRM)	CNRM-CERFACS	1.4 × 1.4°	34; 145; 347; 618; 948; 1324; 1741	Land, coast, sea	Voldoire et al. (2013)

resolution models based on their maximal performance. Since this study focuses in particular on the near-surface representation of the ESMs, which is expected to be resolution-dependent, all ESMs are evaluated based on their original resolution specifications. The reference data are aggregated to the horizontal grid of each ESM and interpolated to the vertical levels of each ESM separately. The horizontal aggregation of the reanalysis grid uses the area-weighted average method. This method calculates the aggregated grid cell as the weighted average of the ERA-Interim grid cells which are covered by the ESM grid cell, proportional to their overlapping area with the ESM grid cell.

Prior to the interpolation of the reference data to each ESM model level height, the model level heights of the ESMs and the reanalysis datasets are calculated. This calculation is based on the hypsometric equation using the surface pressure and the temperature and vertical coordinates of the model levels. The interpolation is performed for each of the ESMs, for each time step and each grid cell separately, and thereby minimizes interpolation errors. Temperature is interpolated using a linear profile, while the interpolation of the wind speed is based on the power law. This is a commonly used approach for wind (Archer & Jacobson 2005, Pryor et al. 2005a), and is typically written as follows:

$$U(z) = U_{\mathrm{ref}} \cdot \left(\frac{z}{z_{\mathrm{ref}}} \right)^{\alpha} \quad (1)$$

For each z-level, α is calculated by using the 2 closest model levels that bracket the z-level (Devis et al. 2013):

$$\alpha = \frac{\ln[U(z+1)/U(z-1)]}{\ln[(z+1)/(z-1)]} \quad (2)$$

In this way, the variability of α in time and height due to changes in surface roughness and atmospheric stability is taken into account (Holt & Wang 2012).

ESM and reference data are split in the following 4 subsets: summer-day, summer-night, winter-day and winter-night, which will hereafter be abbreviated as D_S, N_S, D_W and N_W, respectively. Summer is composed of the months May to September; winter covers the months November to March. Day and night are respectively 12:00 h UTC and 00:00 h UTC. The model evaluation is performed on each subset separately to analyze the diurnal and seasonal dependency of the model performance. For some analyses, day and night samples for all 10 months are combined in what we refer to as the entire dataset.

3. METHODOLOGY

This study evaluates the models based on their ability to represent the PDFs of wind and temperature at climatic time scales. The evaluation is based on the simple and robust score presented by Perkins et al. (2007), which will hereafter be denoted as the PDF score. The score has been frequently used in recent publications on model evaluation (Maxino et al. 2008, Pitman & Perkins 2009, Mao et al. 2010, Brands et al. 2011a, 2011b, Kjellström et al. 2010, Devis et al. 2013).

The score accounts for the similarity between the ESM's PDF and the reference's PDF, as it calculates the cumulative minimum value of 2 distributions for each binned value. PDF scores close to 0 indicate negligible overlap between the reference and the model PDF, while a score of 1 indicates 2 identical PDFs. Apart from precipitation-like mixed probability distributions with clustering around zero, the PDF score has a near perfect linear relation to the often used, but less evidently interpretable, Kolmogorov-Smirnov (KS) test (Brands et al. 2012). Perkins et al. (2007) consider variables with a PDF score below 0.7 as poorly simulated. Indicative PDFs for wind speed and temperature reflecting the 0.7 PDF score are shown in Fig. 2. Performing a downscaling on ESMs with a PDF score below 0.7 would require a large bias-correction. It must be recognized that yielding a PDF score of 0.7, which is applied as threshold in this study, does not necessarily imply that the ESM performs well. For example, in the context of statistical downscaling, a PDF score of 0.7 does not necessarily mean that the GCM performs sufficiently well to be downscaled without prior correction of the bias (Demuzere et al. 2009, Brands et al. 2011). The 0.7 PDF score is stated in this study as the threshold that should be achieved by the ESM data before further using the data in downscaling studies.

The PDF score is preferable to other types of validation methods because it is independent of the distribution. The score is suitable to compare temperature, which is considered to be Gaussian, and wind speeds, which are Weibully distributed. Moreover, the PDF score accounts for errors along the whole distribution (Brands et al. 2011a). In this way, not only the performance of the model to reproduce the mean state but also the frequency of occurrence of rare values is evaluated. On the other hand, the score does not indicate if an error is due to either an over- or underestimation. Therefore, in this study, the PDF score is used to identify remarkable regions (with either bad or good performance), which are then

looked at in more detail by comparing the reference and modeled PDFs in anomaly plots.

The PDF score is calculated for each individual grid cell, model level up to 1.5 km and subset (D_S, N_S, D_W and N_W). Based on the land-sea mask index of the grid cell, the PDF scores are divided into land, coast and sea.

4. RESULTS

The description of the results differentiates between biases acting at small scales and those acting at large scales. Small-scale biases refer to biases acting at individual grid cells, often related to coastal or orographic effects. These are the biases that are related to the resolution of the ESMs and are expected to be altered by downscaling. In contrast, large-scale biases affect a larger region and are less prone to coastal and orographic effects but are susceptible to biases in atmospheric circulation or other large-scale factors. Large-scale biases, as defined in this paper, can originate in the upper-atmosphere or from interactions with the surface. These are the biases that are independent from the resolution of the ESMs and are expected to be retained in the results of the downscaling exercises.

4.1. Wind speed

4.1.1. Performance near the surface

The PDF scores are calculated on the time series of each grid cell and presented in violin plots for the model level closest to 80 m (Fig. 3). Violin plots show the frequency distribution of the PDF scores over land, sea and coastal grid cells. In general, near-surface winds are better simulated over sea than over land and coast and better during day than during night. Also, the difference in average performance between summer and winter is fairly small. ESMs with finer spatial resolutions tend to perform better than lower-resolution models. This is shown by lower median scores and longer violin tails (referring to the lowest PDF scores) for models with coarser horizontal resolutions. Regarding the vertical resolution, the very fine CanESM model performs notably well near the surface. It must be acknowledged, however, that the relationship between the resolution of the models and their performance cannot be fully established from our analysis since this relation can only be tackled via a comparison of runs with the same model at different resolutions. The differences between the various simulations could also arise from other factors, such as the dynamic core of the models. When focusing on the range of the PDF scores (visualized by the tails of the violin plots in Fig. 3), CNRM arises as the model with the lowest spatial dispersion of reliability.

The exact location of the PDF biases in the near-surface wind speed can be derived from the PDF score maps in Fig. 4. These maps show that apart from the land-sea differences, earth's orography also defines the representation of the near-surface wind speed. The lower the resolution of the ESM, the larger the errors over highly orographic areas and coasts. PDF scores down to 0.4 are found for grid cells covering narrow coastal features, like bays, capes, straits or islands, and for grid cells covering orographic terrain. Apart from these locations, the near-surface wind-speed PDF is well rep-

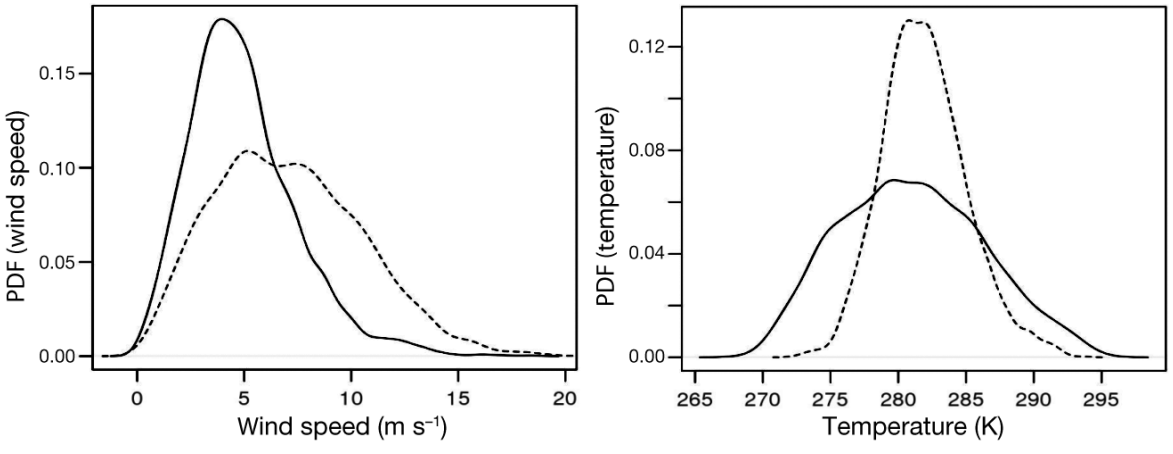

Fig. 2. Examples of probability density function (PDFs) of reanalysis data (solid line) and earth system model data (dashed) for wind speed and temperature, indicative of a PDF score of 0.7

resented for the main part of the domain, also over land. Moreover, the model with the highest horizontal resolution (CNRM) is able to simulate the ~80 m wind-speed PDFs of the entire dataset over the entire domain with a PDF score >0.7. In addition to the small-scale biases, the maps indicate regions with poorer model performance at a larger scale (larger than the individual grid cell scale). Poorly performing large-scale regions are discussed in the following sections, when looking at higher model levels.

4.1.2. Vertical extent of the PDF biases

Large-scale. Figs. 5 & 6 show the lowest level for which all ESMs have PDF scores >0.7 and remain >0.7 up to the highest model level examined (~1500 m/ ~850 hPa) for D_S and D_W respectively. During winter (and to a lesser extent during summer), a large-scale PDF bias is present over the south of Europe (30 to 45° N), most parts of the Mediterranean Sea and parts of the North Atlantic Ocean. To investigate whether this signal is due to a bias in one ESM or to a more common behavior, additional maps of the level at which 50% of the ESMs have a PDF score >0.8 were analyzed (not shown). These maps reveal a similar large-scale feature of low PDF scores over the whole vertical profile for D_W and D_S. The wind PDF bias (according to a 0.8 PDF score threshold) is present in at least half of the ESMs, reaches out over at least 1500 m, and is smaller in summer compared to winter. Because there are well represented levels (with a PDF score > 0.7) below the biased ~1500 m level (indicated by the grey grid cells in Figs. 5 & 6), it is more likely that this PDF bias originates from the atmospheric circulation than through surface interaction effects.

Small-scale. In the south of Europe, the small-scale biases are overshadowed by the east–west stretched large-scale bias zone (Figs. 5 & 6). On the other hand, the surface winds north of 45° N are well represented by all ESMs down to the lowest model level examined (~80 m), with some exceptions for narrow coastal features that are too small to be solved by coarse resolution models (coastal bays, capes, straits or islands). These regions are unskillful up to ~400 m. Winds above high mountain areas are unskillful up to >1500 m in the Alps and the Anatolian highlands.

Based on scatterplots relating the lowest altitude with a PDF score > 0.7 and the grid cell height (not shown), for surface elevation above ~1000 m, the performance of the ESMs decreases with grid cell altitude during D_S and D_W. For grid cells with orography below 1000 m, the grid-cell altitude does not clearly relate to the performance of the ESMs in representing the daytime winds. However, scatterplots for N_S and N_W wind speeds suggest that during the night at lower surface elevations, ESM performance is also influenced by the orography.

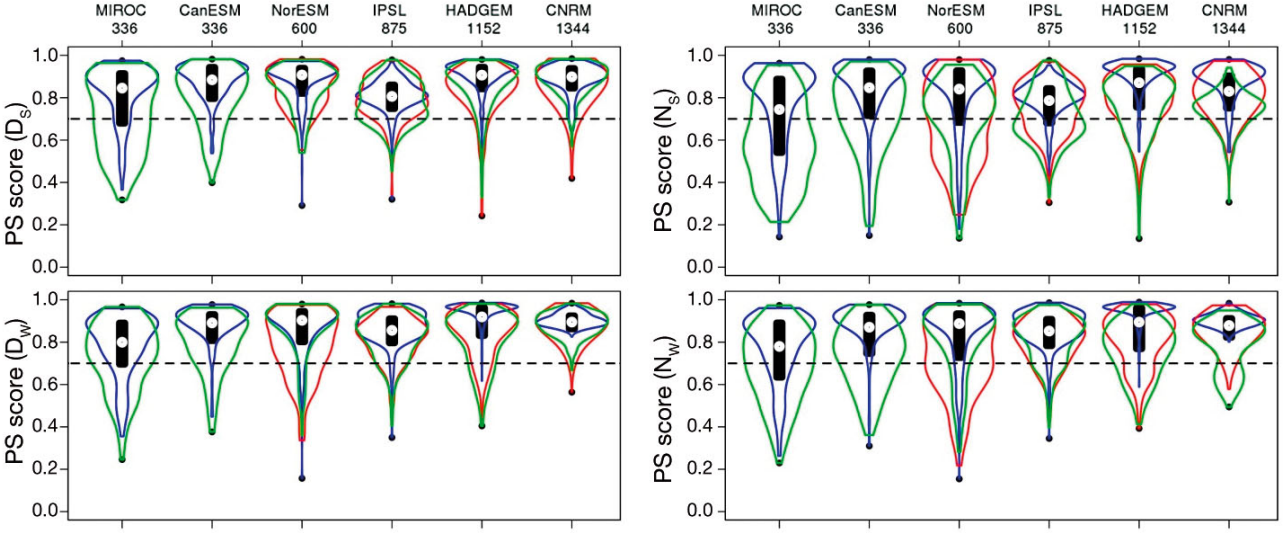

Fig. 3. Violin plots showing the frequency distributions of the probability density function (PDF) scores. The PDF scores are calculated on the time series of the D_S, N_S, D_W and N_W wind speed at ~80 m for sea (blue), coast (red) and land (green) grid cells. The black boxes inside the violin plots represent 50% of all cells, and the white spheres inside the boxes show the median PDF score. Numbers under model names: number of grid cells in the domain. The earth system models are sorted from low (left) to high (right) horizontal resolution (same order as Table 1). Dashed line: 0.7 lower threshold for the PDF score

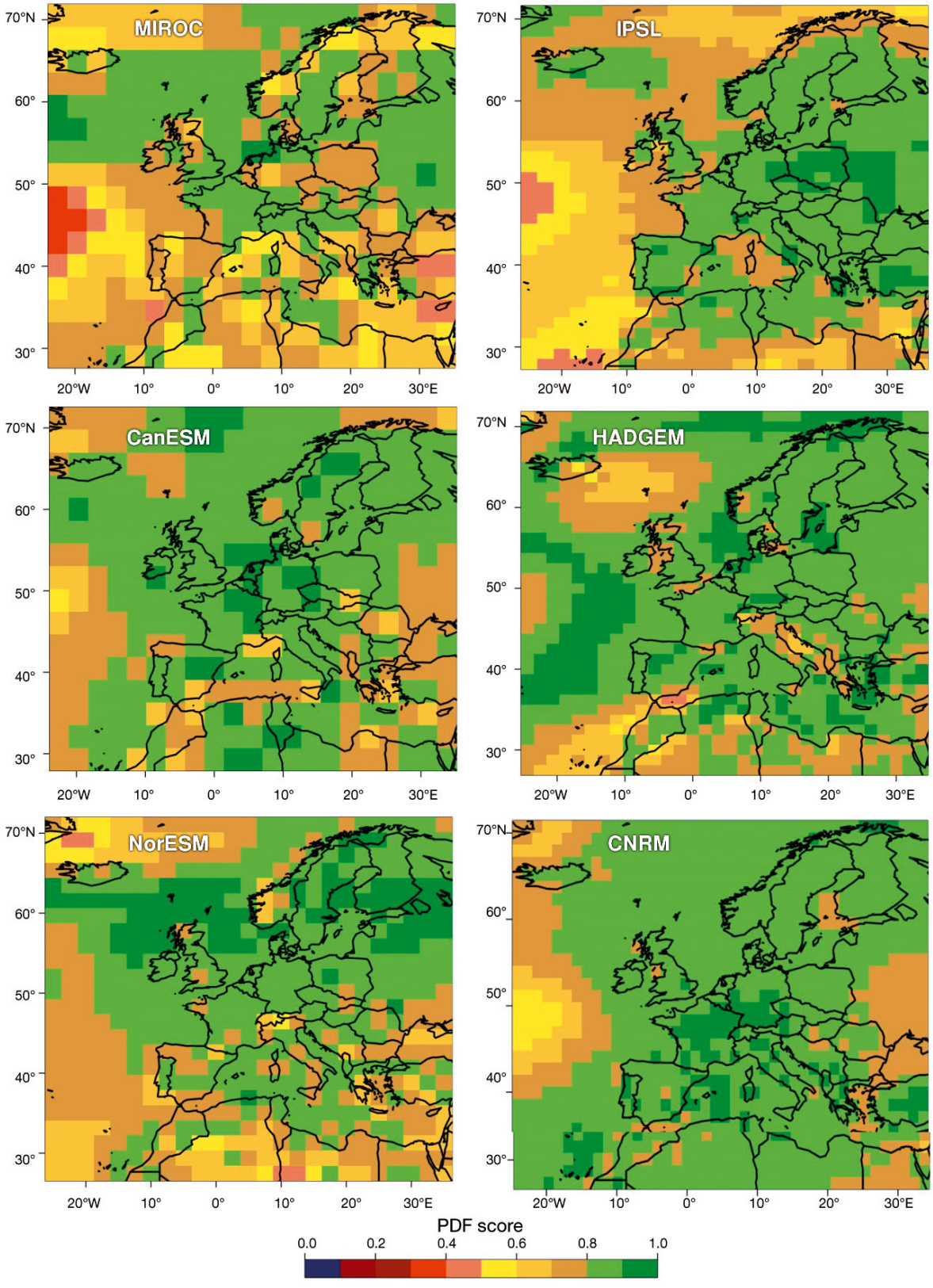

Fig. 4. Probability density function (PDF) scores of the wind speed PDF at ~80 m. The PDF scores are calculated on the time series of the entire dataset. The earth sytem models are sorted from low (top left) to high (bottom right) horizontal resolution

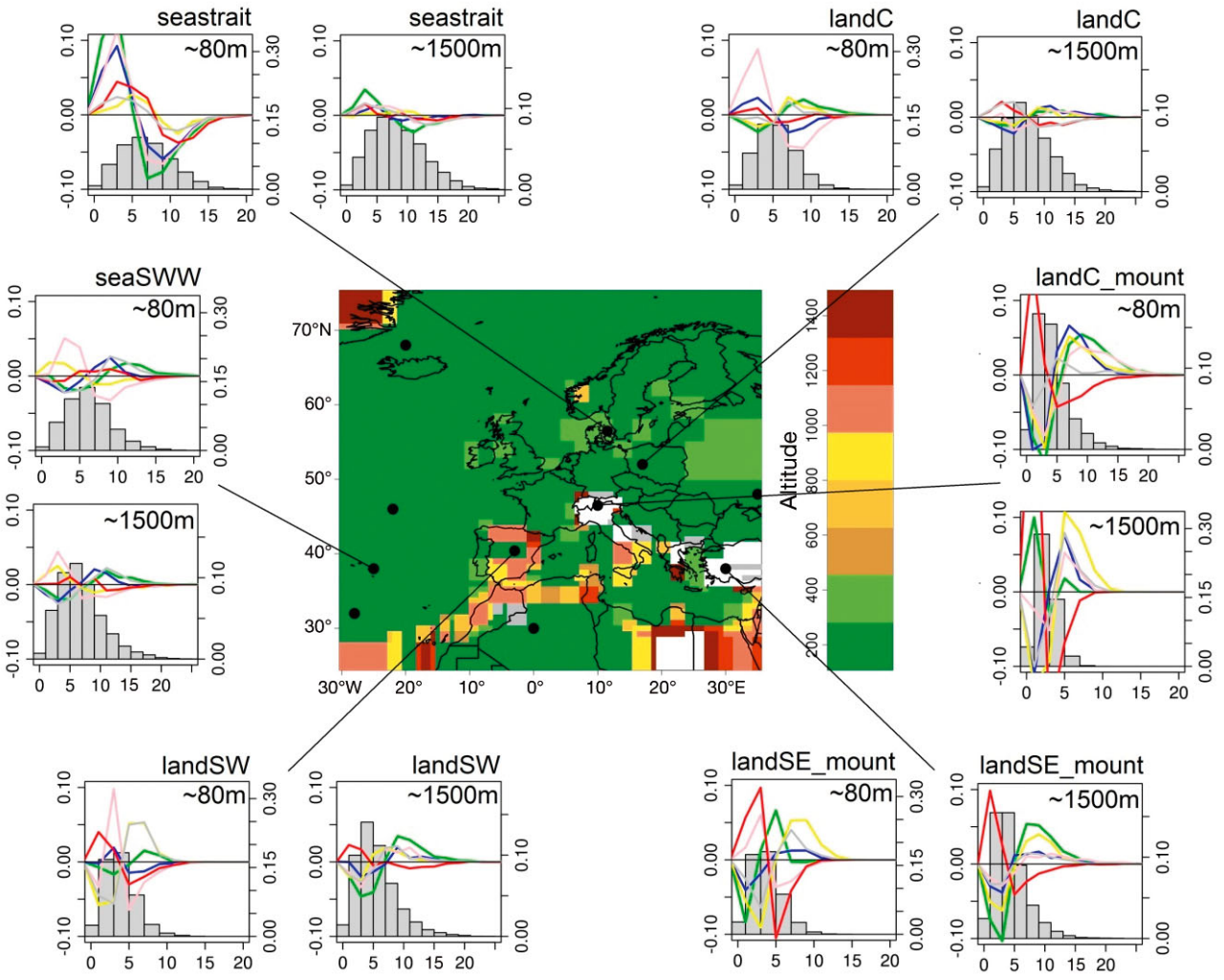

Fig. 5. Map of the altitude of the lowest level for which all the earth system models (ESMs) have probability density function (PDF) scores > 0.7 and remain >0.7 up to ~1500 m in representing the D_S wind speed PDF. Grid cells for which no level is well represented (i.e. no level has PDF score > 0.7 for all ESMs) are plotted in white. Gray grid cells point to the cells which are unskillful at 1500 m but have layers with PDF scores > 0.7 underneath. The graphs surrounding the map depict the probability density for ERA-Interim minus the probability density for the ESM at each bin. The bias at a given bin is hereafter referred to as the anomaly. Anomalies are plotted for the wind speed at ~80 and ~1500 m for the sites 'seastrait', 'landC', 'landC_mount', 'landSE_mount', 'landSW' and 'seaSWW' as indicated in Fig. 1 (in clockwise direction starting from the top left). Anomalies are plotted for MIROC (green), CanESM (blue), NorESM (yellow), ISPL (pink), HADGEM (red) and CNRM (grey) and shown on the left axis of the plots. ERA-Interim reanalysis wind speed histograms are plotted in grey, and their frequency values are shown on the righthand *y*-axis. *x*-axes show wind speed (m s^{-1})

4.1.3. Description of the PDF biases

The anomaly density plots in Figs. 5 & 6 represent the difference between the PDF of the ESM and the PDF of the reanalysis data. A positive/negative peak indicates an over/underestimation of the frequency of occurrence of the wind speeds.

The large-scale east–west stretched wind bias in the south of Europe is characterized by an overestimation of the wind speed. This signal is present above land (Site 'landSW') and sea (Site 'seaSWW'), at ~80 m and ~1500 m and during summer and winter. The anomaly plots indicate that the overestimation of the wind speeds is due to an underestimation of the skewness of the PDFs, especially at higher atmospheric levels, where PDFs are very positively skewed. During summer, the similarity among the various ESMs is smaller; however, the overall signal of overly strong wind speeds in the south of Europe is still present. It is suggested that the overestimation of wind speeds over the Anatolian peninsula (Site 'landSE_mount') is due to biases acting on both scales. On one hand, the ESMs overestimate the winds due to the large-scale east–west stretched wind bias (which is also present

in the upper-levels above the sea surrounding the peninsula, indicating the large-scale character of the bias). On the other hand, ESMs overestimate near-surface winds at the small scale due to the presence of the Anatolian highlands. In contrast, at this site HADGEM underestimates the wind speeds, especially near the surface. Also, at a mountainous location outside of the large-scale wind bias (Site 'landC_mount' in the Alps), HADGEM underestimates the wind speeds, while other ESMs commonly overestimate the winds. At this highly orographic location, the model with the finest resolution (CNRM) outperforms the other ESMs. In contrast, for flat terrain unaffected by large-scale and small-scale biases (Site 'landC'), the difference between the ESMs is minor, and the anomalies are small and inconsistent. Winds above the narrow sea strait between Denmark and Sweden (Site 'seastrait') are commonly underestimated in the lower model layers, but skillful at higher altitudes. In general, anomalies at higher levels in the atmosphere are smaller and vary less from site to site and from ESM to ESM than the anomalies at lower levels.

4.2. Temperature

4.2.1. Performance near the surface

Slightly shorter tails of the violin plots in Fig. 7 for winter than summer indicate that the ESMs are less skillful in representing extreme values during summer (when extreme values are more frequent) than for winter. There is almost no diurnal variation in the performance of the ESMs in simulating the near-surface temperature PDFs, nor do the type of land mask and the horizontal resolution of the ESM have large effects. IPSL and CNRM are not as good as other ESMs in simulating temperatures over sea, especially during the day. CNRM has the least extreme low PDF scores in all situations, and CanESM performs relatively well, despite its coarse horizontal resolution.

The map of PDF scores of the near-surface temperature PDFs (Fig. 8) shows how the performance of the ESMs in simulating the near-surface temperature is dominated by large-scale biases over sea and land. A large-scale PDF bias is present over the North Atlantic Ocean in all the ESMs. This PDF bias is associated with unskillful simulations for all the ESMs (except HADGEM and NorESM). The position of the maximum of the circular bias is situated at 45 to 50° N and 25° W and hardly differs from ESM to ESM. The actual center of the bias might also be situated outside of the study domain. In addition to the North Atlantic near sea surface temperature PDF bias, most ESMs have a low performance in the north of Iceland, and 3 ESMs (IPSL, NorESM and HADGEM) have a PDF bias over the sea in the south-west of the domain. Above land, low PDF scores are present in the east of Europe (except in IPSL), over the northern part of the African continent and over Scandinavia (only for MIROC and IPSL). However, the temperature PDF scores over land are on average higher than over sea and are mostly above the PDF score threshold of 0.7.

At the small scale, the near-surface temperature PDFs are biased above coastal features like bays, capes, straits and islands. However, these coastal effects only led to simulations with a PDF score <0.7 for the coarsest resolution model (MIROC). In contrast, orographic features have little influence on the simulation of the temperature PDFs near the surface.

4.2.2. Vertical extent of the PDF biases

Large scale. The large-scale PDF bias over the North Atlantic Ocean covers the whole vertical profile (Figs. 8 & 9). The bias is shifted more to the south (30 to 40° N) during summer compared to the winter (42 to 52° N), is bigger in winter than in summer and reaches up to >1500 m altitude in its center. The altitude of the lowest skillful level is radially decreasing from the centre, suggesting the origin of the bias in the temperature PDF to be related to the sea surface. The bias in summer temperatures in the east of Europe, in contrast, is present over the whole vertical profile and is therefore suggested to originate from the upper-atmospheric large-scale circulation. In the entire southern part of the domain (the northern African continent, the Mediterranean Sea and the south-east of Europe), ESMs are not able to skillfully represent the summer temperatures over at least the lowest 1500 m. This southern temperature bias is also present during winter, although less extensive. In contrast, in winter, temperatures in the northern part of the domain (north of 75° N) are not skillfully modeled in the lower model layers, varying from near-surface layers up to >1500 m.

Small scale. During summer, temperature PDFs above narrow coastal features can be unskillful up to 600 m. Apart from such regions, small-scale phenomena do not induce unskillful performance of ESMs in simulating the temperature PDFs. A scatterplot relating the altitude of the lowest level with a PDF score > 0.7 and the grid cell elevation indicates that the performance of the ESMs is independent of the orography (not shown).

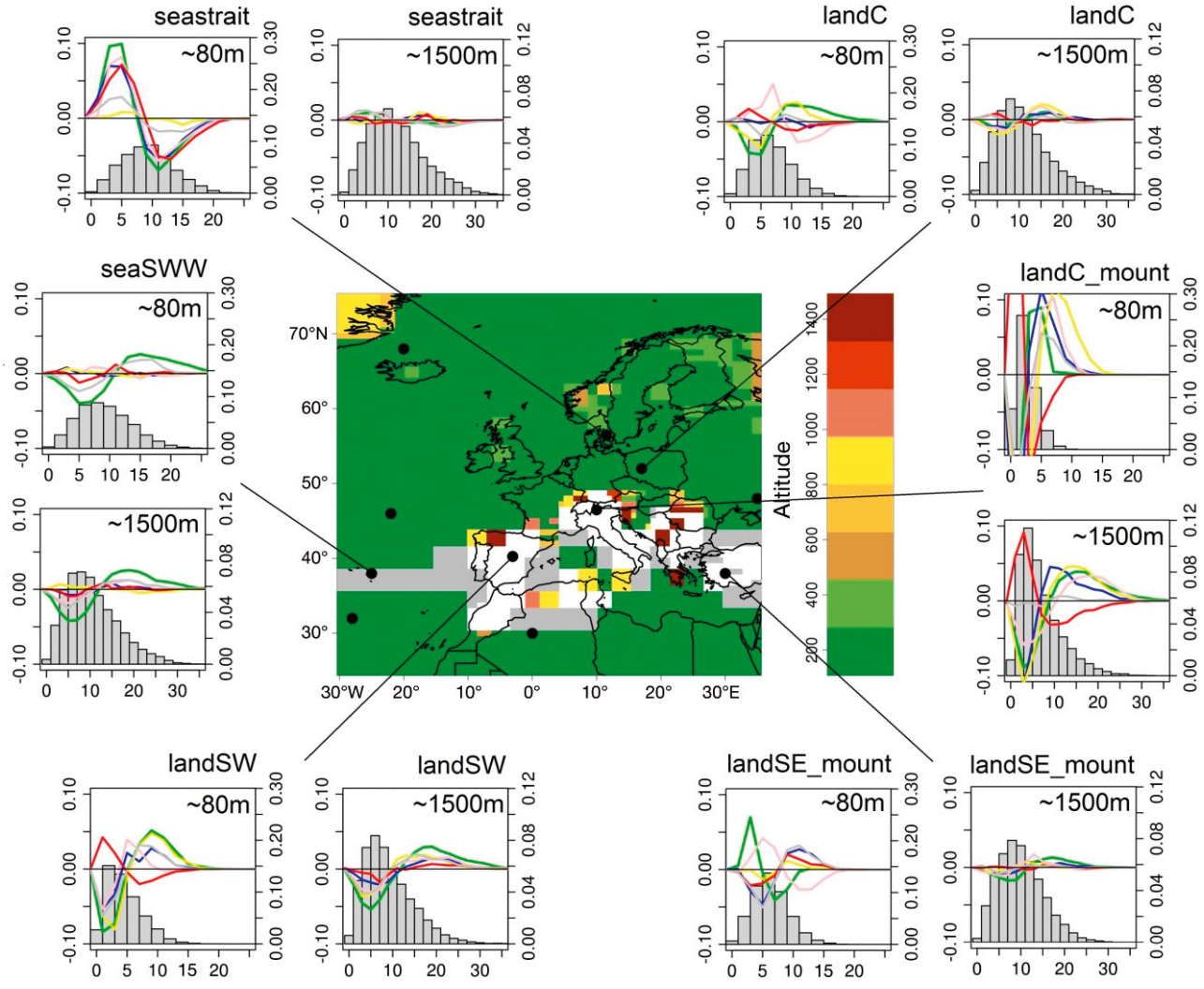

Fig. 6. Same as Fig. 5, but for D_W

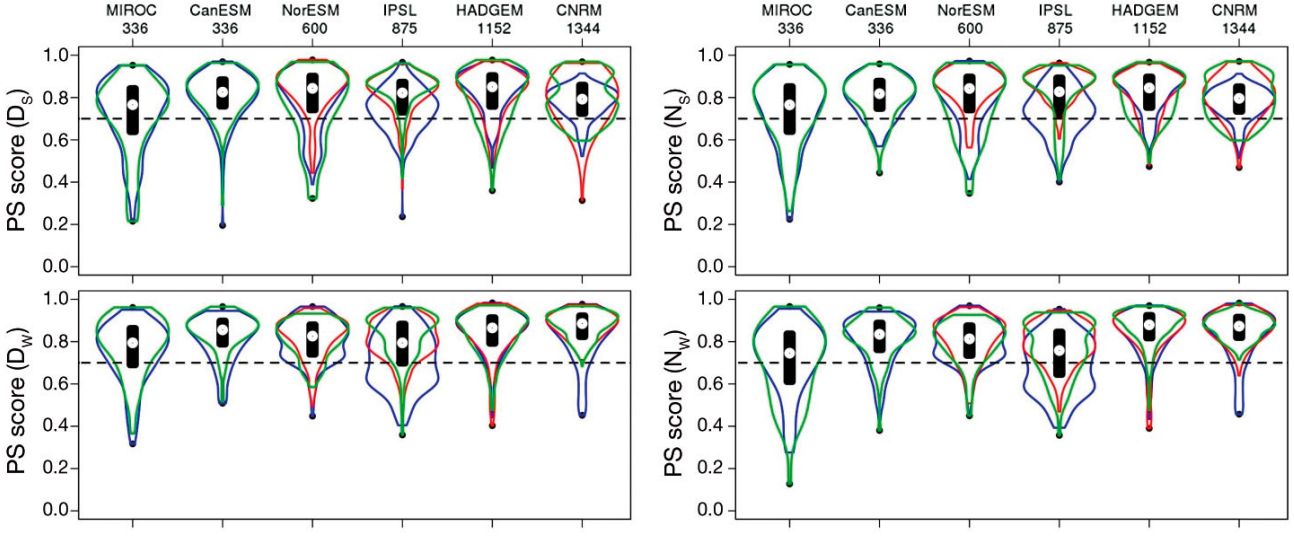

Fig. 7. Same as Fig. 3, but for temperature

4.3. Description of the PDF biases

The anomaly plots for Sites 'seaSW' and 'seaW' in Figs. 9 & 10 show that the large-scale PDF bias over the North Atlantic Ocean is related to overly cold temperatures in the ESMs. The underestimation is smaller at 1500 m than near the sea surface, suggesting a link with the ocean temperature. The low PDF scores north of Iceland (Site 'seaNW') are associated with an underestimation of the skewness of the winter temperature PDF in the lower levels of the atmosphere, leading to overly cold temperatures in the ESMs. Higher in the atmosphere, temperature distributions are more Gaussian, and the PDF bias is smaller. The temperature PDF bias over the sea strait between Denmark and Sweden (Site 'seastrait') is related to temperatures that are overly warm in the ESMs in summer and too cold in winter. However, these small-scale biases in the PDFs are too small to induce unskillful simulations (according to the 0.7 PDF score threshold). In the center of Europe, where temperature PDFs have PDF scores >0.7 down to their lowest levels, the anomaly plots do not show any common signal. Apart from MIROC, the magnitudes of the PDF anomalies are of the same order at ~80 m and at the ~1500 m level. The PDF bias in the summer temperatures in the east of the domain is related to overly warm temperatures in all ESMs and is as large near the surface as it is at ~1500 m. The northern region of the African continent experiences overly cold temperatures down to the surface.

5. DISCUSSION AND CONCLUSION

GCMs are commonly evaluated based on their performance at pressure levels of 1000, 850 and 500 hPa. However, due to their increasing resolution and improved interaction with the surface, GCMs are becoming increasingly realistic in representing variables in the lower atmosphere. This paper focuses on the performance of 6 ESMs in the lowest 1.5 km of the atmosphere over Europe. The evaluation is based on the representation of the PDFs of wind and temperature.

The results indicate 3 types of PDF biases: (1) small-scale PDF biases related to coastal or orographic effects, dependent on the ESMs resolution, acting at individual grid scales and decreasing with height; (2) large-scale PDF biases originating from the interaction with the surface, independent of the ESMs resolution, acting at multiple grid cells independent from coastlines and orography and decreasing with height; and (3) large-scale PDF biases originating from large-scale circulation biases, increasing or constant with height. The first type, the small-scale biases, are expected to be overcome when the ESMs are used in downscaling practices (Gómez-Navarro et al. 2011). The higher the resolution of the RCM and the observations (in the case of dynamical and statistical downscaling respectively), the less their downscaling results will be affected by the small-scale biases in the ESM. In contrast, the large-scale biases in ESMs are expected to affect the downscaling results. RCMs use the ESM data at their boundaries and will replicate the large-scale biases of the ESMs (Flaounas et al. 2013).

Small-scale coastal features like bays, capes, straits and islands lead to unskillful simulations of wind speed and temperature PDFs up to 600 m for the ESMs with the coarsest resolution. These biases are related to land-sea interactions, and propagate to higher levels in summer than in winter. High orography induces overly strong wind speeds in the ESMs. This is likely due to the coarse resolution of the ESMs, which underestimate the topography and roughness and therefore cause the wind speeds to be too high. The vertical extent of this small-scale PDF bias depends on the altitude of the mountains and the time of the day. The results show that during the day, only high mountains (>1 km) cause unskillful simulations, while during the night, wind speeds over lower mountain areas are also poorly modeled. In contrast to wind, temperature is less subject to small-scale effects of orography.

For temperature, the performance of the ESMs is dominated by the presence of large-scale biases. Five of the 6 ESMs are characterized by a large-scale PDF bias originating from the sea surface of the North Atlantic Ocean. This bias is mostly pronounced in winter and is related to a common underestimation of temperature in the ESMs. The bias decreases with height, and is highest at its center. The center is situated around 35° N during summer and around 46° N during winter. Also, more to the north (north of Iceland), a temperature PDF bias, related to overly cold temperatures, originates from the sea surface. The bias is stronger in winter than in summer and induces unskillful PDFs in the ESMs to some 100 m height.

Other large-scale PDF biases originate from the upper-atmospheric large-scale circulation, and do not decrease with height. Such a temperature PDF bias is found in the east of Europe, where temperatures up to at least 1500 m are too warm in summer. In the northern part of the African continent, the whole vertical profile is too cold in both seasons, but

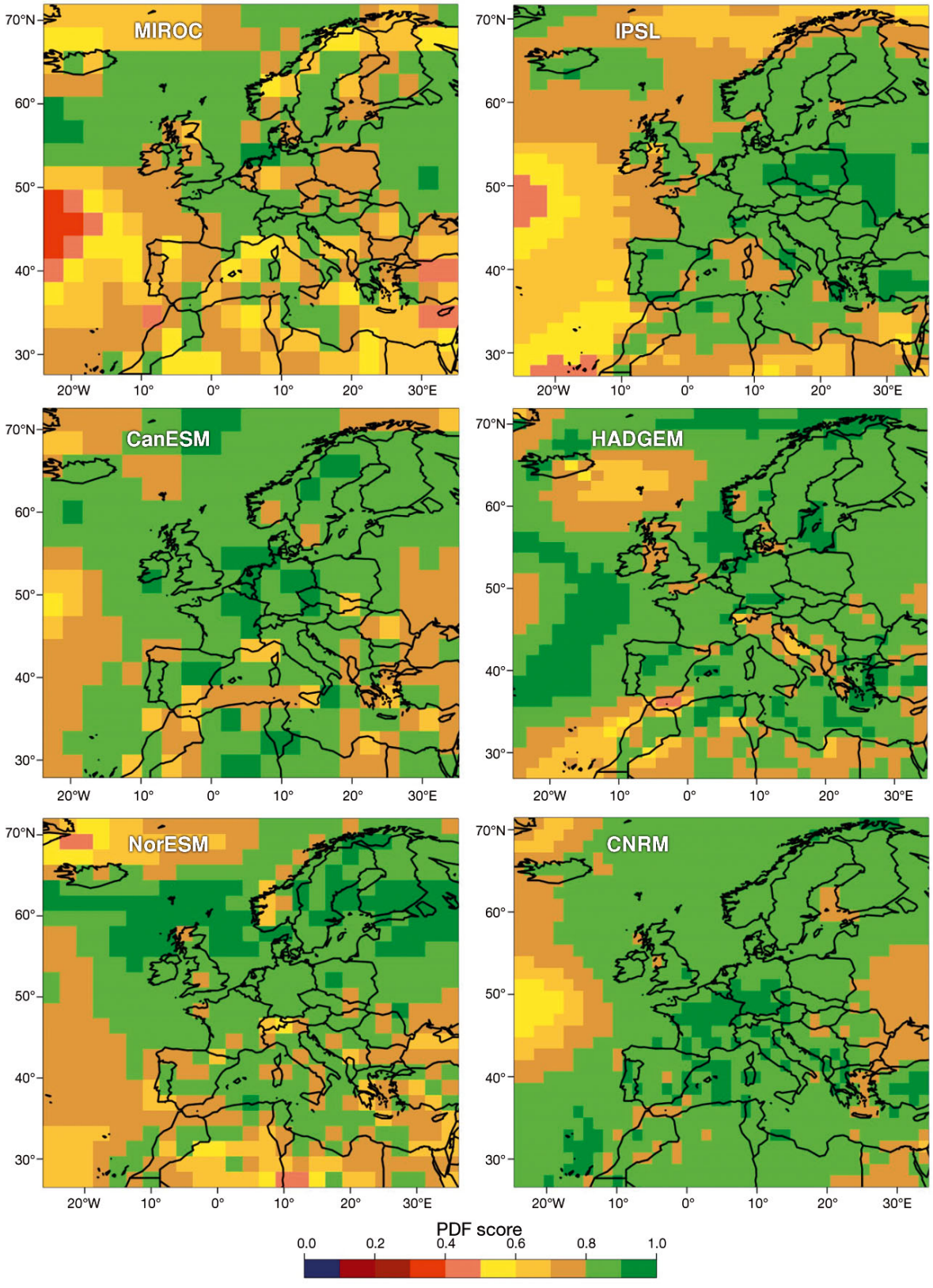

Fig. 8. Same as Fig. 4, but for temperature

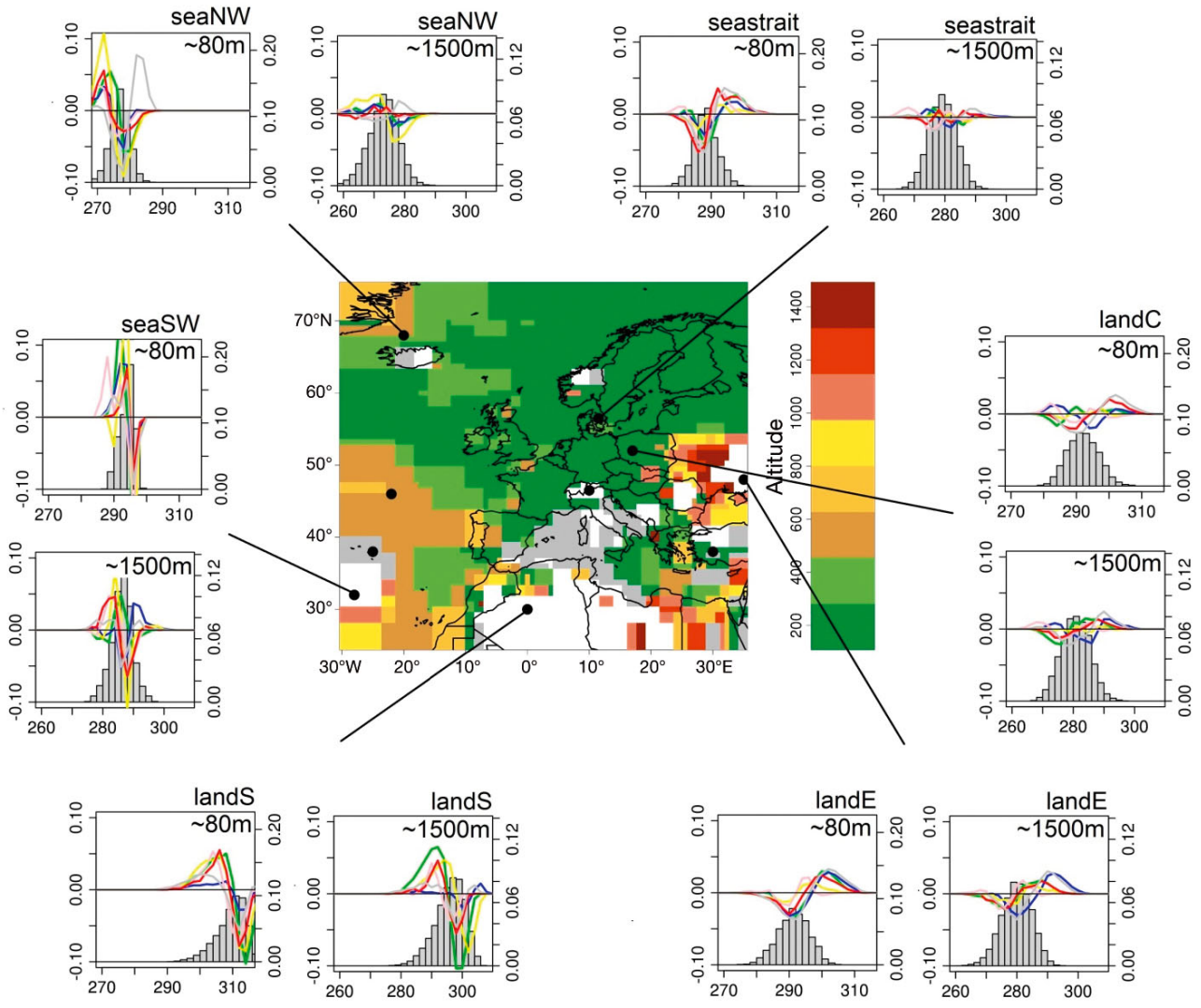

Fig. 9. Same as Fig. 5, but for D_S temperature (K) and for the sites 'seaNW', 'seastrait', 'landC', 'landE', 'landS' and 'seaSW' as indicated in Fig. 1 (in clockwise direction starting from the top left)

most extensive in summer. Apart from the large-scale biases in temperature PDFs, ESMs also show a large-scale east–west stretched bias in the PDF of the wind. This large-scale bias, present in at least half of the ESMs, reflects unskillful wind speeds between 30 and 45° N, and affects the whole vertical profile during winter. During summer, the variability between the ESMs is larger, but the signal of overly strong winds is also present. Down to the surface, layers are slightly better represented than at 1500 m, suggesting an upper-atmospheric origin of the bias. It must be noted that the systematically lower ESM-performance south of 45° N might at least partly be explained by uncertainties in the reanalysis data. As shown by Brands et al. (2012, 2013), the uncertainty in summer temperature, U and V at 850 hPa increases southward of 45° N.

The large-scale east–west stretched bias in wind speed is associated with the bias noticed by Brands et al. (2013) in the 850 hPa westward wind component of the CMIP5 models and by van Ulden & van Oldenborgh (2005) in the CMIP3 models. They argue that during boreal winter and spring, the overly strong westerlies in the Northern Hemisphere mid-latitudes are related to a largely exaggerated latitudinal pressure gradient. Vial & Osborn (2012) found that the CMIP3 models underestimate the frequency and duration of winter-time atmospheric blocking. This underestimation of winter blocking episodes is a common feature of global climate models, which tend to simulate an overly strong North Atlantic jet stream (Scaife et al. 2010).

At the higher levels, HADGEM outperforms the other ESMs, as the large-scale PDF biases are small

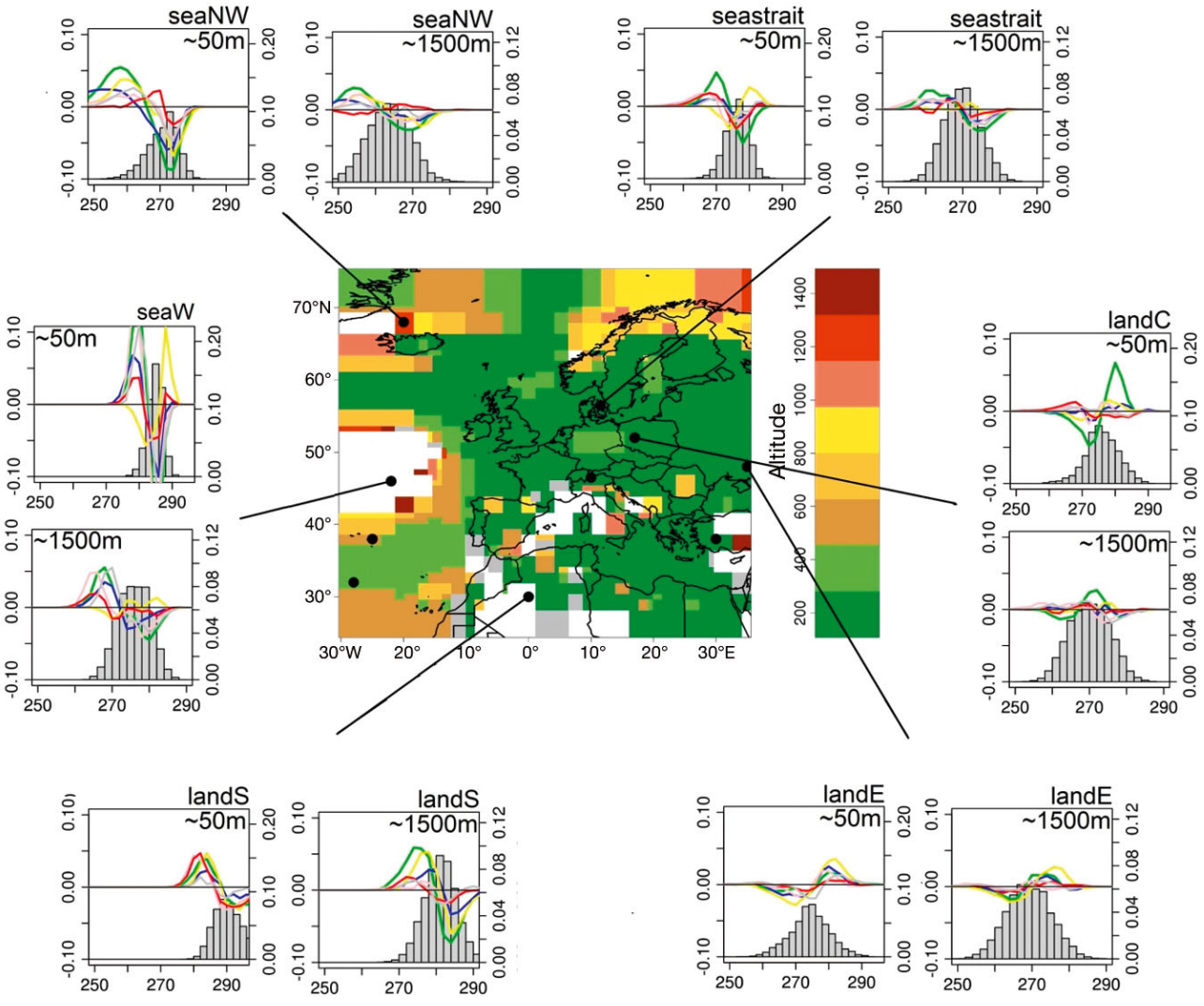

Fig. 10. Same as Fig. 5, but for D_W and for the sites 'seaNW', 'strait', 'landC', 'landE', 'landS' and 'seaW' as indicated in Fig. 1 (in clockwise direction starting from the top left)

or absent. This is in agreement with evaluations based on upper-atmospheric levels (Brands et al. 2013). However, our results show that near the surface, HADGEM mainly underestimates wind speeds. In constrast, CanESM, the model with the finest vertical resolution, performs notably well near the surface, regardless of its coarse horizontal resolution. CNRM was the most consistent model, having the lowest spatial variability in performance. Moreover, the variability among the ESMs is larger for wind than for temperature.

Within the framework of dynamical downscaling, the RCM evaluation study of Imberey et al. (2013) reports PDF skill scores between 0.4 and 0.8 for the monthly mean temperature over Germany as simulated by CLM 2.4.6 and REMO 5.7. It must be noted that the authors used a high-resolution gridded observation dataset as a reference, which might partially explain the lower PDF skill scores of the RCMs compared to the presented scores of the ESMs in this study. In addition, seasonal and regionally mean temperature biases found in the EURO-CORDEX RCM-ensemble evaluation (Kotlarski et al. 2014) are comparable in location and magnitude to the biases in the ESM ensemble of this study. It can be suggested that, since ESM resolutions are increasing and their interaction with the surface is improving, the added value of the RCMs is decreasing, at least in representing seasonally and regionally averaged climatologies. The added value of RCMs is still very important in the representation of the earth's heterogeneity, e.g. in describing urban and coastal effects.

Within the framework of statistical downscaling, the results indicate that the near-surface wind speed

PDFs can also be used as predictors, apart from the standardly used 850 and 500 hPa level variables. In such a way, statistical downscaling models might profit from predictors that are more closely related to the predictands. The signal of temperature is more complicated because the small-scale temperature biases near the surface are overshadowed by large-scale biases, originating from both surface and upper-levels. These large-scale biases should be taken into account when downscaling the ESMs. In general, the height-dependent near-surface evaluation approach that was adopted gives more insight into the origin of large-scale biases, defines up to which altitude ESMs are influenced by small-scale phenomena, and determines the lowest levels for which temperature and wind-speed PDFs are suitable input variables for downscaling models.

Acknowledgements. This research is funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology Flanders (IWT-Flanders). ECMWF is acknowledged for providing operational ERA-Interim data, and the ESG web portals are thanked for providing the ESM datasets.

LITERATURE CITED

Archer CL, Jacobson MZ (2005) Evaluation of global wind power. J Geophys Res 110:D12110, doi:10.1029/2004JD005462

Brands S, Herrera S, San-Martín D, Gutiérrez JM (2011a) Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective. Clim Res 48: 145–161

Brands S, Taboada JJ, Cofino AS, Sauter T, Schneider C (2011b) Statistical downscaling of daily temperatures in the NW Iberian Peninsula from global climate models. Validation and future scenarios. Clim Res 48:163–176

Brands S, Gutiérrez JM, Herrera S, Cofiño AS (2012) On the use of reanalysis data for downscaling. J Clim 25: 2517–2526

Brands S, Herrera S, Fernández J, Gutiérrez JM (2013) How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? Clim Dyn 41: 803–817

Cattiaux J, Douville H, Ribes A, Chauvin F, Plante C (2013a) Towards a better understanding of changes in wintertime cold extremes over Europe: a pilot study with CNRM and IPSL atmospheric models. Clim Dyn 40: 2433–2445

Cattiaux J, Douville H, Yannick P (2013b) European temperatures in CMIP5: origins of present-day biases and future uncertainties. Clim Dyn 41:2889–2907

Chylek P, Li J, Dubey M, Wang M, Lesins G (2011) Observed and model simulated 20th century Arctic temperature variability: Canadian earth system model canESM2. Atmos Chem Phys Discuss 11:22893–22907

Collins WJ, Bellouin N, Doutriaux-Boucher M, Gedney N and others (2011) Development and evaluation of Earth-System mode — HadGEM2. Geosci Model Dev 4:1051–1075

Dee DP, Uppala SM, Simmons AJ, Berrisford P and others (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. QJR Meteorol Soc 137:553–597

Demuzere M, Werner M, van Lipzig NPM, Roeckner E (2009) An analysis of present and future ECHAM5 pressure fields using a classification of circulation patterns. Int J Climatol 29:1796–1810

Devis A, van Lipzig NPM, Demuzere M (2013) A new statistical approach to downscale wind speed distributions at a site in northern Europe. J Geophys Res 118:2272–2283

Dufresne JL, Foujols MA, Denvil S, Caubel A and others (2013) Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Clim Dyn 40:2123–2165

Ehret U, Zehe E, Wulfmeyer V, Warrach-Sagi K, Liebert J, Opinions HESS (2012) Should we apply bias correction to global and regional climate model data? Hydrol Earth Syst Sci 16:3391–3404

Flaounas E, Drobinski P, Bastin S (2013) Dynamical downscaling of IPSL-CM5 CMIP5 historical simulations over the Mediterranean: benefits on the representation of regional surface winds and cyclogenesis. Clim Dyn 40:2497–2513

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metric for climate models. J Geophys Res 113:D06104, doi:10.1029/2007JD008972

Gómez-Navarro JJ, Montávez JP, Jerez S, Jiménez-Guerrero P, Lorente-Plazas R, González-Rouco JF, Zorita E (2011) A regional climate simulation over the Iberian Peninsula for the last millennium. Clim Past 7:451–472

Holt E, Wang J (2012) Trends of wind speed at wind turbine height of 80 m over the contiguous United States using the North American Regional Reanalysis (NARR). J Appl Meteorol Climatol 51:2188–2202

Huth R, Beck C, Philipp A, Demuzere M and others (2008) Classifications of atmospheric circulation patterns: recent advances and applications. Ann NY Acad Sci 1146: 105–152

Imbery F, Plagemann S, Namyslo J (2013) Processing and analysing an ensemble of climate projections for the joint research project KLIWAS. Adv Sci Res 10:91–98

Jakob C (2010) Accelerating progress in global atmospheric model development through improved parameterizations, challenges, opportunities and strategies. Bull Am Meteorol Soc 91:869–875

Jones C, Carvalho LMV (2013) Climate change in the South American monsoon system: present climate and CMIP5 projections. J Clim 26:6660–6678

Kirkevag A, Iversen T, Seland O, Debernard JB, Storelvmo T, Kristjansson JE (2008) Aerosol-cloud climate interactions in the climate model CAM-Oslo. Tellus A 60:492–512

Kjellström E, Boberg F, Castro M, Christensen JH, Nikulin G, Sánchez E (2010) Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. Clim Res 44:135–150

Knutson TR, Sirutis J, Vecchi GA, Garner S and others (2013) Dynamical downscaling projections of twenty-first-century Atlantic hurricane activity: CMIP3 and CMIP5 model-based scenarios. J Clim 26:6591–6617

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758

Kotlarski S, Keuler K, Christensen OB, Colette A and others

(2014) Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. Geosci Model Dev Discuss 7:217–293

Mao J, Shi X, Ma L, Kaiser DP, Li Q, Thornton PE (2010) Assessment of reanalysis daily extreme temperatures with China's homogenized historical dataset during 1979–2001 using probability density functions. J Clim 23:6605–6623

Maxino CC, McAvaney BJ, Pitman AJ, Perkins SE (2008) Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. Int J Climatol 28:1097–1112

Onogi K, Tslttsui J, Koide H, Sakamoto M and others (2007) The JRA-25 reanalysis. J Meteorol Soc Jpn 85:369–432

Perkins SE, Pitman AJ, Holbrook NJ, McAneney JK (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. J Clim 20:4356–4376

Pitman AJ, Perkins SE (2009) Global and regional comparison of Daily 2-m and 1000-hPa maximum and minimum temperatures in three global reanalyses. J Clim 22: 4667–4681

Pryor SC, Schoof JT, Barthelmie RJ (2005a) Empirical downscaling of wind speed probability distributions. J Geophys Res 110:D19109, doi:10.1029/2005JD005899

Roehrig R, Bouniol D, Guichard F, Hourdin F Redelsperger JL (2013) The present and future of the West African monsoon: a process-oriented assessment of CMIP5 simulations along the AMMA transect. J Clim 26:6471–6505

Scaife AA, Woollings T, Knight J, Martin G, Hinton T (2010) Atmospheric blocking and mean biases in climate models. J Clim 23:6143–6152

Schmidli J, Frei C, Vidale PL (2006) Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods. Int J Climatol 26:679–689

Schoetter R, Hoffmann P, Rechid D, Schlünzen KH (2012) Evaluation and bias correction of regional climate model results using model evaluation measures. J Appl Meteorol Climatol 51:1670–1684

Seland O, Iversen T, Kirkevag A, Storelvmo T (2008) Aerosol-climate interactions in the CAM-Oslo atmospheric GCM and investigation of associated basic shortcomings. Tellus Ser A – Dyn Meteorol Oceanol 60:459–491

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498

van Ulden AP, van Oldenborgh GJ (2005) Large-scale atmospheric circulation biases and changes in global climate model simulations and their importance for regional climate scenarios: a case study for West-Central Europe. Atmos Chem Phys Discuss 5:7415–7455

Vial J, Osborn J (2012) Assessment of atmosphere ocean general circulation model simulations of winter northern hemisphere atmospheric blocking. Clim Dyn 39:95–112

Voldoire A, Sanchez-Gomez E, Salas y M'elia D, Decharme B, Cassou C (2013) The CNRM-CM5.1 global climate model: description and basic evaluation. Clim Dyn 40:2091–2121

Watanabe S, Hajima T, Sudo K, Nagashima T and others (2011) MIROC-ESM 2010: model description and basic results of CMIP5 20c3m experiments. Geosci Model Dev Discuss 4:1063–1128

Wilby RL, Hassan H, Hanaki K (1998) Statistical downscaling of hydrometeorological variables using general circulation model output. J Hydrol (Amst) 205:1–19

Yang C, Giese BS (2013) El Niño Southern Oscillation in an ensemble ocean reanalysis and coupled climate models. J Geophys Res 118:4052–4071

Editorial responsibility: Eduardo Zorita, Geesthacht, Germany

Submitted: October 17, 2013; Accepted: May 13, 2014
Proofs received from author(s): August 4, 2014