

Field significance of performance measures in the context of regional climate model evaluation. Part 1: temperature

Martin Ivanov¹  · Kirsten Warrach-Sagi² · Volker Wulfmeyer²

Received: 4 July 2016 / Accepted: 13 March 2017 / Published online: 27 March 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract A new approach for rigorous spatial analysis of the downscaling performance of regional climate model (RCM) simulations is introduced. It is based on a multiple comparison of the local tests at the grid cells and is also known as “field” or “global” significance. New performance measures for estimating the added value of downscaled data relative to the large-scale forcing fields are developed. The methodology is exemplarily applied to a standard EURO-CORDEX hindcast simulation with the Weather Research and Forecasting (WRF) model coupled with the land surface model NOAH at 0.11° grid resolution. Monthly temperature climatology for the 1990–2009 period is analysed for Germany for winter and summer in comparison with high-resolution gridded observations from the German Weather Service. The field significance test controls the proportion of falsely rejected local tests in a meaningful way and is robust to spatial dependence. Hence, the spatial patterns of the statistically significant local tests are also meaningful. We interpret them from a process-oriented perspective. In winter and in most regions in summer, the downscaled distributions are statistically indistinguishable from the observed ones. A systematic cold summer bias occurs in deep river valleys due to overestimated elevations, in coastal areas due probably to enhanced sea breeze circulation, and over large lakes due to the interpolation of water temperatures.

Urban areas in concave topography forms have a warm summer bias due to the strong heat islands, not reflected in the observations. WRF-NOAH generates appropriate fine-scale features in the monthly temperature field over regions of complex topography, but over spatially homogeneous areas even small biases can lead to significant deteriorations relative to the driving reanalysis. As the added value of global climate model (GCM)-driven simulations cannot be smaller than this perfect-boundary estimate, this work demonstrates in a rigorous manner the clear additional value of dynamical downscaling over global climate simulations. The evaluation methodology has a broad spectrum of applicability as it is distribution-free, robust to spatial dependence, and accounts for time series structure.

1 Introduction

Climate change will induce changes not only in temperature statistics but also in the spatial and temporal precipitation patterns. The expected significant societal and environmental impacts call for further assessment and improvement of climate projections (Trenberth et al. 2003; Schär et al. 2004; O’Gorman and Schneider 2009; Hartmann et al. 2013).

Global climate models (GCMs) are the primary source of climate change information. However, the statistics they provide are not reliable on the fine scales required for impact assessment. This fundamental scale gap is the target of the so-called downscaling or regionalisation methods. One-way nesting of regional climate models (RCMs) is the computationally most parsimonious but nevertheless equally well performing physically based downscaling method. Typical RCM resolutions are currently 10–50 km, but simulations with grid resolutions down to ~1 km with explicit resolving of convection are becoming increasingly available

✉ Martin Ivanov
martin.ivanov@geogr.uni-giessen.de

¹ Department of Geography, Climatology, Climate Dynamics and Climate Change, Justus-Liebig University of Gießen, Senckenbergstraße 1, D-35390 Gießen, Germany

² Institute of Physics and Meteorology, University of Hohenheim, Garbenstr. 30, D-70593 Stuttgart, Germany

(Prein et al. 2015). For more details on the state-of-the-art regional climate modelling, the reader is referred to Laprise (2008) and Rummukainen (2010).

A good RCM performance for past climate periods is assumed to be necessary for an adequate performance under different climate conditions in the future. Therefore, RCM hindcast simulations must be evaluated against observation-based products. The evaluation of RCMs requires an observational reference of the same spatial and temporal resolution, which is not always available, and the problem aggravates with the increasing RCM resolution. To minimise biases due to misrepresentation of the large-scale forcing and focus only on RCM-related biases (e.g. due to deficiencies of RCM physics or artefacts of the nesting procedure itself), the evaluation must be performed in a “perfect-boundary” setting, in which an RCM is nested in a global large-scale reanalysis (Christensen et al. 1997; Pan et al. 2001). The perfect-boundary performance is better than the performance the same RCM would have were it driven by a GCM simulation. In this sense, perfect-boundary evaluation yields an upper boundary for RCM skill. The additional information provided by the regional simulation beyond the scales of the driving fields is referred to as added value (Laprise 2008; Di Luca et al. 2012). Due to the data assimilation, reanalyses excel GCM runs and hence are harder to outperform. Therefore, the added value of a reanalysis-driven hindcast is a lower bound of the added value that the same RCM can have in a GCM-driven setting (Prömmel et al. 2010).

Within the EU-funded projects PRUDENCE (Christensen and Christensen 2007) and its follower ENSEMBLES (van der Linden and Mitchell 2009), RCMs with grid resolutions of 20–50 km were applied and evaluated for the European region (e.g. Jacob et al. 2007; Jaeger et al. 2008). The ENSEMBLES project was succeeded by the COordinated Regional climate Downscaling EXperiment (CORDEX) (Giorgi et al. 2009), which focusses on a systematical evaluation of RCM performance in an ensemble of perfect-boundary conditions experiments nested within the ERA-Interim reanalysis (e.g. Dee et al. 2011) for the 1989–2008 period. Within EURO-CORDEX, the European branch of CORDEX, an ensemble of such hindcast simulations has been created in order to provide comparisons at horizontal grid increments of 0.44° (~ 50 km) and 0.11° (~ 12 km).

RCM performance is quantified by statistical measures also known as performance metrics. Relative versions of these metrics allow comparison of RCM skill against that of the large-scale forcing data. The purpose of this study is to introduce a new approach for rigorous spatial analysis of the downscaling performance and develop new relative performance metrics. This is the first of two papers that exemplarily apply the methodology to a standard EURO-CORDEX run

at 0.11° grid resolution with the WRF-NOAH model system (Warrach-Sagi et al. 2013) over the territory of Germany, where high-resolution gridded observation data products are available. The current work, part 1, is devoted to 2 m monthly temperature.

We employ distribution-based evaluation statistics and their relative versions to quantify the downscaling skill relative to the driving reanalysis. Despite being direct objective measures of added value, relative performance metrics are still underapplied. Most previous works estimate the relative skill by comparing domain-aggregated scalar evaluation statistics, which precludes spatial analysis, and/or by visually inspecting the fields of the evaluation statistics for the downscaled and the larger-scale driving data (e.g. Duffy et al. 2006; Feser 2006; Sotillo et al. 2006; Buonomo et al. 2007; Sanchez-Gomez et al. 2009; Prömmel et al. 2010; Di Luca et al. 2012; Kendon et al. 2012; Cardoso et al. 2013; Chan et al. 2013; Pearson et al. 2015; Torma et al. 2015), which is more or less subjective. Relative versions of few distribution-based performance measures were utilised by Winterfeldt and Weisse (2009) and Vautard et al. (2013) only for individual locations, and by Winterfeldt et al. (2011) and Dosio et al. (2015) on a gridpoint basis. Our priority is to study in more detail the spatial structure of model performance as represented by the spatial patterns of grid-cell statistics. We propose general formulae for deriving relative versions of performance metrics. Many of the relative measures are new or applied for the first time in the context of RCM added value analysis.

As performance measures are subject to sampling variability, they should be physically interpreted only after the accompanying sampling uncertainty has been quantified. Within the frequentist approach to statistical inference (Jolliffe 2007) this can be done via confidence intervals, which is the popular approach (e.g. Elmore et al. 2006; Buonomo et al. 2007; Sanchez-Gomez et al. 2009; Kendon et al. 2012, Chan et al. 2013) or more directly via hypothesis testing as in, e.g. Duffy et al. (2006) and Cardoso et al. (2013) as well as in this work. Here, at each grid cell, we estimate the p value for each test statistic in a Monte Carlo framework. The problem of test multiplicity is tackled by determining the “field” significance (e.g. Livezey and Chen 1983; Ventura et al. 2004; Wilks 2006a) as implemented by the false discovery rate (FDR) approach of Benjamini and Hochberg (1995). As FDR is generally more powerful and robust to spatial correlations than alternative multiple comparison methods, it provides a more meaningful spatial pattern of local rejections (Wilks 2006a). The latter comprises spatial configurations of locations at which the values of the evaluation statistics are in breach with the null hypothesis, that is, which are highly unlikely to have occurred by chance. The analysis focusses on these patterns rather than on the magnitudes of the evaluation statistics as

is conventionally the case. Most evaluation studies estimate statistical significance for domain-aggregated scalar measures (e.g. Feser 2006; Sanchez-Gomez et al. 2009; Kendon et al. 2012; Cardoso et al. 2013; Pearson et al. 2015), thus avoiding spatial analysis of the statistical significance. Others estimate statistical significance (grid)point-wise, but take neither multiplicity nor spatial autocorrelation of test statistics into account (e.g. Duffy et al. 2006; Buonomo et al. 2007; Winterfeldt and Weisse 2009; Winterfeldt et al. 2011; Chan et al. 2013, Katragkou et al. 2015). Finally, we suggest a process-oriented interpretation of the spatial and seasonal patterns of statistical significance. To the best of our knowledge, this is the first application of the concept of field significance in the context of RCM evaluation.

The evaluation methodology is described in Section 2. In Section 3, the landscape and temperature climate of the study area as well as the observation and model data are introduced. Results are discussed in Section 4. Section 5 summarises the main findings of the study. Daily precipitation is considered in Ivanov et al. (2017; henceforth referred to as part 2).

2 Evaluation methodology

In line with the objective to rigorously diagnose the skill and added value of the WRF-NOAH model system as a downscaling tool, model output is used directly, without any post-processing. Our simulation is reanalysis-driven, but has only been initialised once; hence, little *deterministic skill in fine scales* is to be expected for daily and monthly statistics due to internal variability (e.g. de Elía et al. 2002; de Elía et al. 2007; Alexandru et al. 2007). Laprise et al. (2008) conclude that the small scales for which the added value of RCMs is to be expected, do not benefit from the extended predictability due to the boundary forcing and state that “for the mission of dynamical downscaling ... it is the statistics, not the specific sequence of weather events that count”. Therefore, we neglect the temporal correspondence between model and observation and only quantify model performance in terms of climate statistics.

2.1 Evaluation statistics

2.1.1 Absolute performance measures

Absolute performance measures compare the downscaling directly to observations.

The overall similarity between the modelled and observed distributions is quantified by the *Perkins score* (Perkins et al. 2007). It measures the common area between the two probability density functions (PDFs), so its perfect value is 1. The histogram bin width is determined by the

Sturges algorithm (Sturges 1926). To achieve a better resolution, the bin size is chosen as the minimum over all bin sizes determined at each grid cell and each season. Thus, we arrive at bin width of 1 °C.

We consider differences between characteristics of simulated and observed distributions, which we refer to as additive *biases*. More specifically, these are the biases of the mean, median, lower and upper quartiles of the distributions. Of course, the perfect value of an additive bias is 0.

The *frequency bias* score for a temperature category is the ratio between the number of events belonging to the category in the downscaling and in the observations. The thresholds of the categories are defined by observed percentiles of the 1980–2009 climatology. In the following, $X\%$ stands for the X th observed climatological percentile. We use the median as a threshold to define the “above normal” ($> 50\%$) category and the quartiles to define the “low” ($< 25\%$), “normal” (25–75%), and “high” ($> 75\%$) temperature categories. The perfect value of these scores is 1.

2.1.2 Relative skill measures

Relative performance measures evaluate the downscaling skill relative to the reanalysis reference.

Absolute performance is quantified by the deviation of the respective evaluation measure from its perfect value. The absolute value of this difference measures the magnitude of the deviation from perfect performance. For a certain performance measure M , consider the difference between such absolute deviations for the reanalysis reference and for the downscaling. The sign of this difference identifies which model is better, i.e. closer to reality with respect to M . The absolute value of the difference quantifies how much closer to reality with respect to M the better model is. So, we propose the following generic relative skill measure:

$$M_{rel} = |M_{ref} - M_{perf}| - |M - M_{perf}|, \quad (1)$$

where M_{rel} is the relative measure, M is the value of the measure for the downscaling, M_{perf} is the perfect value of the measure, and M_{ref} is the value for the reanalysis reference. A value of zero for the relative measure (1) would indicate no difference in performance or zero added value with respect to M . A positive value would indicate that the downscaling is closer to reality than the reanalysis, that is, positive added value. A negative value would lead to the opposite conclusion. We also define the following scaled version of the relative measure:

$$M'_{rel} = 1 - \frac{|M - M_{perf}|}{|M_{ref} - M_{perf}|}. \quad (2)$$

It is dimensionless, with no-skill value of 0, and is bounded from above by 1. Again, positive values indicate that the downscaling is better than the reference and vice versa. This metric is a generalisation of the generic skill score of Jolliffe and Stephenson's (2012), generic skill score, as it is applicable not only to measures that are positively or negatively oriented, but also to measures, the perfect value of which lies somewhere within the range of possible values. All bias metrics introduced in this section are examples for such statistics.

We generate the relative versions of the non-dimensional measures by formula (1) and of the dimensional by formula (2). Thus, our relative metrics are dimensionless, have a no-skill value of 0, their positive values indicate positive added value and vice versa.

2.2 Estimating statistical significance

2.2.1 Local tests

The absolute and relative performance measures allow testing distinct hypotheses locally at the grid cells. The null hypothesis for an absolute test is that observation and downscaling are identically distributed, i.e. that they are drawn from the same population. Ideally, this is what is ultimately expected from a downscaling. For a relative test, the null hypothesis is that the downscaling and the driving large-scale reanalysis are identically distributed. This is the situation of a completely useless downscaling, that is, no added value. The null distributions of the performance measures are unknown and have to be estimated non-parametrically at each grid cell by means of resampling tests (Wilks 2006b). We need precise estimates of the respective p values to be subsequently fed into the field significance procedure (see next subsection). Therefore, we opt for the technique of permutation tests. The expected value of the test statistic under the null hypothesis is 0 for most measures we consider, only for the absolute frequency bias it is 1. Note that we do not directly test whether a performance measure is equal to its expected value, but a rather general null hypothesis concerning the statistical distributions of two data samples.

Each performance measure yields a distinct statistical test with its own size and power for the available sample size. Failure to reject the null hypothesis means that the performance measure is incapable of capturing the distribution differences between the two samples. Formally, the test statistic is not significantly different from its expected value under the null. For an absolute test, the conclusion is that the corresponding distribution characteristic is well reproduced by the downscaling, and for a relative test that the downscaling does not add value to the large-scale reanalysis with regard to this particular distribution characteristic.

In contrast, a performance measure that is sensitive to the nature of the distribution differences between the two samples will yield a statistical test, powerful enough to reject the null hypothesis. In this case, the conclusion is that the distribution differences between the two samples are significant with regard to the distribution attribute tested for. Formally, the test statistic is significantly different from its expected value under the null. For an absolute test, the interpretation is that the distribution attribute is poorly reproduced by the downscaling, and for a relative test that the downscaling and driving reanalysis are not equally good at reproducing the distribution attribute.

Permutation tests are built on the principle of exchangeability of the two samples under the null hypothesis. The principle is only applicable under the assumption that these samples are identically distributed. Figure 1 depicts the routine $PVAL(i, pm)$ that determines the local p value for the performance measure pm at grid cell i for the absolute and relative tests. For example, in the case of a relative skill measure, the downscaled and ERA-Interim data are pooled together and two samples are randomly drawn out of the pool without replacement. The process is implemented after the efficient permutation algorithm suggested in Wilks (2006b). One of the synthetic data batches thus drawn is labelled "WRF-NOAH sample" and the other "ERA-Interim sample". From these two samples and the observations, an artificial value for the respective relative test statistic is calculated. The process is repeated 1999 times to generate the null distribution of the statistic. The nominal value of the statistic, which is computed from the original data samples, is compared against its null distribution to obtain the respective p value.

As the Perkins score cannot exceed its perfect value of 1, we test against the one-tailed alternative of smaller values. The rest of the absolute and all relative tests are implemented two-tailed using the equal-tail bootstrap p value (e.g. Davidson and Mackinnon 2007). To check if the sample size is large enough to give stable results, we changed the random-number seed and repeated the analyses. In no case there were meaningful changes, which indicates that 1999 is an adequate sample size.

This would be a complete description of the resampling if the data were independent. As we work with time series, this is not the case. The distribution of statistical estimators based on dependent data heavily depends on the joint distribution of the observations (Léger et al. 1992). This is why each bootstrap resample of the original data must be a sample from that joint distribution. In effect, the construction of the null distribution of the test statistic must also take into account the serial correlations of the time series. This is achieved by means of block resampling: the artificial data batches at each step are generated by sampling not of individual values, but of temporally contiguous

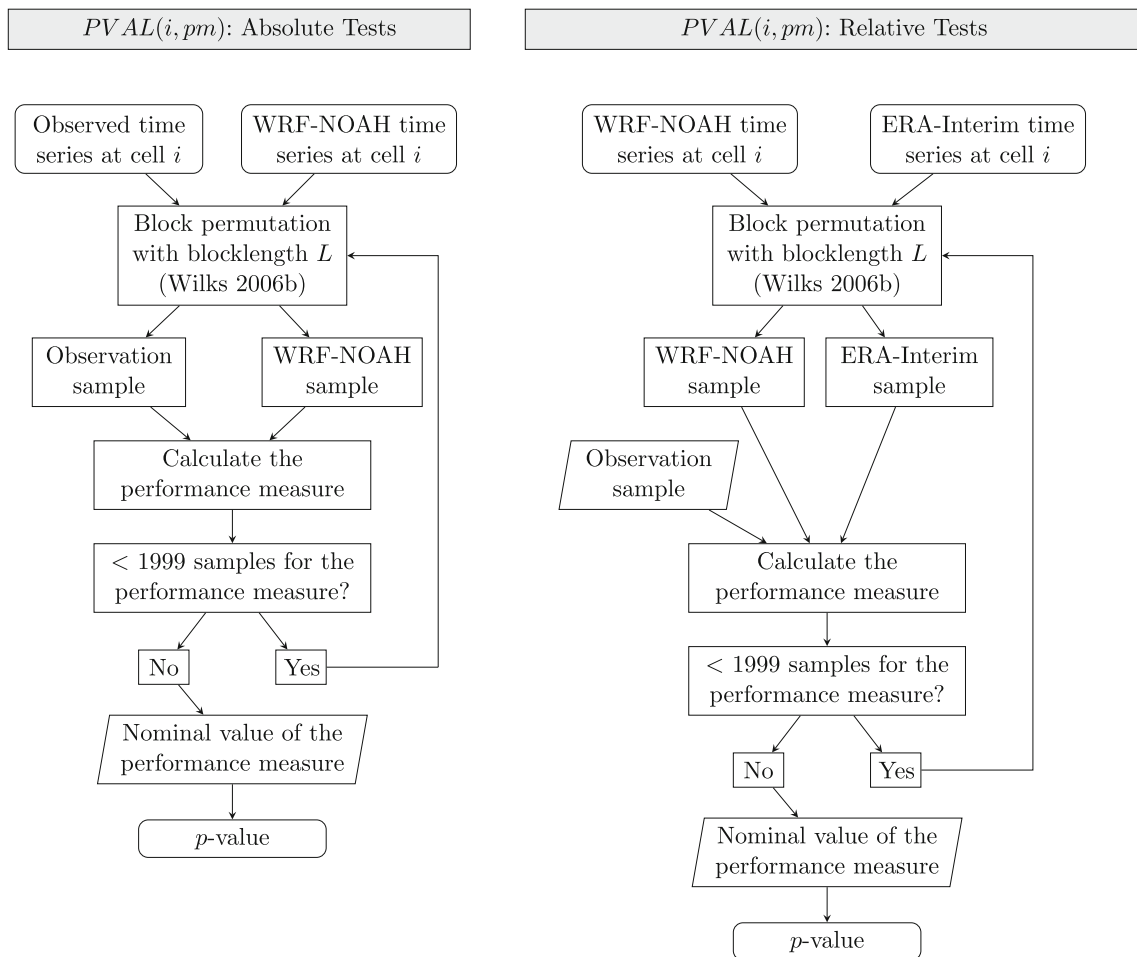


Fig. 1 $PVAL(i, pm)$: the procedure that performs the local permutation test at cell i and determines the p value for the performance measure pm for the absolute (left column) and relative (right column) tests

non-overlapping sequences of a certain length L , called blocks. The blocklength L must be large enough to ensure that the temporal autocorrelation structure in the original series is retained and also that data values separated by a time period of length L or more are essentially independent (Wilks 2006b). Assuming that the interannual autocorrelations are negligible compared to the intraseasonal ones, we choose a blocklength $L = 3$ months.

2.2.2 Field significance

If the local tests at the grid cells were independent, the null hypothesis were in reality true for each of them, and all tests were performed at a significance level α , then the average proportion of the significant tests would converge to α as the number of tests tends to infinity. In geophysical applications, the number of tests on the map is finite and the tests are dependent due to spatial autocorrelations. Each of these effects works to inflate the proportion

of falsely rejected tests. Thus, the expected value of that proportion can become substantially larger than α (Livezey and Chen 1983). As we need interpretable spatial patterns, this is not tolerable. The issue is known as test multiplicity (e.g. Katz and Brown 1991) and is solved by the so-called multiple comparison or field/global significance tests. The null hypothesis of the latter, also called *global* null hypothesis, is that *all* local tests are true. One option is to reject the global null hypothesis when the local rejections exceed a certain number, dependent on the total number of tests K and the degree of spatial dependence (Livezey and Chen 1983). However, the power of this test is reduced as the binary nature of the local tests ignores the strength of local evidence and also because the test statistic takes discrete (integer) values. There is also no indication which local tests are significant (Wilks 2006a). Another option is to use a global test statistic that depends on the magnitudes of the individual p values. It leads either to tests based on the minimum p value or to the false discovery rate (FDR)

approach. Here, we apply the FDR procedure of Benjamini and Hochberg (1995). It controls the FDR, which is the expected proportion of falsely rejected local tests out of all rejected tests. A local test is rejected if its p value does not exceed the FDR threshold:

$$p_{FDR} = \max_{i=1, \dots, K} \{p_{(i)} : p_{(i)} \leq q(i/K)\}, \quad (3)$$

where $p_{(i)}$ is the i th smallest p value and q the desired FDR. A single local rejection warrants global significance. Note that the null hypothesis involves all tests, whereas the alternative is local. The method yields one of the most powerful, yet slightly conservative, multiple comparison tests. As the significance level of the global test is numerically equal to the FDR q , the expected proportion q of false rejections out of all rejections is tightly controlled (Wilks 2006a). This makes the spatial pattern of rejected local tests interpretable. Ventura et al. (2004) showed that the method is robust to spatial autocorrelations, which makes it applicable in an atmospheric science context.

Figure 2 shows a schematic representation of the $FSIG(stat; I, \alpha)$ procedure that determines the field significance at level α of a statistic $stat$ over a spatial domain I . First, the routine $PVAL(i, stat)$ that calculates the local p value for $stat$ at grid cell i (see Fig. 1) is applied for all $i \in I$. This yields a set of p values for $stat$ over the spatial domain I . This set is then fed into the FDR algorithm with nominal FDR level $q = \alpha$ to finally produce a map of

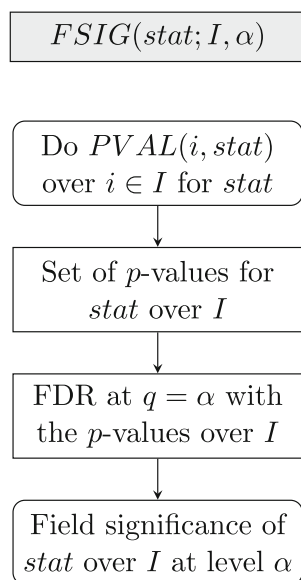


Fig. 2 $FSIG(stat; I, \alpha)$: Generic procedure for determining the field significance at level α of a statistic $stat$ over a spatial domain I . $PVAL(i, stat)$ is the respective procedure for determining the local p value for $stat$ at grid cell $i \in I$ (see Fig. 1), FDR is the False Discovery Rate procedure, and q the nominal FDR level

locally significant tests. All significance tests in this work are performed at the $\alpha = 5\%$ significance level.

Subregional analyses Are the performance measures also significant in each of the four German subregions (see Section 3.1) we consider? Hereafter, the field significance for a subregion and the corresponding test will be termed “subregional significance” and “subregional test”, respectively. In each subregion, the total number of points and the average rank proportionately diminish, so that the expected ratio i/K is the same as in the whole region. However, according to Eq. 3, for a p value to become significant, it is enough that only a *single* p value with a higher rank becomes significant, whereas for a p value to lose significance *all* higher-rank p values must lose significance. So, the probability is higher for a p value to gain than to lose significance. Hence, the total number of significant p values is expected to increase when subregionally tested. Depending on the spatial distribution of the p values, new patterns of significance may occur. In this work, we consider subregional testing only in case it reveals new features and qualitatively modifies results.

3 Climatology of Germany and data sets

3.1 Landscape and climatology of Germany

Landscape Figure 3 shows a topographical $2' \times 2'$ map of Germany with the major landforms and cities labelled. The landscape of Germany can be divided into three distinct parts, from north to south namely North German Lowlands, Central German Uplands, and South Germany. The terrain in the North German Lowlands is flat and mostly below 100 m above mean sea level. The East and North Frisian Islands as well as Germany’s largest island of Rügen are also part of the Lowlands. The Central German Uplands consist of plateaus and low mountain ranges separated by river valleys. Some of the most conspicuous elevations are the Taunus (879 m), Rhön (950 m), Harz (1142 m), Fichtel (1051 m), and Ore Mountains (1215 m) as well as the Thuringian (982 m), Bavarian (1121 m), and Bohemian (1456 m) Forests. South Germany has complex terrain with middle and high mountain ranges separated by river valleys and plateaus. Here belong the Black Forest (1493 m), the Swabian Jura (1015 m), and the Bavarian Alps (2962 m).

An idea about the dominant vegetation types in Germany can be obtained from Fig. 4, which shows the vegetation types in WRF-NOAH. Most of the territory of Germany is croplands, mountains and hills are dominated by mixed and coniferous forests, the areas around the largest cities are urban.

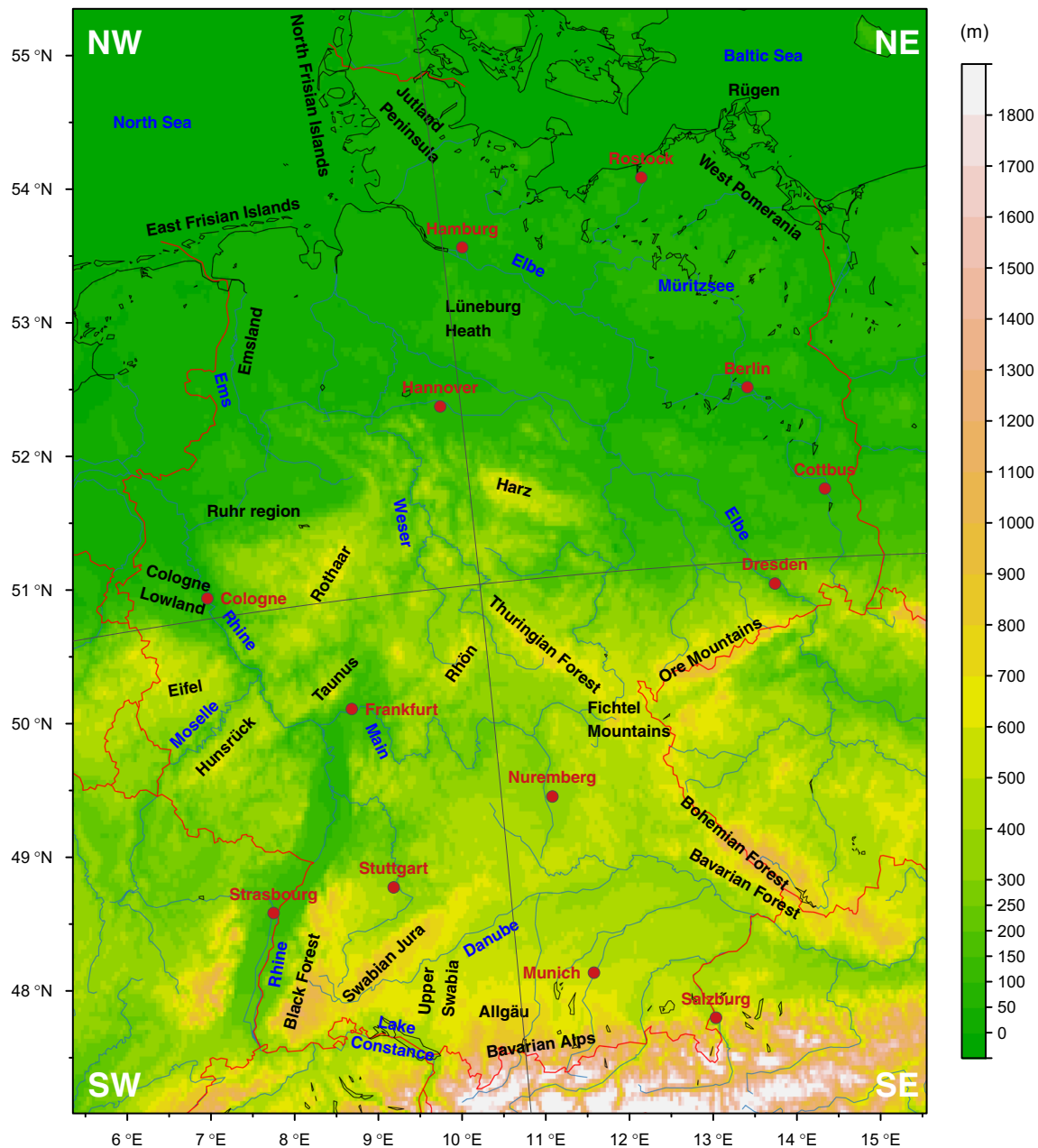


Fig. 3 Topographical map of Germany. The $2' \times 2'$ gridded relief (ETOPO2v2g, <http://www.ngdc.noaa.gov/mgg/fliers/06mgg01.html>) as well as the rivers, shore- and borderlines (GSHHG: the Global Self-consistent, Hierarchical, High-resolution Geography Database, <http://www.ngdc.noaa.gov/mgg/shorelines/shorelines.html>) data are available on the web site of the US National Oceanic and

Atmospheric Administration. Height in metres above mean sea level is plotted in colour scale; cities are displayed as *red dots*, rivers, shore, and borderlines as *blue*, *black*, and *red curves*, respectively. Some major forms of relief are labelled in *black*, water bodies in *blue*, and cities in *red*. The *grey lines* define the NW, NE, SW, and SE German subregions (see text), which are labelled in *white*

Climatology The climatology of mean 2 m air temperature for the 1989–2009 period is shown in Fig. 5. In winter, temperature decreases from north-west ($3\text{--}5\text{ }^{\circ}\text{C}$) to south-east ($-5\text{ to }0\text{ }^{\circ}\text{C}$). In summer, it strongly follows the topography, lowlands and river valleys being warmer ($17\text{--}20\text{ }^{\circ}\text{C}$) and elevated areas cooler ($11\text{--}14\text{ }^{\circ}\text{C}$, till $4\text{--}5\text{ }^{\circ}\text{C}$ at the highest peaks of the Bavarian Alps). The North Sea influence makes

the North Coast relatively warmer in winter and cooler in summer. The temporal variability of monthly temperature, as measured by the standard deviation, is about $1\text{--}2\text{ }^{\circ}\text{C}$ larger in winter than in summer (not shown).

Subregions We divide Germany's projection on the WRF-NOAH grid into four semi-equal rectangular parts to be

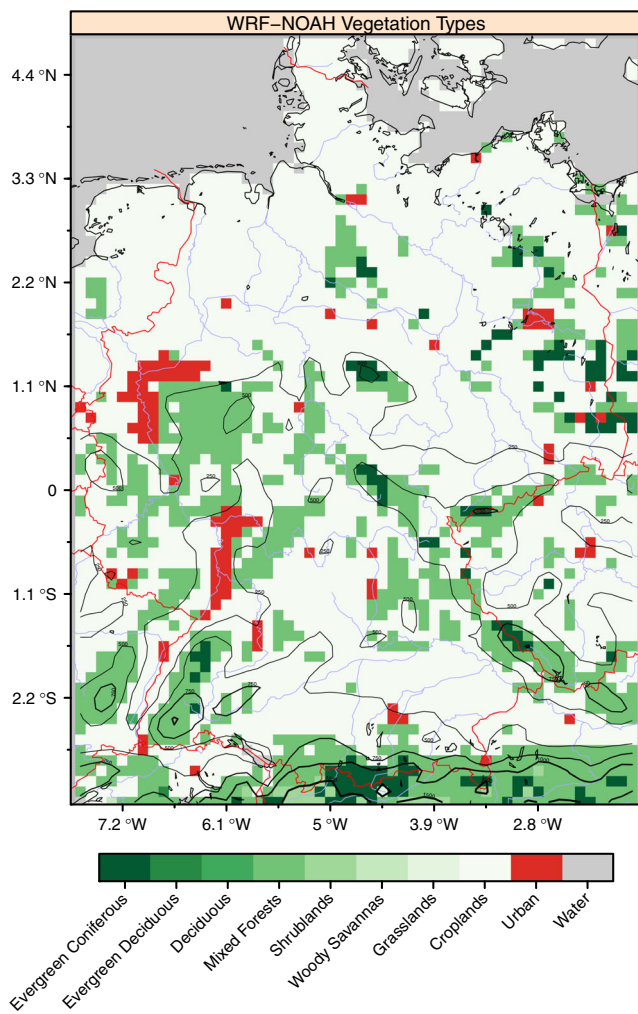


Fig. 4 Vegetation types in WRF-NOAH. *Black contour lines* display the WRF-NOAH orography; rivers, shore-, and borderlines are shown as *blue, black, and red lines*, respectively

studied in more detail. Namely, we use the 4.89° W meridian and 0.55° N parallel, visualised on the geographic projection of Fig. 3 as grey lines, to define north-west (NW), north-east (NE), south-west (SW), and south-east (SE) Germany.

3.2 Data

Observations The observation-based 2 m monthly mean temperature data are a rasterised product of the National Climate Monitoring Department of the German Weather Service (Deutscher Wetterdienst, DWD). They are derived from monthly means of surface air temperature at stations of the DWD network and have a spatial resolution of 1 km (Maier and Müller-Westermeier 2010).

Reanalysis ERA-Interim is the third-generation reanalysis of the European Centre for Medium-Range Weather

Forecasts (ECMWF) for the period 1979–present at a spatial resolution of approximately 0.75° or 79 km (Dee et al. 2011). We use 6-hourly 2 m temperatures, projected on a grid of $0.6^\circ \times 0.6^\circ$ ($\sim 42 \text{ km} \times 55 \text{ km}$) from the original Gaussian reduced grid.

WRF-NOAH simulation The object of this study is a standard EURO-CORDEX evaluation simulation provided by the University of Hohenheim (UHOH) (Warrach-Sagi et al. 2013). The WRF model version 3.3.1 is run with the land surface model NOAH (Chen and Dudhia 2001a, b) for a hindcast evaluation over the period 1987–2009. The model operates one-way nested over the standard EURO-CORDEX domain on a rotated longitude-latitude grid with horizontal resolution of $0.11^\circ \times 0.11^\circ$ (EUR-11). The vertical is described by 50 layers up to 20 hPa. The simulation is driven by the 6-hourly ERA-Interim reanalysis at the lateral boundaries and daily sea surface temperature data also from the reanalysis. The relaxation zone around the model domain is 30 grid cells wide and the time step is 60 s. The physical package includes the Morrison two-moment microphysics scheme (Morrison et al. 2009), the Yonsei University atmospheric boundary layer parametrisation (Hong et al. 2006), the Kain-Fritsch-Eta Model convection scheme (Kain 2004), and the Community Atmosphere Model (CAM) shortwave and longwave radiation schemes (Collins et al. 2004). Soil moisture and temperature profiles were initialised on 1st January 1987 from ERA-Interim after interpolation to the NOAH model. The WRF Preprocessing System (WPS) uses the $30''$ land-cover data from the Moderate Resolution Spectroradiometer (MODIS), classified according to the International Geosphere-Biosphere Programme (IGBP). The soil textures are from the $5'$ data of the Food and Agriculture Organization of the United Nations (UN/FAO). To reduce at least some spin-up effects that may distort the model results, the evaluation starts in the winter of 1989/1990. We use 3-hourly model output fields of 2 m temperature. Note that Vautard et al. (2013) and Kotlarski et al. (2014) evaluated this simulation as a part of the EURO-CORDEX RCM ensemble for Europe on a 25 km scale.

3.3 Data processing

The evaluation is over the territory of Germany and is model-oriented, which ensures a fair performance estimate that takes into account the limited spatial resolution of the model. Correspondingly, the DWD observations and ERA-Interim reanalysis were transformed by quadratic inverse-distance-weighted interpolation to the WRF-NOAH grid using interpolation radius of 11 and 50 km, respectively. The observed and the ERA-Interim temperature data were first reduced to sea level assuming a spatially and temporally uniform lapse rate of 6.5°C/km , then interpolated to

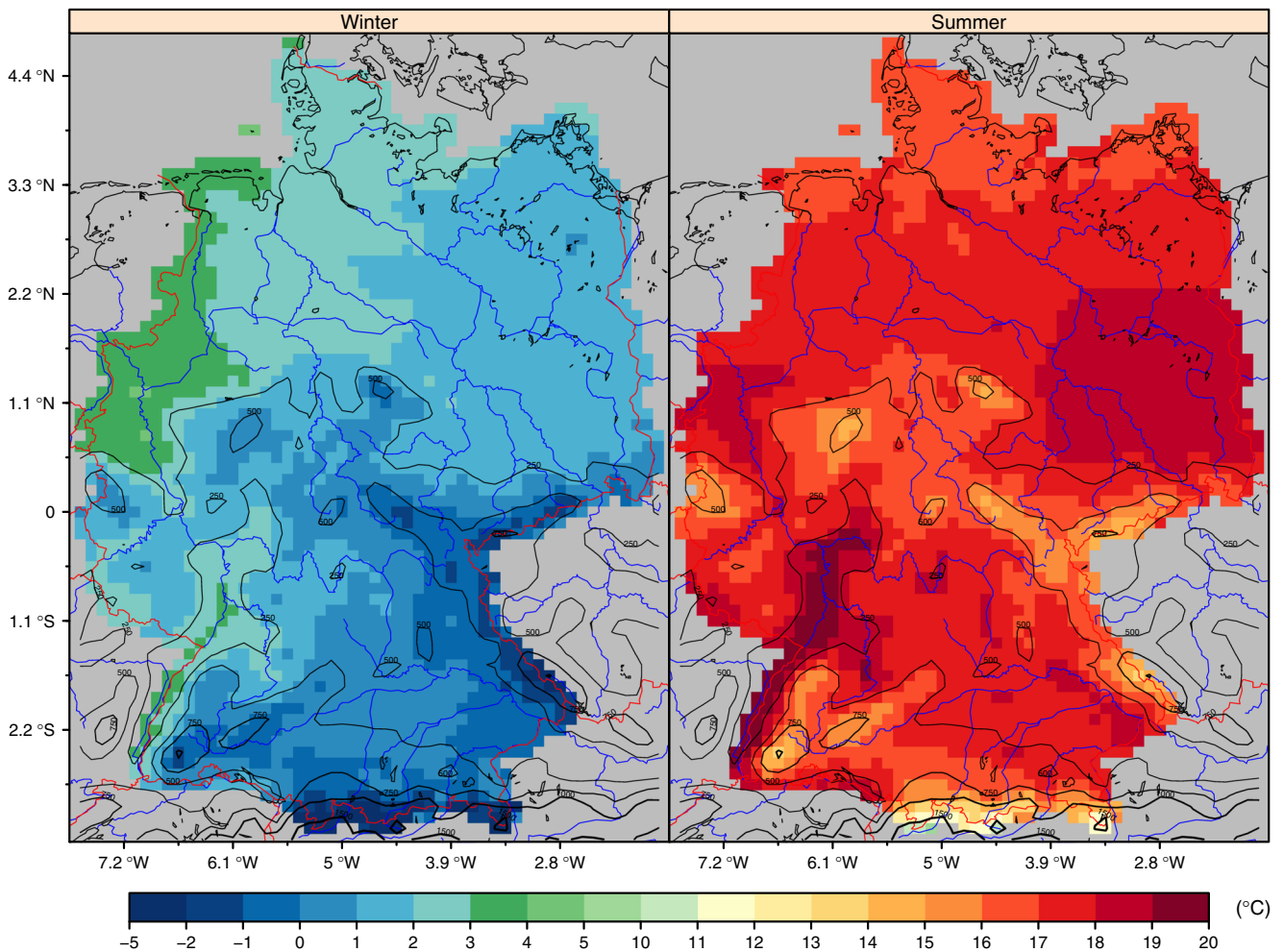


Fig. 5 Mean temperature in winter (left panel) and summer (right panel) for the 1989–2009 period. Black contour lines display the WRF-NOAH orography; rivers, shore-, and borderlines are shown as blue, black, and red lines, respectively

the WRF-NOAH grid, and finally reduced to the WRF-NOAH orography under the same lapse rate assumption. In reality, the average lapse rate in winter is about 5 °C/km and in summer about 7–8 °C/km (e.g. Gao et al. 2012). The average of these lapse rates yields a simple, albeit not optimal, correction that alleviates the elevation dependence of temperature bias. Thus, it enables inter-comparison of different-resolution data and is widely applied (e.g. Jacob et al. 2007; Heikkilä et al. 2011; Vautard et al. 2013; Kotlarski et al. 2014).

After the spatial regridding, from the WRF-NOAH and ERA-Interim outputs we built data sets for mean monthly temperature, from 00 to 00 GMT in the next month. To ensure equal numbers of years for all seasons and to avoid putting months belonging to the same winter season into different years, we investigate the 20-year period 1st December 1989–30th November 2009.

We estimate the added value of the downscaling relative to the large-scale driving reanalysis interpolated on the RCM grid. This preserves the high-resolution climate features,

which are our primary interest. The same approach is followed, e.g. in Sotillo et al. (2006) and Winterfeldt et al. (2011).

4 Results and discussion

In the following, we generalise evaluation results as a (systematic) cold/warm bias when most of the considered temperature characteristics are under-/overestimated. The terms *under-/overprediction* are used as alternative description of the *frequency bias*. For brevity, we only show results for winter (DJF) and summer (JJA). The maps are displayed on the EUR-11 grid of WRF-NOAH, where the evaluation takes place. Note that due to the higher local temporal variability of monthly temperature in winter, the bootstrap sampling variability of the test statistics is also larger in winter. Therefore, in winter, the deviation of a test statistic from its expected value under the null must be correspondingly larger than in summer to be significant.

4.1 Basic diagnostics

Prior to the spatial analysis, to get an overall impression of the model performance, we pool the gridpoint time series together and visually compare the empirical probability density functions (PDFs) of the observations and the model data sets as displayed in Fig. 6.

Results for whole Germany are shown in Fig. 6a. In both seasons, the overall PDFs of WRF-NOAH are clearly closer to reality than the ERA-Interim ones: the Perkins score is 9–10% larger for WRF-NOAH. The performances of both

models are slightly better in winter than in summer (2–3% difference for the Perkins score). The ERA-Interim PDFs are too flat and have too long left tails, particularly in winter; the quantiles are correspondingly underestimated. These problems are apparently solved by the WRF-NOAH downscaling. Nevertheless, WRF-NOAH is also not perfect. In winter, temperatures at the middle part of the distribution are overforecast at the expense of temperatures at the left tail. This indicates overestimation of the quantiles up to about the median. In summer, the WRF-NOAH distribution is slightly shifted to the left: temperatures below the mode

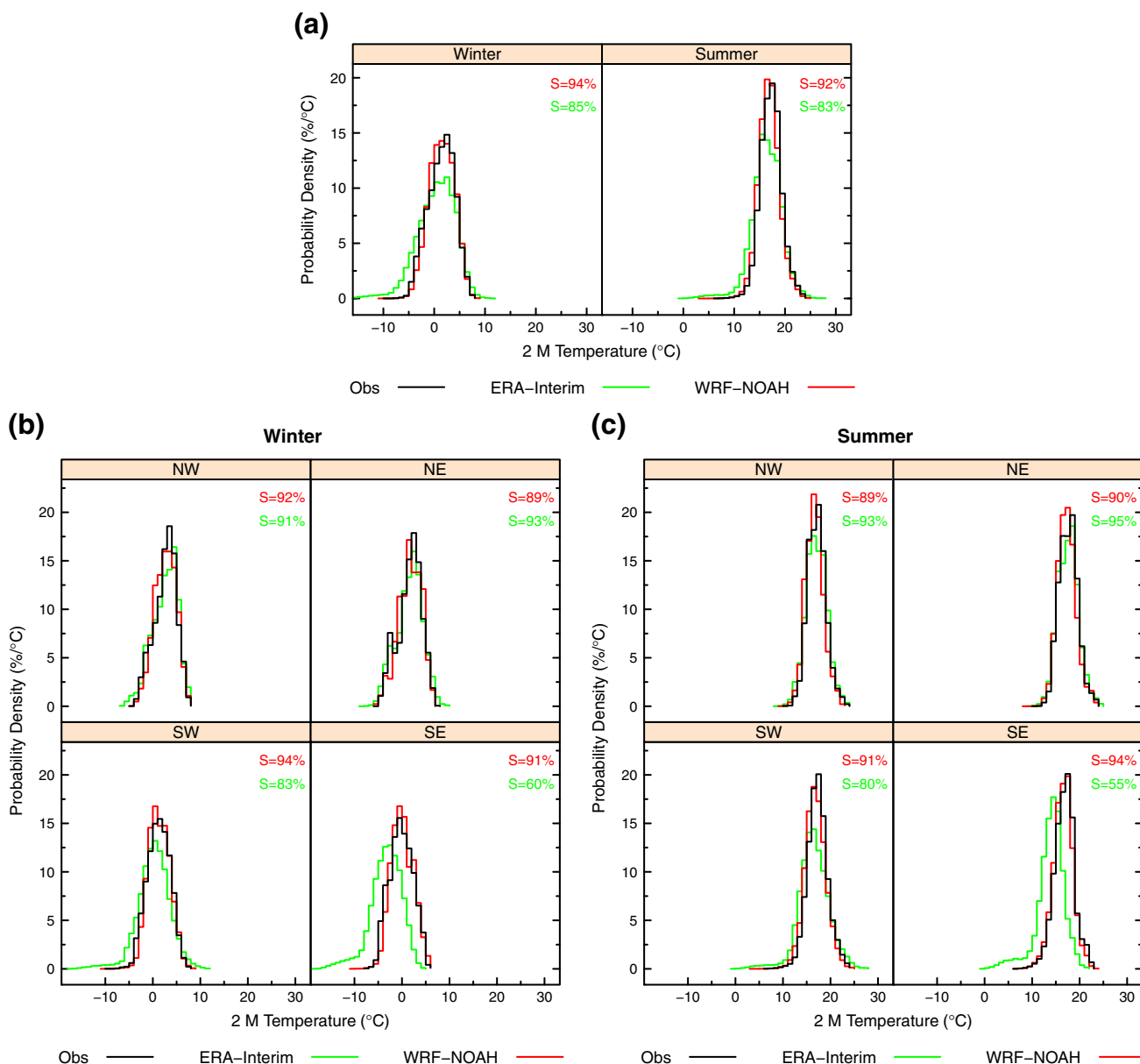


Fig. 6 Probability density functions and Perkins score *S* of monthly mean temperatures for **a** Germany (left panel: winter, right panel: summer), **b** and **c** the German subregions (see text) for winter and summer, respectively, from the DWD observations (black), ERA-Interim

reanalysis (green), and WRF-NOAH downscaling (red). The bin size is 1 °C. The DWD and ERA-Interim temperatures were reduced to WRF elevations assuming a lapse-rate of 6.5 °C/km (see text)

of the distribution are overforecast at the expense of temperatures higher than the mode; temperature quantiles are correspondingly underestimated.

Figure 6b and c show the three PDFs for each of the four German subregions for winter and summer, respectively. In both seasons, the performance of WRF-NOAH in North Germany is comparable to or a bit worse than in South Germany; the Perkins score is about 90%. In contrast, the performance of ERA-Interim in the North is considerably better than in the South; the Perkins score ranges from 90–95% in the North to 50–80% in the South. ERA-Interim overforecasts lower at the expense of intermediate temperature categories in all subregions, the strongest in SE Germany, where the whole density distribution is shifted to the left; temperatures at the right tail are overpredicted by the reanalysis everywhere except in SE Germany. This indicates that ERA-Interim underestimates temperature quantiles, most strikingly in SE Germany. The overall deficiencies of WRF-NOAH are also observable in each subregion. In summer, the overforecasting of lower temperatures at the expense of higher is most pronounced in NW and SW Germany.

4.2 Absolute performance

The only field significant test is that for the bias of the mean in summer, shown in the leftmost panel of Fig. 7. The tests for the biases of the lower quartile, the median, the upper quartile as well as for the frequency bias of the above normal temperature category reject only when tested in subregions

in summer. They are shown in the rest of the panels of Fig. 7 and in Fig. 8, respectively. As seen, the significant grid cells are usually a small number and concentrated in specific regions (river valleys, coastal zones, urban areas, lakes, etc). Thus, they form clear spatial patterns, which can be linked to physical processes known to have the same geographic fingerprint in summer.

There is a systematic cold summer bias that is locally significant in deep river valleys, coastal areas, and over large lakes. Heavily urbanised river valleys and areas around some cities exhibit a systematic and significant warm summer bias. The area of cold bias is larger than that of warm bias, and is the largest in NW and SW Germany. This is consistent with the results from Section 4.1. Note that the warm winter bias seen in Fig. 6 is not field significant. The downscaling is better in winter, when there is no evidence against the null hypothesis of perfect performance. This is also in line with the analysis of the overall PDFs in Section 4.1.

4.2.1 Discussion

The major spatial and seasonal patterns of absolute downscaling performance are now discussed in detail.

- (1) Cold summer bias in deep river valleys like the Danube river valley neighbouring Upper Swabia (Fig. 7, first, third, and fourth panel and Fig. 8), the southernmost part of the Rhine river valley (Fig. 7, first panel), and the Moselle river valley (Fig. 7, first and third panels).

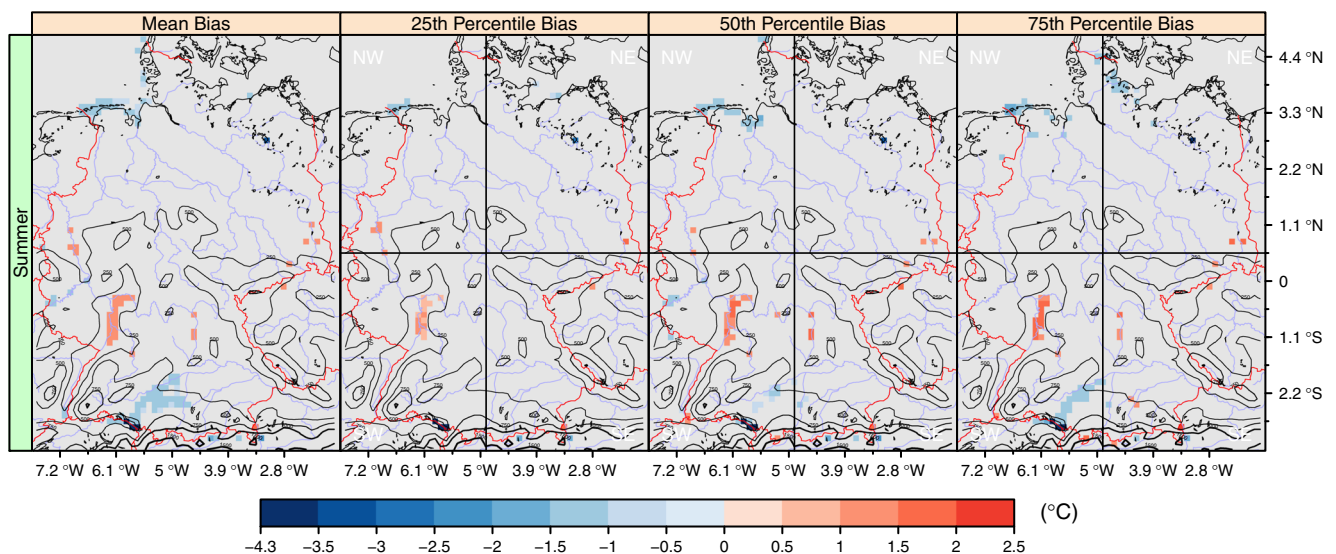


Fig. 7 Biases of selected distributional characteristics for monthly mean temperature in summer, simulated with WRF-NOAH and in comparison to high-resolution observations from the DWD. A grid cell is plotted only if the respective test is locally significant at the 5% level of field significance. The grey straight lines at the 4.89° W meridian

and the 0.55° N parallel define the four German subregions (see text) and indicate subregional testing. DWD temperatures were reduced to WRF-NOAH elevations assuming a lapse rate of 6.5 °C/km (see text). Black contour lines display the WRF-NOAH orography; rivers, shore-, and borderlines are shown as blue, black, and red lines, respectively

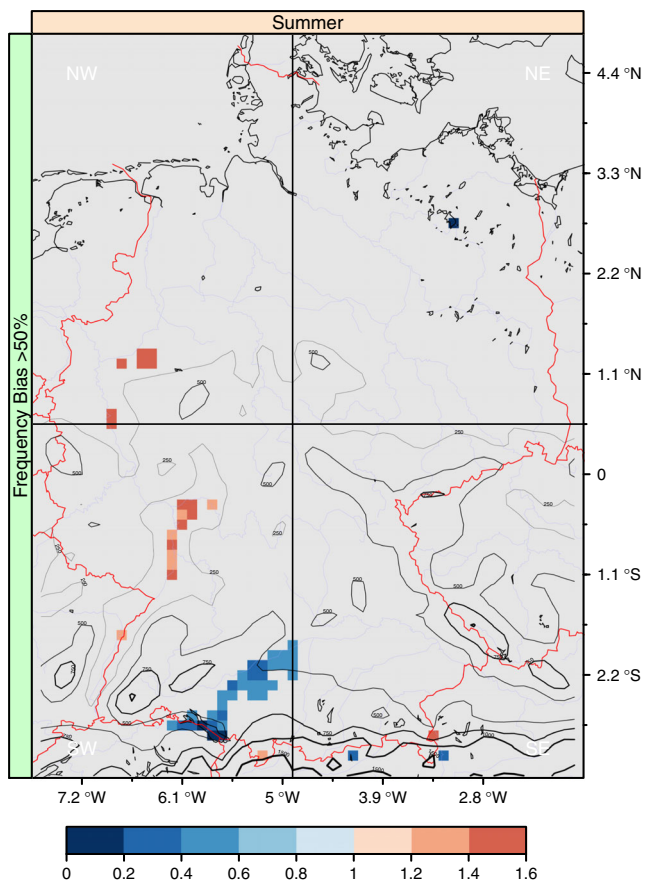


Fig. 8 Same as Fig. 7 but for the frequency bias of the “above normal” category ($X\%$ stands for the X th observed percentile for the 1980–2009 climatological period) of monthly mean temperature

Temperatures are underestimated by 1 to 2 °C and the frequency of above normal temperatures by 40 to 80%.

Figure 9 displays (a) the height differences between the WRF-NOAH and the DWD observation data. As seen, the altitude of deep river valleys is overestimated in the WRF-NOAH model. Summer lapse rates of mean monthly temperature in deep river valleys can substantially exceed the standard atmosphere value of 6.5 °C/km (e.g. Rolland 2003) we use to transform the observations to the WRF-NOAH grid. As a result, the observations are warm biased at the model altitude which leads to a cold model bias. Note that this problem is rooted in the still too coarse WRF-NOAH orography, which necessitates the lapse-rate correction.

- (2) Cold (up to -2 °C) summer bias in the north coast with the Frisian Islands (Fig. 7).

In part 2, we show that the cold summer bias in these areas coincides with a dry bias. A too intense sea breeze circulation is a plausible explanation. However, until further analysis is done, this hypothesis remains speculative.

- (3) Strong cold summer bias over large lakes like lakes Constance and Müritzsee (Figs. 7 and 8). Temperatures are underestimated by 3 to 4.3 °C and the frequency of above normal temperatures by more than 80%.

As seen in Fig. 4, which shows the WRF-NOAH dominant vegetation categories, these lakes are water-covered regions in WRF-NOAH. As the lakes are absent in ERA-Interim, the simulation uses prescribed sea surface temperatures (SSTs) that are spatially interpolated by the WPS on a daily basis from ERA-Interim SSTs. As a result, water temperatures from the north coast are used, which are too cold and lead to nonphysical temperature discontinuities. The adverse effect of spatially interpolating water temperatures when lakes are absent in the large-scale driving data is documented in Mallard et al. (2015).

- (4) Warm summer bias in heavily urbanised river valleys like the Rhine river valley between the Black Forest and Taunus, the Cologne Lowlands, the Ruhr region, and also around some cities like Munich, Stuttgart, Nuremberg, Dresden, Cottbus, as well as the foreign cities of Salzburg and Strasbourg (Figs. 7 and 8). Temperatures are overestimated by up to 2.5 °C and the frequency of above normal temperatures by up to 60%.

As seen in Fig. 4, these areas are indeed “urban” land use category in WRF-NOAH. Generally, climate stations, which are the basis for gridded monthly mean temperature data sets, are located not in urban but in open grassland areas, e.g. close to airports. This can cause a cold bias in the observations. It has already been recognised that with increasing grid resolution urban heat islands need to be accounted for also in gridded observational data (e.g. Fujibe and Ishihara 2010; Gallo and Xian 2014). In Fig. 4, there are urban regions that do not exhibit a warm summer bias, e.g. Berlin, Hamburg, Hannover, Rostock. Note that most of the urban areas with significant warm summer bias are situated in concave topography forms like river valleys and kettles, while the urban areas where this bias is not detected are located in open flat areas. A plausible explanation is that concave topography favours calmer conditions, hence stronger urban heat islands.

- (5) The lack of evidence against the null hypothesis of perfect model performance in winter indicates that the downscaling is overall better in winter than in summer.

Obviously, the effects do not appropriately scale in winter and are masked by the higher variability.

We note that a general cold summer bias over Central Europe has been documented for most members of the 0.11° EURO-CORDEX RCM ensemble (Kotlarski et al. 2014) and also in WRF simulations at 0.44° grid resolution

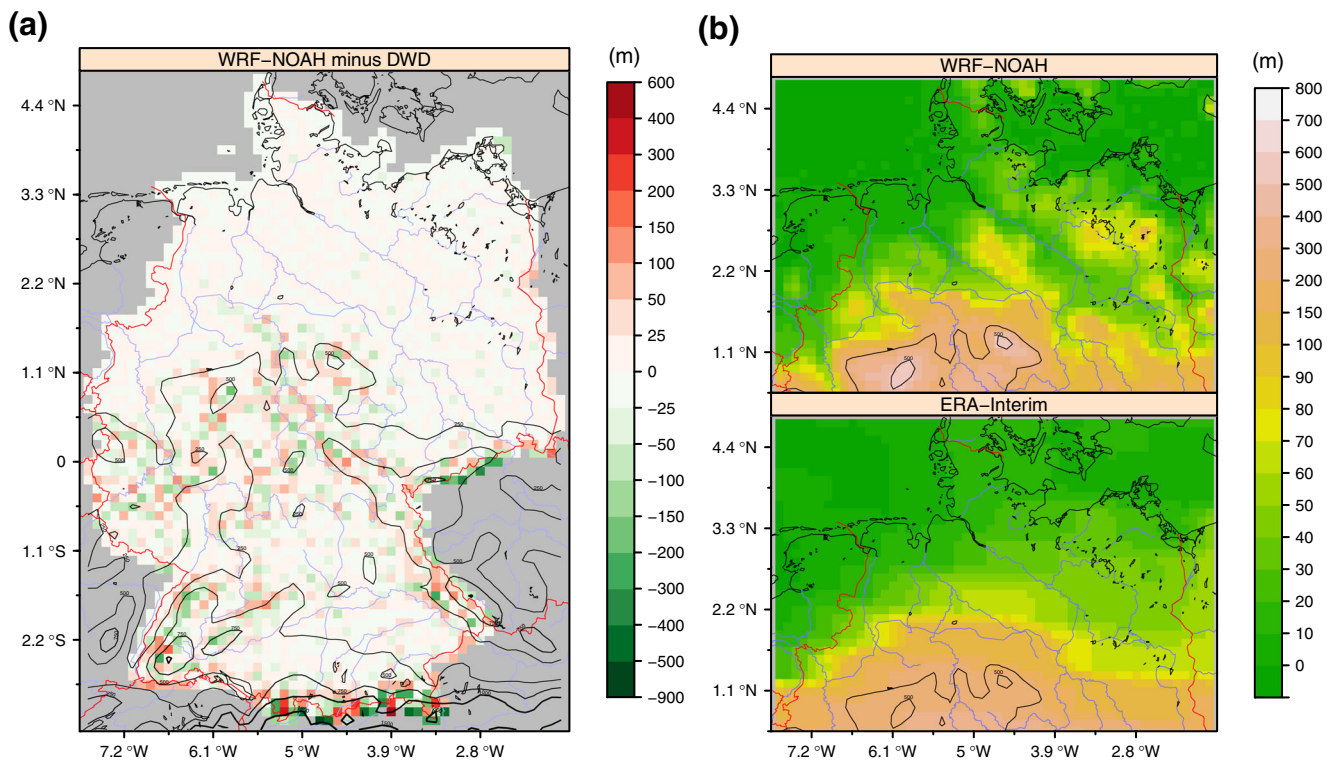


Fig. 9 Heights of orography. **a** Difference in height above mean sea level between the WRF-NOAH and the DWD data. **b** Height above mean sea level of WRF-NOAH (upper panel) and ERA-Interim (lower

panel) for North Germany. Black contour lines display the WRF-NOAH orography; rivers, shore-, and borderlines are shown as blue, black, and red lines, respectively

(Mooney et al. 2013; Katragkou et al. 2015). In these studies, the areas of model biases are large and therefore often not directly interpretable. Katragkou et al. (2015) calculate statistical significance locally but do not account for test multiplicity, which entails that an intolerably large proportion of the significant tests could be due to chance and hence not interpretable. In turn, the methodology we demonstrate accounts for test multiplicity and is robust to spatial autocorrelations. Thus, it objectively picks out the grid cells to interpret, of which only 5% on average are mistaken. As a result, the significantly biased area is spatially confined to specific spatial patterns that immediately point to issues like overestimated altitude of river valleys, a possibly exaggerated sea breeze circulation, and the interpolation of lake temperatures behind the cold summer bias over Germany.

4.3 Relative performance

In contrast to the absolute performance measures, all considered relative measures are field significant. Figure 10 shows the field significance test results for the Perkins skill score (1), the biases of the mean and the quartiles (2) and Fig. 11 for the frequency biases (1). These results reveal the spatial patterns of the significant added value with respect to the different distribution characteristics. Again, these patterns comprise specific geographical areas,

which allows us to link the added value to the respective improved/deteriorated physical processes that have the same geographic and seasonal fingerprint.

The spatial patterns of significant added value are consistent among the measures (includingly the bias of the median, which is not shown for brevity). The downscaling is significantly better than the driving reanalysis over regions of complex orography and in some coastal areas. The strong cold bias of WRF-NOAH in summer deteriorates its relative performance over some coastal areas and the plains of North Germany. Over lake Müritzsee the performance is deteriorated in both seasons. In most of the North German Lowlands, no added value is detectable.

The area, where WRF-NOAH is significantly better than ERA-Interim, is much larger than the area (if any) where the opposite is true, and is the largest in SE Germany. So, overall, WRF-NOAH outperforms the simple spatial interpolation of ERA-Interim, most pronounced in SE Germany and is inferior to it only for some measures locally in North Germany. This is consistent with the conclusions from the analysis of the overall PDFs in Section 4.1.

4.3.1 Discussion

Before we turn to a detailed analysis of relative performance, we briefly review the current state of knowledge

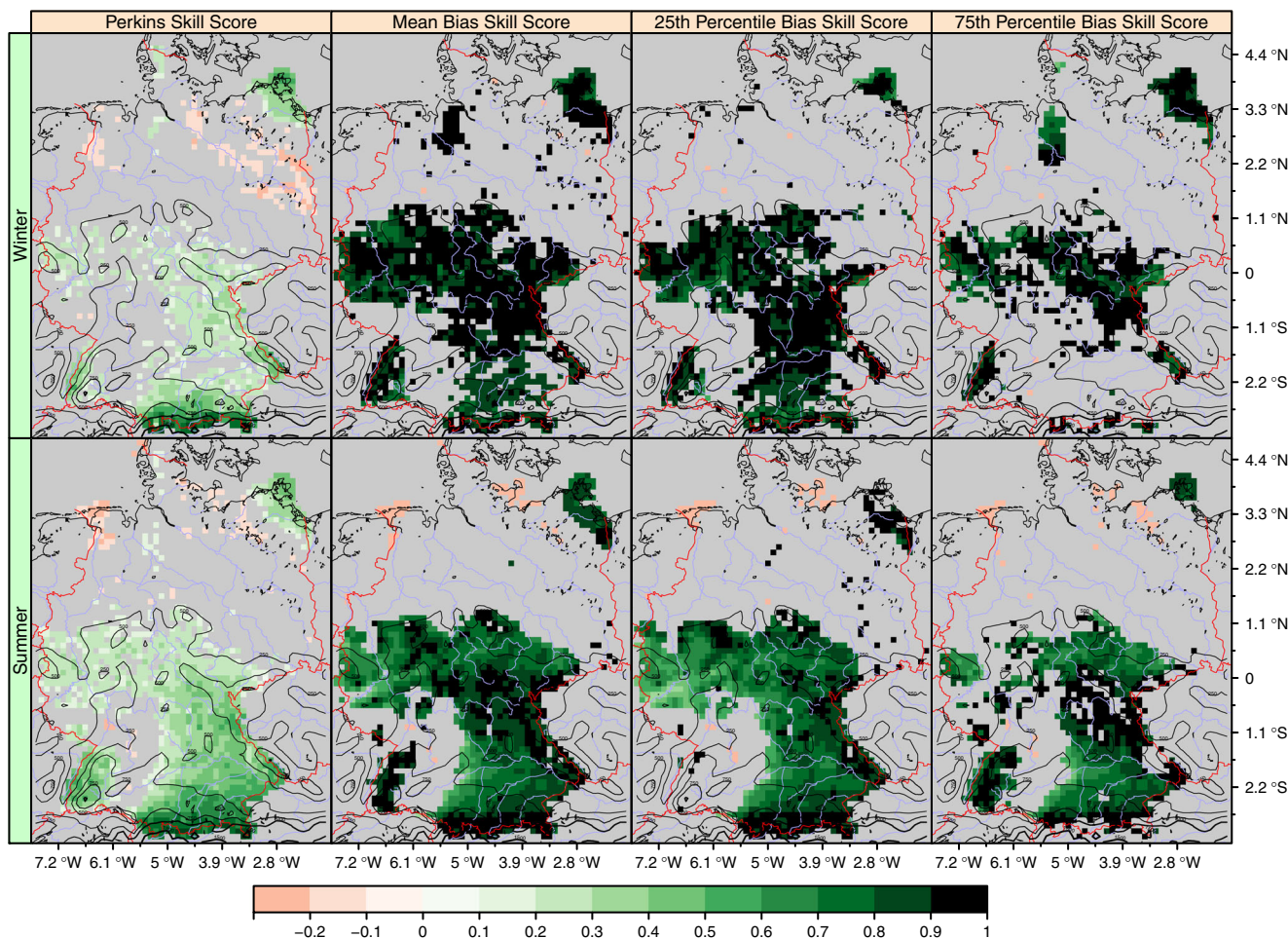


Fig. 10 Selected measures of the added value of the WRF-NOAH downscaling relative to the driving reanalysis with respect to monthly mean temperature in winter (*upper panels*) and summer (*lower panels*). The common observational reference are high-resolution DWD temperatures reduced to WRF-NOAH elevations assuming a lapse-rate

of $6.5\text{ }^{\circ}\text{C}/\text{km}$ (see text). A grid cell is plotted only if the respective test is locally significant at the 5% level of field significance. *Black contour lines* display the WRF-NOAH orography; rivers, shore-, and borderlines are shown as *blue, black, and red lines*, respectively

about the potential and limitations of regional climate modelling with focus on monthly temperature.

In the context of dynamical downscaling, large-scale variability is defined as the variability resolved by the large-scale driving data and fine-scale variability is associated with spatial and temporal scales, not explicitly resolved by the driving data, but resolved by the nested RCM. Clearly, an RCM can potentially add value only with respect to physical variables, climate statistics, regions and seasons, for which there is fine-scale variability. Fine-scale variability, quantified by the variance of the RCM field within a grid box of the driving field, has a stationary and a transient component. The stationary component is associated with small-scale quasi-stationary processes induced by small-scale stationary surface forcings. Therefore, spatially, this variability is present only over regions with localised stationary surface forcings, no matter of the season or temporal scale considered. The transient component

is physically related to small-scale transient processes spontaneously generated through a nonlinear cascade of variance from large to small scales, allowed for by the improved resolution of the thermohydrodynamics of the flow. Therefore, spatially, this variability can potentially be present anywhere in all seasons, but is only detectable at fine temporal scales (Laprise et al. 2008; Di Luca et al. 2012, 2013).

Considering the grid spacings of ERA-Interim and WRF-NOAH as lower limits of the resolved large and small spatial scales, we conclude that in our case fine-scale variability is generated by processes with spatial scales, roughly, between 10 and 100 km. Composite spectral analysis for a broad range of atmospheric fields reveals that the atmospheric processes with dominant spectral power in these spatial scales, have temporal scales approximately between 20 min and 1 day (Di Luca et al. 2012). Hence, monthly averaging should effectively cancel out the transient fine-scale variability, which means that the fine-scale variability is only stationary.

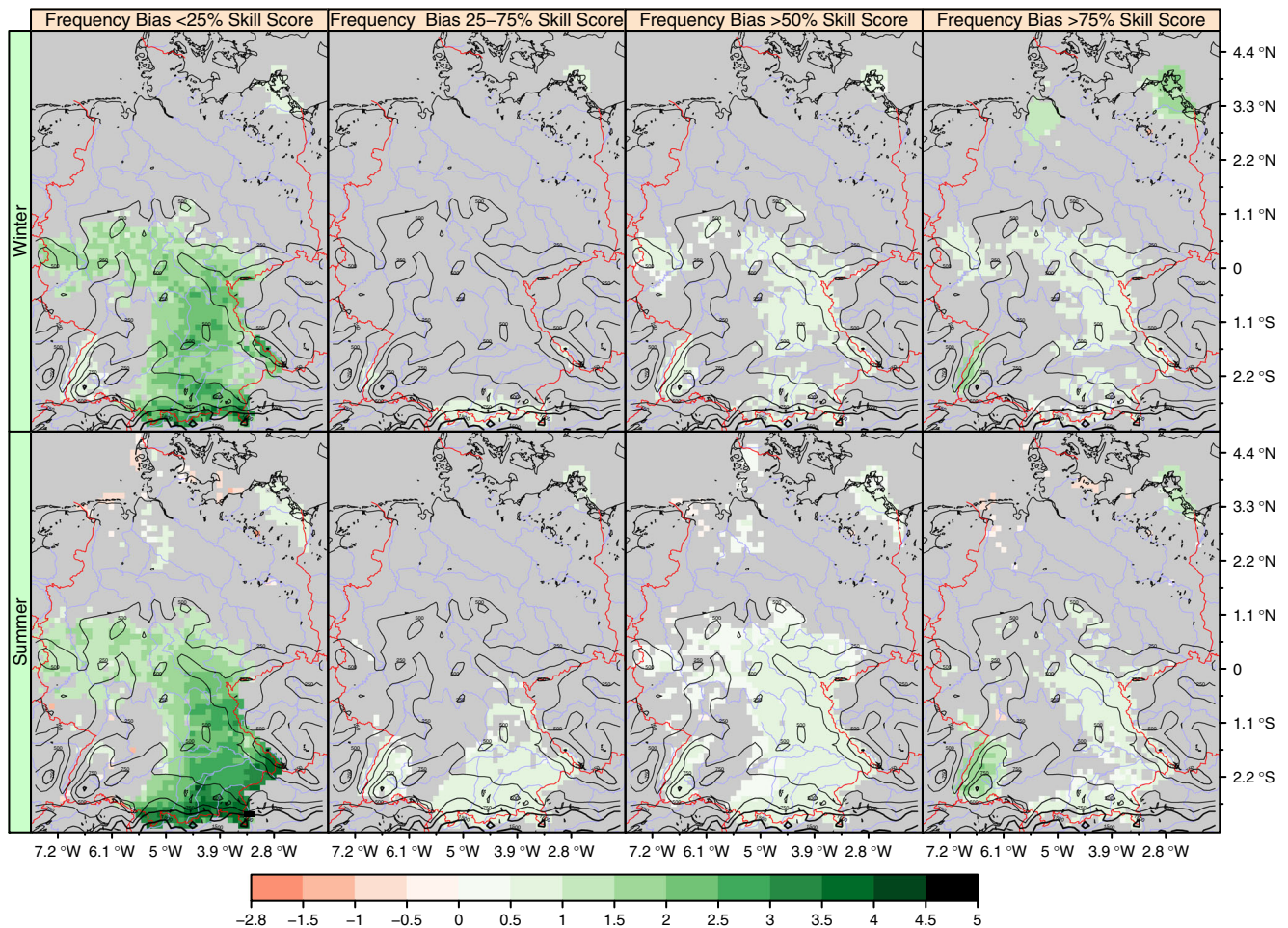


Fig. 11 Same as Fig. 10 but for the relative tests for the frequency biases of the “low”, “normal”, “above normal”, and “high” (from left to right) categories ($X\%$ stands for the X th observed percentile for the 1980–2009 climatological period) of monthly mean temperature

Localised stationary surface forcings for air temperature are found in regions of complex orography due primarily to the strong elevation dependence and in coastal areas due to the differential warming of water and land (Di Luca et al. 2013).

Now, we can proceed to a detailed discussion of the spatial and seasonal patterns of relative performance.

- (1) Positive added value in areas of complex orography like SE Germany, the Black Forest with the contiguous Rhine river valley, the Central German Uplands as well as some coastal areas like West Pomerania with the Rügen Island and parts of the Jutland Peninsula (Figs. 10 and 11).

Actually, for the monthly temperature field, which only has stationary fine-scale variability, positive added value can potentially be expected only in regions of complex topography, where stationary surface forcings for temperature are localised. WRF-NOAH obviously develops appropriate stationary fine-scale climatological features in response to such forcings.

The results are consistent with those of Prömmel et al. (2010). To be more specific, the downscaling alleviates the predominant ERA-Interim cold bias in areas of complex orography like SE Germany and the warm bias in the other mentioned areas (not shown).

- (2) No added value in the North German Lowlands (Figs. 10 and 11); deterioration of the PDFs in some parts of the lowlands like Emsland and the north-eastern lakelands (Fig. 10, first column).

The ERA-Interim driving reanalysis can be considered unbiased on large scales because it assimilates observed 2 m temperatures. As the monthly temperature field only varies on large scales over spatially uniform areas, the RCM does not have much room for improvement there. On the more, even slight deviations should lead to deterioration, i.e. negative added value in such regions. These results are also in line with the conclusions of Prömmel et al. (2010).

- (3) The question arises why the downscaling adds value over the spatially homogeneous region to the west of

the Lüneburg Heath (Fig. 10, first column, second and fourth panels on the upper row, Fig. 11, fourth panel on the upper row, first and third panels on the lower row).

This region is very flat, low, widely open to the North Sea, and surrounded by low hills to the east and to the south (see Fig. 3). As westerlies predominate, it is an isolated area of enhanced marine influence. Figure 9b shows the heights above mean sea level for WRF-NOAH and ERA-Interim for North Germany. Obviously, this topographical structure is caught by WRF-NOAH, but is completely absent in ERA-Interim. We speculate that the added value in this region is due to the inability of the reanalysis to reproduce this “island” of enhanced marine influence. As we show in Part 2, WRF-NOAH also improves the frequency of heavy summer precipitation specifically in this region, which lends further credence to this interpretation.

- (4) Negative added value over some coastal areas like the East Frisian Islands and the east coast of the Jutland Peninsula in summer (Fig. 10, lower row, Fig. 11, first and fourth panels on the lower row).

As already discussed, WRF-NOAH has a cold summer bias in these regions, probably related to an exaggerated sea breeze circulation.

- (5) Negative added value over lake Müritzsee in both seasons (Figs. 10 and 11, fourth panel on the upper row, first, third, and fourth panels on the lower row).

As already discussed, in summer WRF-NOAH has a cold bias over the lake. In winter, it has a warm bias of 1.3–1.5 °C (not shown), which although not significant relative to the observations, leads to a significant deterioration relative to the driving reanalysis. In both seasons, the problem is attributable to the spatial interpolation of water temperatures.

- (6) The area of positive added value does not considerably change with seasons, while that of negative added value tends to be larger in summer (Figs. 10 and 11).

The positive added value is due to the description of localised surface forcings that exist in both seasons; hence, the lack of seasonality is consistent. This also means that the improvement appropriately scales in winter so that it is detectable despite the larger variability. The negative added value is more pronounced in summer, as the smaller bootstrap sampling variability of the test statistics makes deteriorations easier to detect than in winter.

5 Conclusions

A new methodology for rigorous spatial analysis of the downscaling performance of regional climate simulations

is introduced. It is based on a multiple comparison of the local test results by means of the false discovery rate (FDR) approach. Controlling the proportion of falsely rejected tests in a meaningful way and being robust to spatial dependence, the FDR method allows for spatial analysis of the pattern of local rejections. The latter is referred to as the spatial pattern of statistical significance. It includes the locations at which the values of evaluation statistics are highly unlikely to have occurred by chance. A novelty of the study is that it focusses on this pattern rather than on the magnitudes of the evaluation statistics. Indeed, high deviations of the values of evaluation statistics from their expected values under the null are not necessarily statistically significant and small deviations are not necessarily insignificant, because statistical significance depends also on the variability. A small deviation at a location with small variability may be significant, whereas a high deviation at a location with high variability might be insignificant. The sampling uncertainty of the evaluation statistics is rigorously estimated by means of a block permutation procedure that accounts for the time series structure of the data. New quantitative metrics for the added value relative to the driving large-scale field are developed.

The methodology is exemplarily applied to evaluate the winter and summer climatology of monthly mean temperature for the 1990–2009 period from a standard EURO-CORDEX simulation with WRF-NOAH at 0.11° grid resolution over Germany. It objectively selects the interpretable grid cells, of which only 5% on average are misidentified. The specific spatial patterns of statistical significance can be hypothetically linked to physical processes known to have the same geographic and seasonal fingerprints for the respective performance measure.

In most regions, the downscaled distributions are statistically indistinguishable from the observed ones. The still too coarse resolution of orography leads to overestimated altitudes and hence a cold summer bias in deep river valleys. The cold summer bias in the north coastal areas is attributable to an exaggerated sea breeze circulation. Spatial interpolation of water surface temperatures from distant colder sea regions leads to a cold summer bias over large lakes. Strong urban heat islands in heavily urbanised areas located in concave topography forms are not contained in the observations, which makes the downscaling appear systematically too warm in summer. The larger temperature variability in winter masks potential WRF-NOAH biases.

The climatology of mean monthly temperature is improved in both seasons, but only over regions of complex topography. This can be expected for a variable without transient fine-scale variability and a downscaling that generates appropriate stationary fine-scale features. There is no added value of the downscaling in spatially homogeneous areas, because of the negligible fine-scale variability. Moreover, even small biases of the large-scale field lead

to significant local deterioration in such areas. The cold summer and warm winter bias over lake Müritzsee, resulting from the interpolation of water temperatures, is also reflected as negative added value. The total area of positive added value is not seasonally dependent, as it encompasses regions of localised surface forcings that exist throughout the year.

This “perfect-boundary” evaluation suggests that the WRF-NOAH downscaling system generates appropriate fine-scale features in the monthly temperature field over regions of complex topography. As the added value in a climate projection context cannot be smaller than this perfect-boundary estimate, our analysis demonstrates in a rigorous manner the clear additional value of dynamical downscaling over global climate simulations. In part 2, we draw the same conclusion for the downscaling of daily precipitation. The new evaluation methodology has a broad spectrum of applicability to future climate simulations, including ensemble runs, owing to the fact that it is distribution-free, robust to spatial dependence, and accounts for time series structure.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Acknowledgements This work was funded by the Helmholtz Centre for Environmental Research (UFZ) and the Ministry of Science, Research and Arts Baden-Württemberg. The authors gratefully acknowledge the financial support by the German Research Foundation (DFG) in the frame of the integrated research project PAK 346 “Structure and Function of Agricultural Landscapes under Global Climate Change”, which is continued by the Research Unit FOR 1695 “Agricultural Landscapes under Global Climate Change Processes and Feedbacks on a Regional Scale”. Martin Ivanov acknowledges the Water and Earth System Science Competence Cluster (WESS) funded by the Federal Ministry of Education and Research (BMBF) and UFZ Leipzig as well as support by ETH Zürich in the frame of the ELAPSE project (Enhancing Local and Regional Climate Change Projections for Switzerland) funded by the Swiss State Secretariat for Education, Research and Innovation SERI under project number C12.0089.

References

- Alexandru A, de Elía R, Laprise R (2007) Internal variability in regional climate downscaling at the seasonal scale. *Mon Weather Rev* 135(9):3221–3238. doi:10.1175/MWR3456.1
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 57(1):289–300. doi:10.2307/2346101
- Buonomo E, Jones R, Huntingford C, Hannaford J (2007) On the robustness of changes in extreme precipitation over Europe from two high resolution climate change simulations. *Q J Roy Meteor Soc* 133(622):65–81. doi:10.1002/qj.13
- Cardoso RM, Soares PMM, Miranda PMA, Belo-Pereira M (2013) WRF high resolution simulation of Iberian mean and extreme precipitation climate. *Int J Climatol* 33(11):2591–2608. doi:10.1002/joc.3616
- Chan SC, Kendon EJ, Fowler HJ, Blenkinsop S, Ferro CAT, Stephenson DB (2013) Does increasing the spatial resolution of a regional climate model improve the simulated daily precipitation? *Clim Dynam* 41(5-6):1475–1495. doi:10.1007/s00382-012-1568-9
- Chen F, Dudhia J (2001a) Coupling an advanced land surface-hydrology model with the Penn state-NCAR MM5 modeling system. Part I: model implementation and sensitivity. *Mon Weather Rev* 129(4):569–585. doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2
- Chen F, Dudhia J (2001b) Coupling an advanced land surface-hydrology model with the Penn state-NCAR MM5 modeling system. Part II: preliminary model validation. *Mon Weather Rev* 129(4):587–604. doi:10.1175/1520-0493(2001)129<0587:CAALSH>2.0.CO;2
- Christensen HJ, Machehauer B, Jones GR, Schär C, Ruti MP, Castro M, Visconti G (1997) Validation of present-day regional climate simulations over Europe: LAM simulations with observed boundary conditions. *Clim Dyn* 13(7):489–506. doi:10.1007/s003820050178
- Christensen JH, Christensen OB (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Clim Chang* 81(1):7–30. doi:10.1007/s10584-006-9210-7
- Collins WD, Rasch PJ, Boville BA, Hack JJ, McCaa JR, Williamson DL, Kiehl JT, Briegleb B, Bitz C, Lin SJ, Zhang M, Dai Y (2004) Description of the NCAR community atmosphere model (CAM 3.0). NCAR technical note, NCAR/TN-464+STR, <http://www.cesm.ucar.edu/models/atm-cam/docs/description>
- Davidson R, MacKinnon JG (2007) Improving the reliability of bootstrap tests with the fast double bootstrap. *Comput Stat Data An* 51(7):3259–3281. doi:10.1016/j.csda.2006.04.001
- Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentesm M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thépaut JN, Vitart F (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q J R Meteorol Soc* 137(656):553–597. doi:10.1002/qj.828
- Di Luca A, de Elía R, Laprise R (2012) Potential for added value in precipitation simulated by high-resolution nested regional climate models and observations. *Clim Dynam* 38(5-6):1229–1247. doi:10.1007/s00382-011-1068-3
- Di Luca A, de Elía R, Laprise R (2013) Potential for added value in temperature simulated by high-resolution nested RCMs in present climate and in the climate change signal. *Clim Dynam* 40(1-2):443–464. doi:10.1007/s00382-012-1384-2
- Dosio A, Panitz HJ, Schubert-Frisius M, Lüthi D (2015) Dynamical downscaling of CMIP5 global circulation models over CORDEX-Africa with COSMO-CLM: evaluation over the present climate and analysis of the added value. *Clim Dynam* 44(9-10):2637–2661. doi:10.1007/s00382-014-2262-x
- Duffy PB, Arritt RW, Coquard J, Gutowski W, Han J, Iorio J, Kim J, Leung LR, Roads J, Zeledon E (2006) Simulations of present and future climates in the western United States with four nested regional climate models. *J Climate* 19(6):873–895. doi:10.1175/JCLI3669.1
- de Elía R, Laprise R, Denis B (2002) Forecasting skill limits of nested, limited-area models: a perfect-model approach. *Mon Weather Rev* 130(8):1181–1192. doi:10.1175/1520-0493(2002)130<2006:FSLONL>2.0.CO;2

- de Elía R, Caya D, Côté H, Frigon A, Biner S, Giguère M, Paquin D, Harvey R, Plummer D (2007) Evaluation of uncertainties in the CRCM-simulated North American climate. *Clim Dynam* 30(2):113–132. doi:[10.1007/s00382-007-0288-z](https://doi.org/10.1007/s00382-007-0288-z)
- Elmore KL, Baldwin ME, Schultz DM (2006) Field significance revisited: spatial bias errors in forecasts as applied to the Eta model. *Mon Weather Rev* 134(2):519–531. doi:[10.1175/MWR3077.1](https://doi.org/10.1175/MWR3077.1)
- Feser F (2006) Enhanced detectability of added value in limited-area model results separated into different spatial scales. *Mon Weather Rev* 134(8):2180–2190. doi:[10.1175/MWR3183.1](https://doi.org/10.1175/MWR3183.1)
- Fujibe F, Ishihara K (2010) Possible urban bias in gridded climate temperature data over the Japan area. *SOLA* 6:61–64. doi:[10.2151/sola.2010-016](https://doi.org/10.2151/sola.2010-016)
- Gallo K, Xian G (2014) Application of spatially gridded temperature and land cover data sets for urban heat island analysis. *Urban Clim* 8(0):1–10. doi:[10.1016/j.uclim.2014.04.005](https://doi.org/10.1016/j.uclim.2014.04.005)
- Gao L, Bernhardt M, Schulz K (2012) Elevation correction of ERA-Interim temperature data in complex terrain. *Hydrol Earth Syst Sci* 16(12):4661–4673. doi:[10.5194/hess-16-4661-2012](https://doi.org/10.5194/hess-16-4661-2012)
- Giorgi F, Jones C, Asrar GR (2009) Addressing climate information needs at the regional level: the CORDEX framework. *WMO Bull* 58(3):175–183
- Hartmann DL, Klein Tank AMG, Rusticucci M, Alexander LV, Brönnimann S, Charabi Y, Dentener FJ, Dlugokencky EJ, Easterling DR, Kaplan A, Soden BJ, Thorne PW, Wild M, Zhai PM (2013) Observations: atmosphere and surface. In: Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) *Climate change 2013: the physical science basis. Contribution of working group I to the 5th assessment report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, chap 2, pp 159–254
- Heikkilä U, Sandvik A, Sorteberg A (2011) Dynamical downscaling of ERA-40 in complex terrain using the WRF regional climate model. *Clim Dynam* 37(7–8):1551–1564. doi:[10.1007/s00382-010-0928-6](https://doi.org/10.1007/s00382-010-0928-6)
- Hong SY, Noh Y, Dudhia J (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon Weather Rev* 134(9):2318–2341. doi:[10.1175/MWR3199.1](https://doi.org/10.1175/MWR3199.1)
- Ivanov MA, Warrach-Sagi K, Wulfmeyer V (2017) Field significance of performance measures in the context of regional climate model evaluation. Part 2: precipitation. *Theor Appl Climatol*:1–23. doi:[10.1007/s00704-017-2077-x](https://doi.org/10.1007/s00704-017-2077-x)
- Jacob D, Barring L, Christensen JH, de Castro M, Déqué M, Giorgi F, Hagemann S, Hirschi M, Jones R, Kjellström E, Lenderink G, Rockel B, Sánchez E, Schär C, Seneviratne SI, Somot S, van Ulden A, van den Hurk B (2007) An inter-comparison of regional climate models for Europe: model performance in present-day climate. *Clim Chang* 81(1):31–52. doi:[10.1007/s10584-006-9213-4](https://doi.org/10.1007/s10584-006-9213-4)
- Jaeger EB, Anders I, Lüthi D, Rockel B, Schär C, Seneviratne SI (2008) Analysis of ERA40-driven CLM simulations for Europe. *Meteorol Z* 17(4):349–367. doi:[10.1127/0941-2948/2008/0301](https://doi.org/10.1127/0941-2948/2008/0301)
- Jolliffe IT (2007) Uncertainty and inference for verification measures. *Weather Forecast* 22(3):637–650. doi:[10.1175/WAF989.1](https://doi.org/10.1175/WAF989.1)
- Jolliffe IT, Stephenson DB (eds.) (2012) *Forecast verification: a practitioner's guide in atmospheric science*, 2nd edn, Wiley
- Kain JS (2004) The Kain-Fritsch convective parameterization: an update. *J Appl Meteorol* 43(1):170–181. doi:[10.1175/1520-0450\(2004\)043<0170:TKCPAU>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2)
- Katragkou E, García-Díez M, Vautard R, Sobolowski S, Zanis P, Alexandri G, Cardoso RM, Colette A, Fernandez J, Gobiet A, Goergen K, Karacostas T, Knist S, Mayer S, Soares PMM, Pytharoulis I, Tegoulis I, Tsikerdekis A, Jacob D (2015) Regional climate hindcast simulations within EURO-CORDEX: evaluation of a WRF multi-physics ensemble. *Geosci Model Dev* 8(3):603–618. doi:[10.5194/gmd-8-603-2015](https://doi.org/10.5194/gmd-8-603-2015)
- Katz RW, Brown BG (1991) The problem of multiplicity in research on teleconnections. *Int J Climatol* 11(5):505–513. doi:[10.1002/joc.3370110504](https://doi.org/10.1002/joc.3370110504)
- Kendon EJ, Roberts NM, Senior CA, Roberts MJ (2012) Realism of rainfall in a very high-resolution regional climate model. *J Climate* 25(17):5791–5806. doi:[10.1175/JCLI-D-11-00562.1](https://doi.org/10.1175/JCLI-D-11-00562.1)
- Kotlarski S, Keuler K, Christensen OB, Colette A, Déqué M, Gobiet A, Goergen K, Jacob D, Lüthi D, van Meijgaard E, Nikulin G, Schär C, Teichmann C, Vautard R, Warrach-Sagi K, Wulfmeyer V (2014) Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci Model Dev* 7(4):1297–1333. doi:[10.5194/gmd-7-1297-2014](https://doi.org/10.5194/gmd-7-1297-2014)
- Laprise R (2008) Regional climate modelling. *J Comput Phys* 227(7):3641–3666. doi:[10.1016/j.jcp.2006.10.024](https://doi.org/10.1016/j.jcp.2006.10.024)
- Laprise R, de Elía R, Caya D, Biner S, Lucas-Picher P, Diaconescu E, Leduc M, Alexandru A, Separovic L (2008) Challenging some tenets of regional climate modelling. *Meteor Atmos Phys* 100(1–4):3–22. doi:[10.1007/s00703-008-0292-9](https://doi.org/10.1007/s00703-008-0292-9)
- Léger C, Politis DN, Romano JP (1992) Bootstrap technology and applications. *Technometrics* 34(4):378–398. doi:[10.2307/1268938](https://doi.org/10.2307/1268938)
- van der Linden P, Mitchell JFB (2009) ENSEMBLES: Climate change and its impacts: summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK
- Livezey RE, Chen WY (1983) Statistical field significance and its determination by Monte Carlo techniques. *Mon Weather Rev* 111(1):46–59. doi:[10.1175/1520-0493\(1983\)111<0046:SFSASD>2.0.CO;2](https://doi.org/10.1175/1520-0493(1983)111<0046:SFSASD>2.0.CO;2)
- Maier U, Müller-Westermeier G (2010) Verifikation klimatologischer Rasterfelder. *Berichte des Deutschen Wetterdienstes* 235, Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main
- Mallard MS, Nolte CG, Spero TL, Bullock OR, Alapaty K, Herwehe JA, Gula J, Bowden JH (2015) Technical challenges and solutions in representing lakes when using WRF in downscaling applications. *Geosci Model Dev* 8(4):1085–1096. doi:[10.5194/gmd-8-1085-2015](https://doi.org/10.5194/gmd-8-1085-2015)
- Mooney PA, Mulligan FJ, Fealy R (2013) Evaluation of the sensitivity of the weather research and forecasting model to parameterization schemes for regional climates of Europe over the period 1990–95. *J Climate* 26(3):1002–1017. doi:[10.1175/JCLI-D-11-00676.1](https://doi.org/10.1175/JCLI-D-11-00676.1)
- Morrison H, Thompson G, Tatarskii V (2009) Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: comparison of one- and two-moment schemes. *Mon Weather Rev* 137(3):991–1007. doi:[10.1175/2008MWR2556.1](https://doi.org/10.1175/2008MWR2556.1)
- O’Gorman PA, Schneider T (2009) The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *P Natl Acad Sci USA* 106(35):14,773–14,777. doi:[10.1073/pnas.0907610106](https://doi.org/10.1073/pnas.0907610106)
- Pan Z, Christensen JH, Arritt RW, Gutowski WJ, Takle ES, Otieno F (2001) Evaluation of uncertainties in regional climate change simulations. *Journal of Geophysical Research: Atmospheres* 106(D16):17,735–17,751. doi:[10.1029/2001JD900193](https://doi.org/10.1029/2001JD900193)
- Pearson KJ, Shaffrey LC, Methven J, Hodges KI (2015) Can a climate model reproduce extreme regional precipitation events over England and Wales? *Q J R Meteorol Soc* 141(689):1466–1472. doi:[10.1002/qj.2428](https://doi.org/10.1002/qj.2428)
- Perkins SE, Pitman AJ, Holbrook NJ, McAneney J (2007) Evaluation of the AR4 climate models’ simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J Climate* 20(17):4356–4376. doi:[10.1175/JCLI4253.1](https://doi.org/10.1175/JCLI4253.1)

- Prein AF, Langhans W, Fosser G, Ferrone A, Ban N, Goergen K, Keller M, Tölle M, Gutjahr O, Feser F, Brisson E, Kollet S, Schmidli J, van Lipzig NPM, Leung R (2015) A review on regional convection-permitting climate modeling: demonstrations, prospects, and challenges. *Rev Geophys* 53(2):323–361. doi:[10.1002/2014RG000475](https://doi.org/10.1002/2014RG000475). 2014RG000475
- Prömmel K, Geyer B, Jones JM, Widmann M (2010) Evaluation of the skill and added value of a reanalysis-driven regional simulation for Alpine temperature. *Int J Climatol* 30(5):760–773. doi:[10.1002/joc.1916](https://doi.org/10.1002/joc.1916)
- Rolland C (2003) Spatial and seasonal variations of air temperature lapse rates in alpine regions. *J Climate* 16(7):1032–1046. doi:[10.1175/1520-0442\(2003\)016<1032:SASVOA>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<1032:SASVOA>2.0.CO;2)
- Rummukainen M (2010) State-of-the-art with regional climate models. *Wiley Interdiscip Rev Clim Change* 1(1):82–96. doi:[10.1002/wcc.8](https://doi.org/10.1002/wcc.8)
- Sanchez-Gomez E, Somot S, Déqué M (2009) Ability of an ensemble of regional climate models to reproduce weather regimes over Europe-Atlantic during the period 1961–2000. *Clim Dynam* 33(5):723–736. doi:[10.1007/s00382-008-0502-7](https://doi.org/10.1007/s00382-008-0502-7)
- Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability in European summer heatwaves. *Nature* 427(6972):332–336. doi:[10.1038/nature02300](https://doi.org/10.1038/nature02300)
- Sotillo M, Martín M, Valero F, Luna M (2006) Validation of a homogeneous 41-year (1961–2001) winter precipitation hindcasted dataset over the Iberian Peninsula: assessment of the regional improvement of global reanalysis. *Clim Dynam* 27(6):627–645. doi:[10.1007/s00382-006-0155-3](https://doi.org/10.1007/s00382-006-0155-3)
- Sturges HA (1926) The choice of a class interval. *J Am Stat Assoc* 21(153):65–66
- Torma C, Giorgi F, Coppola E (2015) Added value of regional climate modeling over areas characterized by complex terrain—precipitation over the Alps. *J Geophys Res: Atmos* 120(9):3957–3972. doi:[10.1002/2014JD022781](https://doi.org/10.1002/2014JD022781)
- Trenberth KE, Dai A, Rasmussen RM, Parsons DB (2003) The changing character of precipitation. *Bull Amer Meteor Soc* 84(9):1205–1217. doi:[10.1175/BAMS-84-9-1205](https://doi.org/10.1175/BAMS-84-9-1205)
- Vautard R, Gobiet A, Jacob D, Belda M, Colette A, Déqué M, Fernández J, García-Díez M, Goergen K, Güttler I, Halenka T, Karacostas T, Katragkou E, Keuler K, Kotlarski S, Mayer S, van Meijgaard E, Nikulin G, Patarčić M, Scinocca J, Sobolowski S, Suklitsch M, Teichmann C, Warrach-Sagi K, Wulfmeyer V, Yiou P (2013) The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. *Clim Dynam* 41(9-10):2555–2575. doi:[10.1007/s00382-013-1714-z](https://doi.org/10.1007/s00382-013-1714-z)
- Ventura V, Paciorek CJ, Risbey JS (2004) Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J Climate* 17(22):4343–4356. doi:[10.1175/3199.1](https://doi.org/10.1175/3199.1)
- Warrach-Sagi K, Schwitalla T, Bauer HS, Volker-Wulfmeyer (2013) Sustained Simulation Performance 2013. In: Resch MM, Bez W, Focht E, Kobayashi H, Kovalenko Y (eds) A regional climate model simulation for EURO-CORDEX with the WRF model. Springer, pp 147–157. doi:[10.1007/978-3-319-01439-5_11](https://doi.org/10.1007/978-3-319-01439-5_11)
- Wilks DS (2006a) On “field significance” and the false discovery rate. *J Appl Meteorol* 45(9):1181–1189. doi:[10.1175/JAM2404.1](https://doi.org/10.1175/JAM2404.1)
- Wilks DS (2006b) *Statistical Methods in the Atmospheric Sciences*, International Geophysics Series, vol 91, 2nd edn. Elsevier, Cornell University, USA
- Winterfeldt J, Weisse R (2009) Assessment of value added for surface marine wind speed obtained from two regional climate models. *Mon Weather Rev* 137(9):2955–2965. doi:[10.1175/2009MWR2704.1](https://doi.org/10.1175/2009MWR2704.1)
- Winterfeldt J, Geyer B, Weisse R (2011) Using QuikSCAT in the added value assessment of dynamically downscaled wind speed. *Int J Climatol* 31(7):1028–1039. doi:[10.1002/joc.2105](https://doi.org/10.1002/joc.2105)