Evaluation of trends in extreme temperatures simulated by HighResMIP models across Europe

Antonello A. Squintu^{1,2} · Gerard van der Schrier¹ · Else van den Besselaar¹ · Eveline van der Linden^{1,2} · Dian Putrasahan⁶ · Christopher Roberts⁴ · Malcolm Roberts⁵ · Enrico Scoccimarro³ · Retish Senan⁴ · Albert Klein Tank^{1,2,5}

Received: 1 January 2020 / Accepted: 16 December 2020 / Published online: 5 March 2021
© The Author(s) 2021

Abstract

Simulation of past climate is an important tool for the validation of climate models. The comparison with observed daily values allows us to assess the reliability of their projections on climatic extremes in a future climate. The frequency and amplitude of extreme events are fundamental aspects that climate simulations need to reproduce as they have high impacts on economy and society. The ability to simulate them will help policy makers in taking better measures to face climate change. This work aims at evaluating how six models within the High Resolution Model Intercomparison Project reproduce the trends on extreme indices as they have been observed over Europe in the 1970–2014 period. Observed values are provided by the new homogenized version of the E-OBS gridded dataset. The comparison is performed through the use of indices based on seasonal averages and on exceedances of percentile-based thresholds, focusing on six subregions. Winter-average minimum temperature is generally underestimated by models (down to -4°C difference over Italy and Norway) while simulated trends in seasonal averages and extreme values are found to be too cold on Eastern Europe and too warm on Iberia and Southern Europe (e.g. up to a difference of -4% per decade on the number of Cold Nights over Spain). On the other hand the models tend to overestimate summer maximum temperatures averages in the Mediterranean Area (up to $+5^{\circ}\text{C}$ over the Balkans) and underestimate these at higher latitudes. Iberia, Southern and Eastern Europe are simulated with too low trends in average summer temperatures. The simulated trends are too strong on the North West part and too weak on the South East part of Europe (down to $-3\%/\text{decade}$ on the number of Warm Days over Italy and Western Balkans). These results corroborate the findings of previous studies about the underestimation of the warming trends of summer temperatures in Southern Europe, where these are more intense and have more impacts. The high-resolution versions of the models are compared to their lower-resolution counterparts, similar to those used in the CMIP5, showing a slight improvement for the simulation of extreme winter minimum temperatures, while no significant progresses have been found for extreme summer maximum temperatures.

Keywords Climate simulations · Model validation · Extreme values

1 Introduction

Changes in the frequency and intensity of climatic heat extremes have important impacts on sectors such as agriculture, energy demand, transportation industry, health. For this reason a realistic assessment of future climate is fundamental in order to understand the challenges ahead. The climate models used for the production of future projections are also used to simulate the climate of the recent past. This is done by taking as input verified historical observed boundary conditions (e.g. sea surface temperature, land use) and using observed values as internal (artificial and natural,

✉ Antonello A. Squintu
squintu@knmi.nl

¹ Koninklijk Nederlands Meteorologisch Instituut,
Utrechtseweg 297, De Bilt, Netherlands

² Wageningen University, Wageningen, Netherlands

³ CMCC, Bologna, Italy

⁴ ECMWF, Reading, UK

⁵ Met Office Hadley Centre, Exeter, UK

⁶ MPI-M, Hamburg, Germany

such greenhouse gases and aerosol concentration, volcanic eruptions) and external forcings (e.g. solar irradiance). The historical climate simulations are crucial for the validation of projections of the future, as they can be compared with observations. Flato et al. (2014). The Climate Model Inter-comparison Project 3 (CMIP3) (Meehl et al. 2007) and the CMIP5 (Taylor et al. 2012) have contributed to collecting and comparing all the available climate simulations and projections, with a continuous improvement in the use of historical forcings and in the temporal coverage (Taylor et al. 2012; Flato et al. 2014). The CMIP5 experiments showed a clear improvement in model performance on temperature simulations compared to CMIP3 (Flato et al. 2014). More recently the project PRIMAVERA (a European Union Horizon 2020 project) has worked in the frame of the CMIP6 HighResMIP protocol, coordinating a set of experiments designed to assess both standard and enhanced horizontal-resolution simulations in the atmosphere and ocean (with up to 0.25 °C in the atmosphere) Haarsma et al. (2016).

The comparison of these simulations against observations is a powerful tool for the assessment of how the models reproduce the climate under observed conditions and forcings. The evaluations performed in recent years have focused on the intercomparison of the models or on the comparison with reanalyses, observed individual series or, more recently, gridded observational datasets (Flato et al. 2014). The choice of the used reference is fundamental and needs to be performed with care (Sillmann et al. 2013). The use of observations is preferable (Gleckler et al. 2008; Flato et al. 2014) and while the use of individual station data is a very direct approach, it is hampered because the gridbox value of a model represents an area-average whereas the station observation is a point-value. This is particularly problematic for climatic extremes.

This issue is avoided by using gridded observations which remove observational noise and provide a homogeneous spatial distribution (Cornes and Jones 2013). Nevertheless, the gridded datasets, especially on a daily resolution, have been deficient for a long time (Kiktev et al. 2003; Kharin et al. 2005) and the possible comparisons are still limited in space and time (Flato et al. 2014). In addition, the underlying station data may have non-climatic signals due to changes in location, management or observation techniques which can cloud a trend-based evaluation of models. The E-OBS gridded dataset (Haylock et al. 2008; Cornes et al. 2018), used in this study, matches the high spatial and temporal detail of this new generation of models and is based on a set of homogenized station data, see Sect. 2.2 for more details.

The model evaluations performed by Kharin et al. (2005) and, more recently, Sillmann et al. (2013) have shown that climate simulations reproduce time averaged values, like monthly means, better than the extreme ones. However the greatest impacts on economy and society are related to

extreme events rather than averages. Furthermore, while it is generally accepted that climate changes are driven by average values (Scherrer et al. 2005), debate has taken place about the magnitude of the changes in variability (Alexander et al. 2006; Simolo et al. 2010; Morak et al. 2011; Donat and Alexander 2012), which are shown to lead to amplified effects on extreme events (Katz and Brown 1992; Schär et al. 2004) and on the indices used to describe them (Della-Marta et al. 2007). A simple shift in the mean insufficiently explains particular record-breaking events, such as the heatwave of 2003 (Schär et al. 2004).

Early model evaluations (Kharin et al. 2005, 2007; Sterl et al. 2008) have focused on *hard* extremes, whose trends lack significance, since they have a return time of the order of years (Frich et al. 2002). Such indices present a high interannual variability, which makes it difficult to calculate trends on relatively short periods. Furthermore, these indices are very sensitive to quality issues in the observational data, especially if the quality issues occur at a station in a data sparse area. Indices calculated using percentile-based thresholds are good alternatives to focus on climatic extremes. Two examples are TN10p, percentage of days with daily minimum temperature below the 10th percentile and TX90p, percentage of days with daily maximum temperature above the 90th percentile (ETCCDI 2009). In both cases the thresholds are calculated considering the data of the series in a long reference period, usually thirty years. This makes them site specific (the thresholds change for each grid point), and not affected by biases and applicable to any climate (Klein Tank and Können 2003; Kiktev et al. 2003; Sillmann et al. 2014b).

The 5th International Panel on Climate Change (IPCC) assessment report indicated progress for CMIP5 over CMIP3 in the bias of the models (Flato et al. 2014). The multi-model average of CMIP5 presented a general agreement with observations, with the exception of a cold bias in winter months over Northern Europe (Flato et al. 2014). Nevertheless, relevant criticisms have been found when analyzing the trends of average and extreme values. The study of Bhend and Whetton (2013) found a significant underestimation of the trends in summer average temperatures over Southern Europe and the Mediterranean. At the same time, Min et al. (2013) assessed the simulation of TXx (warmest maximum temperature in a year) over North Western Europe performed by regional models. This last work has stressed the fact that important aspects of the occurrence of moderate maximum temperature extremes over North Western Europe are missed by a set of regional models. The poor simulation of the summer extremes in the present climate stimulates questions about the reliability of the same models in their predictions of extreme events in the next decades (Bhend and Whetton 2013). The new homogenized gridded dataset over Europe (referred to as E-OBSv19.0HOM) provides

Table 1 Characteristics of the high resolution models that have been used

Acronym	CMCC	CNRM	EC-Earth	ECMWF	HadGEM3	MPI
Model name	CMCC-CM2-VHR4	CNRM-CM6-1-HR	EC-Earth3P-HR	ECMWF-IFS-HR	HadGEM3-GC31-HM	MPI-ESM1-2-XR
Institute	CMCC	CNRM-CERFACS	EC-Earth-Consortium	ECMWF	MOHC	MPI-M
# of ens. memb.	1	1	3	6	3	1
ens. memb. labels	r1i1p1f1	r1i1p1f2	r1i1p1f1 r2i1p1f1 r3i1p1f1	r1i1p1f1 r2i1p1f1 r3i1p1f1 r4i1p1f1 r5i1p1f1 r6i1p1f1	r1i1p1f1 r1i2p1f1 r1i3p1f1	r1i1p1f1
Long. res. [°]	0.31	0.5	0.49 (regridged)	0.5	0.35	0.47
Lat. res. [°]	0.23	0.5	0.35 (regridged)	0.5	0.23	0.47

Resolutions have been rounded at the second decimal digit. Institute acronyms and corresponding references for the models: Centro euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Cherchi et al. (2019); Centre National de Recherches Météorologiques—Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CNRM-CERFACS), Voldoire et al. (2019); EC-Earth, Haarsma et al. (2020); European Centre for Medium-Range Weather Forecasts (ECMWF), Roberts et al. (2018); MetOffice—Hadley Centre (MOHC), Roberts et al. (2019); Max-Planck-Institut für Meteorologie (MPI-M), Gutzjahr et al. (2019)

a spatial and temporal coverage that allows an evaluation of the global high resolution models that have been made available in the PRIMAVERA project. In contrast to the previous studies on the same topic, the focus will be on the trends of extreme events such as TN10p and TX90p, with a particular attention to seasonal values to some European subregions. This work provides for the first time a complete European-wide assessment of the performance of the considered models in simulating temperature trends, including an assessment of the improvement of the high resolution version in comparison to their lower resolution counterpart. Such an approach is taken following other works that have evaluated the progresses of HighResMIP in the analysis of other phenomena, such as tropical cyclones (Roberts et al. 2020). It is important to stress that this study does not aim at identifying the reasons behind the observed discrepancies. This will be the subject of future projects that will use the results of this work as a starting point for the diagnosis of the criticisms in the models.

2 Data and methods

2.1 Models

The tested simulations have been developed in the framework of the PRIMAVERA Project, that aims at increasing spatial resolution of climate models. Six models have been analyzed at high resolution (HR) (see Table 1) version and in a previously existing lower resolution (from now on referred

to as lower resolution, LR) (see Table 2), focusing on the period from 1970–2014 (see Sect. 2.2) and considering the region enclosed between 22 ° W and 50 ° E and 20 ° N and 76 ° N. The variables considered in this study are minimum temperature (TN) and maximum temperatures (TX) at a daily timescale. Each model taking part in PRIMAVERA has contributed with several experiments, the one that has been used for this work is named "highres-SST-present". This consists of a simulation of the atmospheric conditions over the period 1950–2014, taking observed sea surface temperature, sea-ice concentration and incoming radiation as forcings. Each model has a different spatial resolution and a different number of ensemble members. Tables 1 (HR) and 2 (LR) summarize the characteristics of the used models and the availability of the ensemble members as of 23rd of September, 2019.

The *ECMWF* model has native resolution Tco399 (~ 25 km) for HR and Tco199 (~ 50 km) for LR. In the frame of PRIMAVERA they have been provided in a regridged version, respectively to 0.25° and 0.5° constant latitude-longitude regular grids, more details in Roberts et al. (2018). The *EC-Earth3P* model runs at the resolution TL511 for HR and TL255 for LR on a non-regular latitude-longitude grid. The scripts used for the indices calculation (2.3) require regular grids, making it necessary to regrid the EC-Earth3P model.

Table 2 Characteristics of the low resolution models that have been used

Acronym	CMCC Lowres	CNRM Lowres	Ec-Earth3 Lowres	ECMWF Lowres	HadGEM3 Lowres	MPI Lowres
Model name	CMCC-CM2-HR4	CNRM-CM6-1	EC-Earth3P	ECMWF-IFS-LR	HadGEM3-GC31-LM	MPI-ESM1-2-HR
Institute	CMCC	CNRM-CER-FACS	EC-Earth-Consortium	ECMWF	MOHC	MPI-M
# of ens. memb.	1	1	2	8	5	1
Ens. memb. labels	r1i1p1fl	r1i1p1f2	r1i1p1fl r3i1p1fl	r1i1p1fl r2i1p1fl r3i1p1fl r4i1p1fl r5i1p1fl r6i1p1fl r7i1p1fl r8i1p1fl	r1i1p1fl r1i2p1fl r1i3p1fl r1i14p1fl r1i15p1fl	r1i1p1fl
Long. res. [°]	1.25	1.4	0.98 (regridded)	1	1.88	0.94
Lat. res. [°]	0.94	1.4	0.7 (regridded)	1	1.25	0.94

Resolutions have been rounded at the second decimal digit

2.2 E-OBS

The reference used for the evaluation of the models is the E-OBS for TN and TX (Haylock et al. 2008; Cornes et al. 2018). It comes as a 100-members ensemble, whose spread increases in areas with low station density, indicating a larger uncertainty. In this work only the ensemble mean is considered. E-OBS is based on the station data of the European Climate Assessment & Dataset (ECA&D) (Klein Tank et al. 2002), which collects data of thirteen variables from more than 19,000 stations located in all countries of the European and Mediterranean region. Almost 10,000 of these stations include temperature data. These are provided by National Meteorological and Hydrological Services, universities or private companies and range from late 18th century to current times. However, relocation of stations, instrumentation changes and variations in the surroundings of the meteorological stations affect the quality of ECA&D temperature temporal series related to such stations (and therefore E-OBS), reducing the reliability for temporal analyses. For this analysis, a modified version of E-OBS is constructed based on recent work on the homogenization of the temperature series of ECA&D (Squintu et al. 2019, 2020). These studies describe how a large part of the inhomogeneities have been removed, making it possible to smoothly combine series that belong to neighbouring stations and to gather the data into one long-running homogeneous series, called blended series. This process considerably improves the input data for E-OBS, which becomes a data set of long and homogenized series: a prerequisite for a thorough climatic change assessment (ETCCDI 2009; Jones and Wigley 2010).

Table 3 Number of series which are continuously active during the indicated time interval

Period	# of TN series (% over the total)	# of TX series (% over the total)
Total: no time restriction	3110 (100%)	3086 (100%)
1950–2014	612 (19.7%)	603 (19.5%)
1960–2014	997 (32.1%)	983 (31.8%)
1970–2014	1268 (40.7%)	1248 (40.4%)
1980–2014	1376 (44.2%)	1353 (43.8%)

Percentage is calculated with respect to the total amount of blended homogenized series (first row). As expected, longer periods have lower amount of series which are active during the whole interval

For the purpose of this work, only the blended series that start before 1970 and that stop after 2014 have been considered in the construction of a special version of E-OBS, called E-OBS.hom. This selection aims at having a constant number of blended series contributing to each grid-point, avoiding changes in station density that might introduce inhomogeneities. Table 3 explains that there is not a drastic change in the number of blended series choosing 1970 or 1980 as starting point, thus 1970 has been chosen in order to work with a longer period.

Fig. 1 Difference in the winter average of TN between the HR models and E-OBS. Red indicates overestimation, blue indicates underestimation. Significant differences are indicated by small thin circles for each grid-point, which result in shaded areas

2.3 Data analysis

In this work minimum temperatures and maximum temperatures have been analyzed, focusing on the seasons: winter (December, January and February; DJF), spring (March, April and May; MAM), summer (June, July and August; JJA) and autumn (September, October and November, SON). Furthermore, all the results have been summarized taking means over the whole domain and on six relevant regions: Iberian Peninsula, Southern, Eastern, Western, Central and Northern Europe. The boundaries of these regions can be observed, for example, in Fig. 1. Even though climate phenomena obviously happen across political or statistically determined boundaries, these areas have been identified as those involved by common peculiarities for a large number of the analyzed parameters.

While all the values are available in the tables of Appendices A and B (Tables 4, 5, 6), the figures and the discussion focus on the indices TN10p-DJF and TX90p-JJA. These represent the coldest and warmest events, namely those with highest impact on health, economy and society. After checking the bias of the seasonal averages (e.g. TNavg-DJF and TXavg-JJA), for each grid-point the indices TN10p and TX90p have been calculated on a seasonal level. In all cases the percentile thresholds have been calculated over the 1981–2010 period, making use of the bootstrapping approach introduced by Zhang et al. (2005).

In order to perform a grid-point by grid-point comparison the E-OBS indices have been regridded with a bilinear procedure to the native grid of each model (with the exception of the “substitute” grid used for EC-Earth3, see Table 1), creating six versions of remapped E-OBS for each index.

At this point for each season, dataset and grid-point, the trends on the indices on the 1970–2014 period have been obtained.¹ Calculation of trends has been done following the Sen’s slope method (Sen 1968), which is more robust than a least square approach and does not require the assumption of a normal distribution (Sen 1968; Alexander et al. 2006; ETCCDI (2009)). Some model experiments have been run in ensemble mode and, in order to obtain the ensemble means, for each model the trends on each grid-point related to the ensemble members have been averaged. Each

ensemble mean has been compared to the corresponding E-OBS regridded dataset, taking the difference of the trends on each grid-point. The difference has been considered significant when the 95% interval of each trend on E-OBS and the 95% interval of each corresponding trend on the model don’t overlap. This process, applied to both high and low resolution versions, has allowed us to detect areas in which the models underestimate or overestimate the trends that have been seen in observational datasets.

Finally, the absolute trend bias is defined as the unsigned difference between the trend in the model and the trend in the E-OBS dataset. This operation has been applied to HR and corresponding LR versions of each model in order to compare them. For this purpose, a new temporary dataset has been created using the LR values on the grid resolution of the HR model (LRtoHR). These LRtoHR grid-points have been filled by using the absolute trend bias of the LR grid-point that overlaps with the LRtoHR grid-point. This is done in order to better inspect the local impact of increasing resolution, which would be lost in case the comparison was performed regridding the HR to the LR. The HR and the LRtoHR absolute trend biases have been compared by taking the difference as shown in eq. 1.

$$\begin{aligned} & \text{abstrendbias}_{HR} - \text{abstrendbias}_{LR} \\ &= |\text{trend}_{HR} - \text{trend}_{E-OBS}| - |\text{trend}_{LR} - \text{trend}_{E-OBS}| \quad (1) \end{aligned}$$

If this metric produces a negative value, then the HR absolute trend bias is lower than the one of LR, thus the trend for HR is closer to the observed one, indicating a better performance. On the other side a positive value indicates that the HR’s performance is worse than the corresponding LR. The aim of using absolute trend biases is to assess if the HR trends are closer to the E-OBS trends than corresponding LR trends, independently from the sign of the trend difference. If the comparison was performed with trend biases, it would have only communicated if the HR models simulate warmer (positive result) or colder (negative result) trends than the corresponding LR models, see Eq. 2.

$$\begin{aligned} & \text{trendbias}_{HR} - \text{trendbias}_{LR} \\ &= (\text{trend}_{HR} - \text{trend}_{E-OBS}) - (\text{trend}_{LR} - \text{trend}_{E-OBS}) \quad (2) \\ &= \text{trend}_{HR} - \text{trend}_{LR} \end{aligned}$$

¹ Note that the indices for extreme values used in this work are site specific. This means that, by definition, for each gridpoint of each model the percentage of days in the 1981–2010 period above (below) the 90th (10th) percentile is exactly 10 %. Therefore it is not expected to observe a significant difference between the percentage of values exceeding the thresholds in the models and in the observations. These, if present, would be only related to the larger considered period (1970–2014) and would not carry significant meaning. For this reason the analysis has been conducted directly on the trends, which describe the changes in the distribution shape.

3 Results

3.1 Minimum temperatures

3.1.1 Bias in winter averages

The considered HR models show strong differences in the reproduction of T_{Navg}-DJF, see Fig. 1. The largest mean biases on continental and regional levels are found for CMCC (+2.96 °C), while EC-Earth3, ECMWF and HadGEM3 underestimate the minimum temperature on almost all regions. For all models West presents the largest (or least negative) biases, while within North there is clear contrast between the Norwegian coast and the interior of Sweden and Northern Finland. Strong cold biases are observed over Norway and Italy, and—less pronounced—in the Balkans and in the surrounding regions. MPI and CNRM perform best in terms of mean biases and present considerably lower extension of the shaded area. These are present when the simulated T_{Navg}-DJF is significantly different from the observed one (i.e. absence of overlap between the 95% confidence interval of the two terms of the difference).

3.1.2 Trends in winter averages

Trends on the T_{Navg}-DJF of the models in the 1970–2014 period are compared against the same indices of E-OBS. All models reproduce very well the trends in winter TN. The mean trend biases, Fig. 2, ranges between -0.16 °C per decade (°C/dec) (CNRM) and $+0.02$ °C/dec (ECMWF). This indicates a tendency in simulating lower trends over the continent, especially in East and North, that always show negative biases. Nevertheless recurring positive biases are found over the Kola Peninsula (NorthWestern Russia, 6 models out of 6), together with Iberian, Southern and Central Europe, that present warm biases for all models.

3.1.3 Trends in cold extremes

Trends in winter cold extremes as TN10p-DJF are more challenging and Fig. 3 shows the performance of the HR version of the models compared to the E-Obs (simple difference). While HadGEM3 (mean trend bias: -1.25 %/dec) and, less strongly, ECMWF (mean trend bias: -0.95 %/dec) simulate a lower (thus warmer than the observed) trend of number of days below the 10th percentile in all the regions, CMCC, CNRM and EC-Earth present a contrast between Eastern Europe and the other regions. Average biases in Iberia, South and Center are negative and below continental averages for all models but MPI, furthermore

a considerable number of gridpoints with significant differences are detected, in particular for CMCC, indicating a poor representation of the trends in these areas. The performance of MPI differs from the other models, not showing pronounced patterns, with the exception of having a too strong warming in TN10p-DJF over Sweden and Norway, in common with three other models.

Eventual inconsistencies between the biases in the trends of the average and of the extreme indices allow us to inspect how the models reproduce the changes in the shape of the distribution of the temperatures. In the case of minimum winter temperatures general agreement (excluding MPI) is observed in the overestimation of the trends for Iberia, South and Center. Nevertheless, for models like CMCC and HadGEM3 the presence of significant biases in the TN10p-DJF trends over Iberia and other areas imply that their simulation of cold extremes counts are considerably less events than observed. This reveals the tendency for the simulation of a narrower distribution. Same consideration can be made about Southern and Northern Europe (excluding Finland), where the good representation of the trends of the average contrasts with the general warmer bias related to the trends in the extremes. Finally Eastern Europe presents colder simulated trends for both average and extreme values in four models, while ECMWF and HadGEM3 show disagreement indicating, again, a tendency to a narrower distribution than observed.

The patterns in the trend bias in the HR models can be compared to the low resolution models, whose results are shown in Appendix C. Figure 9 shows that the LR models present similar patterns in the trend biases as the HR models.

Figure 4 presents the difference in absolute trends biases of TN10p-DJF between HR and LR, see Sect. 2.3. Negative values (green) indicate that HR has lower absolute trend bias than LR for that specific grid-point, thus it is performing better. Only CMCC clearly shows an area with worse absolute trend biases over West, North and Center (where very large trends are simulated), which contrasts with the strong performance of the same model over Eastern Europe. Despite this, the mean absolute trend biases over the whole continent are reduced for almost all the models, indicating a general improvement in the description of the cold extremes between low and high resolution. The best improvement is found for HadGEM3 (-0.51 %/dec, especially over Central Europe, -1.12 %/dec), while the only worsening, out of the considered models, is for ECMWF ($+0.17$ %/dec) whose LR version is found to perform the best among the others, see Appendix C. The model with the lowest mean absolute trend bias in high resolution is MPI (0.61 %/dec).

Fig. 2 Difference in the trends of TNavg-DJF for HR models. Red indicates overestimation (warmer simulated trends), blue indicates underestimation (colder simulated trends). Significant differences are indicated by black circles

3.2 Maximum temperatures

3.2.1 Bias in summer averages

The comparison of summer maximum temperatures has started from an evaluation of the biases of the models. Four of them give a mean bias that has a lower absolute value compared to what is observed for summer maximum temperatures, see Fig. 5. CMCC presents a large underestimation ($-3.83\text{ }^{\circ}\text{C}$), similar, but with opposite sign, to the corresponding result for TN. The remaining 5 models show similar patterns, with a warm bias along the Mediterranean and Black Sea coast and a general underestimation over North and Center, together with the Northern part of East. A large overestimation common to all models is found in Northern African regions, influencing the mean bias over Iberia, which has lower values on the Atlantic Coast. Nevertheless these large biases (above $+10\text{ }^{\circ}\text{C}$) can be in part related to the high uncertainty of E-OBS over Morocco and Algeria, due to a lower station density.

3.2.2 Trends in winter averages

The continental average of the difference in trends of TXavg-JJA ranges between $-0.17\text{ }^{\circ}\text{C}/\text{dec}$ (ECMWF) and $+0.03\text{ }^{\circ}\text{C}/\text{dec}$ (CMCC), see Fig. 6. The models tend to slightly underestimate the warming of summer temperatures. This is more evident over Iberia, South and East, whose regional mean biases are always similar or lower than the European mean. Especially in the case of EC-Earth and ECMWF, large areas with significant difference are observed, implying an inaccurate reproduction of the changes in the climate of these areas. On the other hand almost all the models tend to overestimate the trends over Northern Europe (especially Southern Norway).

3.2.3 Trends in warm extremes

Figure 7 shows the difference in trends between model and observations in TX90p-JJA, related to warm temperature extremes. The European mean biases show a large underestimation of the trends for EC-Earth ($-0.73\text{ } \%/ \text{dec}$), ECMWF ($-0.59\text{ } \%/ \text{dec}$) and HadGEM3 ($-0.56\text{ } \%/ \text{dec}$). In all cases stronger trends, consistent with those found for the trends in the averages, are simulated over Northern Europe, in particular Norway and Sweden. This behaviour is partially observed over Center and West and contrasts with the general underestimation of trends over Iberia, South and East, simulated by all models but CMCC. In these areas large significant differences are found in particular for EC-Earth3, ECMWF

and HadGEM3. This aspect (as found for the simple seasonal averages as well) indicates the tendency to reproduce lower trends of warm extremes on the Mediterranean and Black Sea region and slightly larger ones around Northern Sea. In contrast to the other models, CMCC doesn't present this pattern, showing an overestimation of the trends in all regions with the exception of Eastern Europe. These biases are consistent with those observed for the trends of the average over Center and West, indicating a good representation of the changes in the shape of the distribution. On the other hand over North, South and Iberia a discrepancy is detected, which reveals that for CMCC the number of warm events increases faster than the average maximum temperature. The simulation of a wider distribution is common with the other models over Northern Europe. This similarity is not found for Iberia, South and East, whose trends on TX90p-JJA are underestimated, with values that widely appear stronger than those obtained for the average values. This indicates that these models, and especially HadGEM3, simulate a tendency to a narrower distribution of summer maximum temperature over areas, like Mediterranean and Eastern Europe, where the warm extremes reach concerning values.

Figure 8, showing the difference in absolute trend biases between the HR and LR model configuration, does not show a common pattern. Best improvement in the passage from HR to LR is for MPI ($-0.16\text{ } \%/ \text{dec}$), in particular for Central and Eastern Europe. At the same time HadGEM3 presents the most intense worsening ($+0.25\text{ } \%/ \text{dec}$), with larger increases of the mean absolute bias over Iberia and Southern Europe. These findings indicate that the reproduction of trends of warm extremes with High Resolution models hasn't considerably improved over Europe for most of the models.

4 Summary and conclusions

Six models using their High (HR) and Low Resolution (LR) versions have been compared (over the 1970-2014 period) to E-OBS.hom, a version of the gridded dataset E-OBS based on homogenized daily series (each covering at least 1970-2014) of observed temperatures. The analysis has been performed first on the biases of the seasonal averages and of their trends, focusing on winter minimum temperatures (TNavg-DJF) and summer maximum temperatures (TXavg-JJA) and then on two ETCCDI (ETCCDI 2009) defined indices. These are the percentage of days with minimum temperatures below the 10th percentile of winter values ('cold nights', TN10p-DJF) and the percentage of days with maximum temperatures exceeding the 90th percentile of summer

Fig. 3 Difference in trends of TN10p-DJF between the HR models and E-OBS. Red (blue) indicates an underestimation (overestimation) of the trend, related to a warmer (colder) trend in the model

values ('warm day-times', TX90p-JJA). The percentile thresholds have been calculated using the 1981–2010 period. After the calculation of the trends of the considered indices, for those models with more than one ensemble member (see Table 1), the ensemble mean has been calculated. For each grid-point, average values and trends in the models have been compared to observations and an assessment is made of the difference between the HR and LR model versions. The results have been aggregated over six regions: Iberia, South, East, West, Center and North, see Figs. 1, 2, 3, 4, 5, 6, 7. These areas have been chosen to highlight recurrent behaviours and to allow a thorough analysis of the model performances in different geographical and climatic contexts.

For both winter-mean TN and summer-mean TX strong biases have been found in the simulations, with the strongest ones in all regions for CMCC. This model shows mean bias of $+2.96\text{ }^{\circ}\text{C}$ ~ for TNavg-DJF up to $+4.96\text{ }^{\circ}\text{C}$ ~ in South and $-3.83\text{ }^{\circ}\text{C}$ ~ for TXavg-JJA down to $-4.89\text{ }^{\circ}\text{C}$ ~ in North, indicating an underestimation of the amplitude of the seasonal cycle all over the continent.

On the contrary, the other models present smaller biases in the continental average, with regional anomalies. In particular, the biases of maximum summer temperature show a common North-South gradient in the bias, with warmer values along the European coasts of the Mediterranean and the Black Sea (up to $+4.06\text{ }^{\circ}\text{C}$ over Iberia for MPI). This may be related to excessive moisture in Northern Europe and a lack of moisture in the Southern sector (Seneviratne et al. 2006; van Oldenborgh et al. 2009; Lorenz et al. 2010). Furthermore, a recent work Boé et al. (2020) connects larger simulated temperatures to lower evapotranspiration over the sea. Such phenomenon has direct effects on specific and relative humidity, implying a reduction in cloud cover and in precipitation, thus soil moisture.

At the same time, the simulated trends of TXavg-JJA overestimate those observed on Northern and, less often, Western Europe. This result differs from the underestimation of the trends over Southern Europe (e.g. $-0.33\text{ }^{\circ}\text{C}/\text{dec}$ for ECMWF, which reduces by a half the observed trend of $+0.66\text{ }^{\circ}\text{C}/\text{dec}$, see Table 4). This pattern confirms the findings of Bhend and Whetton (2013), that detected an underestimation of trends of average summer TX over Europe in CMIP5.

Such a gradient is found, with larger intensity, in the analysis of the biases for the trends on extreme summer maximum temperatures. In particular three models (EC-Earth, ECMWF and HadGEM3) strongly underestimate the increase of days above the 90th percentile. The observed trend over Southern Europe ($+2.92\text{ }^{\circ}\text{C}/\text{dec}$, see Table 4) is halved by these models, whose biases are respectively

$-1.35\text{ }^{\circ}\text{C}/\text{dec}$, $-1.64\text{ }^{\circ}\text{C}/\text{dec}$, $-1.83\text{ }^{\circ}\text{C}/\text{dec}$, see Table 6. Similar anomaly is found for ECMWF in spring: $-1.14\text{ }^{\circ}\text{C}/\text{dec}$, more than a half of the observed $+1.81\text{ }^{\circ}\text{C}/\text{dec}$. Nevertheless, these are not isolated cases: the other models (except CMCC) generally underestimate the trends of TX90p on Southern and Eastern Europe for all seasons. This reveals on one side the simulation of narrower seasonal temperature distributions than are observed and, more strikingly, a lack of simulated warm events over an area which is extremely sensitive to the increased strength and length of heatwaves (Della-Marta et al. 2007; Simolo et al. (2010; Squintu et al. 2019).

In Southern Europe and Iberia, the combination of an excessively large negative bias in summer maximum temperatures with a too weak increase in the seasonal average and with a much weaker (compared to observations) increase in the extreme indices points to issues in the representation of soil moisture in the models. In a climate which is too warm the soil can be expected to lack more moisture than in cooler conditions due to enhanced evaporation. Once the soil is dry the radiation balance is shifted to a state where sensible heat is dominant over latent heat. Under boundary conditions where the incoming energy flux (due to increase of green house gases) raises, this implies a further increase in sensible heat and surface warming. Nevertheless in conditions of moist soil, the simulated warming trend in temperatures would be even stronger, due to the shift from latent, thus getting close to the observed conditions (Seneviratne et al. 2006; van Oldenborgh et al. 2009; Lorenz et al. 2010; Min et al. 2013).

When considering the results for winter minimum temperatures common patterns are found among the models. The most evident one is the underestimation of the average winter minimum temperatures over Italy and Norway. These are likely to be related to poor representation of winter minimum temperatures in mountain areas such the Alps, the Apennines and the Scandinavian Mountains. At the same time winter minimum temperatures are overestimated in the plain regions of the north of Sweden and Finland, this is probably connected to a lack of snow coverage simulated by the models, as suggested by van Oldenborgh et al. (2009) and Diffenbaugh et al. (2013).

All the models present an overestimation of the trends in average winter TN over Southern and Central Europe and an underestimation over Eastern Europe (excluding the Kola peninsula), Fig. 2. These colder biases on the trends ($-0.41\text{ }^{\circ}\text{C}$ for CMCC and CNRM) present almost the same amplitude of the observed trend ($0.51\text{ }^{\circ}\text{C}$, Table 4), indicating that these models forecast a very weak increase of winter TN in these regions. This might be linked to an underestimation

Fig. 4 Difference in absolute trend bias of TN10p-DJF between HR and LR models. Red (green) pixels indicate an increase (decrease) of the absolute trend bias, thus a better (worse) performance

Fig. 5 Same as Fig. 1 but for TX-JJA

Fig. 6 Same as Fig. 2 but for TX-JJA

Fig. 7 Difference in trends of TX90p-DJF between the considered models and E-OBS. Red(blue) indicates an overestimation (underestimation) of the trend, related to a warmer (colder) trend

Fig. 8 Same as Fig. 4 but for TX90p-JJA

Table 4 Observed trends on average (°C per decade) and extreme values (TN10p and TX90p, % per decade) over whole Europe and each of the regions for all the seasons, E-Obs data

Region	Season	TNavg-DJF [°C/dec]	TN10p-DJF [%/dec]	TXavg-JJA [°C/dec]	TX90p-JJA [%/dec]
Europe	DJF	+ 0.35	− 0.56	+ 0.26	+ 1.13
	MAM	+ 0.31	− 1.41	+ 0.50	+ 1.78
	JJA	+ 0.39	− 2.46	+ 0.47	+ 1.62
	SON	+ 0.43	− 1.84	+ 0.41	+ 1.73
Iberia	DJF	+ 0.10	− 0.56	+ 0.20	+ 1.02
	MAM	+ 0.54	− 3.70	+ 0.66	+ 2.35
	JJA	+ 0.47	− 3.34	+ 0.50	+ 1.96
	SON	+ 0.37	− 2.41	+ 0.22	+ 0.87
South	DJF	+ 0.04	+ 0.15	+ 0.16	+ 1.43
	MAM	+ 0.34	− 1.76	+ 0.51	+ 1.81
	JJA	+ 0.48	− 3.36	+ 0.66	+ 2.92
	SON	+ 0.45	− 2.41	+ 0.40	+ 2.01
East	DJF	+ 0.51	− 0.80	+ 0.32	+ 1.17
	MAM	+ 0.22	− 0.63	+ 0.43	+ 1.48
	JJA	+ 0.41	− 2.34	+ 0.51	+ 1.78
	SON	+ 0.48	− 1.90	+ 0.47	+ 1.94
West	DJF	+ 0.14	− 0.16	+ 0.23	+ 1.58
	MAM	+ 0.42	− 2.30	+ 0.59	+ 2.11
	JJA	+ 0.33	− 2.35	+ 0.35	+ 0.99
	SON	+ 0.40	− 1.74	+ 0.33	+ 1.33
Center	DJF	+ 0.13	− 0.28	+ 0.18	+ 1.19
	MAM	+ 0.30	− 1.39	+ 0.56	+ 1.86
	JJA	+ 0.36	− 2.38	+ 0.52	+ 2.02
	SON	+ 0.29	− 1.40	+ 0.36	+ 1.54
North	DJF	+ 0.54	− 0.95	+ 0.31	+ 0.72
	MAM	+ 0.34	− 1.32	+ 0.49	+ 1.84
	JJA	+ 0.34	− 1.97	+ 0.30	+ 0.60
	SON	+ 0.46	− 1.81	+ 0.46	+ 1.86

of the reducing trend of snow coverage compared to the observations (van Oldenborgh et al. 2009). Furthermore, a recent study (Dai et al. 2019) has found a strict connection between sea-ice loss and warmer trends in winter temperatures at high latitudes, the Arctic Amplification. The values in Table 4 confirm that Eastern and Northern Europe have experienced the largest warming trends in winter average and extreme values. Hence, a poor simulation of the Arctic Amplification can be the reason of the colder simulated trends in the areas around the Baltic Sea and the Arctic Ocean (ignoring the Kola issue).

The most relevant anomaly for the trends in the extremes is found for the warm biases over South (− 1.88%/dec for HadGEM3), Center (− 2.29%/dec for CMCC) and Iberia (− 2.29%/dec for CMCC). In these regions the observed trends range between (− 0.56%/dec and + 0.15%/dec, Table 4). Thus, the model trends of the cold extremes are significantly stronger than the observed ones, revealing an

excessive narrowing tendency in the distribution of winter minimum temperatures. This relevant anomaly is not detected in the other seasons, only Central Europe presents warmer simulated trends in almost all cases.

The too warm simulated trends on the peninsula of Kola are found for TNavg-DJF and TN10p-DJF (as an underestimation of the number of days below the 10th percentile) and is related to E-OBS station density issues. The only series with observed values in the area (Krasnoshelye) starting before 1970 has missing data between 1972 and 1980. The interpolation of data coming from series in surrounding stations, in the case of TN, brings to higher values in the 1972–1980 compared to the following years, introducing a too cold trend that doesn't take place in the models. This behaviour, limited to only one series, motivates the ECA&D group to work on further data collection and in increasing the station density in this and other areas. This will enable

Table 5 Difference between the trends of TN10p in the models and in the E-Obs for each region and each season

Region	Season	CMCC	CNRM	EC-Earth3	ECMWF	HadGEM3	MPI
Europe	DJF	+ 0.99	− 0.26	− 0.36	− 0.65	− 1.25	+ 0.06
	MAM	− 0.74	− 0.86	− 0.25	− 0.35	− 0.42	− 0.68
	JJA	− 0.72	− 0.05	− 0.07	− 0.23	+ 0.86	− 0.55
	SON	− 0.61	+ 0.04	− 0.43	− 0.37	− 0.03	+ 0.06
Iberia	DJF	− 2.68	− 0.82	− 1.47	− 1.15	− 2.08	+ 0.26
	MAM	+ 0.87	+ 2.29	+ 1.23	+ 1.64	+ 1.86	+ 1.57
	JJA	+ 0.36	+ 0.89	+ 0.72	+ 0.62	+ 1.56	+ 0.48
	SON	+ 0.37	− 0.25	− 0.23	+ 0.25	+ 0.07	− 0.32
South	DJF	− 0.95	− 1.20	− 1.15	− 1.31	− 1.88	− 0.32
	MAM	− 0.20	− 0.44	+ 0.07	+ 0.22	− 0.26	− 1.24
	JJA	− 0.51	+ 0.43	+ 0.73	+ 0.31	+ 1.82	− 0.85
	SON	− 0.86	+ 0.34	− 0.08	− 0.59	+ 0.16	− 0.59
East	DJF	+ 0.24	+ 0.28	+ 0.20	− 0.53	− 1.05	+ 0.30
	MAM	− 1.12	− 1.06	− 0.88	− 1.07	− 1.03	− 1.24
	JJA	− 0.45	+ 0.23	− 0.22	− 0.14	+ 0.90	− 0.85
	SON	− 0.67	+ 0.39	− 0.37	− 0.41	− 0.13	+ 0.59
West	DJF	− 2.63	− 0.13	− 1.02	− 0.76	− 1.12	− 0.05
	MAM	− 0.51	+ 0.17	+ 0.56	+ 0.69	+ 0.92	+ 0.38
	JJA	− 2.04	− 0.38	− 0.06	− 0.46	+ 0.98	− 0.62
	SON	− 0.17	− 0.46	− 0.48	− 0.16	+ 0.30	− 0.68
Center	DJF	− 2.29	− 0.32	− 0.46	− 0.82	− 1.28	+ 0.34
	MAM	− 1.52	+ 0.17	− 0.20	− 0.15	− 0.40	− 0.79
	JJA	− 1.87	− 0.38	− 0.24	− 0.54	+ 1.12	− 1.40
	SON	− 1.96	− 0.46	− 0.60	− 0.90	− 0.06	− 0.67
North	DJF	− 1.02	− 0.58	− 0.15	− 0.21	− 0.97	− 0.40
	MAM	− 0.62	− 2.13	− 0.27	− 0.70	− 0.96	− 1.20
	JJA	− 0.48	− 0.96	− 0.44	− 0.74	− 0.18	− 0.29
	SON	− 0.18	− 0.41	− 0.67	− 0.22	− 0.12	+ 0.11

an increase in the quality of the interpolation and avoid such criticisms.

The combination of the results for TN and TX indicates that around the Mediterranean and in Central Europe the trends in the percentage of events below the 10th percentile and above the 90th are underestimated for a relevant number of regions and models. This implies that the tails of the distribution are simulated to get closer to each other faster than observed. Thus, in these areas the distribution of simulated daily temperatures becomes narrower compared to the distribution of observed daily temperatures, underestimating the intensity of the extremes, especially the warm ones. As aforementioned, several issues in the models can be the reasons behind this (soil moisture, sea surface evapotranspiration). The diagnosis of such problems is not in the scope of this paper, which only aims at identifying eventual issues in the simulations.

As a last step, the analysis of the absolute trend bias evolution in the models from LR to HR does not show a general improvement. Each model presents different patterns and diverse behaviour in terms of change of mean absolute trend bias. Nevertheless this index decreases for TN10p-DJF in almost all models (but ECMWF), indicating a better improvement compared to what is found for TX90p-JJA, where only 3 methods slightly improve and the other ones are worsening by up to + 0.25%/dec.

Finally, it appears that the new high resolution models, even though they do not significantly increase or decrease their absolute bias on the trends of the extremes, still have some problems especially on the area of the Mediterranean. In this region the most serious discrepancy to observations is the large underestimation of the increasing trends of warm extremes. Considering the high economic and societal vulnerability of these areas to very warm events in summer and

Table 6 Difference between the trends of TX90p in the models and in the E-Obs for each region and each season

Region	Season	CMCC	CNRM	EC-Earth3	ECMWF	HadGEM3	MPI
Europe	DJF	+ 0.36	− 0.77	− 0.13	− 0.44	− 0.16	− 0.28
	MAM	+ 0.45	− 0.44	− 0.67	− 0.87	− 0.71	− 0.61
	JJA	+ 0.53	− 0.29	− 0.73	− 0.59	− 0.56	0.00
	SON	+ 0.26	− 0.69	− 0.39	− 0.11	− 0.29	− 0.19
Iberia	DJF	+ 1.51	+ 0.16	+ 0.53	+ 0.08	+ 0.01	+ 0.79
	MAM	− 0.39	− 1.47	− 1.05	− 1.44	− 1.55	− 1.19
	JJA	+ 0.95	− 0.33	− 0.43	− 0.38	− 1.21	+ 0.37
	SON	+ 1.60	+ 0.84	+ 0.70	+ 0.23	+ 0.63	+ 1.08
South	DJF	− 0.70	− 0.28	− 0.15	− 0.41	− 0.01	− 0.32
	MAM	+ 0.50	− 0.22	− 0.42	− 1.14	− 0.41	− 0.49
	JJA	+ 0.38	− 0.41	− 1.35	− 1.64	− 1.83	− 1.39
	SON	+ 0.92	− 1.02	− 0.30	− 0.33	− 0.27	− 0.89
East	DJF	− 0.41	− 1.34	− 0.64	− 0.68	− 0.10	− 0.91
	MAM	+ 0.98	− 0.30	− 0.15	− 0.54	− 0.60	− 0.68
	JJA	− 0.20	− 0.94	− 0.89	− 0.88	− 0.79	− 0.21
	SON	− 0.17	− 1.69	− 0.72	− 0.15	− 0.50	− 0.65
West	DJF	+ 1.05	− 0.96	− 0.02	− 0.55	− 0.56	− 0.03
	MAM	− 0.30	− 0.47	− 1.29	− 1.15	− 1.37	− 1.01
	JJA	+ 1.49	+ 0.98	− 0.34	− 0.21	+ 0.20	+ 0.53
	SON	− 0.69	+ 0.53	− 0.37	− 0.32	− 0.36	− 0.38
Center	DJF	+ 1.12	− 0.26	+ 0.09	− 0.37	− 0.52	− 0.27
	MAM	+ 0.38	− 0.22	− 0.90	− 0.86	− 0.73	− 0.23
	JJA	+ 1.29	− 0.35	− 1.18	− 0.90	− 0.98	0.00
	SON	+ 0.77	− 0.26	− 0.30	+ 0.21	+ 0.18	+ 0.27
North	DJF	+ 0.99	− 0.66	+ 0.28	− 0.23	− 0.03	+ 0.21
	MAM	+ 0.27	− 0.51	− 1.12	− 0.97	− 0.40	− 0.35
	JJA	+ 0.77	+ 0.37	− 0.17	+ 0.32	+ 0.68	+ 0.65
	SON	− 0.39	− 0.26	− 0.38	− 0.18	− 0.90	− 0.13

the importance of the prediction of heatwaves intensity and frequency for the next decades, it is fundamental to improve the simulation of these phenomena and of their projections in future decades.

It's not in the purpose of this work to pick the best model or to uphold the reliability of the models in their simulations or in their projections to the future. Nevertheless, for most of the models the number of gridpoints that present significant difference between simulated and observed trends is relatively low, even though not ideal. This allows us to affirm that, notwithstanding relevant biases in the seasonal averages, the trends in the average values are trustworthy and the trends in the extremes can roughly describe the general tendencies. Nevertheless the need of serious improvements on the simulation of temperature variability and of its consequences on extreme events is clear.

A Trends of the observed values for all regions and seasons

This appendix collects the values of the trends on the indices that have been analyzed in this study, so that the reader can be informed on the trends that the models are expected to reproduce. The trends on the average values are in °C per decade, while the extreme indices are reported in % per decade. Hence, the values cannot be directly compared to detect changes in variability. The only possible inferences can be derived from eventual discrepancies in the tendency towards colder or warmer climate, which happens only for the winter minimum temperature in Iberia. In general the average trends on TX appear larger than TN for all the seasons, excluding DJF for North and East where the Arctic Amplification causes larger trends for TN. As for the extreme values, the nearly gaussian shape of the temperature distribution imply that a shift (e.g. to warmer values) of

Fig. 9 Difference in the trends of TX10p-DJF between the LR models and E-OBS. Red (blue) indicates that the models have more negative (positive) trends in the number of days above the 10th percentile, related to warming (cooling) trends. Circles indicate significant differences

Fig. 10 Difference in the trends of TX90p between the LR models and E-OBS. Red (blue) indicates that the models have more positive (negative) trends in the number of days above the 90th percentile, related to warming (cooling) trends. Circles indicate significant differences

the distribution would produce trends that are larger in the increase of warm events than the decrease of cold events (Hartmann et al. 2013). Therefore, since the changes in the two indices have the same order of magnitude in almost all the regions and seasons, it's possible to affirm that the cold tail of the distribution is warming faster than the warm tail, especially in summer.

B Model biases of the trends for all regions and seasons

The main text of this work focused on the trends of winter TN10p and summer TX90p. Here all the differences between simulated and observed trends for the other seasons are reported.

In the case of TN10p a positive (negative) trend indicates more (less) cold events and thus a colder (warmer) trends than observed. Negative biases are slightly more present, except for South and Iberia, similarly to what found for winter TN10p.

In the case of TX90p a positive (negative) trend indicates more (less) warm events and thus a warmer (colder) trends than observed. Excluding CMCC, almost all models present mainly negative differences, with summer values almost always as the most negative in regions as South, Centre and East.

C Performance of the models in their low resolution version

The LR version of the models tend to underestimate the trend in TN10p-DJF. Figure 9 displays that all the continental mean biases are negative. This indicates a general overestimation of the warming trend, with a few exceptions for some models in certain regions, e.g. for MPI. Only for a small part of the grid-points the difference is significant, i.e. the 95% confidence level ranges of the models and E-OBS trends don't superimpose.

Performance of LR models on trends of TX90p-JJA present different patterns. In general trends on South and East are underestimated (colder trends than observed) by all models. Nevertheless the continental mean biases indicate that, contrarily to what is seen for the other models, CMCC and CNRM slightly overestimate the trends (Fig. 10).

Acknowledgements We acknowledge funding from the PRIMAVERA project, funded by the European Union's Horizon 2020 programme under Grant Agreement No. 641727. The E-OBS observational dataset is made available through the European Climate Assessment & Dataset (www.ecad.eu).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alexander LV, Zhang X, Peterson TC, Caesar J, Gleason B, Klein Tank A, Haylock M, Collins D, Trewin B, Rahimzadeh F et al (2006) Global observed changes in daily climate extremes of temperature and precipitation. *J Geophys Res Atmos* 111(D5):1–22
- Bhend J, Whetton P (2013) Consistency of simulated and observed regional changes in temperature, sea level pressure and precipitation. *Clim Change* 118(3–4):799–810
- Boé J, Somot S, Corre L, Nabat P (2020) Large discrepancies in summer climate change over Europe as projected by global and regional climate models: causes and consequences. *Clim Dyn* 54(5):2981–3002
- Brown S, Caesar J, Ferro CA (2008) Global changes in extreme daily temperature since 1950. *J Geophys Res Atmos* 113(D5):1–11
- Cherchi A, Fogli PG, Lovato T, Peano D, Iovino D, Gualdi S, Masina S, Scoccimarro E, Materia S, Bellucci A et al (2019) Global mean climate and main patterns of variability in the cmcc-cm2 coupled model. *J Adv Model Earth Syst* 11(1):185–209
- Cornes R, Jones P (2013) How well does the era-interim reanalysis replicate trends in extremes of surface temperature across Europe? *J Geophys Res Atmos* 118(18):10–262
- Cornes RC, van der Schrier G, van den Besselaar EJ, Jones PD (2018) An ensemble version of the e-obs temperature and precipitation data sets. *J Geophys Res Atmos* 123(17):9391–9409
- Dai A, Luo D, Song M, Liu J (2019) Arctic amplification is caused by sea-ice loss under increasing co2. *Nat Commun* 10(1):1–13
- Data C (2009) Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation. World Meteorological Organization
- Della-Marta PM, Haylock MR, Luterbacher J, Wanner H (2007) Doubled length of western european summer heat waves since 1880. *J Geophys Res Atmos* 112(D15): 1–11
- Diffenbaugh NS, Schrer M, Ashfaq M (2013) Response of snow-dependent hydrologic extremes to continued global warming. *Nat Clim Change* 3:379–384. <https://doi.org/10.1038/NCLIMATE1732>
- Donat MG, Alexander LV (2012) The shifting probability distribution of global daytime and night-time temperatures. *Geophys Res Lett* 39(14):1–5
- Flato G, Marotzke J, Abiodun B, Braconnot P, Chou SC, Collins W, Cox P, Driouech F, Emori S, Eyring V, et al. (2014) Evaluation of climate models. In: Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, pp 741–866
- Frich P, Alexander LV, Della-Marta P, Gleason B, Haylock M, Tank AK, Peterson T (2002) Observed coherent changes in climatic extremes during the second half of the twentieth century. *Clim Res* 19(3):193–212
- Giorgi F (2006) Climate change hot-spots. *Geophys Res Lett* 33(8):1–4

- Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. *J Geophys Res Atmos* 113(D6):1–20
- Gutjahr O, Putrasahan D, Lohmann K, Jungclauss JH, Storch JSv, Brüggemann N, Haak H, Stössel A (2019) Max Planck Institute earth system model (mpi-esm1. 2) for the high-resolution model intercomparison project (highresmip). *Geosci Model Dev* 12(7):3241–3281
- Haarsma R, Acosta M, Bakhshi R, Bretonnière PA, Caron LP, Castrillo M, Corti S, Davini P, Exarchou E, Fabiano F, Fladrich U, Fuentes Franco R, García-Serrano J, von Hardenberg J, Koenigk T, Levine X, Loana Meccia V, van Noije T, van den Oord G, Palmeiro FM, Rodrigo M, Ruprich-Robert Y, Le Sager P, Tourigny E, Wang S, van Weele M, Wyser K (2020) HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR—description, model computational performance and basic validation. *Geosci Model Dev* 13(8):3507–3527
- Haarsma RJ, Roberts MJ, Vidale PL, Senior CA, Bellucci A, Bao Q, Chang P, Corti S, Fučkar NS, Guemas V et al (2016) High resolution model intercomparison project (highresmip v1. 0) for cmip6. *Geosci Model Dev* 9(11):4185–4208
- Hartmann D, Tank A, Rusticucci M (2013) Working group i contribution to the IPCC fifth assessment report. *Climatic Change* pp 31–39
- Haylock M, Hofstra N, Tank AK, Klok E, Jones P, New M (2008) A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J Geophys Res Atmos* 113(D20):1–12
- Jones PD, Wigley T (2010) Estimation of global temperature trends: what's important and what isn't. *Clim Change* 100(1):59–69
- Katz RW, Brown BG (1992) Extreme events in a changing climate: variability is more important than averages. *Clim change* 21(3):289–302
- Kharin VV, Zwiers FW, Zhang X (2005) Intercomparison of near-surface temperature and precipitation extremes in amip-2 simulations, reanalyses, and observations. *J Clim* 18(24):5201–5223
- Kharin VV, Zwiers FW, Zhang X, Hegerl GC (2007) Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *J Clim* 20(8):1419–1444
- Kiktev D, Sexton DM, Alexander L, Folland CK (2003) Comparison of modeled and observed trends in indices of daily climate extremes. *J Clim* 16(22):3560–3571
- Klein Tank A, Können G (2003) Trends in indices of daily temperature and precipitation extremes in Europe, 1946–99. *J Clim* 16(22):3665–3680
- Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Milate M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westernmeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Forland E, Mielus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnár T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander L, Petrovic P (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int J Climatol* 22(12):1441–1453
- Lorenz R, Jaeger EB, Seneviratne SI (2010) Persistence of heat waves and its link to soil moisture memory. *Geophys Res Lett* 37:L09703. <https://doi.org/10.1029/2010GL042764>
- Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JF, Stouffer RJ, Taylor KE (2007) The wcrp cmip3 multimodel dataset: a new era in climate change research. *Bull Am Meteorol Soc* 88(9):1383–1394
- Min E, Hazeleger W, Van Oldenborgh G, Sterl A (2013) Evaluation of trends in high temperature extremes in north-western Europe in regional climate models. *Environ Res Lett* 8(1):014011
- Morak S, Hegerl G, Kenyon J (2011) Detectable regional changes in the number of warm nights. *Geophys Res Lett* 38(17):1–5
- Roberts CD, Senan R, Molteni F, Boussetta S, Mayer M, Keeley SP (2018) Climate model configurations of the ecmwf integrated forecasting system (ecmwf-ifs cycle 43r1) for highresmip. *Geosci Model Dev* 11(9):3681–3712
- Roberts MJ, Baker A, Blockley EW, Calvert D, Coward A, Hewitt HT, Jackson LC, Kuhlbrodt T, Mathiot P, Roberts CD et al (2019) Description of the resolution hierarchy of the global coupled hadgem3-gc3. 1 model as used in cmip6 highresmip experiments. *Geosci Model Dev* 12(12):4999–5028
- Roberts MJ, Camp J, Seddon J, Vidale PL, Hodges K, Vanniere B, Mecking J, Haarsma R, Bellucci A, Scoccimarro E, Caron LP, Chauvin F, Terray L, Valcke S, Moine MP, Putrasahan D, Robert C, Senan R, Zarzycki C, Ullrich P, Yamada Y, Mizuta R, Kodama C, Fu D, Zhang Q, Danabasoglu G, Rosenbloom N, Wang H, Wu L (2020) Projected future changes in tropical cyclones using the CMIP6 HighResMIP multimodel ensemble. *Geophys Res Lett* 47(14):1–12
- 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
- Scherrer SC, Appenzeller C, Liniger MA, Schär C (2005) European temperature distribution changes in observations and climate change scenarios. *Geophys Res Lett* 32(19):1–5
- Sen PK (1968) Estimates of the regression coefficient based on Kendall's tau. *J Am Stat Assoc* 63(324):1379–1389
- Seneviratne SI, Lüthi D, Litschi M, Schär C (2006) Land-atmosphere coupling and climate change in Europe. *Nature* 443:205–209. <https://doi.org/10.1038/nature05095>
- Sillmann J, Kharin V, Zhang X, Zwiers F, Bronaugh D (2013) Climate extremes indices in the cmip5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J Geophys Res Atmos* 118(4):1716–1733
- Sillmann J, Donat MG, Fyfe JC, Zwiers FW (2014) Observed and simulated temperature extremes during the recent warming hiatus. *Environ Res Lett* 9(6):064023
- Sillmann J, Kharin V, Zwiers F, Zhang X, Bronaugh D, Donat M (2014) Evaluating model-simulated variability in temperature extremes using modified percentile indices. *Int J Climatol* 34(11):3304–3311
- Simolo C, Brunetti M, Maugeri M, Nanni T, Speranza A (2010) Understanding climate change-induced variations in daily temperature distributions over Italy. *J Geophys Res Atmos* 115(D22):1–12
- Squintu AA, van der Schrier G, Brugnara Y, Klein Tank A (2019) Homogenization of daily temperature series in the European climate assessment and dataset. *Int J Climatol* 39(3):1243–1261
- Squintu AA, van der Schrier G, van den Besselaar EJ, Cornes RC, Klein Tank AM (2020) Building long homogeneous temperature series across Europe: a new approach for the blending of neighboring series. *J Appl Meteorol Clim* 59(1):175–189
- Stainforth DA, Chapman SC, Watkins NW (2013) Mapping climate change in European temperature distributions. *Environ Res Lett* 8(3):034031
- Sterl A, Severijns C, Dijkstra H, Hazeleger W, van Oldenborgh GJ, van den Broeke M, Burgers G, van den Hurk B, van Leeuwen PJ, van Velthoven P (2008) When can we expect extremely high surface temperatures? *Geophys Res Lett* 35(14):1–5
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of cmip5 and the experiment design. *Bull Am Meteorol Soc* 93(4):485–498
- van Oldenborgh GJ, Drijfhout S, Van Ulden A, Haarsma R, Sterl A, Severijns C, Hazeleger W, Dijkstra H et al (2009) Western Europe is warming much faster than expected. *Clim Past* 5(1):1–12
- Voltaire A, Saint-Martin D, Sénési S, Decharme B, Alias A, Chevalier M, Colin J, Guérémy JF, Michou M, Moine MP et al (2019)

Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *J Adv Model Earth Syst.* <https://doi.org/10.1029/2019MS001683>

Zhang X, Hegerl G, Zwiers FW, Kenyon J (2005) Avoiding inhomogeneity in percentile-based indices of temperature extremes. *J Clim* 18(11):1641–1651

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.