ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED
PUBLISHING™

LETTER • OPEN ACCESS

Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods

To cite this article: Thordis L Thorarinsdottir *et al* 2020 *Environ. Res. Lett.* **15** 124041

View the article online for updates and enhancements.

You may also like

- Predictable quantum efficient detector based on *n*-type silicon photodiodes
Timo Dönsberg, Farshid Manoocheri, Meelis Sildoja et al.
- Simulations of a predictable quantum efficient detector with PC1D
Jarle Gran, Toomas Kübarsepp, Meelis Sildoja et al.
- Predictable quantum efficient detector: II. Characterization and confirmed responsivity
Ingmar Müller, Uwe Johannsen, Ulrike Linke et al.

This content was downloaded from IP address 5.135.140.155 on 21/11/2024 at 19:24

Environmental Research Letters

LETTER

Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods

Thordis L Thorarinsdottir[1], **Jana Sillmann**[2], **Marion Haugen**[1], **Nadine Gissibl**[3] and **Marit Sandstad**[2]

[1] Norwegian Computing Center, Oslo, Norway
[2] Center for International Climate and Environmental Research, Oslo, Norway
[3] Department of Mathematics, Technical University of Munich, Munich, Germany

E-mail: thordis@nr.no

OPEN ACCESS

RECEIVED
26 August 2020

REVISED
22 October 2020

ACCEPTED FOR PUBLICATION
4 November 2020

PUBLISHED
11 December 2020

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Keywords: climate model evaluation, performance measure, temperature extremes, testing equal performance, integrated quadratic distance, proper divergence functions

Supplementary material for this article is available online

Abstract

Reliable projections of extremes by climate models are becoming increasingly important in the context of climate change and associated societal impacts. Extremes are by definition rare events, characterized by a small sample associated with large uncertainties. The evaluation of extreme events in model simulations thus requires performance measures that compare full distributions rather than simple summaries. This paper proposes the use of the integrated quadratic distance (IQD) for this purpose. The IQD is applied to evaluate CMIP5 and CMIP6 simulations of monthly maximum and minimum near-surface air temperature over Europe and North America against both observation-based data and reanalyses. Several climate models perform well to the extent that these models' performance is competitive with the performance of another data product in simulating the evaluation set. While the model rankings vary with region, season and index, the model evaluation is robust against changes in the grid resolution considered in the analysis. When the model simulations are ranked based on their similarity with the ERA5 reanalysis, more CMIP6 than CMIP5 models appear at the top of the ranking. When evaluated against the HadEX2 data product, the overall performance of the two model ensembles is similar.

1. Introduction

Current climate projections indicate a significant warming of the hottest days and the coldest nights in all land areas of the world already under low emission scenarios (Hoegh-Guldberg *et al* 2018), and even more severe increases are projected for higher emission scenarios (Sillmann *et al* 2013b). Impact studies, for instance for the health, agriculture or energy sector, often use climate model projections as input to estimate possible impacts of increasing temperatures for informing adaptation and mitigation decision-making (e.g. Orlov *et al* (2019)). Reliable projections of near-surface air temperature (SAT) extremes by climate models become more and more important in this context. The performance of climate models is assessed on the basis of their historical simulations for the recent past, which are forced by observed greenhouse gas concentrations, sulfate and volcanic aerosol, stratospheric ozone and solar luminosity variations as outlined in the protocols of the Coupled Model Intercomparison Project (CMIP) now being in its 6th phase (Eyring *et al* 2016).

Climate model evaluation has traditionally been performed by comparing summary statistics that are derived from simulated model output and corresponding observed quantities using, for instance, the root mean squared error (RMSE) or mean bias (Flato *et al* 2014). Both RMSE and mean bias compare averages over time and/or space, ignoring the variability, or the uncertainty, in the underlying values. However, a quantification of the uncertainty in the model simulations is a critical and challenging task (Knutti *et al*

© 2020 The Author(s). Published by IOP Publishing Ltd

2003, Tebaldi and Knutti 2007, Palmer 2012). As a consequence, climate models should be evaluated by comparing the probability distribution of model output to the corresponding distribution of observed data (Guttorp 2011, Thorarinsdottir *et al* 2013), particularly when evaluating extremes. By definition, extremes are simultaneously highly variable and rare. Mean values and similar summary statistics may therefore not provide sufficient information to properly evaluate the underlying processes (Maraun *et al* 2017).

Traditionally, probabilistic model evaluation has been applied to the setting where a prediction given by a probability distribution is compared against an observation given by a single value (Gneiting and Raftery 2007). When the aim of the evaluation is to compare and rank competing models, it is essential that the expected optimal performance is obtained for the true data generating process. This decision-theoretic condition encourages transparent and careful assessment. Performance measures fulfilling this property are called proper scoring rules and are considered essential in scientific and managerial practice in various application fields, including economics and meteorology (Winkler and Murphy 1968, Gneiting and Raftery 2007, Bröcker and Smith 2007, Armantier and Treich 2013). Thorarinsdottir *et al* (2013) extended the framework of proper scoring rules to proper divergence functions for comparing probability distributions of model output against corresponding probability distributions of observed data. The two concepts are tightly linked in that every proper scoring rule is associated with a proper divergence function.

Special care is required for model evaluation with respect to extremes, see e.g. Sippel *et al* (2015). A common procedure is to select a small extreme subset of all observed events and evaluate the model's performance based on its ability to simulate only these specific events. However, without adjusting for the event selection process, the evaluation will favor whichever model is most likely to generate the extremes, even if this model serverely overestimates the occurrence rate (Lerch *et al* 2017). An alternative approach is to define a new variable that represents the extremes of interest and evaluate the full distribution of this variable.

In order to capture extreme temperature events in model simulations, the Expert Team of Climate Change Detection and Indices (ETCCDI) has defined a set of widely used indices for climate extremes (Zhang *et al* 2011). These indices are based on daily data and characterize moderate but robust large-scale extreme events. The ETCCDI indices have proven useful in the analysis of observations (Donat *et al* 2013), the evaluation of global climate models (IPCC 2013, Sillmann *et al* 2013a), and the projection of changes in climate extremes (Tebaldi *et al* 2006, Sillmann *et al* 2013b).

For evaluation of ETCCDI indices, we propose to use the integrated quadratic distance (IQD) (Thorarinsdottir *et al* 2013, Thorarinsdottir and Schuhen 2018) to compare distributions of simulated indices to the corresponding distributions from a data product. The IQD is the proper divergence associated with the proper continuous ranked probability score (CRPS) (Hersbach 2000, Gneiting and Raftery 2007). It has previously been used by Vrac and Friederichs (2015) and Yuan *et al* (2019) to evaluate statistical bias-correction and downscaling approaches.

Many different data products exist, both purely observation-based products as well as reanalyses that merge physical model simulations and observations. These products commonly show systematic differences, particularly in mountainous and sparsely observed regions (Eum *et al* 2014, Lussana *et al* 2018). We thus argue that a model performs well if its performance is competitive with the performance of an alternative data product, for example, a reanalysis. To assess this, we apply a testing framework from the economic literature for comparing model performance (Diebold and Mariano 1995). Note that this is conceptually different from using a statistical test to directly compare the empirical distribution of the model output to the corresponding observational distribution (Von Storch and Zwiers 2003, Orskaug *et al* 2011, Baker and Taylor 2016). In general, requiring a model to perform competitively with e.g. a reanalysis when both are compared against the same observational-based product is a weaker condition than requiring equality in distribution of model output and observations.

The remainder of the paper is organized as follows. The next section 2 introduces the extreme SAT indices that form the basis for the model evaluation as well as the various data products and climate models used in the analysis. Section 3 introduces the concept of proper divergence measures, the specific performance measure used in the analysis and the statistical test for comparing model performance. The results of the analysis are presented in the following section 4 with a discussion and conclusions provided in the final section 5.

2. Data sets and extreme indices

In this study, we evaluate climate model simulations of extreme SAT indices over North America and Europe, respectively, against observational and reanalysis data products for the time period 1979-2005. We focus on these two regions because they have the most complete observational data coverage. The specific indices and data sets are described below with further information, including access information, given in the supplementary information (https://stacks.iop.org/ERL/15/124041/mmedia).

2.1. Extreme indices

We analyze a set of indices defined by the ETCCDI that are derived from daily minimum and maximum SAT (TN and TX, respectively), measured in °C. Specifically, we focus on monthly minimum SAT (TNn), monthly maximum SAT (TXx) and monthly SAT range (TXx-TNn). The last quantity is also referred to as extreme temperature range (ETR) in Donat *et al* (2013), and we will use this notation in the following. We further build seasonal distributions by combining values for the Boreal summer months June, July and August, or values for the Boreal winter months December, January and February. Specifically, we consider distributions of TXx and ETR in summer, and distributions of TNn and ETR in winter.

2.2. Data sets

2.2.1. Observation-based Data

The gridded HadEX2 data set of observation-based indices (Donat *et al* 2013) allows comparison between model-simulated and observed indices. HadEX2 indices are calculated directly from station observations and then interpolated to a global grid of size 96×73, which results in a spatial scale mismatch with indices calculated from model output because the latter represents area (grid box) averages rather than point values, see Donat *et al* (2014). Similarly, the order of operation is important when extreme indices on a grid are derived from station observations. Specifically, the values tend to be more extreme if the extreme indices are first calculated for the station time series before the values are interpolated to a grid (Donat *et al* 2014).

Indices from model simulations and reanalyses were interpolated to the 3.75° (longitude) × 2.5° (latitude) grid of the HadEX2 data set to facilitate comparison. Furthermore, a mask was applied to all models and HadEX2 to exclude regions where HadEX2 data coverage is insufficient (i.e. where annual indices were available in fewer than 38 of the 40 years in the time period 1971–2010). Note that the spatial coverage in the HadEX2 data set varies among the different indices (Donat *et al* 2013). The more recent version of this data set, HadEX3 (Dunn *et al* 2020), is included as a model simulation for comparative assessment. HadEX3 originally has a resolution of $1.875° \times 1.25°$, corresponding to 192×144 grid cells. The spatial coverage of HadEX2 and HadEX3 is not identical, resulting in approximately 2.5% missing grid cells for North America.

For a detailed analysis of three grid cells in North America, we additionally consider an observation-based data set that is only available for North America. This data set, called ANUSPLIN + Livneh, is based directly on station observations in Canada (McKenney *et al* 2011) and the continental United States (Livneh *et al* 2013), which are combined with bilinear interpolation at the border; see Whan and Zwiers (2017) for more information.

2.2.2. Reanalyses

Reanalyses data are more readily comparable with model simulations due to their gridded output, complete global spatial coverage and similarity of scales represented. Although reanalyses are essentially observationally constrained model output, variables that are directly assimilated in the reanalysis forecast model are typically closer to observations. SAT fields such as those used for the indices calculation here are classified as 'type B' variables (Kalnay *et al* 1996), because the forecast model has substantial influence on the reanalyzed values and subsequently the simulated SAT extremes in the reanalysis are not constrained by observations. In this study, we compute indices for three reanalyses: ERA5 (Hersbach and Dee 2016), ERA-Interim (Dee *et al* 2011) and NCEP-DOE Reanalysis 2 (NCEP-2) (Kanamitsu *et al* 2002). ERA5 is downloaded on a regular $0.25° \times 0.25°$ grid of size 1440×721 and ERA-Interim is downloaded on a regular $0.75° \times 0.75°$ grid of size 480×241. The NCEP-2 reanalysis data set is available on a 192×94 Gaussian grid. In addition to evaluating the model simulations against HadEX2, the simulations are evaluated against the ERA5 reanalysis for comparison. The other two reanalysis, ERA-Interim and NCEP-2, are treated as model simulations for comparative assessment.

2.2.3. Climate model data

We evaluate 18 CMIP6 models and 30 CMIP5 models, see tables 1-4 in the supplementary material for further information, including model names, institutions and grid resolutions. We analyze in total 73 ensemble members for CMIP5, where the number of runs varies from one to five for each model. For CMIP6 there is only one model with several runs and we analyze in total 22 ensemble members. The analysis is based on the historical simulations of the CMIP5 and CMIP6 models employing historical changes in the atmospheric composition reflecting both anthropogenic and natural sources (Taylor *et al* 2012).

3. Evaluation methods

3.1. General properties

We compare a model simulation and a data product (i.e. HadEX2) by comparing the corresponding empirical cumulative distribution functions (ECDFs) of an extreme index over the entire time period 1979–2005 in each grid cell on a common grid. Specifically, we employ a divergence, or a distance, function d that compares two univariate distribution functions F and G, and returns a numeric value $d(F, G) \geq 0$ summarizing their differences with $d(F, G) = 0$ if $F = G$. More generally, a lower value indicates a smaller difference between F and G. Regional differences between two data sets are summarized by the average divergence over all grid cells in the region,

$$\frac{1}{N}\sum_{i=1}^{N} d(F_i, G_i), \tag{1}$$

where $i = 1, \ldots, N$ is the grid cell index.

For divergences, propriety (or the expected optimal performance of the true data generating process) is defined as follows (Thorarinsdottir *et al* 2013). Assume that $\widehat{G}_{(k)}$ is the ECDF of k values $y_1, \ldots, y_k$ that are independent realizations with distribution G. The divergence function d is k-proper if

$$\mathbb{E}_G d(G, \widehat{G}_{(k)}) \leq \mathbb{E}_G d(F, \widehat{G}_{(k)}), \tag{2}$$

for all distributions F and G, where $\mathbb{E}_G$ denotes the expected value with respect to G. This property should hold for any value of k in which case d is called a proper divergence function.

3.2. Integrated quadratic distance

In the evaluation, we employ the integrated quadratic distance (IQD),

$$d(F, G) = \int_{-\infty}^{+\infty} (F(x) - G(x))^2 \mathrm{d}x, \tag{3}$$

which fulfills all the conditions above (Thorarinsdottir *et al* 2013) while also comparing the full distributions. To demonstrate the IQD, figure 1 shows two example comparisons where ECDFs based on normally distributed samples of size 81 are compared, a situation somewhat corresponding to our application[4]. In the first example (figure 1, left), both samples have a variance of 4 with means equal to 0 and 1; in the second example (figure 1, right) the joint mean value is 0 while the variances equal 1 and 4. The IQD calculates the squared area between the two ECDFs (area indicated in gray in the figures) and the resulting values are 0.13 for the left example and 0.10 for the right example.

In comparison, the squared error (the squared difference between the mean values) is 1.10 for the example on the left in figure 1 and 0.01 for the example on the right. That is, performance evaluation based on the RMSE would detect only a minor difference between the two samples on the right while a substantial difference would be assigned to the two samples on the left.

3.3. Assessing the significance of the results

To compare the performance of a model simulation against that of a reanalysis, we apply a computationally efficient permutation test relying on resampling (Good 2013, Möller *et al* 2013). When evaluating against a data product with distribution G, the comparative performance of two models F^1 and F^2 under the divergence d equals

$$c = \frac{1}{N}\sum_{i=1}^{N} (d(F_i^1, G_i) - d(F_i^2, G_i)), \tag{4}$$

where $i = 1, \ldots, N$ is the grid cell index. If $c < 0$, the average divergence over all grid cells in the region is smaller for F^1 which then performs better overall, while F^2 is better if $c > 0$.

The permutation test is based on resampling copies of c with the labels of F^1 and F^2 swapped for a random subset of grid cells. That is, the index set $1, \ldots, N$ is randomly split in two sets, S_1 and S_2, and a permutation of c is calculated as

$$c^p = \frac{1}{|S_1|}\sum_{i \in S_1} (d(F_i^1, G_i) - d(F_i^2, G_i)) + \frac{1}{|S_2|}\sum_{i \in S_2} (d(F_i^2, G_i) - d(F_i^1, G_i)), \tag{5}$$

where $|S_1|$ and $|S_2|$ is the number of grid cells in S_1 and S_2, respectively, with $|S_1| + |S_2| = N$. Under the null hypothesis, F^1 and F^2 perform equally well and c cannot be distinguished from permutations of the type c^p. By considering the rank of c within a set of permutations, a test is obtained. Specifically, we sample 1000 random permutations and say that the performance of F^1 and F^2 is significantly different if the p-value is less than 0.05.

4. Results

4.1. Comparison with HadEX2

We first present results where reanalyses, an alternative observational product and CMIP model simulations are compared against the observational product HadEX2. Figure 2 shows the model rankings for summer TXx and winter TNn over North America. HadEX3 is very similar to HadEX2 and the reanalyses ERA5 and ERA-Interim are quite similar, while the NCEP-2 reanalysis performs poorly; for winter TNn only 7 out of 48 climate model simulations perform worse than NCEP-2. Nine climate models perform competitively with either ERA5 or ERA-Interim for summer TXx and 14 models for winter TNn. Notably, only the CMIP6 model CNRM-ESM2-1 performs competitively with either reanalysis for both summer TXx and winter TNn. Four CMIP6 models show particularly poor performance for winter TNn due to being too cold.

An example of a more detailed analysis is given in figure 3, focusing on winter TNn, the two reanalyses ERA-Interim and NCEP-2, and the CMIP5 model HadGEM2-CC. Out of these, ERA-Interim shows the best performance, with NCEP-2 performing slightly better than HadGEM2-CC. Interestingly, the range of scores across individual grid cells is largest for

[4] We analyze 27 years of data (1979-2005) and we have three observations per season, i.e. June, July and August for summer and December, January and February for winter.

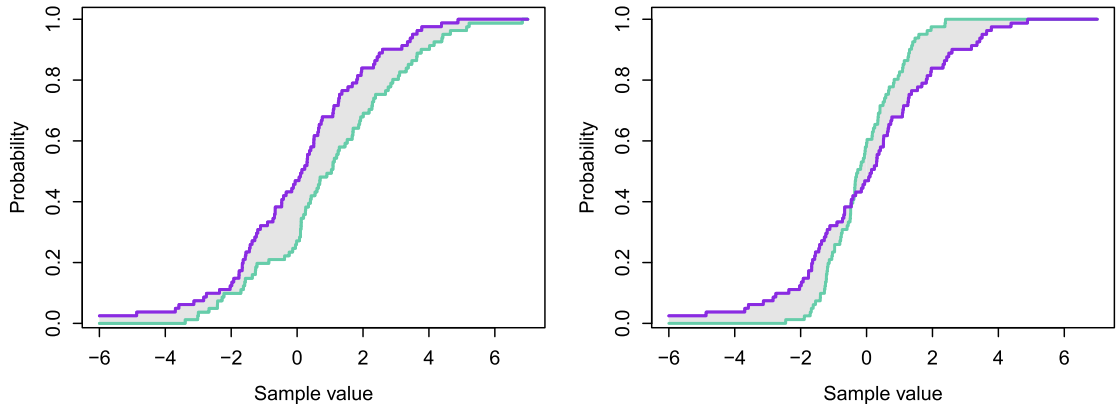

Figure 1. ECDFs of samples from two normal distributions with either different means but same spread (left), or same mean but different spreads (right). The IQD performance metric calculates the squared area (indicated in gray) between the two distributions, see the main text for more details.

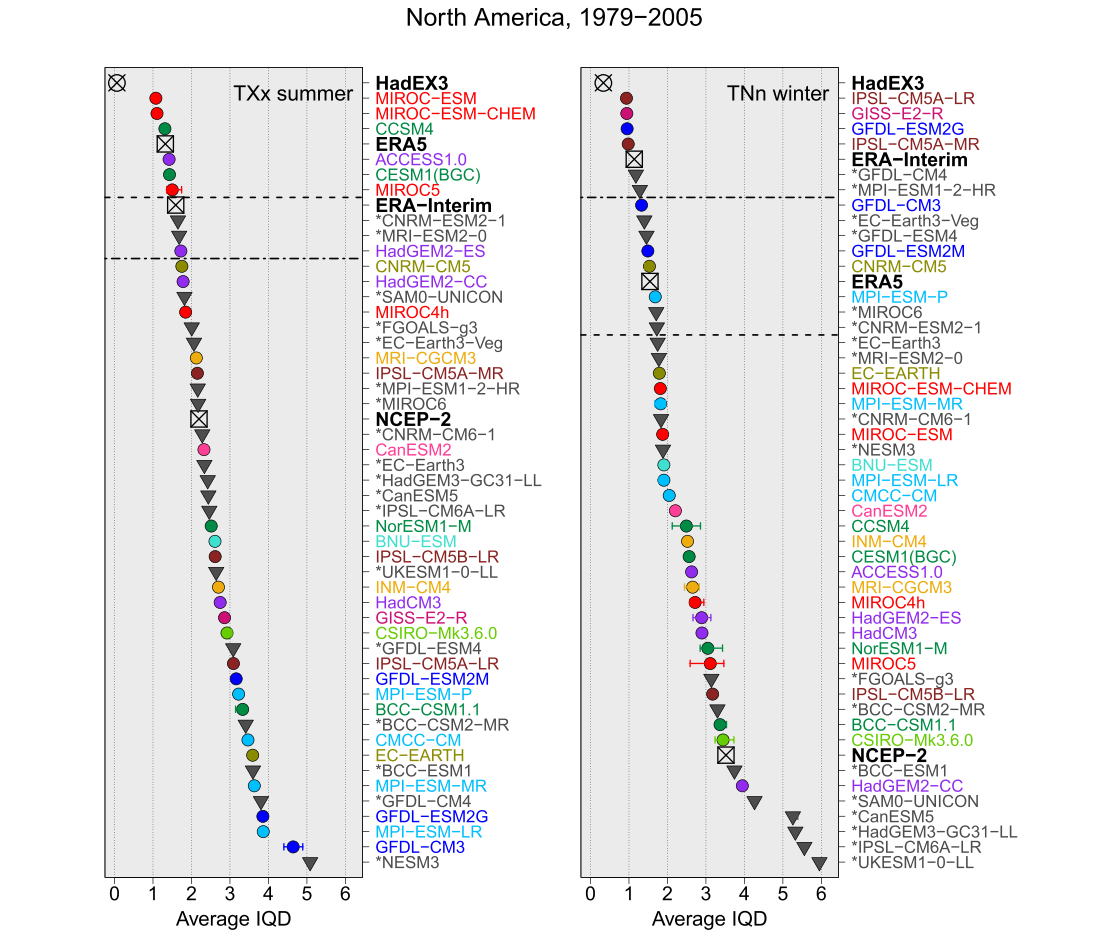

Figure 2. Average IQD over grid points in North America for an evaluation against HadEX2 for TXx distributions in summer (left) and TNn distributions in winter (right) from 1979-2005: Reanalyses (squares with (x), observation-based data sets (circles with (x), CMIP6 models (gray triangles) and CMIP5 models (filled circles). The models are ranked with the best performing model at the top. CMIP5 models are sorted in model families by color according to Knutti *et al* (2013), CMIP6 models are indicated with a star and reanalyses/data products in bold. If a model has multiple runs, the spread across the runs is indicated with a bar. Horizontal lines indicate the 5% significance level of testing equal performance to ERA5 (dashed) and ERA-Interim (two-dash).

ERA-Interim, while NCEP-2 and HadGEM2-CC have much larger areas where the performance is poor. ERA-Interim mainly diverges from HadEX2 along the coast, indicating that differences between the two data sets may be related to differences in model grids and land-sea masks. NCEP-2 additionally differs from HadEX2 in western regions with higher elevations and in the eastern part of Canada. For HadGEM2-CC,

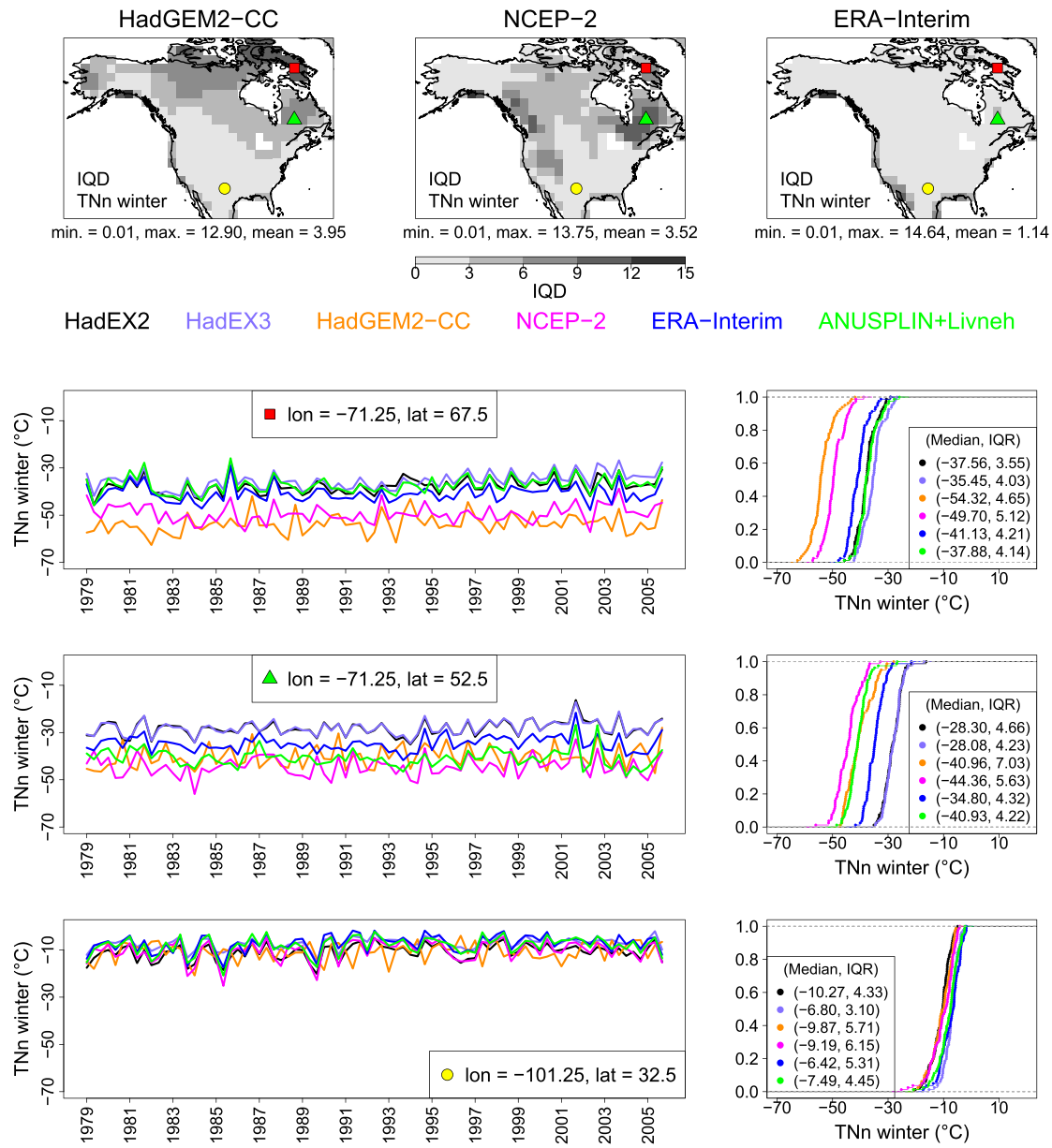

Figure 3. Top panel: IQD scores for evaluation against HadEX2 for winter TNn from 1979-2005 across grid locations in North America for the model/reanalysis simulations listed on top of each plot. Minimum (best), maximum (worst) and mean IQD values over the area are given below each plot. Panels two to four: Time series and distribution functions over the study period for six different data sets at three locations indicated on the maps in the top panel, with the median and the interquartile range (IQR) of each distribution. The name of each data set is given above panel two in the corresponding color used in the plots below.

substantial differences are largely concentrated in the higher latitudes.

Figure 3 also shows the winter TNn time series over the study period and the corresponding distributions at three locations: on the Canadian Baffin Island (red square) where HadGEM2-CC has a high IQD value, in the Canadian province of Quebec (green triangle) where NCEP-2 has a high IQD value, and in the US state of Texas (yellow circle) where all three simulations get a low IQD value. For comparison, we have also included the time series of the observation-based data sets ANUSPLIN+Livneh and HadEX3. In the grid point located in Texas, all the distributions are quite similar, with the TNn values minimally warmer for HadEX2 and HadEX3 than the other data sets. Similarly, in the other two locations, HadEX2 and HadEX3 also yield the warmest TNn values. At the Quebec location, even the observation-based data sets show significant differences, with the HadEX2 median roughly 13°C warmer than the ANUSPLIN+Livneh median. Furthermore, the distributions from the observation-based products have relatively small spread (as measured by the interquartile range) compared to the distributions from the other data sets.

Results for Europe, corresponding to those for North America shown in figure 2, are shown in figure 4. As for North America, we observe that ERA5 has slightly stronger similarities with HadEX2 than ERA-Interim for summer TXx, while the opposite

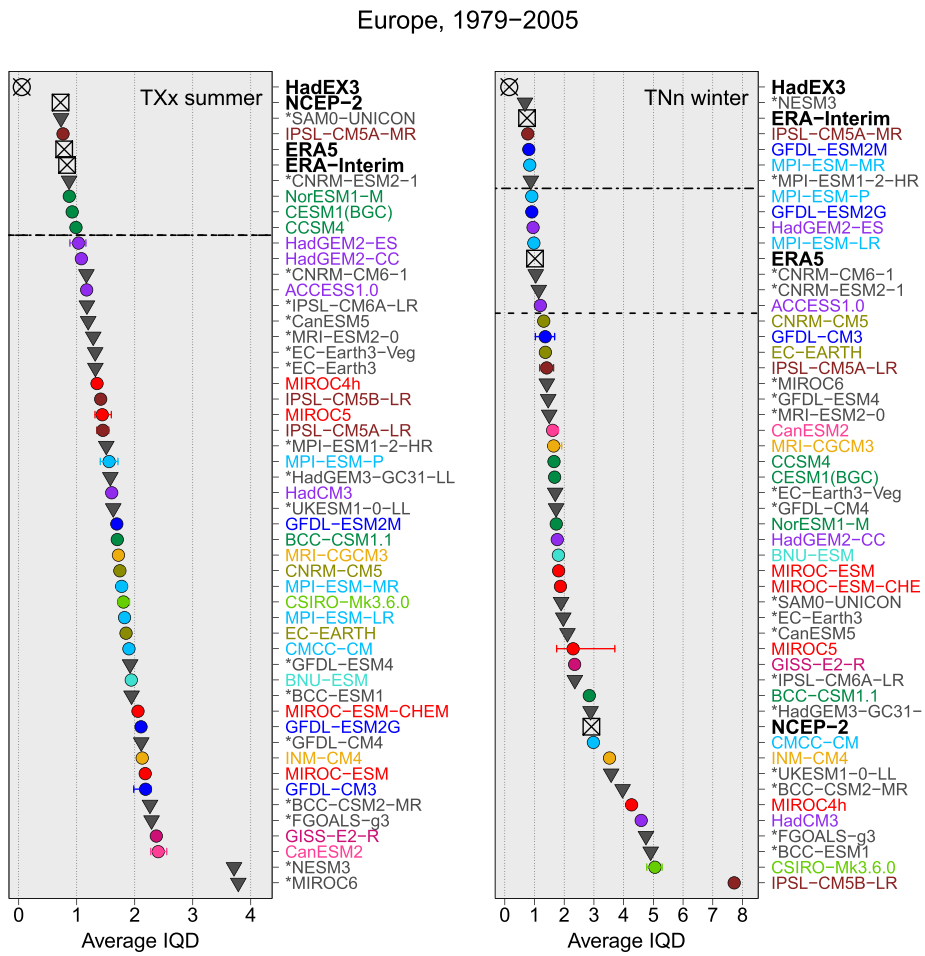

Figure 4. Average IQD over grid points in Europe for an evaluation against HadEX2 for TXx distributions in summer (left) and TNn distributions in winter (right) from 1979-2005: Reanalyses (squares with (x), observation-based data sets (circles with (x), CMIP6 models (gray triangles) and CMIP5 models (filled circles). The models are ranked with the best performing model at the top. CMIP5 models are sorted in model families by color according to Knutti *et al* (2013), CMIP6 models are indicated with a star and reanalyses/data products in bold. If a model has multiple runs, the spread across the runs is indicated with a bar. Horizontal lines indicate the 5% significance level of testing equal performance to ERA5 (dashed) and ERA-Interim (two-dash). In the figure on the left the lines are on top of each other.

holds for winter TNn. Excluding the highly-related HadEX3, the NCEP-2 reanalysis is the data set most similar to HadEX2 for summer TXx in Europe. However, it ranks 41 out of 51 for winter TNn, which is similar to its performance for winter TNn in North America. Here, six models perform competitively with ERA5 and ERA-Interim in the case of summer TXx and 12 models perform competitively with at least one of these reanalyses when winter TNn is considered. Two models, the CMIP6 model CNRM-ESM2-1 and the CMIP5 model IPSL-CM5A-MR perform well for both summer TXx and winter TNn. Notably, the CMIP6 model NESM3 ranks first for winter TNn and second last for summer TXx. NESM3's poor performance for summer TXx is due to it being too cold (overall about 6 °C colder than HadEX2), an effect that is also observed for North America, cf figure 2. The CMIP6 model MIROC6, however, produces too warm summer TXx in Europe (overall about 6°C warmer than HadEX2), an effect that is not seen in North America.

Figure 5 shows the spread of IQD scores for winter TNn over grid cells in Europe for the NCEP-2 and ERA-Interim reanalyses as well as the poor-performing CMIP5 model IPSL-CM5B-LR. ERA-Interim has a mean IQD of 0.75 with the largest IQD values appearing in coastal regions where ERA-Interim is slightly warmer than HadEX2. Overall, ERA-Interim is about 1°C warmer than HadEX2. While the NCEP-2 reanalysis is also slightly too warm in coastal zones, its values are too cold over most of the region and overall about 5°C too cold, resulting in a mean IQD of 2.92. The IPSL-CM5B-LR model, on the other hand, is too cold overall and particularly in the northern half of the region with an average negative bias of approximately 12°C and mean IQD of 7.72. More generally, a comparison of the spatial patterns in the top and bottom panels of figure 5 shows that while the spatial patterns are similar for each data set, they are not identical, emphasizing that the IQD evaluates both the center and the spread of the distributions.

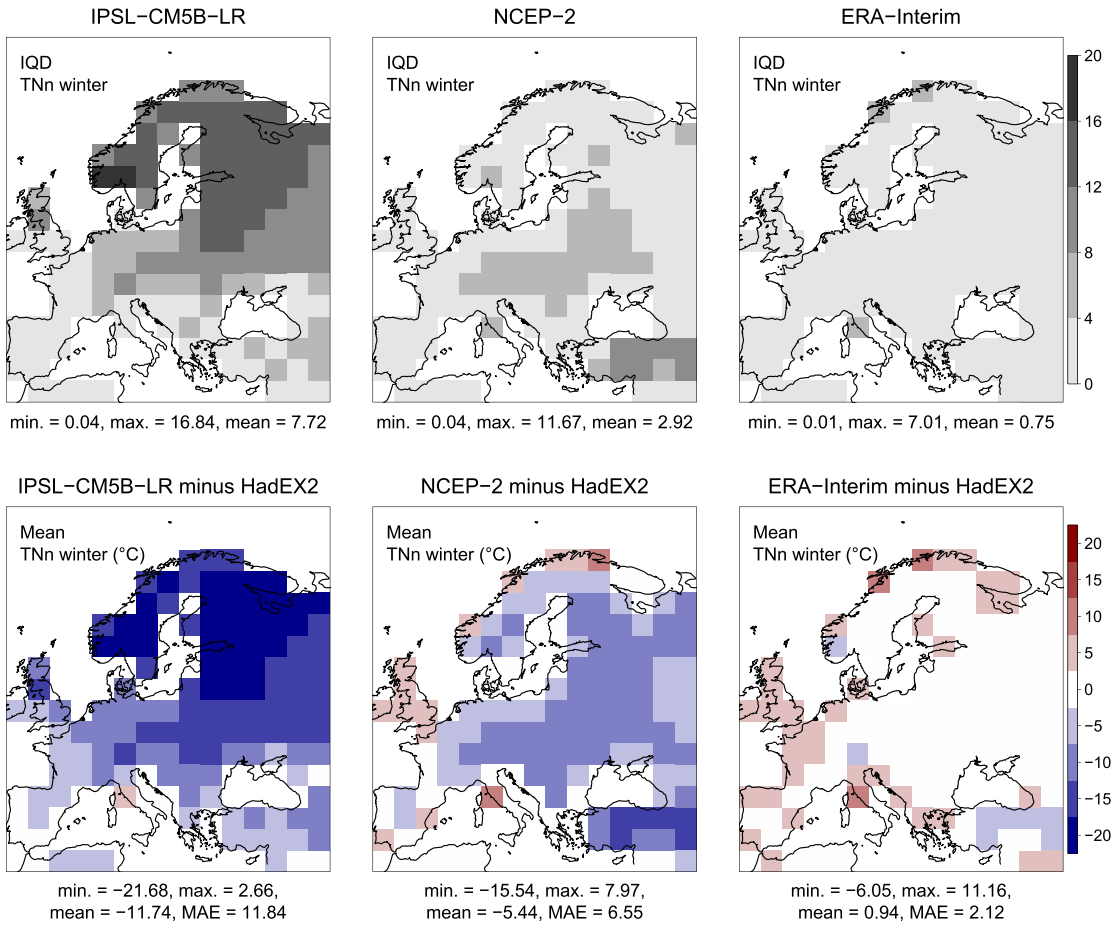

Figure 5. Top panel: IQD scores for TNn in winter across grid locations in Europe for three data sets, with minimum (best), maximum (worst) and mean IQD values over the area given below the plots. Bottom panel: Difference in average winter TNn (in °C) across grid locations in Europe for the three data sets compared against HadEX2. Minimum, maximum and mean differences across the region are given below the plots together with the mean absolute error (MAE).

Results for monthly ETR are given in section 2 of the supplementary material. Here, the same variable is assessed in both summer and winter, resulting in an overall more consistent ranking. Two models, the CMIP6 models EC-Earth3 and EC-Earth3-Veg, show performance comparable to at least one of ERA5 or ERA-Interim for both seasons in both regions.

4.2. Comparison with ERA5

In this section, we analyze the sensitivity of the results to the choice of reference dataset. Figure 6 shows the model rankings over North America when the data sets are compared against ERA5 instead of HadEX2 using the same grid resolution as in figure 2 (i.e. the HadEX2 grid). ERA5 compared against itself yields an IQD score of zero and for both variables, ERA-Interim is the data set most similar to ERA5. Only one model performs comparably to ERA-Interim for summer TXx and no model for winter TNn. Furthermore, there is a notable change in the ranking compared to figure 2. Here, there is a concentration of CMIP6 models obtaining either the lowest or the highest ranks. For Europe, there is similarly a concentration of CMIP6 models at the top of the rank list, see figure 7. Further, two models perform comparatively to ERA-Interim for summer TXx in Europe and three for winter TNn. Several climate models rank higher than HadEX3, especially for winter TNn where HadEX3 lands approximately in the middle of the pack.

Corresponding results for ETR are shown in section 3 of the supplementary material. Here, the CMIP6 models also rank somewhat better than when compared against HadEX2. The CMIP6 models EC-Earth3 and EC-Earth3-Veg perform comparably to ERA-Interim for summer in North America and both seasons in Europe. While these two models have a particularly high spatial resolution, see table 3 in the supplementary material, they are also highly related to the ERA reanalysis models. Additionally, the performance of the CMIP6 models GFDL-CM4 and GFDL-ESM4 is comparable to that of ERA-Interim for both seasons in Europe.

The results in figures 2 and 6 are based on the same underlying data and can thus be compared. The IQD scores are generally lower in figure 6, indicating that the model simulations are overall more similar to ERA5 than to HadEX2. For instance, for summer

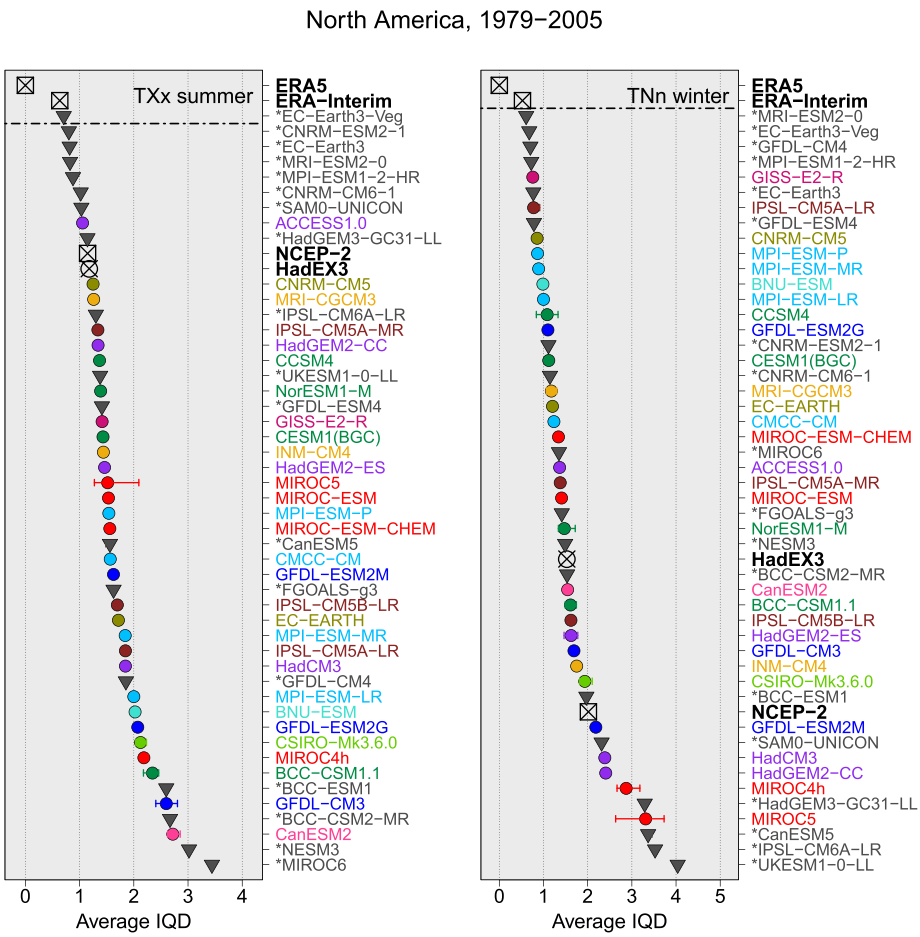

Figure 6. Average IQD over grid points in North America for an evaluation against ERA5 on the HadEX2 grid for TXx distributions in summer (left) and TNn distributions in winter (right) from 1979–2005: Reanalyses (squares with (x)), observation-based data sets (circles with (x)), CMIP6 models (gray triangles) and CMIP5 models (filled circles). The models are ranked with the best performing model at the top. CMIP5 models are sorted in model families by color according to Knutti *et al* (2013), CMIP6 models are indicated with a star and reanalyses/data products in bold. If a model has multiple runs, the spread across the runs is indicated with a bar. Two-dash horizontal lines indicate the 5% significance level of testing equal performance to ERA-Interim.

TXx, 15 climate models have an IQD score lower than 2 when compared against HadEX2, while this holds for 38 climate models when compared against ERA5. The European results in figures 4 and 7 are, however, more alike. Similar patterns are observed for the ETR, see the supplementary material.

Both ERA5 and the CMIP6 models exist on a finer grid than HadEX2. Section 4 of the supplementary material shows the model rankings when the CMIP6 model simulations are compared against ERA5 on the CMIP6 model grid. Comparing these results to those in figure 6 and 7, we see that the two evaluations yield very similar, albeit not identical, model rankings. Note that IQD scores cannot be directly compared across two grid resolutions. Distributions at different resolutions may not present the same physical processes, making it impossible to separate the confounding effects of intrinsic predictability and model performance (Gneiting and Raftery 2007).

5. Discussion and conclusions

A comprehensive evaluation of climate models requires performance measures that are simultaneously flexible and specific. We propose that climate model simulations should be evaluated by comparing distributions of model output to corresponding distributions of observational or reanalysis data products. Specifically, we propose to use the integrated quadratic distance (IQD) score, as it fulfills essential decision-theoretic properties for ranking competing models and testing equality in performance, while also assessing the full distribution. The IQD is here used to evaluate simulations of surface air temperature (SAT) extremes. However, its applicability extends to any univariate weather variable.

We evaluate seasonal distributions of SAT extremes, specifically monthly minimum and maximum SAT as well as monthly temperature range, for the time period 1979–2005 over North America

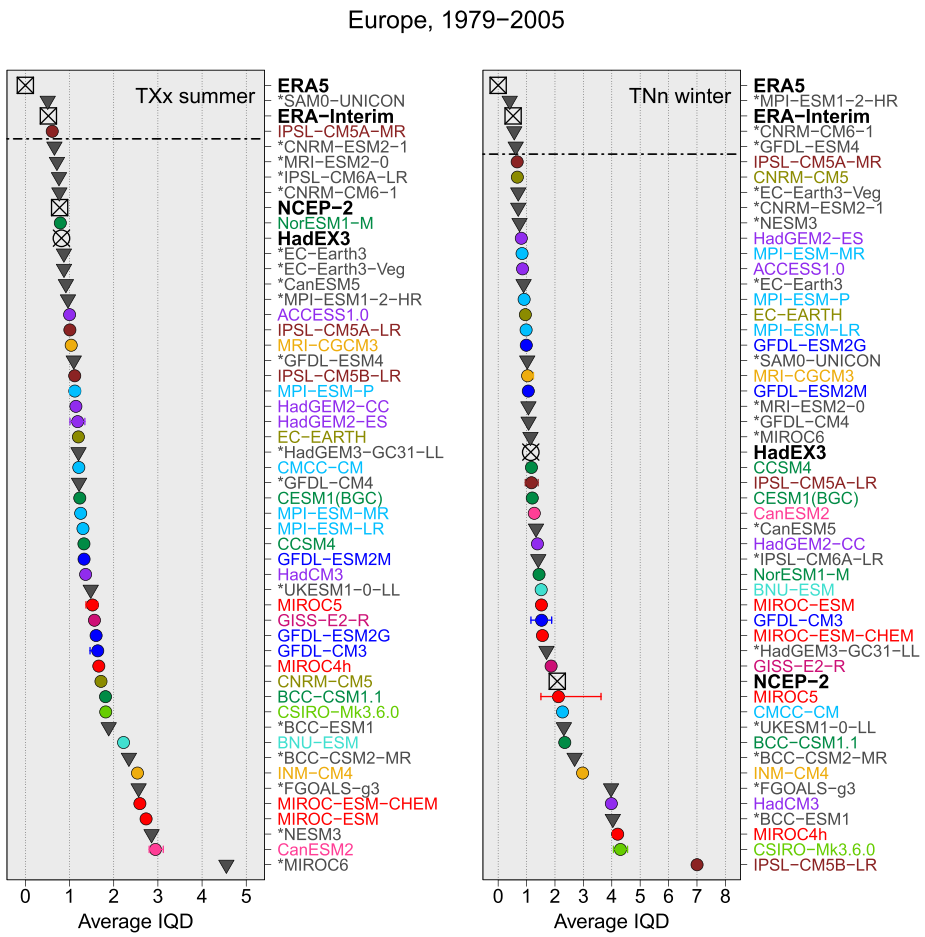

Figure 7. Average IQD over grid points in Europe for an evaluation against ERA5 on the HadEX2 grid for TXx distributions in summer (left) and TNn distributions in winter (right) from 1979-2005: Reanalyses (squares with (x)), observation-based data sets (circles with (x)), CMIP6 models (gray triangles) and CMIP5 models (filled circles). The models are ranked with the best performing model at the top. CMIP5 models are sorted in model families by color according to Knutti *et al* (2013), CMIP6 models are indicated with a star and reanalyses/data products in bold. If a model has multiple runs, the spread across the runs is indicated with a bar. Two-dash horizontal lines indicate the 5% significance level of testing equal performance to ERA-Interim.

and Europe. We compare climate model simulations from 48 different CMIP5 and CMIP6 models, three different reanalysis data sets and two observational data sets. For the CMIP5 models, the results are displayed by model families as defined by Knutti *et al* (2013). There is a general tendency for models that belong to the same family to show similar skill. However, the degree of similarity varies across variables and regions. Multiple runs are evaluated for 17 out of 30 CMIP5 models, using 2–5 runs per model as listed in table 2 of the supplementary material. In most cases, different runs from the same model yield nearly identical results. The most notable exception here is the CMIP5 model MIROC5 where the spread in skill is large, in particular in winter.

There is not a notable difference between the model generations CMIP5 and CMIP6 when the model simulations are compared against HadEX2. However, the CMIP6 models show a better agreement with ERA5 than CMIP5 models, with a few exceptions. Overall, the climate models show higher skill when compared against ERA5 than when compared against HadEX2. As HadEX is based on station observations while ERA5 is a gridded reanalysis product, it is to be expected that the extremes in HadEX may be more extreme than those in ERA5 (e.g. Donat *et al* (2014)). Comparisons in section 5 of the supplementary material show that this is indeed the case, except for the northern half of North America for winter TNn where, somewhat unexpectedly, the opposite is observed.

The models are evaluated against HadEX2 for four variables, or indices, and two regions (a 4×2 set of comparison); 23 out of 30 CMIP5 models and 12 out of 18 CMIP6 models show performance comparable to that of either ERA5 or ERA-Interim in at least one of those comparisons. However, the degree of agreement may vary substantially between variables and regions; the overall best performing model, the CMIP6 model CNRM-ESM2-1, is competitive with the reanalysis in five out of eight evaluations. This suggests that care should be exercised when extrapolating performance results in a specific setting to other, potentially unrelated, applications. When the datasets are compared against ERA5, no single model performs competitively with ERA-Interim

across both regions and seasons. An important factor here is that these two reanalysis products are highly related.

The IQD score is a general and easily implemented performance measure for comparing distributions of climate model simulations to corresponding distributions of observed data, as opposed to comparing point estimates with or without confidence intervals. As the comparison requires either the interpolation of station observations to a grid, or the use of gridded reanalysis products, the comparison should be performed with multiple truths, if possible.

Acknowledgments

The work of Thordis L Thorarinsdottir, Jana Sillmann and Marion Haugen was supported by the Research Council of Norway through project number 243953 "Physical and Statistical Analysis of Climate Extremes in Large Datasets" (ClimateXL). Jana Sillmann and Marit Sandstad are further supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 820655 (EXHAUSTION) and by the Belmont Forum Collaborative Research Action on Climate, Environment, and Health, supported by the Norwegian Research Council (Contract No. 310672) (HEATCOST). The data were shared on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway.

This study contains modified Copernicus Climate Change Service Information 2019. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus Information or Data it contains. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the involved climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

The source code for the performance analysis is implemented in the statistical programming language `R` (R Core Team 2019). The relevant functions are available on GitHub at http://github.com/NorskRegnesentral/IQD.

ORCID iD

Thordis L Thorarinsdottir
https://orcid.org/0000-0001-6702-0469

References

Armantier O and Treich N 2013 Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging *Eur. Econ. Rev.* **62** 17–40

Baker N C and Taylor P C 2016 A framework for evaluating climate model performance metrics *J. Clim.* **29** 1773–82

Bröcker J and Smith L A 2007 Scoring probabilistic forecasts: The importance of being proper *Weather Forecast.* **22** 382–8

Dee D P *et al* 2011 The ERA-Interim reanalysis: configuration and performance of the data assimilation system *Q. J. R. Meteorol. Soc.* **137** 553–97

Diebold F X and Mariano R S 1995 Comparing predictive accuracy *J. Bus. Econ. Stat.* **13** 253–63

Donat M G, Sillmann J, Wild S, Alexander L V, Lippmann T and Zwiers F W 2014 Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets *J. Clim.* **27** 5019–35

Donat M *et al* 2013 Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: the HadEX2 dataset *J. Geophys. Res.: Atmos.* **118** 2098–118

Dunn R J *et al* 2020 Development of an updated global land in situ-based data set of temperature and precipitation extremes: HadEX3 *J. Geophys. Res.: Atmos.* **125** e2019JD032263

Eum H-I, Dibike Y, Prowse T and Bonsal B 2014 Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the athabasca watershed, canada *Hydrol. Process.* **28** 4250–71

Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization *Geosci. Model Dev. (Online)* **9** 1937–58

Flato G *et al* 2014 Evaluation of climate models *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* eds T F Stocker *et al* (Cambridge: Cambridge University Press) pp 741–866

Gneiting T and Raftery A E 2007 Strictly proper scoring rules, prediction and estimation *J. Am. Stat. Assoc.* **102** 359–78

Good P 2013 *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses* (Berlin: Springer)

Guttorp P 2011 The role of statisticians in international science policy *Environmetrics* **22** 817–25

Hersbach H 2000 Decomposition of the continuous ranked probability score for ensemble prediction systems *Weather Forecasting* **15** 559–70

Hersbach H and Dee D 2016 ERA5 reanalysis is in production *ECMWF Newslett.* **147** 5–6

Hoegh-Guldberg O *et al* 2018 Intergovernmental Panel on Climate Change, chapter Impacts of 1.5°C Global Warming on Natural and Human Systems *Global Warming of 1.5°C. An Ipcc Special Report on the Impacts of Global Warming of 1.5°C Above pre-Industrial Levels and Related Global Greenhouse gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development and Efforts to Eradicate Poverty.*

IPCC 2013 *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: Cambridge University Press) p 1535

Kalnay E *et al* 1996 The NCEP/NCAR 40-year reanalysis project *Bull. Am. Meteorol. Soc.* **77** 437–70

Kanamitsu M, Ebisuzaki W, Woollen J, Yang S-K, Hnilo J, Fiorino M and Potter G 2002 NCEP-DOE AMIP-II Reanalysis (R-2) *Bull. Am. Meteorol. Soc.* **83** 1631–44

Knutti R, Masson D and Gettelman A 2013 Climate model genealogy: generation CMIP5 and how we got there *Geophys. Res. Lett.* **40** 1194–9

Knutti R, Stocker T, Joos F and Plattner G-K 2003 Probabilistic climate change projections using neural networks *Clim. Dyn.* **21** 257–72

Lerch S, Thorarinsdottir T L, Ravazzolo F and Gneiting T 2017 Forecaster's dilemma: extreme events and forecast evaluation *Stat. Sci.* **32** 106–27

Livneh B, Rosenberg E A, Lin C, Nijssen B, Mishra V, Andreadis K M, Maurer E P and Lettenmaier D P 2013 A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states: update and extensions *J. Clim.* **26** 9384–92

Lussana C, Saloranta T, Skaugen T, Magnusson J, Tveito O E and Andersen J 2018 senorge2 daily precipitation, an observational gridded dataset over norway from 1957 to the present day *Earth System Sci. Data* **10** 235

Maraun D *et al* 2017 Towards process-informed bias correction of climate change simulations *Nat. Clim. Change* **7** 764–73

McKenney D W *et al* 2011 Customized spatial climate models for north america *Bull. Am. Meteorol. Soc.* **92** 1611–22

Möller A, Lenkoski A and Thorarinsdottir T L 2013 Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas *Q. J. R. Meteorol. Soc.* **139** 982–91

Orlov A, Sillmann J, Aaheim A, Aunan K and de Bruin K 2019 Economic losses of heat-induced reductions in outdoor worker productivity: a case study of Europe *Econ. Disasters Clim. Change* **3** 191–211

Orskaug E, Scheel I, Frigessi A, Guttorp P, Haugen J, Tveito O and Haug O 2011 Evaluation of a dynamic downscaling of precipitation over the norwegian mainland *Tellus* A **63** 746–56

Palmer T 2012 Towards the probabilistic earth-system simulator: a vision for the future of climate and weather prediction *Q. J. R. Meteorol. Soc.* **138** 841–61

R Core Team 2019 *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing) www.R-project.org/

Sillmann J, Kharin V V, Zwiers F, Zhang X and Bronaugh D 2013 Climate extremes indices in the cmip5 multimodel ensemble: part 2. Future climate projections *J. Geophys. Res.: Atmospheres* **118** 2473–93

Sillmann J, Kharin V, Zhang X, Zwiers F and Bronaugh D 2013 Climate extremes indices in the cmip5 multimodel ensemble: part 1. Model evaluation in the present climate *J. Geophys. Res.: Atmos.* **118** 1716–33

Sippel S, Zscheischler J, Heimann M, Otto F E, Peters J and Mahecha M D 2015 Quantifying changes in climate variability and extremes: pitfalls and their overcoming *Geophys. Res. Lett.* **42** 9990–8

Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98

Tebaldi C, Hayhoe K, Arblaster J M and Meehl G A 2006 Going to the extremes. An intercomparison of model-simulated historical and future changes in extreme events *Clim. Change* **79** 185–211

Tebaldi C and Knutti R 2007 The use of the multi-model ensemble in probabilistic climate projections *Phil. Trans. R. Soc.* A **365** 2053–75

Thorarinsdottir T L, Gneiting T and Gissibl N 2013 Using proper divergence functions to evaluate climate models *SIAM/ASA J. Uncertain. Quant.* **1** 522–34

Thorarinsdottir T L and Schuhen N 2018 Verification: assessment of calibration and accuracy *Statistical Postprocessing of Ensemble Forecasts* (Amsterdam: Elsevier) pp 155–86

Von Storch H and Zwiers F W 2003 *Statistical Analysis in Climate Research* (Cambridge: Cambridge University Press)

Vrac M and Friederichs P 2015 Multivariate–intervariable, spatial and temporal–bias correction *J. Clim.* **28** 218–37

Whan K and Zwiers F 2017 The impact of enso and the nao on extreme winter precipitation in North America in observations and regional climate models *Clim. Dyn.* **48** 1401–11

Winkler R L and Murphy A H 1968 'Good' probability assessors *J. Appl. Meteorol.* **7** 751–8

Yuan Q, Thorarinsdottir T L, Beldring S, Wong W K, Huang S and Xu C-Y 2019 New approach for bias correction and stochastic downscaling of future projections for daily mean temperatures to a high-resolution grid *J. App. Meteorol. Climatol.* **58** 2617–32

Zhang X, Alexander L, Hegerl G C, Jones P, Tank A K, Peterson T C, Trewin B and Zwiers F W 2011 Indices for monitoring changes in extremes based on daily temperature and precipitation data *Wiley Interdiscip. Rev.: Climate Change* **2** 851–70