Title

Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 1, model evaluation

Authors

Wehner, Michael
Gleckler, Peter
Lee, Jiwoo

Publication Date

2020-12-01

DOI

10.1016/j.wace.2020.100283

Peer reviewed

Weather and Climate Extremes 30 (2020) 100283

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: http://www.elsevier.com/locate/wace

Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 1, model evaluation

Michael Wehner [a,*], Peter Gleckler [b], Jiwoo Lee [b]

[a] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[b] Lawrence Livermore National Laboratory, Livermore, CA, USA

ARTICLE INFO

Keywords:
Uncertainty in extreme precipitation
Sample size
Long-period return values
Generalized extreme value theory

ABSTRACT

Using a non-stationary Generalized Extreme Value statistical method, we calculate selected extreme daily temperature and precipitation indices and their 20 year return values from the CMIP5 and CMIP6 historically forced climate models. We evaluate model performance of these indices and their return values in replicating similar quantities calculated from gridded land based daily observations. We find that at their standard resolutions, there are no meaningful differences between the two generations of models in their quality of simulated extreme daily temperature and precipitation.

1. Introduction

Much attention has been paid over the last 30 years to incorporating additional processes relevant to the climate system and their changes into climate models. Significant resources have also been directed to improving the quality of model simulations as compared to available observations over the recent past (Flato et al., 2013). Model evaluation is enabled by the *historical* simulation protocols of the Coupled Model Intercomparison Project (CMIP), which is currently in its 6th incarnation (CMIP6; Eyring et al., 2016). The *historical* simulations typically span the mid-19th century to the recent past (2005 for CMIP5 and 2015 for CMIP6) and are forced by realistic greenhouse gas, sulphate aerosol, stratospheric ozone, volcanic aerosol and solar luminosity variations based on observations. This paper focuses on evaluating the average and long period return values of extreme daily precipitation and temperature produced by the fully coupled climate models in the CMIP5 and CMIP6 projects with freely varying atmosphere, ocean, land surface and sea ice submodels but not with the biogeochemistry submodels of Earth System Models. A companion paper utilizes the same extreme value statistical methods to compare projections of average and long period return values of extreme daily precipitation and temperature in the CMIP5 and CMIP6 models. While the bulk of the model intercomparison and evaluation literature focuses on mean quantities and dominant variability characteristics, extreme precipitation and temperature have been assessed and compared for the CMIP3 and CMIP5 model submissions (Sillmann et al., 2013). More recently, evaluation of extreme precipitation in the higher resolution CMIP6 models has also been performed (Bador et al., 2020). Many of these studies examined some or all of the indices of the Expert Team on Climate Change Detection Indices (ETCCDI).[1] While these 27 indices were devised to aid in climate change detection and attribution, they are convenient model evaluation variables as they were chosen based on available global observations (Zhang et al., 2011).

We focus on 20-year return values of daily extreme temperature and precipitation as such rare events can have much higher impacts on human and natural systems than the annual or seasonal averages of the ETCCDI indices. In a stationary climate, 20 year return values may be experienced only 3 or 4 times in a normal human lifetime and would likely be remembered anecdotally in tall tales told to grandchildren. However, in a warming climate, hot and wet extremes of a specified magnitude generally become more common and cold extremes less common (Kharin et al., 2013). As a consequence, a non-stationary 20-year return value at any given time is better thought of as an event that has a 5% chance of occurring during that particular year.

2. Data and methods

In retrospect, not all 27 indices are useful for global purposes. In particular, those based on hard thresholds may be extremely rare in some parts of the globe and very common in other parts. This illustrates

* Corresponding author.
E-mail address: mfwehner@lbl.gov (M. Wehner).
[1] http://etccdi.pacificclimate.org/.

https://doi.org/10.1016/j.wace.2020.100283
Received 28 December 2019; Received in revised form 9 June 2020; Accepted 8 September 2020
Available online 10 September 2020

the subjective nature of observed and simulated climate extreme value analyses and precludes a single definition of what constitutes "extreme". In this paper, we use 4 temperature and 1 precipitation indices of the ETCCDI set to evaluate and compare CMIP6 and CMIP5 model simulations under their respective "historical" forcing scenarios of the mid-19th to early 21st century periods.

We deliberately choose the ETCCDI indices that are "block" extrema. "Hot days" are represented by *TXx*, the annual maximum of the daily maximum temperature. "Warm nights" are represented by *TNx*, the annual maximum of the daily minimum temperature. "Cool days" are represented by *TXn*, the annual minimum of the daily maximum temperature. "Cold nights" are represented by *TNn*, the annual minimum of the daily minimum temperature. "Wet days" are represented by *Rx1day*, the annual seasonal maxima of daily total precipitation.

We evaluate these extreme variables and their return values by comparing CMIP5/6 model simulations to gridded observations on the observational products' grids. Annual maxima and minima of daily maximum and minimum temperature are obtained directly from the HadEx3 database at a resolution of 2.5^{o}x3.75° (Dunn, 2020). Annual maxima of daily precipitation are obtained by extraction from the REGEN (Rainfall Estimates on a Gridded Network) gridded daily precipitation at a resolution of 1^{o}x1^{o} (Contractor et al., 2019). While both gridded observational products have values only over land, REGEN fills in unobserved land regions by ordinary block Kriging, whereas HadEx3 assigns missing values over areas without quality station data. Both datasets use station data from a wide variety of sources. REGEN provides daily precipitation data from which the block extrema can be extracted directly. HadEx3 on the other hand, calculates the extrema at the stations first followed by the gridding process, as some of the raw station data cannot be made available by agreement with the station data sources. This difference in the order of operations can have a noticeable effect on the gridded extreme values (Donat et al., 2013; Gervais et al., 2014), as it is not guaranteed that all stations within a grid cell will have extrema on the same day and would generally result in higher gridded extreme values than if daily gridded quantities are constructed first (Risser et al., 2019). However, this effect is likely larger for precipitation than for temperature due to the more complex spatial structure associated with precipitation extreme events. Gridding procedures, choice of station data sources and quality control methods also contribute to uncertainty in gridded land based observational products. Remote sensing via satellite can offer a spatially complete dataset, but their relationship to ground based estimates of extreme precipitation is not close (Timmermans et al., 2019) probably due either to retrieval algorithms or temporal sampling limitations. For the CMIP5 and CMIP6 models, block extrema are extracted from the daily variables and masked when the model's native land area fraction is less than 0.5.

Twenty year return values are calculated using a nonstationary Generalized Extreme Value (GEV) distribution using $ln(CO_2)$ as a covariate in the location parameter. This choice of statistical model motivates the selection of the block extrema ETCCDI variables over those defined by percentile exceedances. While Peaks over Threshold (POT) extreme value methods may be a more efficient use of the limited extreme sample data in a stationary setting, non-stationary thresholds (Acero et al., 2010; Kyselý et al., 2010; Roth et al., 2012; Solari et al., 2017) are not as straightforward as covariate GEV methods (Coles, 2001; Katz, 2010). Uncertainty in both the average and the return values of extreme temperature and precipitation statistics is strongly dependent on available sample sizes (Wehner et al., 2020). For this study, we aimed to use all available model output from all the individual realizations to reduce uncertainty. This necessitates a non-stationary statistical approach. Of these, the Maximum Likelihood Estimate technique of fitting GEV parameters is the most well exercised in the literature (Easterling et al., 2016) Other choices are possible and equally or even more valid. However, as will be seen, model errors are substantially larger than uncertainties in the methodology. In all model cases, the sample sizes are much larger than 20 years, the return period of interest in this study. Observational sample sizes are smaller but more than twice this return period.

$ln(CO_2)$ is chosen as a physically based covariate as it has long been known to force global mean temperature changes (Arrhenius, 1896). While global or local temperature could also provide a useful and physically motivated non-stationary covariate, the annual average of $ln(CO_2)$ was chosen as it isolates most of the anthropogenic components to climate change without any significant internal variability. The CMIP5/6 experimental protocols specify atmospheric CO_2 as an external forcing agent for the *historical* simulations and hence it is the same for each model, simplifying the analysis. Furthermore, given this physical connection to the anthropogenic variations of extreme temperature and precipitation, usage of much longer portions of the available datasets than in previous GEV-based analyses can be justified (Risser et al., 2019; Risser and Wehner, 2017). The resulting GEV estimates of long period return values are a function of $ln(CO_2)$ and can be easily estimated for any given year by using the appropriate value for that year. In a certain sense, this choice of non-stationary covariate then can be interpreted of as a (weakly) non-linear time covariate. Again, other choices of covariate are possible and can serve different purposes. For instance, choice of the aforementioned global or local temperature would implicitly incorporate other natural and external forcings in addition to $ln(CO_2)$. However, the purpose of this paper is to quantify errors in extreme precipitation and temperature. Use of a temperature based covariate would incorporate model differences (and errors) into the statistical methodology and could tend to hide differences in the extreme indices. The choice of $ln(CO_2)$ as the covariate for this study is deliberately made so that the statistical models are the same across climate models. Arguably, in the companion paper about projections of extremes at selected global warming levels (Wehner, 2020), global mean temperature would be appropriate as climate sensitivity would then be removed.

In this study, we have introduced a linear covariate into the GEV location parameter only, fitting the scale and shape parameters as constants. Due to the uncertainties in estimating the shape parameters, most previous studies also keep it stationary (Cooley et al., 2007). However, in some locations, the quality of the fitted distribution may or may not be improved by a non-stationary scale parameter and/or a nonlinear covariate dependence in the location parameter. However, in the interest of simplicity and consistent with the relevant detection and attribution studies of the human influence on extreme temperature (Brown et al., 2008; S.-K. Seung-Ki Min et al., 2013; Zwiers et al., 2011) and precipitation (Min et al., 2011; Westra et al., 2012; Zhang et al., 2013), we have not added these additional testing requirements.

A more complete discussion of non-stationary covariate choices and their implications for isolating both anthropogenic and natural variations in extreme values can be found in (Risser et al., 2020). We have not performed goodness of fit analyses for each model. However, in our previous studies, this particular choice of non-stationary statistical model performs adequately. For a detailed discussion of some of the statistical issues concerning sample size see Appendix C of Risser et al. (2019).

Fitting GEV distribution parameters was done using a Maximum Likelihood Estimates (MLE) procedure in the climextRemes software package (Paciorek et al., 2018), a python and R library built upon the extRemes library (Gilleland and Katz, 2016, 2011) and available at https://cran.r-project.org/web/packages/climextRemes/index.html. The covariate appears linearly in the GEV location parameter as $\mu(t) = \mu_0 + \mu_1 ln(CO_2)$. Hence, there are four fitted parameters, the two components of the location parameter, μ_0,μ_1, the scale parameter σ and the shape parameter ξ. In order to reduce the statistical uncertainty in fitting GEV distributions and in calculating the averages of the selected extreme variables, the entire time series of all available complete realizations of the historical simulations were used. With some exceptions CMIP5 models were run from 1851 to 2005 and CMIP6 models were run from 1851 to 2015. The two observational products do not span such a large time interval. Daily precipitation from REGEN is available from 1950 to

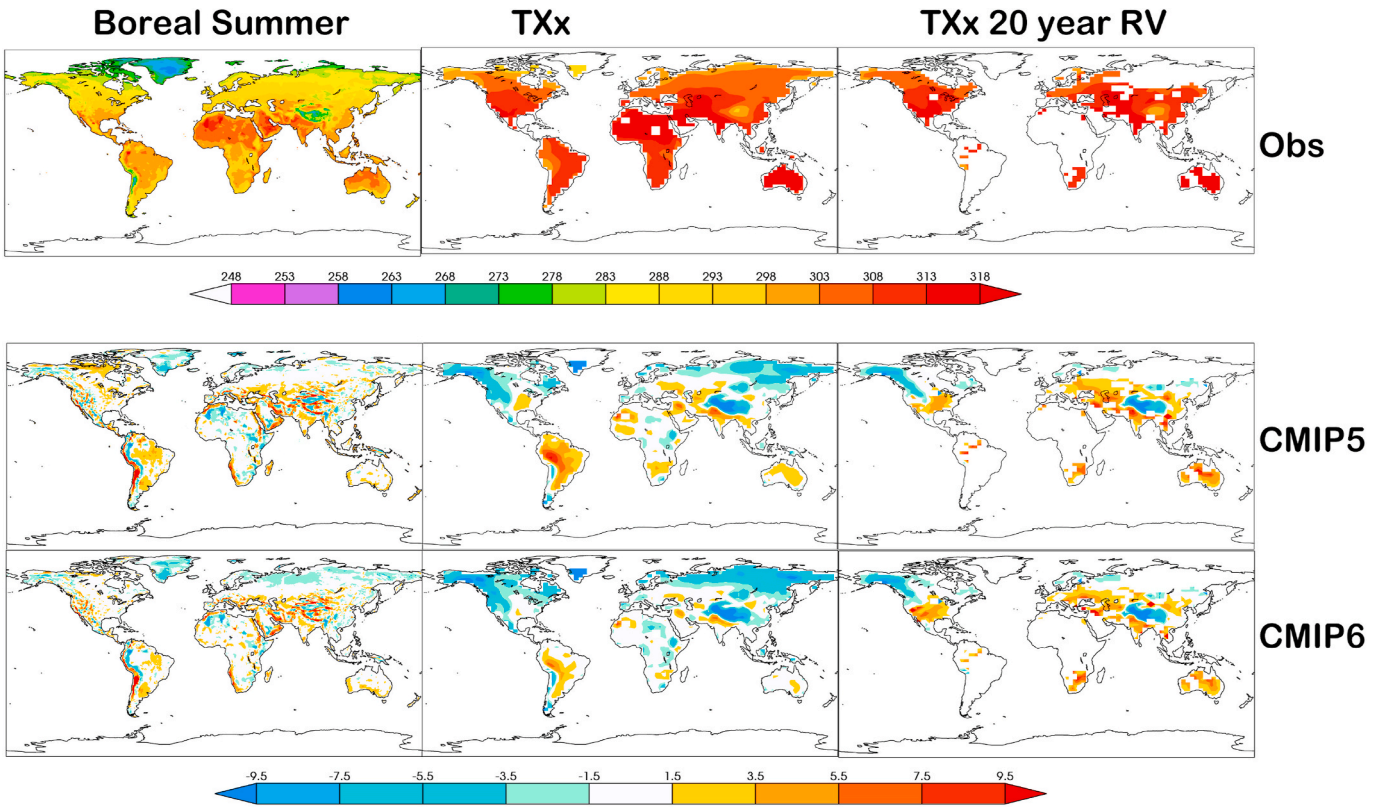

Fig. 1. Hot Days. Left column: Boreal summer average temperature. Middle column: *TXx*. Right column: Twenty year return value of *TXx*. Top row: 1961–2004 averages. Middle and bottom row: CMIP5 and CMIP6 multi-model average minus observations. Units: (°C).

2013. Although the HadEx3 temperatures extrema dataset runs from 1901 to 2018, our starting date for fitting the GEV parameters was chosen as 1960 to reduce the amount of missing data. To compare estimates of both the 7 block extrema and their twenty year return values between these observational products and the CMIP5/6, we choose to average all performance metrics over the longest common period between the observations and two historical CMIP ensembles, 1961–2004. It is important to note that while the data that goes into the GEV estimates of return values is not the same between observations and models, that additional data only serves to improve the fit of the GEV and reduce statistical uncertainty due to the length of the data records and to note that the comparison is over the same period. We also note that non-stationary return values calculated in this manner are temporally smoothed over both internal and externally forced variations and it would be appropriate to perform model evaluation at a midpoint of this period. We chose to average return values over this period simply for clarity in the comparing to the corresponding mean climatology and average extreme value indices.

Although changes in both temperature (Kim et al., 2015; Seung-Ki, Min et al., 2013) and precipitation (Min et al., 2011; Zhang et al., 2013) extremes have been attributed at the global scale and large regional scale (King et al., 2016; Wang et al., 2017; Zwiers et al., 2010) to anthropogenic changes to the composition of the atmosphere, we do not evaluate model performance of simulating trends due to the high natural variability at sub-continental scales (Kay et al., 2015).

As model errors in simulating extreme temperatures and precipitation rates may be affected by their errors in simulating the average temperature and precipitation climatology, we also present a brief comparison of mean state errors to extreme value errors. As the HadEx3 does not contain the underlying daily data nor an alternate method to calculate mean temperature, we use the GHCNCAMS gridded 2 m temperature dataset as a reference for observations of mean temperature (Fan and van den Dool, 2008). We note this and similar global products provide monthly averages of daily mean temperature, which is usually approximated as the average of the daily maximum and minimum temperatures rather than the separate averages of these individual daily extrema. However, annual average precipitation climatologies are calculated directly from the REGEN daily values providing a more direct comparison between mean and extreme precipitation errors.

For the multi-model averages, relationships between the error structures for each average extreme value and the appropriate mean climatology are discussed as well as the relationships between average extreme value errors and return value errors. Relationships between the error structures of the CMIP5 and CMIP6 multi-model averages are also discussed for each variable.

In Sections 3 and 4, we present performance metrics of average annual extreme values and 20-year return values in both tabular and graphic forms with the difference between observations and the multi-model average statistics shown as spatial maps. In addition to this multi-model analyses, results from individual models are being made publicly available as part of the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Simulation Summaries.[2] As new CMIP6 models are added to the CMIP database, results will be continually updated to this online resource where "performance portraits" (Gleckler et al., 2008) provide an interactive gateway to maps of individual models from which our evaluation statistics were derived. The analysis software used to produce our extreme value analysis are publicly available as part of the PCMDI Metrics Package (PMP; Gleckler et al., 2016).

We use Taylor Diagrams (Taylor, 2001) to provide statistical summaries for each model as well as the multi-model averages. Taylor diagrams use the pattern correlation between models and observational products as the angular dimension, and normalized

[2] https://pcmdi.llnl.gov/research/metrics.

Table 1a
Extreme temperature Taylor's modified skill scores. *TXx, TNn, TNx, TXn* and their 20-year return values for the CMIP6 models. Row labelled "cmip6" is calculated from the multi-model mean indices. Models are arranged from highest skill averaged over the 4 temperature return value indices to the lowest. Failure to converge in large regions are noted as "fail".

Model	TXx	TXx Return value	TNn	TNn Return value	TNx	TNx Return value	TXn	TXn Return value
cmip6	**0.70**	**0.47**	**0.97**	**0.89**	**0.70**	**0.50**	**0.98**	**0.92**
NorESM2-LM	0.68	0.51	0.97	0.90	0.70	0.55	0.98	0.91
NorESM2-MM	0.66	0.51	0.97	0.90	0.69	0.55	0.98	0.90
NESM3	0.69	0.51	0.98	0.90	0.62	0.52	0.98	0.91
GFDL-CM4	0.69	0.47	0.97	0.91	0.69	0.53	0.98	0.92
GFDL-ESM4	0.69	0.49	0.97	0.90	0.68	0.52	0.98	0.91
MPI-ESM-1-2-HAM	0.69	0.47	0.98	0.90	0.68	0.52	0.98	0.91
ACCESS-CM2	0.70	0.51	0.96	0.90	0.70	0.51	0.98	0.89
CNRM-CM6-1-HR	0.70	0.51	0.95	0.88	0.69	0.51	0.97	0.90
GISS-E2-1-G	0.64	0.50	0.97	0.88	0.65	0.54	0.98	0.88
SAM0-UNICON	0.67	0.49	0.96	0.87	0.68	0.55	0.98	0.88
EC-Earth3-Veg	0.71	0.50	0.97	0.89	0.71	0.50	0.98	0.90
TaiESM1	0.71	0.53	0.97	0.86	0.63	0.53	0.98	0.87
CNRM-CM6-1	0.69	0.50	0.95	0.88	0.68	0.51	0.98	0.90
FGOALS-g3	0.59	0.50	0.96	0.87	0.65	0.52	0.98	0.88
INM-CM4-8	0.67	0.48	0.96	0.90	0.71	0.51	0.98	0.89
ACCESS-ESM1-5	0.65	0.46	0.97	0.91	0.69	0.49	0.98	0.90
FGOALS-f3-L	0.63	0.50	0.96	0.86	0.71	0.53	0.98	0.88
MIROC-ES2L	0.70	0.47	0.97	0.90	0.76	0.49	0.98	0.91
AWI-CM-1-1-MR	0.65	0.45	0.98	0.89	0.64	0.50	0.98	0.92
HadGEM3-GC31-LL	0.69	0.51	0.95	0.86	0.67	0.51	0.98	0.88
HadGEM3-GC31-MM	0.70	0.49	0.96	0.87	0.68	0.50	0.98	0.89
BCC-ESM1	0.62	0.51	0.95	0.85	0.67	0.52	0.98	0.88
CNRM-ESM2-1	0.69	0.46	0.96	0.88	0.69	0.51	0.98	0.90
UKESM1-0-LL	0.68	0.50	0.95	0.85	0.66	0.51	0.98	0.88
BCC-CSM2-MR	0.60	0.46	0.95	0.86	0.66	0.53	0.98	0.88
MIROC6	0.57	0.44	0.97	0.85	0.71	0.51	0.98	0.88
MPI-ESM1-2-LR	0.68	0.43	0.98	0.87	0.68	0.47	0.98	0.91
MPI-ESM1-2-HR	0.67	0.45	0.98	0.86	0.65	0.45	0.98	0.91
INM-CM5-0	0.68	0.38	0.96	0.88	0.71	0.52	0.98	0.87
CanESM5	0.65	0.49	0.92	0.82	0.61	0.45	0.97	0.88
NorCPM1	0.66	0.42	0.96	0.85	0.67	0.45	0.98	0.86
IPSL-CM6A-LR	0.68	0.37	0.95	0.80	0.64	0.43	0.98	0.86
MRI-ESM2-0	0.64	fail	0.97	0.86	0.61	fail	0.98	0.89
EC-Earth3	0.71	0.45	0.97	0.78	0.70	fail	0.98	0.90
CESM2-WACCM	0.72	fail	0.98	0.88	0.66	0.52	0.98	0.90

(simulated/observed) spatial standard deviation as the radial dimension (Taylor, 2001). Exploiting a geometric relationship between the basic definitions of a centered root mean square error (RMSE), correlation and standard deviation, a byproduct of the Taylor Diagram is that the distance between a model result and the reference data on the abscissa is the centered RMSE. In the event that a model has identical spatial variance as the reference data, its data point will lie on a radial arc of unity. If a model is perfectly correlated with the reference data, its data point will lie on the abscissa. If both are true, the RMSE will also be zero and the model data will coincide with the reference data point on the abscissa. As a complement to the routine statistics shown on the Taylor diagram, Taylor's modified skill (Wehner, 2013) is presented in tables across models for each extreme variable and its return value. This measure scales the centered RMSE to reduce the possibility of models' skill being artificially inflated simply by data smoothing.

3. Temperature- hot days

In the mid and high latitudes, the hottest day of the year (*TXx*) generally occurs in the summer. The top row of Fig. 1 shows the 1961–2004 average of observed boreal summer temperatures from GHCNCAMS, (left), *TXx* (center)and the 20-year return value of *TXx* (right) from HadEx3 over land. In this and similar following figures, the middle row shows the difference (model minus observation) for the CMIP5 multi-model average while the bottom row shows the same for the CMIP6 multi-model average. Only models where a 20-year return value can be calculated over land are included (see Tables 1a and 1b). Models were generally excluded if they did not provide both daily data and a land/sea mask. All results use the same set of models for consistency of comparison. Missing data over land in the observed 20-year return values, is usually a result of a lack of convergence in the Maximum Likelihood Estimates (MLE) used in the calculation of the GEV coefficients and is shown as white in the top row of Fig. 1. In the HadEx3 data, this usually is seen in regions where the data record is incomplete and hence shorter. While all post-1960 data from the HadeX3 product was used, in the less well-observed areas of South America, Africa and Northern Asia, records may not span the entire 1960–2018 period. HadEx3 data was used as is, with no effort made to impose a minimum non-missing data length. However, the presence of severe outliers cannot be ruled out. Indeed, certain models exhibited a lack of convergence in isolated individual cells primarily located in hot and dry regions despite sample sizes exceeding 150 years. Two CMIP6 models failed to converge over most of the Northern Hemisphere. This is likely an indication of unrealistic model behavior but further investigation is needed.

While the global range of errors in Fig. 1 is similar for warm mean and extreme temperatures, no clear relationship exists between them with errors of opposite sign often the case. The centered pattern correlation (Mitchell et al., 2001) between average boreal summer temperature and *TXx* errors is small for both generations of models with values of 0.07 for the CMIP5 and 0.03 for the CMIP6 multi-model averages. *TXx* errors and its return value errors exhibit similar behavior in some parts of the world but are of opposite sign in the well observed parts of the North America and Europe causing the centered pattern correlation between them to be low with values of 0.03 for the CMIP5 and 0.05 for the CMIP6 multi-model averages.

The spatial pattern of average boreal summer temperature and *TXx* errors between the CMIP5 and CMIP6 multi-model averages are very

Table 1b
Extreme temperature Taylor's modified skill scores. *TXx, TNn, TNx, TXn* and their 20-year return values for the CMIP5 models. Row labelled "cmip5" is calculated from the multi-model mean indices. Models are arranged from highest skill averaged over the 4 temperature return value indices to the lowest. Models with inadequate data are marked "miss".

Model	TXx	TXx Return value	TNn	TNn Return value	TNx	TNx Return value	TXn	TXn Return value
cmip5	**0.68**	**0.50**	**0.97**	**0.89**	**0.69**	**0.48**	**0.98**	**0.91**
CMCC-CMS	0.64	0.50	0.97	0.89	0.65	0.54	0.98	0.91
GFDL-ESM2M	0.69	0.49	0.97	0.91	0.71	0.52	0.98	0.91
MIROC4h	0.62	0.50	0.96	0.91	0.68	0.52	0.97	0.91
CMCC-CM	0.63	0.51	0.97	0.89	0.64	0.53	0.98	0.90
MPI-ESM-P	0.66	0.49	0.98	0.90	0.65	0.53	0.98	0.91
ACCESS1-0	0.71	0.52	0.96	0.88	0.70	0.53	0.98	0.89
GFDL-ESM2G	0.69	0.47	0.97	0.91	0.71	0.54	0.98	0.91
IPSL-CM5A-MR	0.65	0.52	0.96	0.91	0.63	0.47	0.97	0.91
CESM1-BGC	0.64	0.53	0.96	0.88	0.66	0.50	0.98	0.89
MPI-ESM-LR	0.66	0.49	0.98	0.89	0.66	0.51	0.98	0.91
CMCC-CESM	0.61	0.49	0.97	0.88	0.67	0.54	0.98	0.88
inmcm4	0.66	0.51	0.92	0.89	0.68	0.51	0.97	0.88
MPI-ESM-MR	0.66	0.46	0.97	0.89	0.65	0.52	0.98	0.91
bcc-csm1-1-m	0.59	0.50	0.95	0.87	0.63	0.52	0.98	0.89
BNU-ESM	0.68	0.48	0.96	0.87	0.73	0.53	0.98	0.89
CCSM4	0.64	0.53	0.96	0.88	0.66	0.47	0.98	0.89
bcc-csm1-1	0.63	0.51	0.95	0.86	0.68	0.50	0.98	0.89
IPSL-CM5A-LR	0.65	0.46	0.96	0.90	0.63	0.50	0.97	0.89
MIROC-ESM-CHEM	0.54	0.45	0.96	0.90	0.70	0.49	0.97	0.91
CanESM2	0.61	0.50	0.94	0.89	0.60	0.48	0.98	0.87
MIROC5	0.65	0.49	0.97	0.86	0.70	0.51	0.98	0.88
NorESM1-M	0.68	0.48	0.96	0.88	0.69	0.49	0.98	0.89
IPSL-CM5B-LR	0.66	0.48	0.95	0.87	0.65	0.50	0.97	0.88
MIROC-ESM	0.54	0.44	0.96	0.90	0.70	0.47	0.97	0.91
CSIRO-Mk3-6-0	0.63	0.43	0.95	0.84	0.65	0.48	0.97	0.87
GISS-E2-R	miss	miss	0.97	0.89	0.63	0.54	miss	miss
GISS-E2-H	miss	miss	0.97	0.89	0.67	0.53	miss	miss
GFDL-CM3	miss	miss	0.97	0.91	0.70	0.51	miss	miss
ACCESS1-3	miss	miss	0.97	0.91	0.68	0.49	miss	miss
MRI-CGCM3	0.70	0.50	miss	miss	miss	miss	0.98	0.89
MRI-ESM1	miss	miss	0.96	0.88	0.71	0.51	miss	miss
CNRM-CM5	miss	miss	0.96	0.88	0.64	0.47	miss	miss
HadCM3	miss	miss	0.93	0.85	0.63	0.35	miss	miss
HadGEM2-ES	miss	miss	0.96	0.90	0.70	0.30	miss	miss

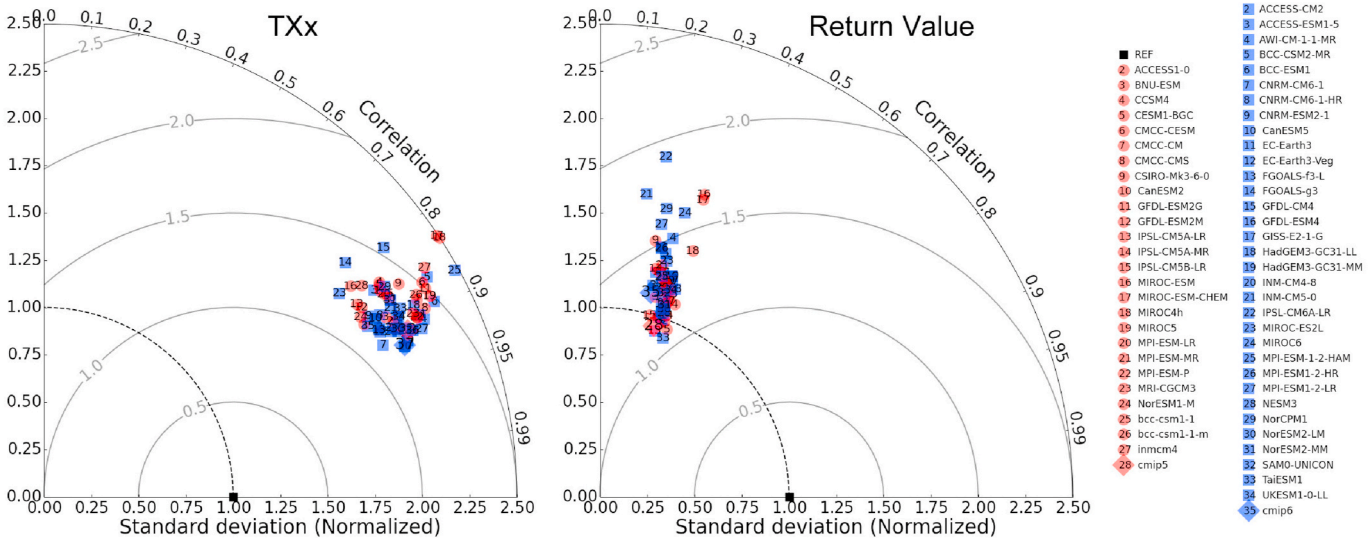

Fig. 2. Taylor diagram measuring model performance of simulating *TXx* (left) and its 20-year return value (right). The radial axis is normalized standard deviation while the angular axis is the pattern correlation. The reference data set is HadEX3 (black square). The concentric circles show the models' centered RMSE. CMIP5 models are shown in red. CMIP6 models are shown in blue. Multi-model averages are denoted as "cmip5" and "cmip6" in the legend. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

similar with centered pattern correlation values exceeding 0.95. However, despite some similarities in the CMIP5 and CMIP6 return value errors in Fig. 1, this metric of pattern similarity is very near zero, probably due to the much larger difference in local errors. Indeed, the root mean square difference between CMIP5 and CMIP6 multi-model average return value errors is 10 times larger than it is for *TXx* errors.

The Taylor plots of Fig. 2 show model performance in simulating *TXx* (left) and its return value (right) as measured by pattern correlation with

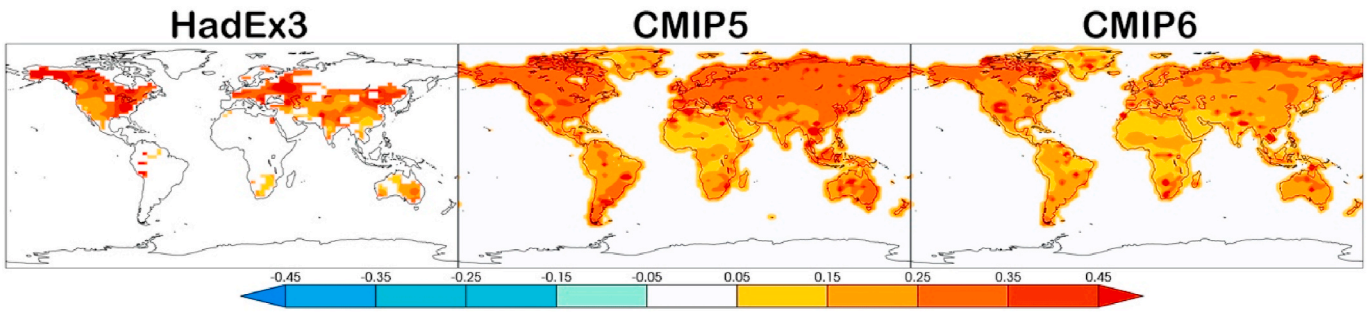

Fig. 3. Observed and multi-model average 1961–2004 standard error calculated from the delta method in fitted twenty year return value of *TXx*. Left: HadEx3 observations. Middle: CMIP5. Right: CMIP6. Units: ^{o}C.

HadEX3 and spatial standard deviation normalized by HadEX3. The centered RMSE is shown by the concentric circles surrounding the reference point. The 2nd and 3rd columns of Tables 1a and 1b shows the individual model performance as measured by the modified Taylor skill (Wehner, 2013) for hot days. In the tables, models are ordered within their generation from highest to lowest skill averaged over the 4 temperature return value indices. The multi-model average (denoted *cmip5* and *cmip6*) are constructed by first regridding each models' average *TXx* and return value to the 2.5^{o}x3.75° HadEX3 grid, performing the multi-model average and finally calculating the error statistics.

The CMIP5 and CMIP6 models are tightly clustered in the *TXx* Taylor diagram with the spatial standard deviation (radial distance from origin) of all models being more than double that of the reference data. Nevertheless, the pattern correlation is high, between 0.8 and 0.92 for all models, although some of the CMIP5 models are at the lower end of this range. Thus, while the pattern of the simulated spatial distribution in *TXx* corresponds reasonably well with HadEX3, its variation in amplitude is excessive. The two generations of models are even more tightly clustered in the *TXx* 20-year return value Taylor diagram. Although centered RMSE of the return value cluster is not very different, pattern correlation is very much degraded for the return value at between 0.2 and 0.4. As a result, Taylor skill, which ranges from 0.6 to 0.7 for *TXx* for most models is degraded for its return values to a range of 0.45–0.5. The substantial degradation in the return values pattern correlation is partly due to a small comparison region (due to the omission of poorly sampled regions of HadEx3) and to somewhat larger errors than *TXx*.

Taylor's modified *TXx* skill ranges from 0.6 to 0.7 for most models. Despite the low pattern correlation between the CMIP5 and CMIP6, *TXx* 20-year return values errors shown in Fig. 1, Taylor's modified *TXx* return value skill is very similar between model generations for this variable with most models close to 0.5. Overall, the performance of the CMIP5 and CMIP6 multi-model average in simulating *TXx* and its return value are very similar.

This lack of a relationship between the errors in simulated warm temperatures across rarities likely reflects errors in the different responsible physical mechanisms. It is known that land surface moisture feedbacks influence temperature extremes (Lorenz et al., 2016), as do the duration of blocking events (Zschenderlein et al., 2019). The relative importance of errors in these and other relevant heatwave mechanisms

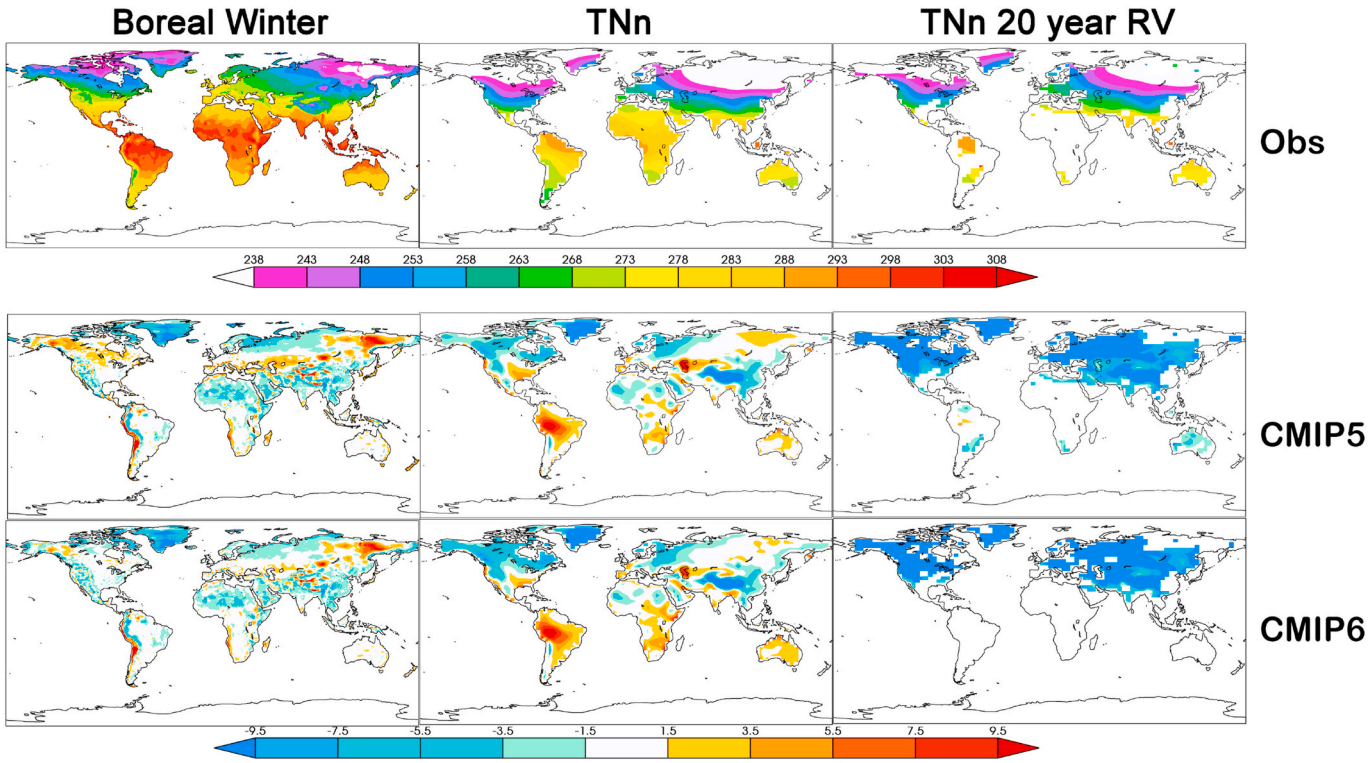

Fig. 4. Cold Nights. Left column: Boreal winter average temperature. Middle column: *TNn*. Right column: Twenty year return value of *TNn*. Top row: 1961–2004 averages. Middle and bottom row: CMIP5 and CMIP6 multi-model average minus observations. Units: ^{o}C.

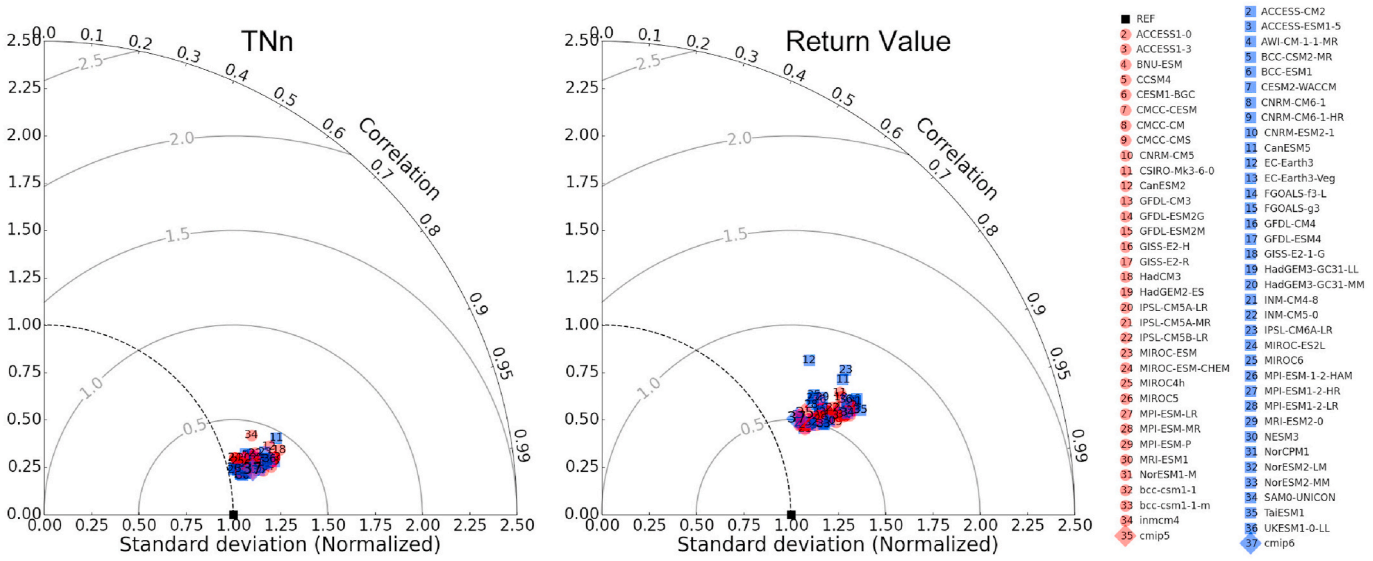

Fig. 5. Taylor diagram measuring model performance of simulating *TNn* (left) and its 20 year return value (right). The radial axis is normalized standard deviation while the angular axis is the centered pattern correlation. The reference data set is HadEX3 (black square). The concentric circles show the models' centered RMSE. CMIP5 models are shown in red. CMIP6 models are shown in blue. Multi-model averages are denoted as "cmip5" and "cmip6" in the legend. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

may vary across the models but the similarity in the error structures between the CMIP5 and CMIP6 in Figs. 1 and 2 suggest common causes.

Uncertainty in long period return value estimates from the limited data of the tail of the available sample can be considerable. In (Wehner et al., 2020), several methods of estimating this sampling uncertainty are compared. For large sample sizes, all methods considered produced similar uncertainty estimates. For small samples, the delta method, an asymptotic formula for the standard error of a function of maximum likelihood estimates (Coles, 2001) was found to be a more accurate estimation of the true uncertainty for simulated precipitation from a climate model of the CMIP5 generation as well as much computationally less intensive than bootstrapping approaches. The single realization of the observational datasets is a relatively small sample for calculating 20-year return values with periods of roughly 45–70 years. However, the non-stationary GEV approach permits using the much longer 150+ year simulation datasets. Furthermore, using the entire available ensemble increases the parent dataset size by several factors, especially for the more mature CMIP5 experiments. Fig. 3 shows this standard error of the 20-year *TXx* return value for the HadEX3 observations on the left averaged over 1961–2004. The middle and right maps show this standard error averaged over the CMIP5 and CMIP6 models respectively.

As a result of these large datasets, enabled by the non-stationary covariate, the uncertainty in the 20-year *TXx* return value estimates due to GEV fit uncertainty is much smaller than the model errors discussed above. A larger source of uncertainty in quantifying model error stems from observational uncertainties. However, a complete discussion of differences between observational products, either for temperature or precipitation, is outside the scope of this study but is an active area of research.

4. Temperature - cold nights

In the mid and high latitudes, the coldest day of the year (*TNn*) generally occurs in the winter. The top row of Fig. 4 shows the 1961–2004 average of observed boreal winter temperatures from GHCNCAMS, (left), *TNn* (center) and the 20 year return value of *TNn* (right) from HadEx3 over land while the middle and lower rows show the multi-model average errors. Note that the color scale on observations is 10C lower than Fig. 1 but remains the same for the mean model errors. The area of missing data, again reflecting a lack of MLE convergence from the HadEX3 data is significantly reduced from that of hot days. This is most easily visualized by comparing the white areas over land in the return value error maps of Figs. 1 and 3 as there is no comparison to be made in missing data regions. Lack of convergence was not an issue for any of the models.

In many of the well observed areas, simulated average *TNn* is colder than observed. Simulated 20-year *TNn* return values are everywhere too cold with much larger errors for both model generations than in average *TNn*. Seasonal mean cold temperature errors are weakly related to average extreme cold night temperature errors with pattern correlation values of 0.25 for CMIP5 and CMIP6. The pattern correlation between average *TNn* errors and return value errors is also low with values of about 0.4 for both generations of models.

As for hot seasons and *TXx*, the spatial pattern of errors for average boreal winter temperature and *TNn* errors are again very similar between CMIP5 and CMIP6 with pattern correlation values between them exceeding 0.95. *TNn* return value errors are weakly related between CMIP5 and CMIP6 with a pattern correlation value of 0.4. *TNn* return value uncertainty (not shown) is much smaller than the errors shown in Fig. 4.

The Taylor diagrams in Fig. 5 show that model performance in simulating *TNn* (left) and its return value (right) is very tightly clustered. The 4th and 5th columns of Table 2 shows that the modified Taylor skill for *TNn* exceeds 0.95 and 0.85 for its return value for nearly every model. *TNn* centered RMSE is lower than 0.5 for all models and 0.75 for its return value for most models despite the large biases of Fig. 4. Normalized standard deviation is greater than one for each model for both measures of cold nights but not higher than 1.25 for *TNn* and 1.5 for its return value. Unlike the case for hot days (Fig. 2), the simulated patterns of 20-year return values for cold nights are highly correlated with the reference data. As for hot days, there is little difference between CMIP5 and CMIP6 multi-model average cold night performance metrics.

Mid-latitude winter cold snaps typically occur on clear nights favoring outgoing longwave radiative cooling. As the underlying radiative physics is well understood, the overly cool northern hemisphere *TNN* return values in Fig. 4 may be the result of errors in the frequency and duration of winter blocking which are known to be an important process (Sillmann et al., 2011). It is also been shown that block statistics can be influenced by horizontal resolution (Schiemann et al., 2017) which is essentially unchanged between CMIP5 and CMIP6. While a systematic

Table 2
Extreme precipitation Taylor's modified skill scores relative to the REGEN gridded observations. Annual *Rx1day* and 20 year return values for the CMIP6 (left group) and CMIP5 (right group) models. Top rows labelled "cmip6" and "cmip5" are calculated from the multi-model mean and are shown in bold font. Within each group, models are arranged from highest averaged annual return value skill to the lowest.

	Rx1day	return value		Rx1day	return value
cmip6	**0.87**	**0.77**	**cmip5**	**0.84**	**0.73**
CNRM-CM6-1-HR	0.84	0.80	CMCC-CM	0.82	0.78
CESM2	0.83	0.79	CESM1-FASTCHEM	0.80	0.78
CESM2-WACCM	0.82	0.77	CCSM4	0.80	0.78
CNRM-CM6-1	0.81	0.69	CESM1-BGC	0.80	0.77
CNRM-ESM2-1	0.81	0.70	HadGEM2-ES	0.80	0.65
NorESM2-MM	0.81	0.70	ACCESS1-3	0.79	0.69
ACCESS-ESM1-5	0.79	0.69	ACCESS1-0	0.79	0.64
MRI-ESM2-0	0.79	0.69	CNRM-CM5	0.79	0.77
TaiESM1	0.78	0.76	CMCC-CMS	0.77	0.52
GFDL-CM4	0.78	0.78	MPI-ESM-MR	0.76	0.49
UKESM1-0-LL	0.77	0.69	MPI-ESM-P	0.74	0.47
HadGEM3-GC31-LL	0.77	0.68	MPI-ESM-LR	0.74	0.47
GFDL-ESM4	0.76	0.77	MIROC4h	0.73	0.71
ACCESS-CM2	0.76	0.69	MIROC5	0.70	0.69
NorESM2-LM	0.76	0.55	MRI-CGCM3	0.70	0.72
HadGEM3-GC31-MM	0.75	0.60	IPSL-CM5A-MR	0.69	0.63
FGOALS-f3-L	0.75	0.77	GFDL-CM3	0.67	0.64
SAM0-UNICON	0.74	0.69	bcc-csm1-1	0.67	0.75
MIROC6	0.74	0.72	NorESM1-M	0.66	0.51
CESM2-FV2	0.73	0.59	IPSL-CM5A-LR	0.64	0.52
INM-CM5-0	0.73	0.70	GFDL-ESM2M	0.62	0.60
EC-Earth3	0.73	0.64	IPSL-CM5B-LR	0.61	0.46
EC-Earth3-Veg	0.73	0.64	bcc-csm1-1-m	0.60	0.68
IPSL-CM6A-LR	0.72	0.72	CMCC-CESM	0.54	0.29
CanESM5	0.71	0.72	MIROC-ESM-CHEM	0.44	0.36
INM-CM4-8	0.71	0.60	MIROC-ESM	0.44	0.35
BCC-ESM1	0.70	0.73	inmcm4	0.38	0.20
FGOALS-g3	0.69	0.75			
NESM3	0.68	0.42			
NorCPM1	0.67	0.55			
MPI-ESM1-2-HR	0.66	0.40			
MPI-ESM1-2-LR	0.61	0.33			
MPI-ESM-1-2-HAM	0.60	0.26			
MIROC-ES2L	0.55	0.44			
BCC-CSM2-MR	0.49	0.54			

exploration of blocking statistics errors is outside the scope of this study, there is evidence of improvement in the CMIP6 models (Schiemann et al., 2020). The relationship of this process to errors in both hot and cold temperature extremes is worthy of further analyses.

Errors and Taylor diagrams for simulated warm nights (*TNx*) and cool days (*TXn*) are shown in the Appendix. Modified Taylor skill is shown in columns 6–9 of Tables 1a and 1b for these extreme temperature metrics.

5. Daily precipitation

The top row of Fig. 6 shows the 1961–2004 average of observed mean annual precipitation, (left), average *Rx1day* and the 20-year return value of *Rx1day* from the REGEN database. Model errors in the middle and bottom rows are shown as percent errors. This choice tends to emphasize dry regions, whereas portraying absolute errors would highlight wet areas. Lack of convergence in return calculations was minimal and confined to isolated cells, often in dry regions.

Errors in simulated annual mean precipitation are more closely related to errors in extreme daily precipitation than any of the temperature extreme indices. The pattern correlation between annual mean percent error and annual *Rx1day* percent error is 0.88 for the CMIP5 and CMIP6 multi-model averages. Furthermore, annual *Rx1day* errors are closely related to return value errors with pattern correlations between them of 0.81 for the CMIP5 and 0.83 for the CMIP6 multi-model averages.

The relationship between model generations is also much closer for extreme precipitation errors than it is for extreme temperature errors. The pattern correlation between the two multi-model averages is 0.99 for *RX1day* errors and 0.88 for *RX1day* return value errors. It is also high for annual mean precipitation at 0.98.

Fig. 7 shows the 1961–2004 averaged standard error of the 20-year annual *Rx1day* return value for the REGEN observations calculated from the delta method. The standard error is normalized by the 1961–2004 average return value and is shown as a percentage. The middle and right maps show this standard error averaged over the CMIP5 and CMIP6 models respectively. This uncertainty is generally much less than the errors shown in Fig. 6.

Taylor diagrams for *Rx1day* and its return value are shown in Fig. 8. Table 2 shows Taylor's modified skill for average annual *Rx1day* and its return value. Models are ordered within their generation from return value skill to lowest. Both generations of models exhibit wider ranges of Taylor skill than they do for extreme temperatures and are not as tightly clustered in the Taylor diagrams. Centered RMSEs of annual *Rx1day* and its return value are between 0.5 and 1.0 for most models but a few CMIP5 models and a single CMIP6 models are outliers with centered RMSE values exceeding unity. However, extreme daily precipitation Taylor skill is generally not as degraded from the average annual extreme to the return value as it is for extreme daily temperatures and for some models is improved. Note the spread in the Rx1day Taylor diagram results is largely due to large inter-model differences in the amplitude of the spatial pattern, with some models substantially underestimating the pattern amplitude while for others it is overestimated. Unlike temperature extremes, both pattern correlation and Taylor skill for the multi-model averages are better than any single model for both Rx1day and its return value for both generations of climate models. Overall model performance in simulating wet days is similar across both generations of models from the best models to the poor performers.

While the main purpose of this paper is to compare the relative performance of the CMIP5 and CMIP6 generations of climate models, the actual performance of the models is highly dependent on the reference data set evaluated against. Fig. 9 shows the multi-model errors in average annual *Rx1day* as measured by REGEN, HadEx3 and a reanalysis product, ERA5 (Hersbach et al., 2020). While *Rx1day* from the REGEN and ERA5 products are constructed in the same manner as in the models (i.e. constructed as annual maximum of gridded daily precipitation totals), the HadEx3 product is not. Due to station data availability limitations, HadEx3 is constructed by calculating the annual maxima at the individual stations followed by the gridding procedure. Chen and Knutson (2008) and Gervais et al. (2014) argue that such products are inappropriate for model evaluation. Rather, products such as REGEN (e.g. gridded daily station precipitation totals) more closely resemble the conserved quantities that models actually produce. Indeed, Gervais et al. (2014) demonstrated that extreme precipitation indices constructed by gridding station indices produce larger estimates than from gridded station daily totals. This then explains why the models in Fig. 9 are assessed to be consistently drier when compared to HadEx3 than to REGEN. We note that this is particularly apparent in North America and Western Europe, even though the raw data entering into the two products are based on very similar dense station networks. While HadEx3 assigns poorly observed land areas as "missing data", REGEN fills in these regions with a statistical algorithm (Contractor et al., 2019).

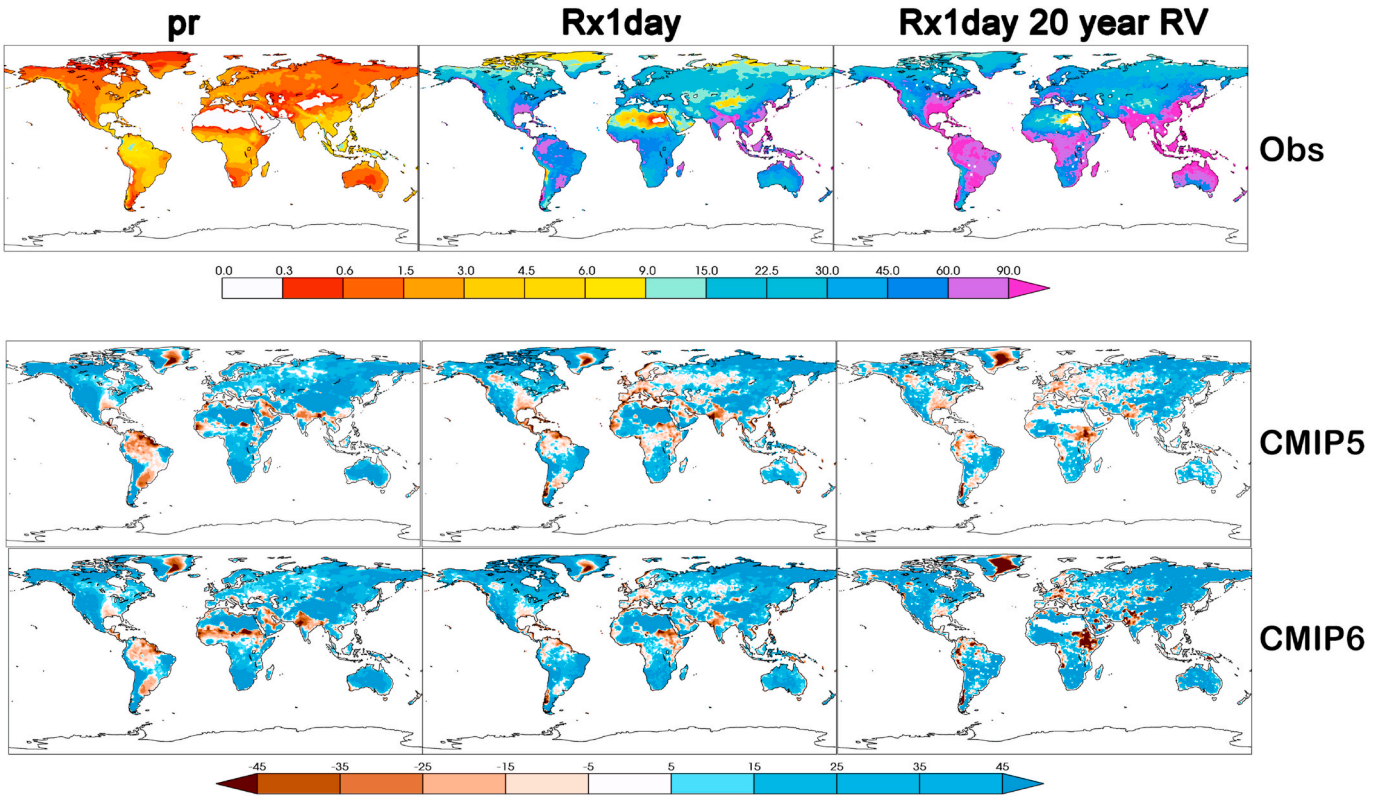

Fig. 6. Wet days. Left column: Annual average precipitation. Middle column: Annual average *Rx1day*. Right column: Twenty year return value of annual *Rx1day*. Top row: 1961–2004 averages. Units: mm/day. Middle and bottom row: CMIP5 and CMIP6 multi-model average minus observations. Units: percent.

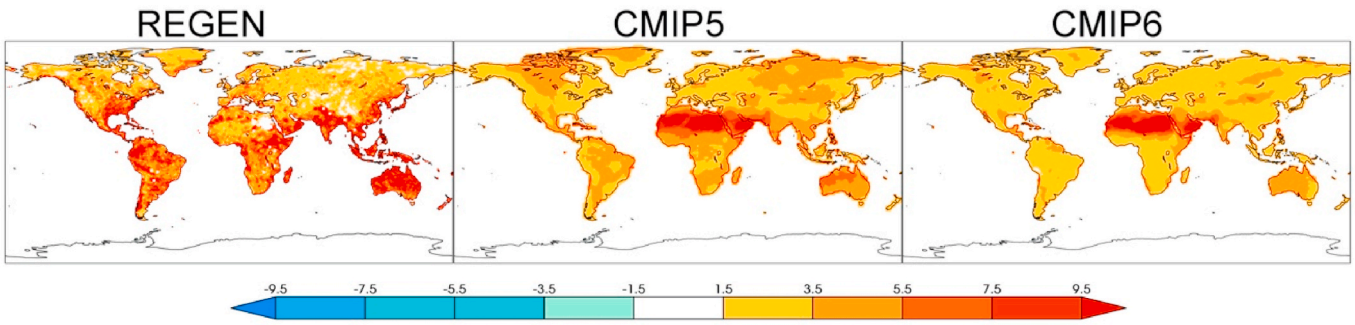

Fig. 7. 1961–2004 average standard error calculated from the delta method in fitted twenty year return value of annual *Rx1day*. Left: CMIP5 multi-model average. Right: CMIP6 multi-model average. Units = %.

Reanalysis products have also been used in the evaluation of simulated temperature and precipitation with the added benefit of coverage over the oceans (Sillmann et al., 2013). Products such as ERA5 do not directly assimilate precipitation, although some such as the North American Regional Reanalysis (NARR) do (Mesinger et al., 2006). However, they do assimilate moisture and its transport, which are clearly important as noted in the discussion of Fig. 6. However, reanalyses are model products, however highly constrained, and exhibit their own parameterization errors. Hence, there are both similarities and differences in the model errors when compared to REGEN and ERA in Fig. 9, illustrating part of the effects of observational uncertainties in evaluating model performance.

6. Discussion

We have quantified simulated errors in extreme temperature and precipitation in the two most recent generations of climate models, CMIP5 and CMIP6. We analyzed the annual average and 20 year return value of 4 extreme daily temperature indices, *TXx* (hot days), *TNn* (cold nights), *TNx* (warm nights), *TXn* (cool days) and an extreme annual daily precipitation index, wet days (*Rx1day*). The annual average indices, also interpretable as the 1 year return values, are the input into a non-stationary Generalized Extreme Value (GEV) statistical model to produce estimates of the much rarer 20 year return values. The non-stationarity of the GEV model permits usage of longer input data sets than in previous quasi-stationary approaches, yielding uncertainties in long period return values that are much smaller than climate model errors themselves. Model performance in simulating the rarer extremes is generally substantially degraded from the simulation of less rare extremes.

For the temperature indices, the pattern of winter extremes (*TNn, TXn*) is subtantially better than the pattern of summer extremes (*TXx, TNx*) but the magnitudes of the errors are larger. Extreme temperature errors bear little resemblance to seasonal mean temperature errors for

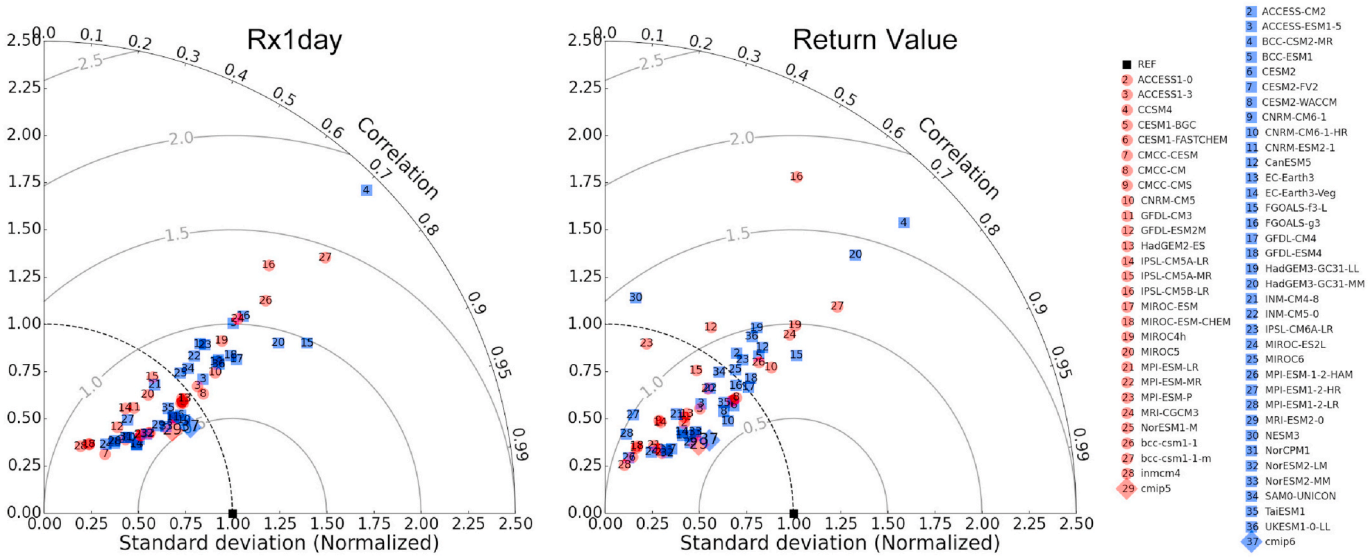

Fig. 8. Taylor diagram measuring model performance of simulating annual *Rx1day* (left) and its 20 year return value (right). The radial axis is normalized standard deviation while the angular axis is the centered pattern correlation. The reference data set is REGEN (black square). The concentric circles show the models' centered RMSE. CMIP5 models are shown in red. CMIP6 models are shown in blue. Multi-model averages are denoted as "cmip5" and "cmip6" in the legend. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

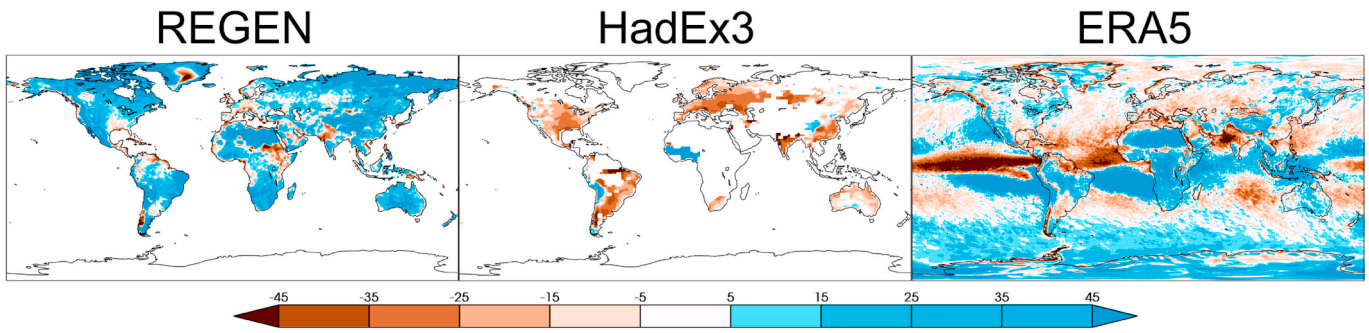

Fig. 9. CMIP6 annual *Rx1day* errors averaged over 1961–2004 as calculated from three reference data sets on the reference grid. Left: REGEN. Center: HadEx3. Right: ERA5.

either the CMIP5 or the CMIP6 ensemble. Furthermore, only a weak relationship between model errors in simulated annual temperature extreme metrics and model errors in corresponding 20-year return values is found. This implies that the causes of model mean temperature errors are different from some of the causes of model extreme temperature error. It also implies that mean state model errors do not influence the distribution of simulated extreme temperatures uniformly.

The percent errors of simulated annual mean precipitation, average *Rx1day* and return values are remarkably similar in pattern and magnitude suggesting that errors in water vapor transport are important to both the simulated mean state and to extremes. Errors in the annual average *Rx1day* and its 20 year return value are more closely related and return value skill is not so degraded as for extreme temperatures.

Although statistical uncertainty in the estimation of return values is low, uncertainty in the observational products may not be and has not been fully addressed here. Of particular concern is that algorithms used to infill poorly observed regions may not adequately treat the far tail of the distribution of daily data. This is not an issue for the HadEX3 datasets but could be for the REGEN precipitation dataset. However, the process of gridding station data introduces biases (Donat et al., 2014; Gervais et al., 2014; Risser et al., 2019) and products that are based on gridded extreme indices are not ideal for model evaluation.

Among other usages, model output is often used for the projection of future climate change and/or the attribution of climate change that has already occurred. Bias correction is a popular but potentially misleading strategy to force model output to more closely resemble observations. In particular, as the errors between mean state and extreme temperatures are weakly anti-correlated, simply bias correcting the distribution of daily values by mean state bias would make corrected simulated extreme temperatures worse. Similarly, simply bias correcting the distribution of annual extrema by the average error may also degrade corrected 20 year return values when the correlation between errors is low. In these cases, quantile bias correction (Jeon et al., 2016) of the 20 year return value itself is appropriate. While not considered here, the possibility of time dependence in the errors is very real. This could come from anthropogenic and/or natural modes of variability and be highly localized. A more detailed analysis of the effect of relevant covariates on observed precipitation return values will be presented in Risser et al. (2020).

Before any decision to utilize model simulated quantities, whether bias corrected or not, an arbitrary value judgement must be made as to whether the model and/or experimental design is "fit for the purpose" intended. While the model error metrics presented here can provide some guidance to fitness decisions, other value judgements about the appropriateness of performance metrics and how to use them must be made. Also, the credibility of the observed databases used as reference standards must also be considered even in well observed regions (Gibson et al., 2019). Performance metrics can also form a basis for model skill

weighting but the process is inherently arbitrary (Sanderson et al., 2017).

The analysis presented here reveals that no single CMIP5 nor CMIP6 model stands out as distinctly superior across either temperature or precipitation extremes. The range of model performance in simulating temperature extremes is comparable between the two generations of climate models and little difference in the performance of multi-model average simulations of annual average temperature and precipitation extremes or their 20 year return values. While we have not fully explored the effect of observational uncertainty on the selected extreme temperature and precipitation metrics, these general conclusions about the similarity between the CMIP5 and CMIP6 simulations are robust.

Author statement

Michael Wehner: Conceptualization, writing, data preparation, statistical Methodology, visualizations, Peter Gleckler: intercomparison Methodology, Jiwoo Lee: visualizations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The research performed at Lawrence Berkeley National Laboratory was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE340AC02-05CH11231. The research performed at Lawrence Livermore National Laboratory was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC52-07NA27344. Support from the Regional and Global Climate Modeling program is gratefully acknowledged. This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

This work used resources at the National Energy Research Supercomputer Center (NERSC) at the Lawrence Berkeley National Laboratory. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, coordinated and promoted CMIP5 and CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wace.2020.100283.

Appendix

In the mid and high latitudes, the warmest night of the year (*TNx*) generally occurs in the summer(Fig. 1, left column). The top row of Fig. A1shows *TNx* and its 20 year return value from the HadEX3 dataset and the middle and bottom rows show the CMIP5 and CMIP6 multi-model average errors. Relationships between these error maps are very similar to those of hot days with no substantial pattern correlation between boreal summer seasonal temperature errors and warm night errors and little correlation between *TNx* errors and return value errors. Like hot days, the pattern correlation between the CMIP5 and CMIP6 multi-model averages is high for both *TNx* errors (0.98) but very low for return value errors.

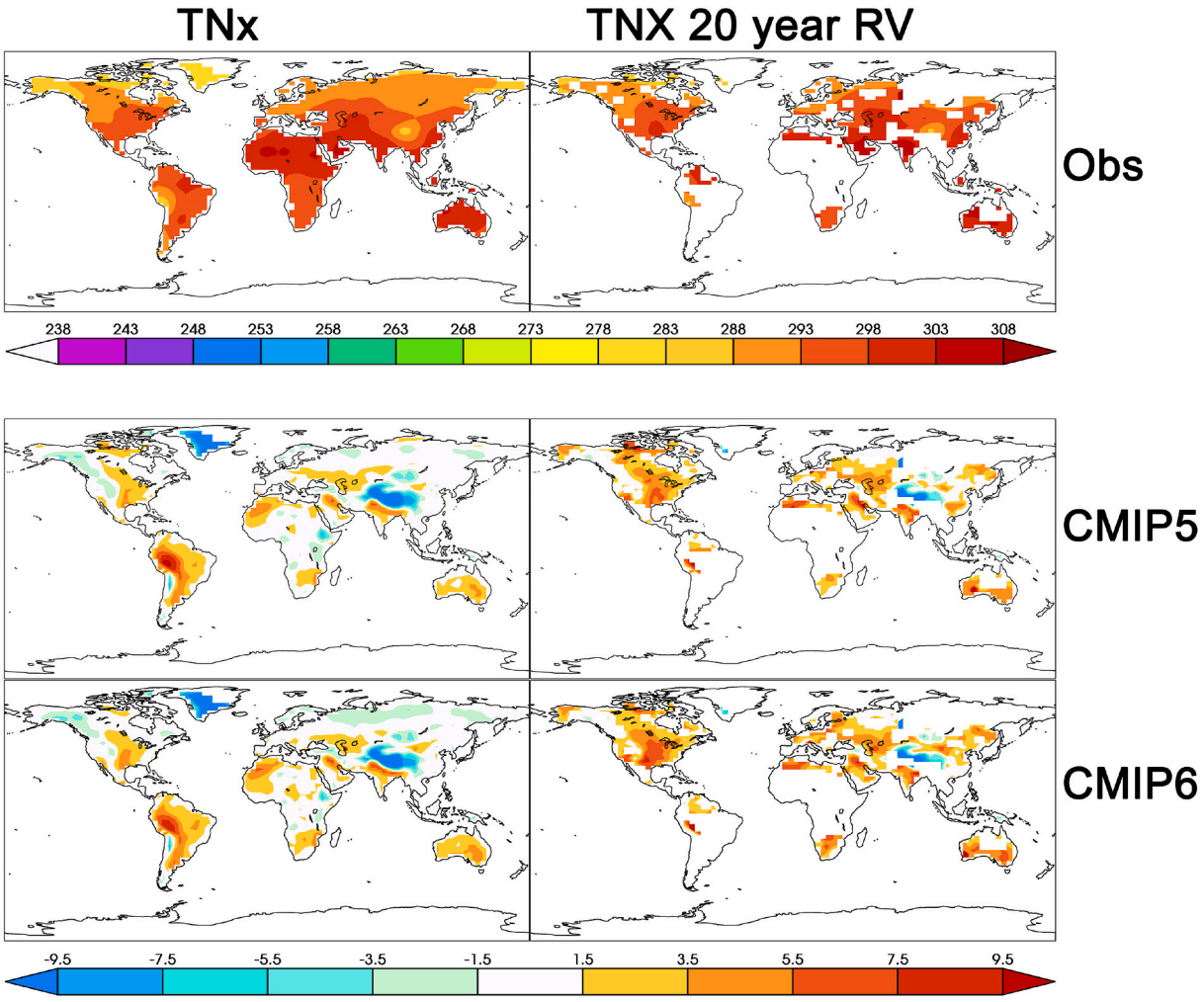

Fig. A1. Warm Nights (°C). Left column: *TNx*. Right column: Twenty year return value of *TNx*. Top row: 1961–2004 averages. Middle and bottom row: CMIP5 and CMIP6 multi-model average minus observations.

Model performance metrics in simulating warm nights are similar to hot days. The Taylor diagrams in Figure A2 show that model performance in simulating *TNx* (left) and its return value (right) are very tightly clustered except for a few outliers in the return value diagram. As for *TXx*, pattern correlation is significantly degraded for the return value compared to the average annual value with all values below about 0.4. Columns 6 and 7 of Table 1a,b shows the individual models' Taylor skill for *TNx* and its return value. Performance of most of the models is tightly clustered although both generations have some poor performers as return value error outliers. Except for these outliers, Taylor skill and centered RMSE for warm nights span similar ranges for CMIP5 and CMIP6 ensembles and the multi-models average performance are about the same.

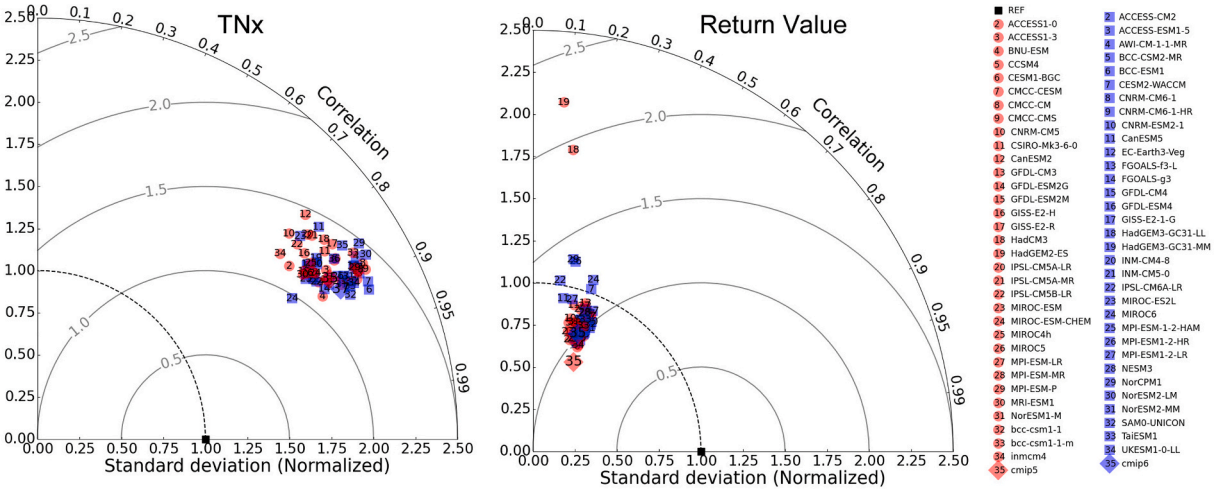

Fig. A2. Taylor diagram measuring model performance of simulating *TNx* (left) and its 20-year return value (right). The radial axis is normalized standard deviation while the angular axis is the centered pattern correlation. The reference data set is HadEX3 (black square). The concentric circles show the models' centered RMSE.

CMIP5 models are shown in red. CMIP6 models are shown in blue. Multi-model averages are denoted as "cmip5" and "cmip6" in the legend.

In the mid and high latitudes, the coolest day of the year (*TXn*) generally occurs in the winter (Fig. 3, left column). The top row of Fig. A2 and Fig. A3 shows *TXn* and its 20 year return value from the HadEX3 dataset and the middle and bottom rows show the CMIP5 and CMIP6 multi-model average errors. Relationships between these error maps are very similar to those of cold nights with weak centered pattern correlation (~0.25) between boreal winter average temperature errors and TXn errors. While the CMIP5 multi-model average *TXn* errors exhibits a moderate centered pattern correlation to its return values (0.55), the CMIP6 multi-model average does not. As for cold nights, the centered pattern correlation between the CMIP5 and CMIP6 multi-model averages is high for *TXn* errors (0.97) but not for return value errors.

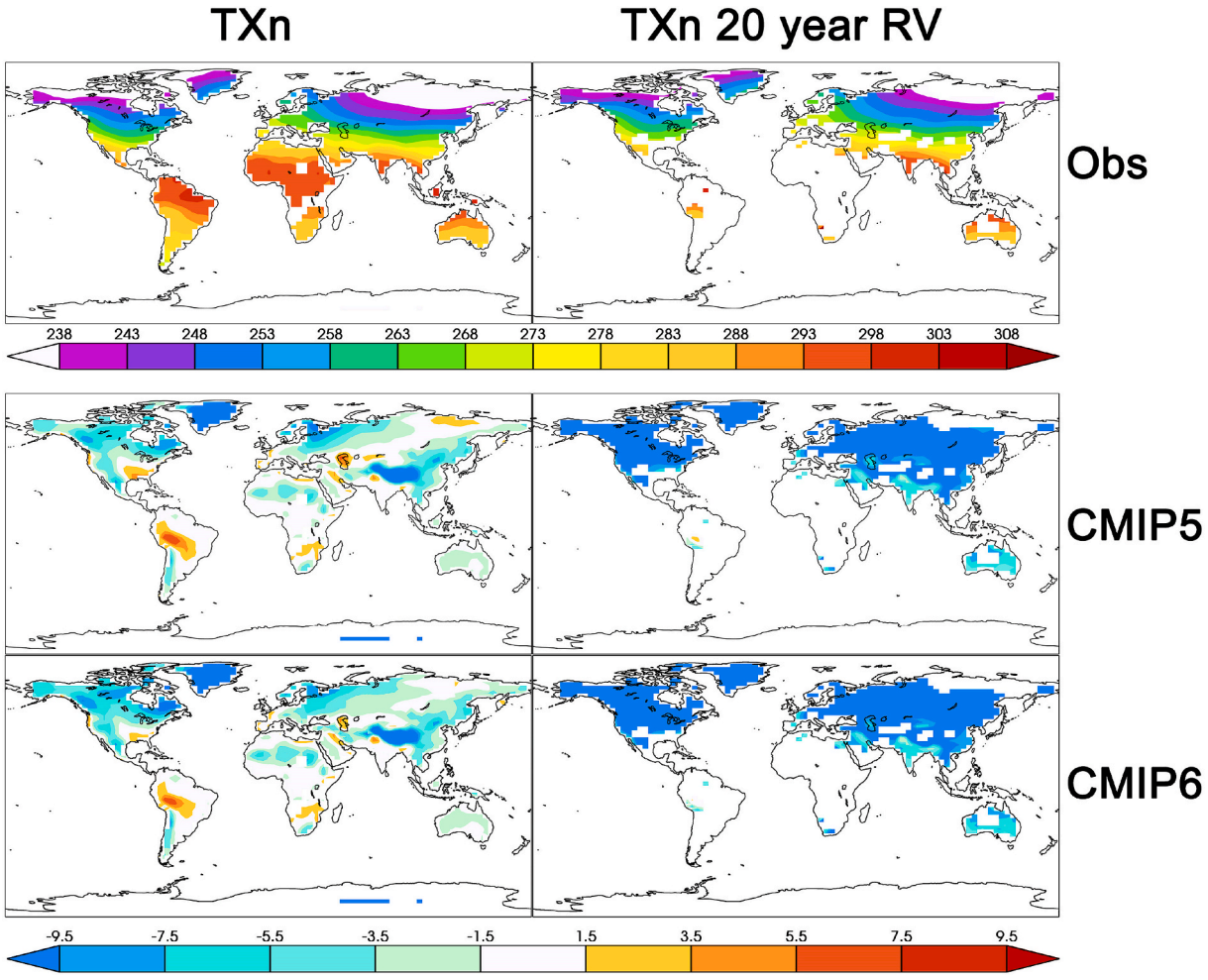

Fig. A3. Cool days (°C). Left column: *TXn*. Right column: Twenty year return value of *TXn*. Top row: 1961–2004 averages. Middle and bottom row: CMIP5 and CMIP6 multi-model average minus observations.

Model performance metrics in simulating cool days are similar to cold nights. The Taylor diagrams in Fig. A4 show that model performance in simulating *TXn* (left) and its return value (right) are the most tightly clustered of any of the performance metrics considered in this study despite a very cold bias. The range of both centered RMSE and Taylor skill is also the smallest and the difference between performance of the CMIP5 and CMIP6 multi-model average small. In general, model performance metrics are slightly better for cool days than for cold nights.

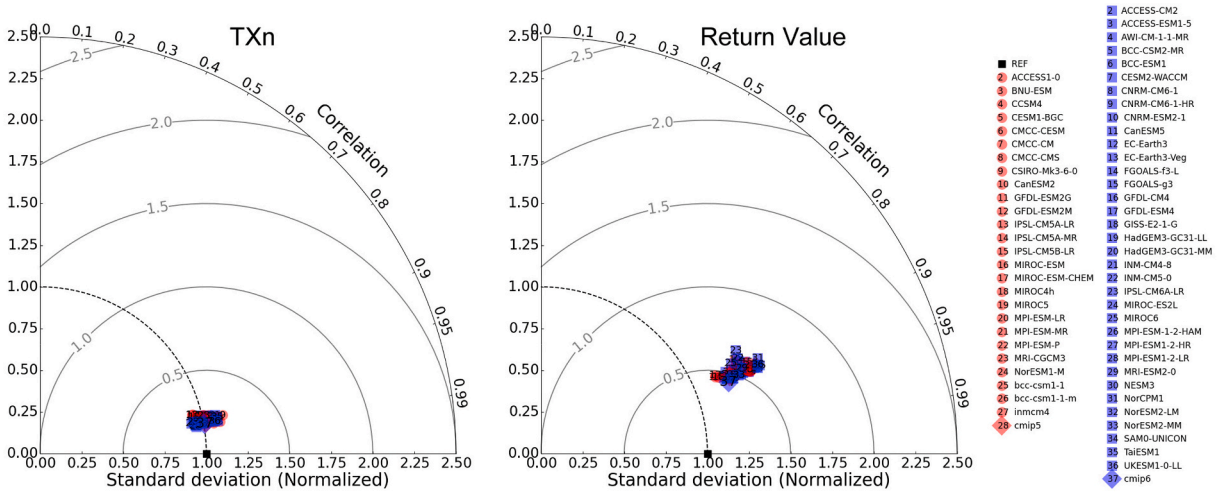

Fig. A4. Taylor diagram measuring model performance of simulating *TXn* (left) and its 20 year return value (right). The radial axis is normalized standard deviation while the angular axis is the centered pattern correlation. The reference data set is HadEX3 (black square). The concentric circles show the models' centered RMSE. CMIP5 models are shown in red. CMIP6 models are shown in blue. Multi-model averages are denoted as "cmip5" and "cmip6" in the legend.

References

Acero, F.J., García, J.A., Gallego, M.C., 2010. Peaks-over-Threshold study of trends in extreme rainfall over the iberian Peninsula. J. Clim. 24, 1089–1105. https://doi.org/10.1175/2010JCLI3627.1.

Arrhenius, S., 1896. On the influence of carbonic acid in the air upon the temperature of the ground. Philos. Mag. J. Sci 41, 237–276.

Bador, M., Boé, J., Terray, L., Alexander, L.V., Baker, A., Bellucci, A., Haarsma, R., Koenigk, T., Moine, M.-P., Lohmann, K., Putrasahan, D.A., Roberts, C., Roberts, M., Scoccimarro, E., Schiemann, R., Seddon, J., Senan, R., Valcke, S., Vanniere, B., 2020. Impact of higher spatial atmospheric resolution on precipitation extremes over land in global climate models. J. Geophys. Res. Atmos. n/a, e2019JD032184. https://doi.org/10.1029/2019JD032184.

Brown, S.J., Caesar, J., Ferro, C.A.T., 2008. Global changes in extreme daily temperature since 1950. J. Geophys. Res. Atmos. 113 https://doi.org/10.1029/2006JD008091.

Chen, C.-T., Knutson, T., 2008. On the verification and comparison of extreme rainfall indices from climate models. J. Clim. 21, 1605–1621. https://doi.org/10.1175/2007JCLI1494.1.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. Springer.

Contractor, S., Donat, M.G., Alexander, L.V., Ziese, M., Meyer-Christoffer, A., Schneider, U., Rustemeier, E., Becker, A., Durre, I., Vose, R.S., 2019. Rainfall Estimates on a Gridded Network (REGEN) – A global land-based gridded dataset of daily precipitation from 1950–2013, 2019 Hydrol. Earth Syst. Sci. Discuss. 1–30. https://doi.org/10.5194/hess-2018-595.

Cooley, D., Nychka, D., Naveau, P., 2007. Bayesian spatial modeling of extreme precipitation return levels. J. Am. Stat. Assoc. 102, 824–840. https://doi.org/10.1198/016214506000000780.

Donat, M.G., Alexander, L.V., Yang, H., Durre, I., Vose, R., Caesar, J., 2013. Global land-based datasets for monitoring climatic extremes. Bull. Am. Meteorol. Soc. 94, 997–1006.

Donat, M.G., Sillmann, J., Wild, S., Of, L.V.A.J., 2014. Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets, 2014 J. Clim. 27 (13), 5019–5035.

Dunn, R., 2020. Development of an updated global land in-situ-based dataset of temperature and precipitation extremes: HadEX3. J. Geophys. Res. Atmosph. 125, e2019JD032263 https://doi.org/10.1029/2019JD032263.

Easterling, D.R., Kunkel, K.E., Wehner, M.F., Sun, L., 2016. Detection and attribution of climate extremes in the observed record. Weather Clim. Extrem 11. https://doi.org/10.1016/j.wace.2016.01.001.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E., 2016. Overview of the coupled model intercomparison project Phase 6 (CMIP6) experimental design and organization. Geosci. Model Dev 9, 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016.

Fan, Y., van den Dool, H., 2008. A global monthly land surface air temperature analysis for 1948–present. J. Geophys. Res. Atmos. 113 https://doi.org/10.1029/2007JD008470.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., Rummukainen, M., 2013. Evaluation of climate models. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), 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, Cambridge, United Kingdom and New York, NY, USA, pp. 741–866. https://doi.org/10.1017/CBO9781107415324.020.

Gervais, M., Tremblay, L.B., Gyakum, J.R., Atallah, E., 2014. Representing extremes in a daily gridded precipitation analysis over the United States: impacts of station density, resolution, and gridding methods. J. Clim. 27, 5201–5218. https://doi.org/10.1175/JCLI-D-13-00319.1.

Gibson, P.B., Waliser, D.E., Lee, H., Tian, B., Massoud, E., 2019. Climate model evaluation in the presence of observational uncertainty: precipitation indices over the contiguous United States. J. Hydrometeorol. 20, 1339–1357. https://doi.org/10.1175/JHM-D-18-0230.1.

Gilleland, E., Katz, R.W., 2016. extRemes 2.0: an extreme value analysis package in R. J. Stat. Software 1 (8). https://doi.org/10.18637/jss.v072.i08.

Gilleland, E., Katz, R.W., 2011. New software to analyze how extremes change over time. Eos, Trans. Am. Geophys. Union 92, 13–14. https://doi.org/10.1029/2011EO020001.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. n/a 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Jeon, S., Paciorek, C.J., Wehner, M.F., 2016. Quantile-based bias correction and uncertainty quantification of extreme event attribution statements. Weather Clim. Extrem 12, 24–32. https://doi.org/10.1016/j.wace.2016.02.001.

Katz, R.W., 2010. Statistics of extremes in climate change. Climatic Change 100, 71–76. https://doi.org/10.1007/s10584-010-9834-5.

Kay, J.E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J.M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., Vertenstein, M., 2015. The community Earth system model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. Bull. Am. Meteorol. Soc. 96, 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1.

Kharin, V.V., Zwiers, F.W., Zhang, X., Wehner, M., 2013. Changes in temperature and precipitation extremes in the CMIP5 ensemble. Climatic Change 119. https://doi.org/10.1007/s10584-013-0705-8.

Kim, Y.-H., Min, S.-K., Zhang, X., Zwiers, F., Alexander, L.V., Donat, M.G., Tung, Y.-S., 2015. Attribution of extreme temperature changes during 1951–2010. Clim. Dynam. 46, 1769–1782. https://doi.org/10.1007/s00382-015-2674-2.

King, A.D., Black, M.T., Min, S.-K., Fischer, E.M., Mitchell, D.M., Harrington, L.J., Perkins-Kirkpatrick, S.E., 2016. Emergence of heat extremes attributable to anthropogenic influences. Geophys. Res. Lett. 43, 3438–3443. https://doi.org/10.1002/2015GL067448.

Kyselý, J., Picek, J., Beranová, R., 2010. Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. Global Planet. Change 72, 55–68. https://doi.org/10.1016/j.gloplacha.2010.03.006.

Lorenz, R., Argüeso, D., Donat, M.G., Pitman, A.J., van den Hurk, B., Berg, A., Lawrence, D.M., Chéruy, F., Ducharne, A., Hagemann, S., Meier, A., Milly, P.C.D., Seneviratne, S.I., 2016. Influence of land-atmosphere feedbacks on temperature and precipitation extremes in the GLACE-CMIP5 ensemble. J. Geophys. Res. Atmos. 121, 607–623. https://doi.org/10.1002/2015JD024053.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., Shi, W., 2006. North American regional reanalysis. Bull. Am. Meteorol. Soc. 87, 343–360. https://doi.org/10.1175/BAMS-87-3-343.

Min, S.-K., Zhang, X., Zwiers, F., Shiogama, H., Tung, Y.-S., Wehner, M., 2013. Multimodel detection and attribution of extreme temperature changes. J. Clim. 26 https://doi.org/10.1175/JCLI-D-12-00551.1.

Min, Seung-Ki, Zhang, X., Zwiers, F., Shiogama, H., Tung, Y.-S., Wehner, M., 2013. Multimodel detection and attribution of extreme temperature changes. J. Clim. 26, 7430–7451. https://doi.org/10.1175/JCLI-D-12-00551.1.

Min, S.-K., Zhang, X., Zwiers, F.W., Hegerl, G.C., 2011. Human contribution to more-intense precipitation extremes. Nature 470, 378.

Mitchell, J.F.B., Karoly, D.J., Hegerl, G.C., Zwiers, F.W., Allen, M.R., Marengo, J., 2001. Detection of climate change and attribution of causes. In: Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., Johnson, C.A. (Eds.), Climate Change 2001: the Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 695–738.

Paciorek, C.J., Stone, D.A., Wehner, M.F., 2018. Quantifying statistical uncertainty in the attribution of human influence on severe weather. Weather Clim. Extrem 20, 69–80. https://doi.org/10.1016/j.wace.2018.01.002.

Risser, M., O'Brian, J.P., O'Brien, T., Patricola, C.M., Wehner, M.F., 2020. Quantifying the impact of climate variability on in situ measurements of extreme daily precipitation. J. Clim.

Risser, M.D., Paciorek, C.J., O'Brien, T.A., Wehner, M.F., Collins, W.D., 2019. Detected changes in precipitation extremes at their native scales derived from in situ measurements. J. Clim. 32, 8087–8109. https://doi.org/10.1175/JCLI-D-19-0077.1.

Risser, Mark D., Paciorek, Christopher J., Wehner, Michael F., O'Brien, Travis A., Collins, W.D., 2019. A probabilistic gridded product for daily precipitation extremes over the United States. Clim. Dynam. https://doi.org/10.1007/s00382-019-04636-0.

Risser, M.D., Wehner, M.F., 2017. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during hurricane harvey. Geophys. Res. Lett. 44 (12), 412–457. https://doi.org/10.1002/2017GL075888, 464.

Roth, M., Buishand, T.A., Jongbloed, G., Klein Tank, A.M.G., van Zanten, J.H., 2012. A regional peaks-over-threshold model in a nonstationary climate. Water Resour. Res. 48 https://doi.org/10.1029/2012WR012214.

Sanderson, B.M., Wehner, M., Knutti, R., 2017. Skill and independence weighting for multi-model assessments. Geosci. Model Dev 10. https://doi.org/10.5194/gmd-10-2379-2017.

Schiemann, R., Athanasiadis, P., Barriopedro, D., Doblas-Reyes, F., Lohmann, K., Roberts, M.J., Sein, D., Roberts, C.D., Terray, L., Vidale, P.L., 2020. The representation of Northern Hemisphere blocking in current global climate models. Weather Clim. Dyn. Discuss 2020, 1–21. https://doi.org/10.5194/wcd-2019-19.

Schiemann, R., Demory, M.-E., Shaffrey, L.C., Strachana, J., Vidale, P.L., Mizielinski, M. S., Roberts, M.J., Wehner, M.F., Jung, T., 2017. The resolution sensitivity of Northern Hemisphere blocking in four 25-km atmospheric global circulation models. J. Clim. 30 https://doi.org/10.1175/JCLI-D-16-0100.1.

Sillmann, J., Croci-Maspoli, M., Kallache, M., Katz, R.W., 2011. Extreme cold winter temperatures in Europe under the influence of North Atlantic atmospheric blocking. J. Clim. 24, 5899–5913. https://doi.org/10.1175/2011JCLI4075.1.

Sillmann, J., Kharin, V., Zhang, X., Zwiers, F.W., Bronaugh, D., 2013. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. J. Geophys. Res. Atmos. 118, 1716–1733. https://doi.org/10.1002/jgrd.50203.

Solari, S., Marta, E., José, P.M., Losada, M., 2017. Peaks over Threshold (POT): a methodology for automatic threshold estimation using goodness of fit p-value. Water Resour. Res. 53, 2833–2849. https://doi.org/10.1002/2016WR019426.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106, 7183–7192. https://doi.org/10.1029/2000JD900719.

Timmermans, B., Wehner, M., Cooley, D., O'Brien, T., Krishnan, H., 2019. An evaluation of the consistency of extremes in gridded precipitation data sets. Clim. Dynam. 52, 6651–6670. https://doi.org/10.1007/s00382-018-4537-0.

Wang, Z., Jiang, Y., Wan, H., Yan, J., Zhang, X., 2017. Detection and attribution of changes in extreme temperatures at regional scale. J. Clim. 30, 7035–7047. https://doi.org/10.1175/JCLI-D-15-0835.1.

Wehner, M.F., 2013. Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. Clim. Dynam. 40, 59–80. https://doi.org/10.1007/s00382-012-1393-1.

Wehner, Michael, 2020. Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 2, projections of future change. Weather Clim. Extr. in press.

Wehner, M.F., Duffy, M., Risser, M.D., Paciorek, C.J., Stone, D.A., Pall, P., Krishnan, H., 2020. On the uncertainty of long-period return values of extreme daily precipitation. Adv. Stat. Climatol. Meteorol. Oceanogr.

Westra, S., Alexander, L.V., Zwiers, F.W., 2012. Global increasing trends in annual maximum daily precipitation. J. Clim. 26, 3904–3918. https://doi.org/10.1175/JCLI-D-12-00502.1.

Zhang, X., Alexander, L., Hegerl, G.C., Jones, P., Tank, A.K., Peterson, T.C., Trewin, B., Zwiers, F.W., 2011. Indices for monitoring changes in extremes based on daily temperature and precipitation data. WIREs Clim. Chang 2, 851–870. https://doi.org/10.1002/wcc.147.

Zhang, X., Wan, H., Zwiers, F.W., Hegerl, G.C., Min, S.-K., 2013. Attributing intensification of precipitation extremes to human influence. Geophys. Res. Lett. 40, 5252–5257. https://doi.org/10.1002/grl.51010.

Zschenderlein, P., Fink, A.H., Pfahl, S., Wernli, H., 2019. Processes determining heat waves across different European climates. Q. J. R. Meteorol. Soc. 145, 2973–2989. https://doi.org/10.1002/qj.3599.

Zwiers, F.W., Zhang, X., Feng, Y., 2011. Anthropogenic influence on long return period daily temperature extremes at regional scales. J. Clim. 24, 881–892. https://doi.org/10.1175/2010JCLI3908.1.