

RESEARCH ARTICLE

10.1029/2017JD027992

Key Points:

- Model weighting slightly reduces summer warming signal over central North America
- More than one predicting diagnostics should be used to inform the weighting
- Shortwave radiation trend, mean precipitation, and SST variability are possible constraints on projections of summer maximum temperature

Supporting Information:

- Supporting Information S1

Correspondence to:

R. Lorenz,
ruth.lorenz@env.ethz.ch

Citation:

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *Journal of Geophysical Research: Atmospheres*, 123, 4509–4526. <https://doi.org/10.1029/2017JD027992>

Received 31 OCT 2017

Accepted 4 APR 2018

Accepted article online 16 APR 2018

Published online 10 MAY 2018

©2018. The Authors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America

Ruth Lorenz¹, Nadja Herger², Jan Sedláček¹, Veronika Eyring^{3,4}, Erich M. Fischer¹, and Reto Knutti¹

¹Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland, ²ARC Center of Excellence for Climate System Science and Climate Change Research Center, UNSW Australia, Sydney, New South Wales, Australia, ³Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany, ⁴Institute of Environmental Physics (IUP), University of Bremen, Bremen, Germany

Abstract Uncertainties in climate projections exist due to natural variability, scenario uncertainty, and model uncertainty. It has been argued that model uncertainty can be decreased by giving more weight to those models in multimodel ensembles that are more skillful and realistic for a specific process or application. In addition, some models in multimodel ensembles are not independent. We use a weighting approach proposed recently that takes into account both model performance and interdependence and apply it to investigate projections of summer maximum temperature climatology over North America in two regions of different sizes. We quantify the influence of predicting diagnostics included in the method, look at ways how to choose them, and assess the influence of the observational data set used. The trend in shortwave radiation, mean precipitation, sea surface temperature variability, and variability and trend in maximum temperature itself are the most promising constraints on projections of summer maximum temperature over North America. The influence of the observational data sets is large for summer temperature climatology, since the observational and reanalysis products used for absolute maximum temperatures disagree. Including multiple predicting diagnostics leads to more similar results for different data sets. We find that the weighted multimodel mean reduces the change in summer daily temperature maxima compared to the nonweighted mean slightly (0.05–0.45 °C) over the central United States. We show that it is essential to have reliable observations for key variables to be able to constrain multimodel ensembles of future projections.

1. Introduction

Climate projections are associated with uncertainties that need to be well quantified for impact studies, policymakers, and the general public. While we have high confidence that global mean temperature is increasing, we are less certain about the rate, the magnitude of this increase, and changes at local and regional scales. Uncertainties are particularly large for climate extremes (Kharin et al., 2013; Sillmann et al., 2013). Climate extremes are strongly influenced by climate variability, which is superimposed on long-term trends, making model evaluation and attribution difficult (Fischer & Knutti, 2014; Perkins & Fischer, 2013). Temperature extremes in North America are related to anomalies in large-scale circulation (Horton et al., 2016; Meehl & Tebaldi, 2004) and can be enhanced by land surface processes (Lorenz et al., 2016; Seneviratne et al., 2013; Teng et al., 2016). While the large-scale circulation patterns, which are important for heat extremes, are modeled reasonably well (Loikith & Broccoli, 2015), biases and uncertainties regarding land surface processes are substantial. Merrifield and Xie (2016) conclude that many models overestimate the influence from the land surface on temperature and also misplace the region where the influence from the land surface is most important. Similarly, Mueller and Seneviratne (2014) found that biases in evapotranspiration can be linked to biases in temperature. They identified a tendency for negative biases in evapotranspiration and positive biases in temperature for boreal summer in North America among other regions in many models contributing to the Coupled Model Intercomparison Project phase 5 (CMIP5, Taylor et al., 2012). Donat et al. (2017) found that spatial patterns, where hot extremes are increasing more than global average temperatures in the CMIP5 models, are inconsistent with observations, except over Europe. Also, Sippel et al. (2017) showed that some models

in the CMIP5 archive encounter too frequent phases with small evapotranspiration and high temperatures. Excluding those models from the CMIP5 multimodel mean resulted in reduced biases and lower absolute values in extreme temperature projections. Hence, models that overestimate the influence of the land surface on the atmosphere are more likely to have positive absolute temperature and variability biases in the historical period. Christensen and Boberg (2012) found a connection between positive temperature biases in the warm period of the year and larger temperature change into the future. Therefore, models with larger positive biases in absolute temperature and variability in the historical period are more likely to overestimate future temperature change. Hence, by constraining temperature biases in the historical period, we can potentially be more confident in our projections into the future.

More generally, uncertainties in climate projections originate from three main sources: (1) natural variability, (2) model uncertainty, which is a combination of the fact that models can never completely represent the reality and differences in how the models approximate it, and (3) scenario uncertainty (Hawkins & Sutton, 2009). It is not possible to reduce the uncertainty due to natural variability (Deser et al., 2012) on time scales more than a few years, but we can estimate how large the natural variability itself is and can therefore account for it. The scenario uncertainty is also difficult to reduce since societal decisions, innovation, and technological progress do not follow physical laws but are largely choices that society makes. In the Intergovernmental Panel for Climate Change (IPCC) Assessment Reports this is taken into account by creating several possible scenarios which span a range of plausible economic pathways. Based on the corresponding greenhouse gas (and other) emissions, climate models make projections. This leaves us with model uncertainty as our best starting point, if we want to reduce the uncertainty in future projections and increase the reliability thereof.

Multimodel ensembles help to understand the uncertainty in climate projections by providing multiple estimates of future climate. It is common to calculate an arithmetic multimodel mean when working with multimodel ensembles, giving each model (or in a few situations each ensemble member) the same weight (e.g., Collins et al., 2013). However, multiple studies have argued that this is not necessarily the best choice, at least in cases where we know that some models might be more reliable than others (e.g., Annan & Hargreaves, 2011; Knutti, Abramowitz, et al., 2010; Knutti, Furrer, et al., 2010; Knutti et al., 2013). In addition, large multimodel ensembles, such as CMIP, increasingly contain multiple versions of the same or almost the same model, only differing in resolution, additional components, or single components that have been replaced. An arithmetic mean implies that each sample is independent and should get the same weight. This needs to be reconsidered given that multimodel ensembles contain near replicates of the same model that are not independent anymore (Knutti et al., 2010; Knutti et al., 2013; Masson & Knutti, 2011). Several methods have been proposed to take into account performance and/or interdependence when calculating multimodel averages (e.g., Abramowitz & Bishop, 2015; Abramowitz et al., 2008; Herger et al., 2018; Karpechko et al., 2013; Sanderson et al., 2015a, 2015b; Tebaldi et al., 2006; Waugh & Eyring, 2008). Most of these approaches are rather complex, and so far no agreement was reached about which method should be used in general, how model dependence should be defined, and how it could best be accounted for.

The approach here is based on a method published by Knutti et al. (2017). Following Sanderson et al. (2015a, 2015b, 2017), Knutti et al. (2017) proposed a method how climate model projections can be weighted based on performance and interdependence and applied their concept to projections of Arctic sea ice extent and temperature. The weighted mean reduced the spread in the projections of when the Arctic will be ice free. However, multiple caveats were mentioned, such as the difficulty to choose relevant diagnostics to inform the method and the determination of two parameters required to obtain model weights (see section 2.3). Here we apply this method to weight maximum temperature (tasmax) projections and study the sensitivity of the results by using different observational data sets, diagnostics, ways to determine which diagnostics to use, and the choice of the analyzed region. The selected diagnostics are based on seasonal means from which we calculate climatologies, standard deviations, and trends. By investigating these aspects we can further test the advantages and disadvantages of the method and identify potential arising problems. Section 2 describes the data sets used and the methods applied. Section 3 presents all results. First, we discuss how we choose the diagnostics relevant for maximum temperature and its changes into the future. Then we weight the multimodel ensemble. We weight based on its performance in representing not only maximum temperature in the historical period but also the other diagnostics identified in the first step, as well as interdependence. We present the results of the weighting method for two differently defined overlapping regions over North America for projections of mean summer maximum temperature. The fourth section discusses these results and the last section contains a summary and conclusions.

Table 1
Variables and Diagnostics Tested for Correlations With TasmaxCLIM for Future Change (Column 3) and Over Historical Period (Column 5) and Corresponding p values (Columns 4 and 6) Over North America (NAM)

Variable	Diagnostics	Δ tasmaxCLIM	p value	tasmaxCLIM	p value
Daily maximum temperature	tasmaxCLIM	0.36	0.003	1.00	0.00
Total precipitation	prCLIM	-0.56	0.00	-0.63	0.00
Longwave upward radiation	rlusCLIM	0.46	0.00	0.74	0.00
Shortwave downward radiation	rsdsCLIM	0.48	0.00	0.86	0.00
Surface specific humidity	hussCLIM	-0.24	0.05	0.34	0.004
Sea level pressure	pslCLIM	-0.05	0.707	0.43	0.00
Latent heat flux	hflsCLIM	-0.57	0.00	-0.59	0.00
Sea surface temperature	tosCLIM	-0.17	0.161	0.11	0.355
Daily maximum temperature	tasmaxSTD	0.41	0.001	0.25	0.042
Total precipitation	prSTD	-0.41	0.00	-0.62	0.00
Longwave upward radiation	rlusSTD	0.47	0.00	0.29	0.016
Shortwave downward radiation	rsdsSTD	0.11	0.357	-0.16	0.207
Surface specific humidity	hussSTD	0.27	0.025	0.00	0.991
Sea level pressure	pslSTD	-0.05	0.697	-0.38	0.002
Latent heat flux	hflsSTD	0.24	0.046	0.04	0.719
Sea surface temperature	tosSTD	0.51	0.00	0.07	0.553
Daily maximum temperature	tasmaxTREND	0.3	0.013	0.46	0.00
Total precipitation	prTREND	-0.05	0.711	0.1	0.433
Longwave upward radiation	rlusTREND	0.36	0.003	0.49	0.00
Shortwave downward radiation	rsdsTREND	0.6	0.00	0.28	0.024
Surface specific humidity	hussTREND	-0.1	0.413	0.19	0.13
Sea level pressure	pslTREND	-0.24	0.054	0.08	0.533
Latent heat flux	hflsTREND	0.02	0.899	0.02	0.882
Sea surface temperature	tosTREND	0.15	0.216	0.32	0.008

2. Data and Methods

2.1. Global Climate Model Data From CMIP5

The CMIP5 archive contains model output from global climate models (GCMs) and Earth system models from multiple institutions from all over the world. We use historical model simulations (1850–2005) as well as Representative Concentration Pathway (RCP) 8.5 (Meinshausen et al., 2011) projections (2006–2100). We use all available model runs, including multiple initial conditions ensembles. Up to 40 models with up to 12 ensemble members were included in the analysis, for tasmax a total of 89 runs were available (see Table S1 in the supporting information for all the models and ensembles used and which variables are available for which runs). We regrid data from all models (as well as the observational data sets described below) to the same $2.5^\circ \times 2.5^\circ$ grid using the Climate Data Operators bilinear regridding algorithm. CMIP5 data can be obtained from <http://cmip-pcmdi.llnl.gov/cmip5/>.

2.2. Evaluation Data Sets

2.2.1. MERRA-2 Reanalysis

The Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2, <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/>) provides data from 1980 to 2014. MERRA-2 includes satellite as well as conventional weather observations and is produced with the goal of an updated version from MERRA including new observations which could not be assimilated in the older MERRA system. To improve land surface hydrology it does not use model based precipitation but observations-corrected precipitation. It is also the first reanalysis which includes aerosol measurements from space and their interactions. MERRA-2 provides all the variables we investigate in this study (see Table 1 for all variables used).

2.2.2. ERAinterim Reanalysis

ERAinterim (ERAint) is a global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Dee et al., 2011). It provides reanalysis data from 1979 until the present, we use data until 2014 in this study. It is a widely used reanalysis product and was produced as a new atmospheric reanalysis to replace the older ERA-40. ERAint uses a new model (IFS Cy31r2) and an updated assimilation method (4D-Var). The main goals for ERAint were to improve the representation of the hydrological cycle, the quality of the stratospheric circulation, and the consistency in time of the reanalyzed fields. It provides all the variables we needed for this study except surface specific humidity (huss).

2.2.3. HadGHCND Gridded Maximum Temperature Data Set

We use the HadGHCND gridded daily temperature data set (Caesar et al., 2006) derived from near-surface maximum and minimum temperature observations. HadGHCND covers the period from 1951 to the present on a 2.75° latitude \times 3.75° longitude grid. It was designed for the analysis of climate extremes and the evaluation of climate models. Note that the data coverage is varying in time. We used HadGHCND to evaluate daily maximum temperature (tasmax).

2.2.4. GPCP Gridded Precipitation Data Set

The Global Precipitation Climatology Project (GPCP) Version-2.3 (<http://www.esrl.noaa.gov/psd/data/gridded/data.gpcp.html>) precipitation data set is derived from a combination of satellite and rain gauge measurements (Adler et al., 2003). GPCP is available as global, monthly analysis of surface precipitation at $2.5^\circ \times 2.5^\circ$ resolution from 1979 to the present. GPCP has been shown to agree well with ground-based observations (Ma et al., 2009; Pfeifroth et al., 2013). We used GPCP to evaluate total precipitation (pr).

2.2.5. CERES EBAF Satellite Radiation Data Set

The NASA "Clouds and the Earth's Radiant Energy System" (CERES EBAF Surface Ed2.8) data set provides satellite-based estimates of surface radiative fluxes. This data set was specifically created for evaluation of climate models (<http://ceres-tool.larc.nasa.gov>). It includes surface all-sky downwelling shortwave and longwave radiation, surface upwelling shortwave and longwave radiation, and estimates for clear-sky radiation from March 2000 to December 2015. Kato et al. (2013) found that biases over land were on average between -1.7 and 4.7 W/m^2 for downward shortwave and between -1.0 and -2.5 W/m^2 for downward longwave radiation. We used CERES EBAF to evaluate longwave upward radiation (rlus) and shortwave downward radiation (rsds).

2.2.6. HadISST Global Sea Ice and Sea Surface Temperature Data Set

We use the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST) monthly data set. This data set provides global data from 1871 to the present on a 1° resolution (Rayner, 2003). HadISST uses mainly ship track data until 1981 and a blend of adjusted satellite-derived and in situ data afterward. The main problems were noticed in polar regions, due to low data availability and the sea ice analysis, and the equatorial Pacific where the data set experiences nonrobust trends (National Center for Atmospheric Research Staff (Eds), 2017). The data set was developed to improve on earlier sea surface temperature data sets to be used for driving atmospheric models and also to evaluate coupled atmosphere-ocean models. We used HadISST to evaluate sea surface temperature (SST or tos).

2.3. Calculation of Diagnostics and Motivation for Diagnostic Choices

The diagnostics we use can easily be calculated from the standard CMIP5 output, namely, climatologies (CLIM), standard deviations (STD), and linear trends (TREND) based on seasonal data over historical as well as future time periods for the variables listed in Table 1. We focus on summer (June to August, JJA) and average variables over these months before we concatenate them over the respective time periods.

The historical time period is determined by the availability of observational and reanalysis products and is generally 1980–2014. An exception to this is the CERES EBAF radiation data, which are only available from 2000 to 2015. To calculate the observation-model distances, we therefore only use 2000–2014 for CERES and the CMIP5 models. Calculating the trend for rsds and rlus over this shorter period is freighted with uncertainty and should be taken with caution. As future time period we chose the same number of years (35) as for the historical period but at the end of the century (2065–2099).

We do not calculate climate extremes indices but investigate daily maximum temperatures in JJA in general. The diagnostics are calculated for two regions, North America (NAM, 25.0°N – 50.0°N and 140.0°W – 55.0°W) and Central North America (CNA, 28.566°N – 50.0°N and 105.0°W – 85.0°W), and only data from within these regions over land are taken into account in the analysis. The only exception is sea surface temperature (tos)

which is included over the NAM ocean region in both cases. Hence, we only include local to regional processes and assume remote or even global processes, which are relevant for the region, to be represented reasonably in the models if they represent the regional diagnostics well. Since we focus on the summer season, when large-scale circulation and advection play a less dominant role compared to winter (e.g., Loikith & Broccoli, 2014, 2015), this is a reasonable assumption for this study.

An important factor influencing summer daytime temperature is how much radiation is received at the Earth's surface. In North America, extreme temperatures have been identified with a combination of two more factors: hot air masses displaced from their usual location and strong subsidence causing adiabatic warming (Grotjahn et al., 2016; Loikith & Broccoli, 2012; Meehl & Tebaldi, 2004). Unfortunately, not all CMIP5 models provide the output necessary to calculate geopotential height (gph) at a certain pressure level. Hence, using gph as a potential predicting variable would significantly decrease our sample size (Loikith & Broccoli, 2015). We included sea level pressure (psl) as a potential describing variable, but none of the diagnostics turned out to be relevant for our target diagnostics in our region (see section 3.1). How much of the energy reaching the surface is translated into sensible versus latent heat flux is modulated by land surface conditions (e.g., dryness and vegetation) and can influence temperatures (e.g., Seneviratne et al., 2010). Based on the hypothesis that land surface processes play a role for summer temperature in North America (e.g., Dirmeyer et al., 2013; Koster et al., 2006; Lorenz et al., 2016; Merrifield & Xie, 2016) we would expect evapotranspiration (represented by the latent heat flux, hfls) to be relevant for tasmaxCLIM and/or tasmaxSTD. As we will see in section 3.1, hflsCLIM is highly correlated with prCLIM. This is not surprising since hfls depends a lot on moisture availability, which is strongly influenced by precipitation. Earlier studies have shown that antecedent precipitation can be used as a proxy for evapotranspiration and influences extreme temperatures in certain regions (e.g., Hirschi & Seneviratne, 2010; Mueller & Seneviratne, 2012; Perkins et al., 2015). Using precipitation instead of evapotranspiration also has the advantage of more reliable observations (Wang & Dickinson, 2012). Variability in tos, in particular, phases of the El Niño–Southern Oscillation, have also been shown to influence extreme temperatures over America (Alexander et al., 2009; Arblaster & Alexander, 2012) and are, therefore, a logical potential predicting diagnostic.

2.4. Multimodel Weighting Method

The weighting method we use is defined by Knutti et al. (2017) which again is based on Sanderson et al. (2015a, 2015b). The main equation calculates a weight w_i for each model run in the ensemble, taking into account model performance in the numerator and model dependence in the denominator:

$$w_i = \frac{e^{-\frac{D_i^2}{\sigma_D^2}}}{1 + \sum_{j \neq i}^M e^{-\frac{S_{ij}^2}{\sigma_S^2}}} \quad (1)$$

with D_i being the distance of model i to observations, σ_D the parameter which determines how strongly model performance is weighted, M the number of model runs, S_{ij} the distance between models i and j , and σ_S the parameter that determines how strongly model similarity is weighted. Hence, some choices need to be made before we can calculate w_i : (1) How do we measure model performance? (2) How do we measure model similarity? (3) How strongly do we weight for model performance (σ_D)? (4) When do we consider models to be similar (σ_S)? In addition, since the method allows to take into account multiple diagnostics relevant for the projected quantity, we need to decide (5) which diagnostics to take into account, over which region, time period, etc.

1. Model performance can be measured using skill scores or other metrics that compare model output to observations. As in Knutti et al. (2017) we use root-mean-square error (RMSE) as a performance measure in this study. We tested one other skill score (defined and used in Perkins et al., 2007) and found that it makes very little difference overall for this particular use case and therefore stuck to RMSE. Observational data sets also have some uncertainty that needs to be considered. We perform the analysis with multiple reanalysis and observational data sets to test the sensitivity of the results to the data set used.
2. Model similarity can be measured in similar ways to model performance. Again, we use RMSE but between all the model pairs. If a few models have relatively small biases and are near the observations, there is the danger of them being downweighted unjustifiedly, because they are close and, therefore, judged as being similar. We investigated this issue using multidimensional scaling (MDS) of the distance matrix (Figure S1 for HadGHCND as an example, more details on MDS can be found in the supporting information).

No models are close enough to observations to be in danger of being considered similar. Nevertheless, this has to be taken care of once models become better and closer to observational estimates. In such a case, the model-model distance needs to be put into context with the model-observation distance, for instance, by normalizing the model-model distances by the model-observation distance of model i .

3. The σ_D determines how strongly models are being downweighted if they are far away from observations.
4. The σ_S determines a typical distance when two models are considered to be similar. The larger this parameter, the farther away two models are allowed to be in order to still be considered almost identical. Here we estimate σ_D and σ_S based on the same principle as Knutti et al. (2017) using a perfect model test. One model is assumed to be the truth and we predict the truth using all the other models (only one initial ensemble member per model is used). For every model as truth we calculate the weighted multimodel mean for 41×41 σ combinations. The best parameter combination should predict a distribution of the target diagnostic, which is neither too narrow nor too wide. A small value for σ_D , for example, results in aggressive weighting and possibly overconfident results, while a large value converges to equal weighting (see Knutti et al., 2017 for a discussion). Hence, we choose the parameters so that the model as truth lies within the 10th–90th percentile range 80% of the time. For more than two diagnostics this leads to $\sigma_D \approx 0.5$ and $\sigma_S \approx 0.6$ (see supporting information Figure S2 for results of the perfect model test). In principle, σ_D and σ_S should be chosen separately for different numbers of diagnostics but they are estimated to be very similar in our approach. Only when using one diagnostic that σ_D should be higher. For simplicity, we use only one value per parameter throughout the study. Having very similar models in the ensemble, such as from the same institution, could make the perfect model test work better than it should. We tested this by removing obvious duplicates and using only one model per institution (supporting information Figure S2c). This leads to lower σ values ($\sim \sigma_D = 0.4$ and $\sigma_S = 0.5$), which would lead to slightly more aggressive weighting. There are also dependencies across institutes, and we also calculated the σ parameters only using models that, to our knowledge, are mostly independent (supporting information Figure S2d), which suggests a smaller σ_D . However, this ensemble is small (12 models), and we might have excluded too much information or still have similar models in the ensemble, and it is not obvious where to draw the line. Since the suggested parameters do not vary much between all these cases, the results will not be strongly affected by the set of models in the perfect model test and our σ values should not lead to too aggressive weighting.
5. Whenever we use more than one diagnostic, we calculate the distances D_i and S_{ij} for all diagnostics and then use a normalized average difference across all of them. The distances are normalized by their median. To decide which diagnostics to use, we check if there is a correlation between the change in the target diagnostic (tasmax climatology or standard deviation) into the future (see Table 1 for all the tested diagnostics and correlations with tasmaxCLIM). This ensures that there is a statistical relationship with the projected change into the future, in addition to the physical relationships touched upon in section 2.3. We use linear regressions between the different diagnostics as well as statistical feature selection methods (see supporting information section S2 for more details) to determine the relevant diagnostics.

2.5. Error Index I^2

We use the error index I^2 based on Baker and Taylor (2016) to compare nonweighted and weighted multimodel means. This error index can measure combined errors, that is, over multiple variables, compared to the observed climate. However, we only evaluate our target variable tasmax (maximum temperature) over the evaluation period 1980–2014 in this study. In a first step, the normalized error is calculated as the difference between weighted multimodel mean and observed mean

$$e_w^2 = \sum_n \left[A_n \frac{(\langle \bar{S}_n \rangle - \bar{o}_n)^2}{\sigma_n^2} \right] \quad (2)$$

where $\langle \bar{S}_n \rangle$ is the weighted multimodel climatology for our target variable per grid point n . The parameter \bar{o}_n is the observed climate from the corresponding variable and σ_n^2 is the observed interannual variability from the same variable. Then the normalized errors are area weighted and regionally averaged. The corresponding e_{eq}^2 is calculated for the nonweighted multimodel mean, and the averaged errors from the weighted multimodel mean are scaled relative to e_{eq}^2 :

$$I^2 = \frac{e_w^2}{e_{eq}^2} \quad (3)$$

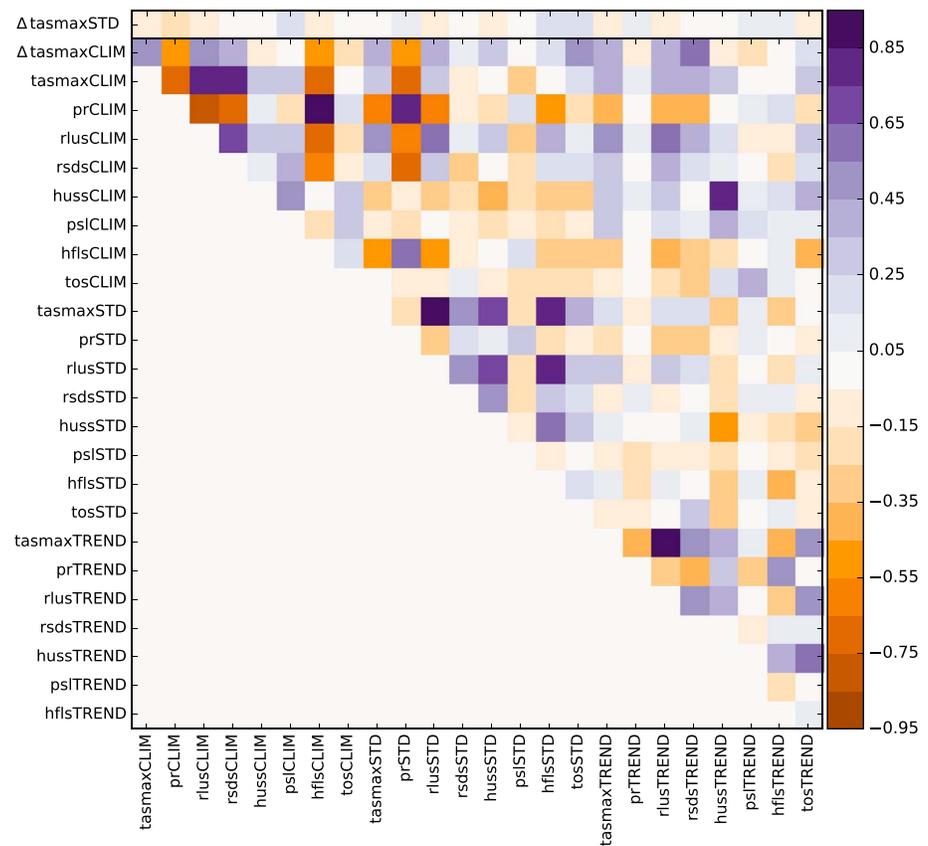


Figure 1. Spearman rank correlation coefficients between different diagnostics for North America region. Time period for all models is historical (1980–2014) except for Δ s which correspond to changes from historical to future (2065–2099) period.

Baker and Taylor (2016) use R^2 to evaluate performance metrics, based on the metric's ability to improve metric-weighted ensemble mean simulations. While we do not evaluate our RMSE metric itself, this index can also be useful to evaluate if the weighted mean is improved compared to the nonweighted mean.

3. Results

3.1. Correlations and Linear Regression Analysis

We use correlations and linear regressions to determine which diagnostics are relevant for the future change in maximum temperature climatology (tasmaxCLIM) and variability (tasmaxSTD). Based on correlation coefficients we preselect a subset out of a total 24 diagnostics shown in Table 1 and Figure 1. This decreases the dimension in our data set and is necessary for some of the feature selection methods presented in supporting information section S2 because not all of them can handle collinearity between features well or would randomly choose one of the correlated features. Table 1 lists all tested diagnostics and the correlation values of the area means over North America (NAM, defined as 25°N–50°N and 140°W–55°W) with tasmaxCLIM and its change into the future.

Figure 1 shows the correlation coefficients for Δ tasmaxSTD (first row) and Δ tasmaxCLIM (second row) with historical values of all other tested diagnostics including tasmaxCLIM and tasmaxSTD. Thereby, we identify the diagnostics relevant for the future changes of tasmaxCLIM and tasmaxSTD (high correlations desired). Figure 1 also shows which diagnostics are highly correlated and, thus, do not provide additional information and might cause problems because of collinearity (high correlations not desired). We perform our preselection by (1) removing diagnostics that are not correlated with Δ tasmaxCLIM and Δ tasmaxSTD. We define “not correlated” as not statistically significant at a 0.05 level (Δ tasmaxCLIM) and 0.1 level (Δ tasmaxSTD; see Table 1 for exact correlations and p values of tasmaxCLIM). Note that we choose a higher p level for tasmaxSTD since correlations are in general smaller for tasmaxSTD; (2) removing one (the second mentioned) of the very strongly correlated (≥ 0.85) diagnostics pairs (tasmaxCLIM and rsdsCLIM, prCLIM and hfisCLIM, tasmaxSTD

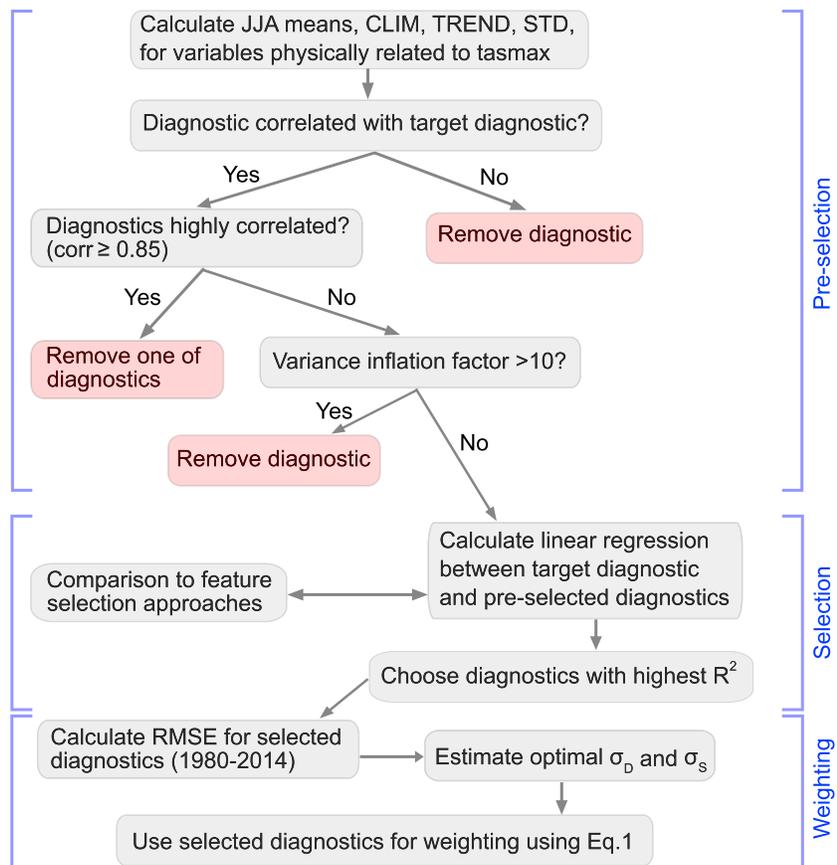


Figure 2. Flow chart showing all the steps in the analysis including the steps undertaken for the preselection of the diagnostics. JJA = June to August; CLIM = climatologies; TREND = linear trend; STD = standard deviation; RMSE = root-mean-square error.

and rluSTD, hussCLIM and hussTREND, and tasmaxTREND and rluTREND). In addition, we check the variance inflation factor ($VIF = \frac{1}{1-R^2}$, see supporting information section S3 for further details) and remove further diagnostics in case $VIF > 10$ ($\approx R > 0.95$, which is the case for rluCLIM here, which is correlated to tasmaxCLIM). We use this step in addition to removing diagnostics with correlations > 0.85 to ensure a lot of independent information from our potential predicting diagnostics and to further reduce the dimensions of the data set, even though collinearity does not pose a problem for the weighting method itself. Note that removing correlated predictors might not always be beneficial. In some cases, one might want to include correlated diagnostics if they represent different physical processes or can be constrained by different observational data sets. See Figure 2 for a flow chart of all the steps of the analysis we undertook in the example here.

For Δ tasmaxCLIM (1) leads to hussCLIM, psCLIM, tosCLIM rdsSTD, psSTD, prTREND, hussTREND, psITREND, hflsTREND, and tosTREND being removed. For tasmaxSTD none of the correlations are statistically significant on the 0.1 level. Hence, we are not able to determine relevant diagnostics for constraining the change in maximum temperature variability and stop the analysis of Δ tasmaxSTD at this point. For constraining Δ tasmaxCLIM (2) leaves the following diagnostics to consider: tasmaxCLIM, prCLIM, tasmaxSTD, prSTD, hussSTD, hflsSTD, tosSTD, tasmaxTREND, rdsTREND. Out of these diagnostics we choose one to six diagnostics to inform the weighting method.

The scatter plots in Figures 3 and 4 show the linear regressions as lines based on the ordinary least squares (OLS, gray) and Theil-Sen (black) regression models as well as the corresponding R^2 values. The dependency of the R^2 values to the chosen regression model is generally small, and the linear fits are similar as long as there are no outliers. The first two columns show the correlations with the change in tasmaxCLIM in the future period, whereas the third and fourth columns show the correlations between historical values. High correlations and large R^2 values ensure that the respective diagnostics are relevant for our target diagnostic.

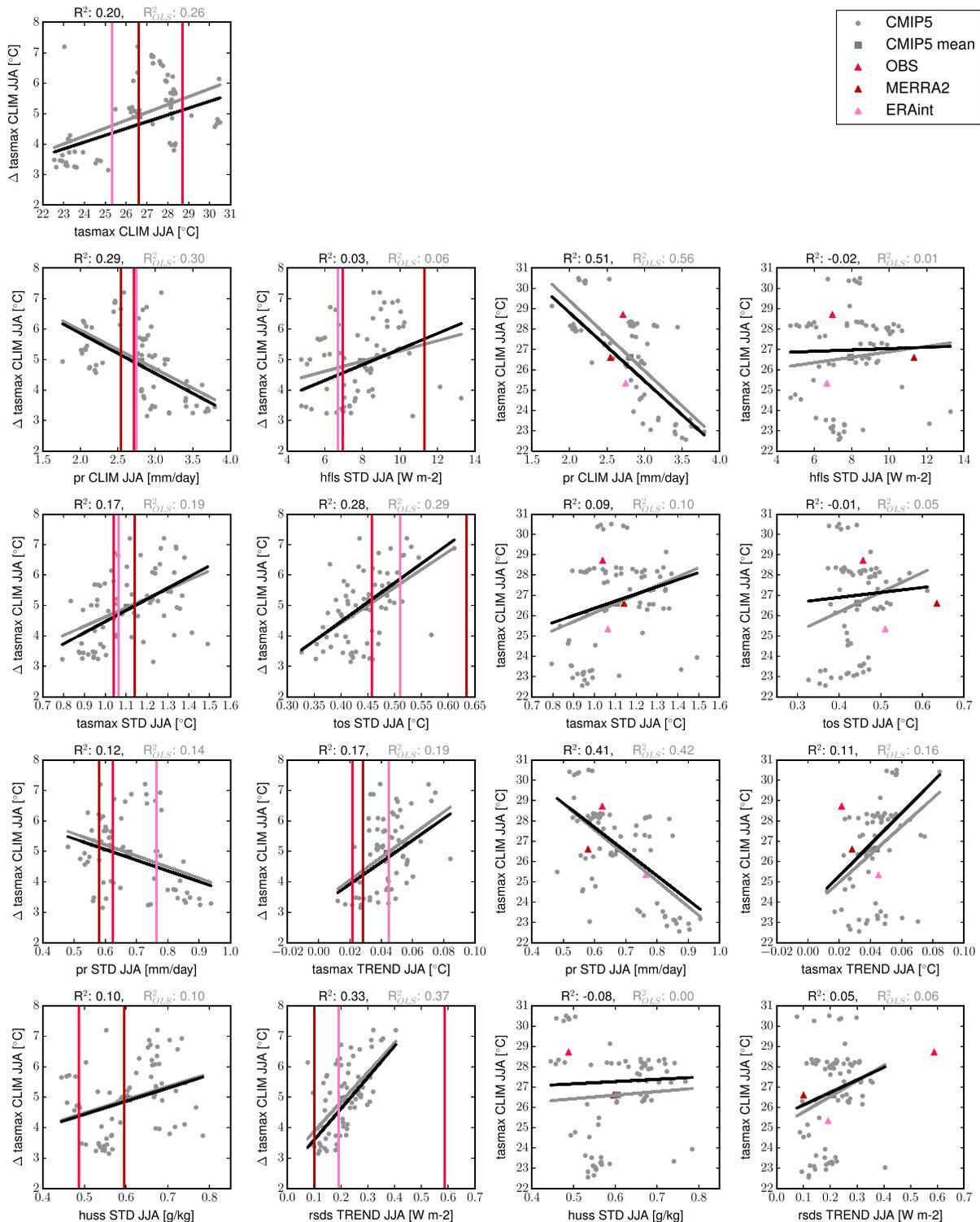


Figure 3. Linear regressions for tasmaxCLIM in the North America region. Grey lines and R^2 correspond to Ordinary Least Squares regression, while black lines and R^2 correspond to Theil-Sen regression method. The red lines (first two columns) and the red triangles (third and fourth columns) are the observational and reanalysis estimates (lines because we do not have observations for the delta terms).

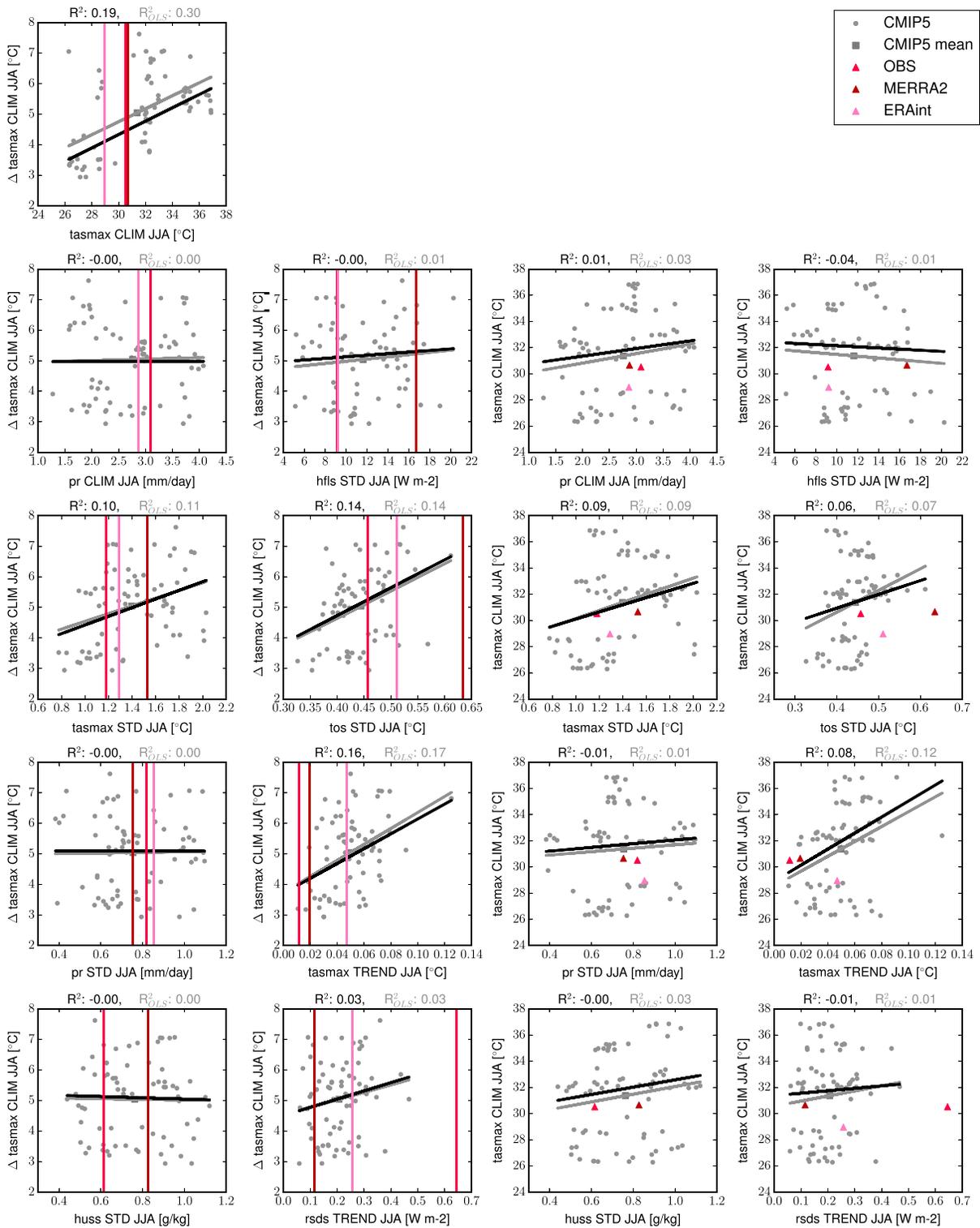


Figure 4. As in Figure 3 but for the Central North America region.

We also include estimates from observations and reanalysis data sets (red lines in the first two columns and triangles of the same colors in the third and fourth columns) showing for which diagnostics the observational and reanalysis products agree, and therefore providing an estimate of observational uncertainty. For rdsTrend the observational data sets are relatively far apart, in particular, for the small CNA region. Because the CERES observations cover a shorter time period, the trend was calculated for a shorter time period in this case, which could lead to the discrepancy between the data sets. Including the observational estimates also shows whether the models do a reasonable job at simulating these diagnostics in the historical period. The approach to choose the diagnostics does not formally take into account observational uncertainty. However, we choose variables with more reliable observations (e.g., tasmax, pr) over variables with less reliable observations (e.g., hfls) or a smaller number of estimates in our preselection when high correlations between diagnostics occur.

The change in tasmaxCLIM in NAM shows the highest correlations with: rdsTrend, prCLIM, tos STD, tasmaxCLIM, tasmaxSTD, and tasmaxTrend (Figure 3, columns 1 and 2). In addition, prCLIM is also strongly correlated with tasmaxCLIM in the historical period, and the linear regression gives an R^2 of 0.5. We also tested more sophisticated methods such as Lasso (Tibshirani, 1996), Random Forest (Breiman, 2001), or Stability selection (Meinshausen & Bühlmann, 2010; supporting information Text S2) to determine which diagnostics to include in our weighting method (supporting information Table S2). For Δ tasmaxCLIM these methods would lead to a similar choice of diagnostics for our analysis as by our expert judgment method (which is similar to the linear correlation, Corr., method in Table S2). While not all of the methods come up with the same order of importance of the diagnostics, they suggest the same four to five diagnostics as being the most important ones. However, we only use these methods to test the sensitivity of the results to the method used; we do not actually use the numbers in Table S2. For CNA the R^2 values are generally lower and Table S3 suggests a different order of diagnostics to include. Nevertheless, the statistical methods suggest similar diagnostics to be the most important ones. Therefore, the included predicting diagnostics for Δ tasmaxCLIM are as follows: tasmaxCLIM, rdsTrend, prCLIM, tosSTD, tasmaxSTD, and tasmaxTrend. The explained variance for the change in tasmaxCLIM by these diagnostics is up to 60% for NAM (see Table S4) and between 20% and 40% for CNA (Table S5), depending on the regression model used and the number of diagnostics included. Hence, there is an upper limit by how much these diagnostics are potentially able to constrain the change in tasmaxCLIM.

3.2. Weighting CMIP5 Projections for Tasmax Summer Climatology

We use equation (1) to obtain a weighted mean of summer tasmaxCLIM over North America. Figure 5 shows time series of tasmax for the nonweighted (black) and weighted (red) multimodel mean over the larger NAM region. First, we include tasmaxCLIM as predictor only (a–c) and then add rdsTrend as the second diagnostic (d–f). Finally, we also add the other predictors successively, which is illustrated in supporting information Figure S3. In the left column we use MERRA2, in the middle ERAint, and in the right column HadGHCND, and CERES (OBS) as observational data sets. Note that due to the short time period of the CERES data, the constraint based on rdsTrend from CERES should be interpreted with caution. The difference in absolute mean values between using the MERRA2 and ERAint reanalysis versus HadGHCND is due to the missing data over the great lakes in HadGHCND. The weighted multimodel mean is increased when using MERRA2 and OBS, while it is decreased when using ERAint. The first row shows the largest decrease in spread, when only tasmaxCLIM itself is used as predicting diagnostic. In the row below, the decrease in spread is smaller and is further decreased when adding additional diagnostics (see Figure 6). When the spread is decreased it is mostly at the high end for MERRA2 and ERAint, while for OBS the high and the low ends are changed.

Figure 6 shows the difference in mean and spread between weighted and nonweighted multimodel mean at the end of the century (2081–2100) for one to six diagnostics included to inform the weighting. As already seen in Figure 5, using one diagnostic shows the largest decrease in spread. Adding additional diagnostics to inform the weighting decreases the difference in mean as well as in spread. In NAM for MERRA2 and OBS the increase in mean is around 1 °C when using one diagnostic, and then this difference decreases and flattens out at around +0.5 °C when using five to six diagnostics. For ERAint the difference in mean is negative for one diagnostic and increases when adding more diagnostics to inform the weighting. For MERRA2 the spread is decreased for up to two included diagnostics, for ERAint the spread is decreased for up to three included diagnostics, and for OBS the spread is noticeably decreased only for one included diagnostic. For CNA the change in spread is even smaller than for NAM (Figure 6b). As soon as more than one diagnostic is included the spread is not decreased anymore and in some cases the spread is even increased in the weighted ensemble.

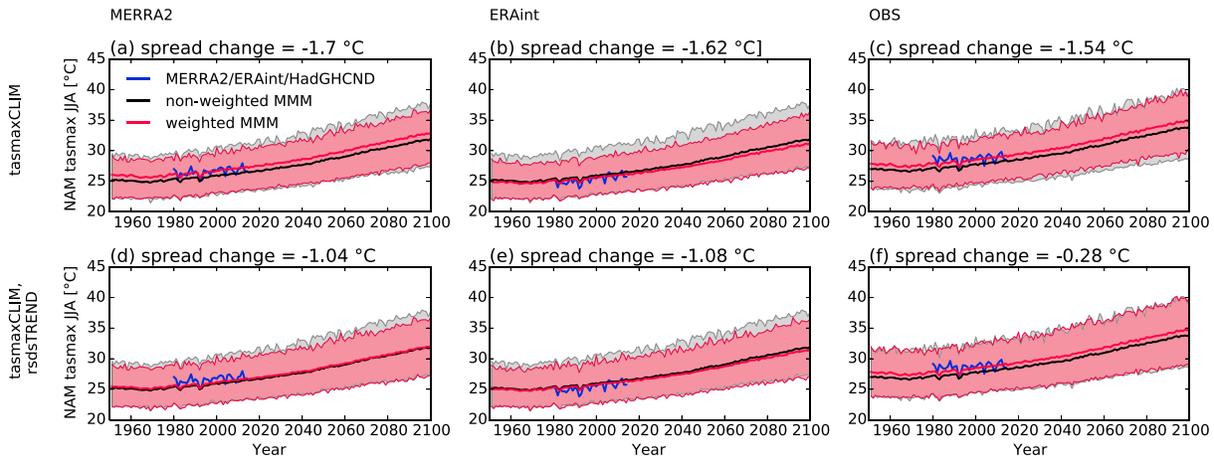


Figure 5. Weighted and nonweighted tasmax summer mean time series for North America. In the left column Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) is used as reference data set, in the middle column ERAinterim (ERAint), and in the right column HadGHCND/GPCP/CERES/HadISST. The change in spread is calculated over the last 20 years of the time series only and in °C.

The differences in the weighted multimodel mean depend on the data set used. Using MERRA2 leads to an increase, using OBS results in differences $< \pm 0.5$ °C, while using ERAint leads to a decrease when using one to two diagnostics and an increase when using three and more diagnostics.

Figures 7 and 8 show maps of the differences in Δ tasmaxCLIM (2065–2100 minus 1980–2014) between the weighted ensemble mean and the nonweighted ensemble mean. The differences between the weighted multimodel mean and the nonweighted multimodel mean Δ tasmaxCLIM is small, generally < 0.5 °C. In NAM all data sets suggest a smaller warming in central United States in the weighted multimodel mean when using more than three diagnostics (Figures 7j–7r), even if in some cases these differences are very small. In some cases there is a tendency for an enhanced warming in the northern part of the domain in the weighted multimodel mean. This enhanced warming is largest when using only tasmaxCLIM itself and MERRA2 or OBS (ERAint suggests a smaller warming in this case, Figures 7a–7c). These differences are similar when using CNA (Figure 8). However, Figure 8c shows a different sign than Figure 7c. Hence, depending on the region, the weighted multimodel mean change can be smaller or larger than the nonweighted multimodel mean change.

A way of evaluating the weighted versus nonweighted multimodel means is the error index I^2 described in section 2.5. The smaller the I^2 the larger the improvement in the weighted multimodel mean compared to the nonweighted for the target variable tasmax in the historical period compared to a certain data set ($I^2 = 1$ means no improvement). The error index is indicated in each panel at the top in Figures 7 and 8. For NAM I^2 is smaller than 1 in 14 out of 18 panels. For CNA I^2 is smaller than 1 in 9 out of 18 panels. For NAM all $I^2 > 1$ occur when using ERAint and more than two diagnostics (Figures 7h, 7k, 7n, and 7q). In these cases the weighted tasmaxCLIM moves farther away from the ERAint climatology over the historical period.

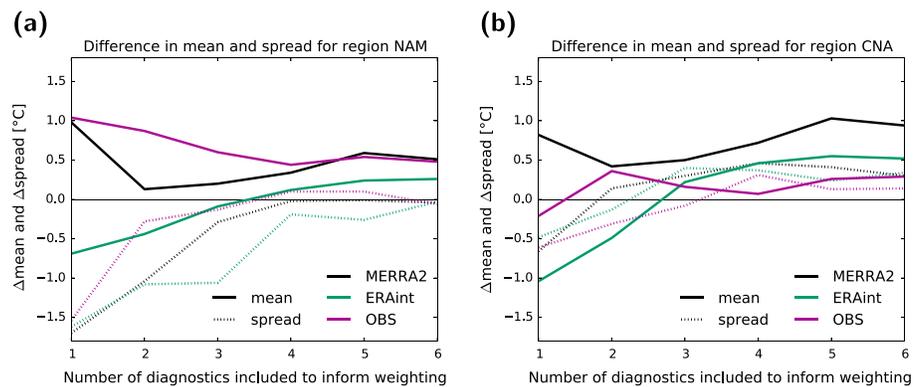


Figure 6. Difference in multimodel mean and spread in the last 20 years of the time series (2081–2100) between the weighted and nonweighted ensembles. Mean and spread differences are shown as function of the number of diagnostics used (a) for region North America (NAM), (b) for region Central North America (CNA).

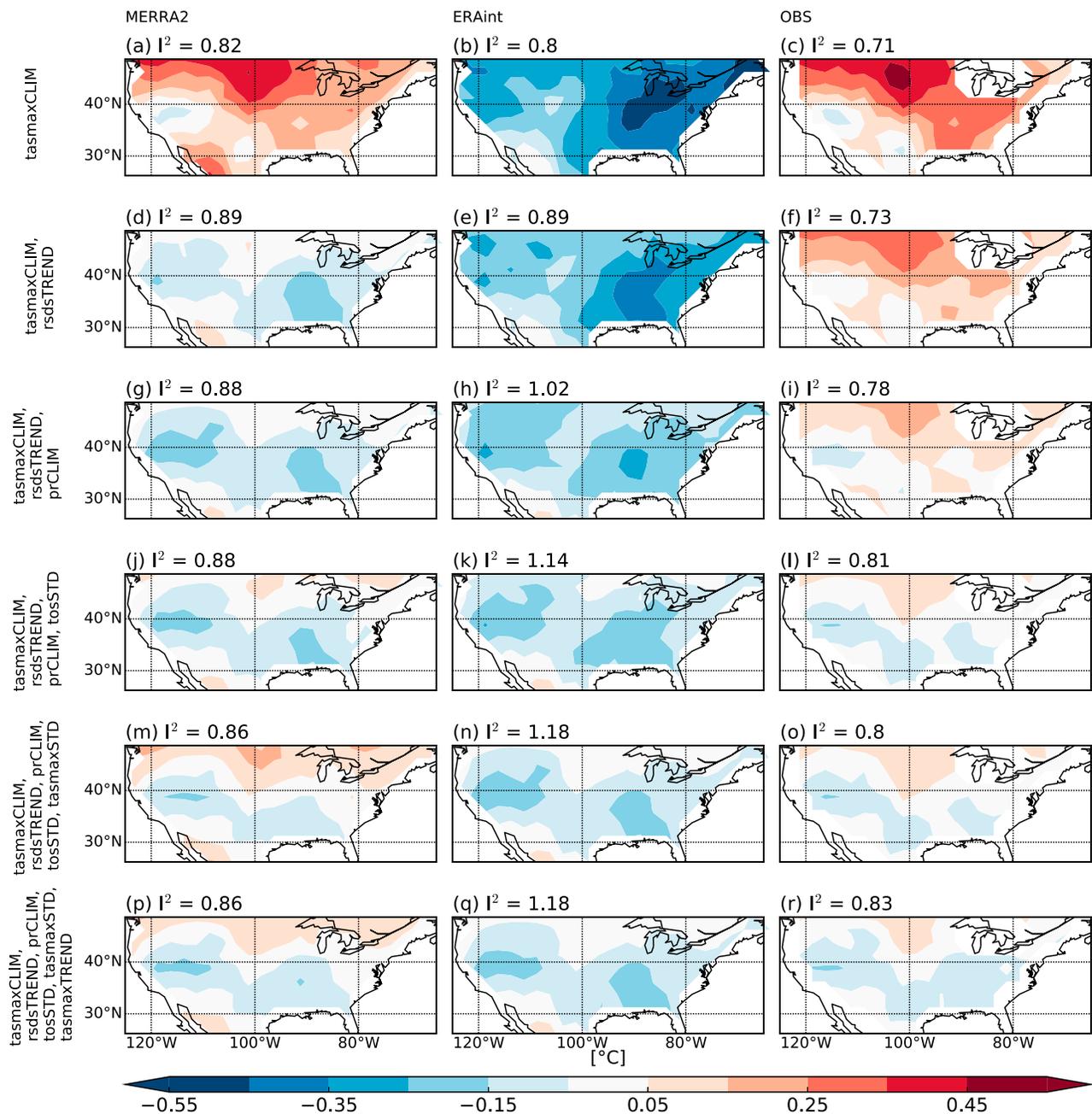


Figure 7. Difference between weighted and nonweighted $\Delta tasmxCLIM$ maps for North America. In the left column Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) is used as reference data set, in the middle column ERAinterim (ERAint) and in the right column HadGHCND/GPCP/CERES/HadISST.

For CNA this is the same, when using ERAint the weighted multimodel mean is only improved when using up to two diagnostics. In addition, using OBS data only improves the weighted multimodel mean when using tasmxCLIM itself. Hence, the choice of the region matters and can influence whether we find the weighting to lead to improvements in the historical period compared to observations. However, the smaller CNA region only includes a few grid points and is probably too small for this analysis.

There are also large differences between the data sets, and the choice of data set will influence the results. Using three to six predicting diagnostics, the pattern in the difference between the weighted and non-weighted multimodel mean changes become consistent over all data sets, even though for ERAint I^2 does not suggest an improvement. For CNA I^2 is smallest for a decrease in $\Delta tasmxCLIM$ using ERAint (Figures 8b

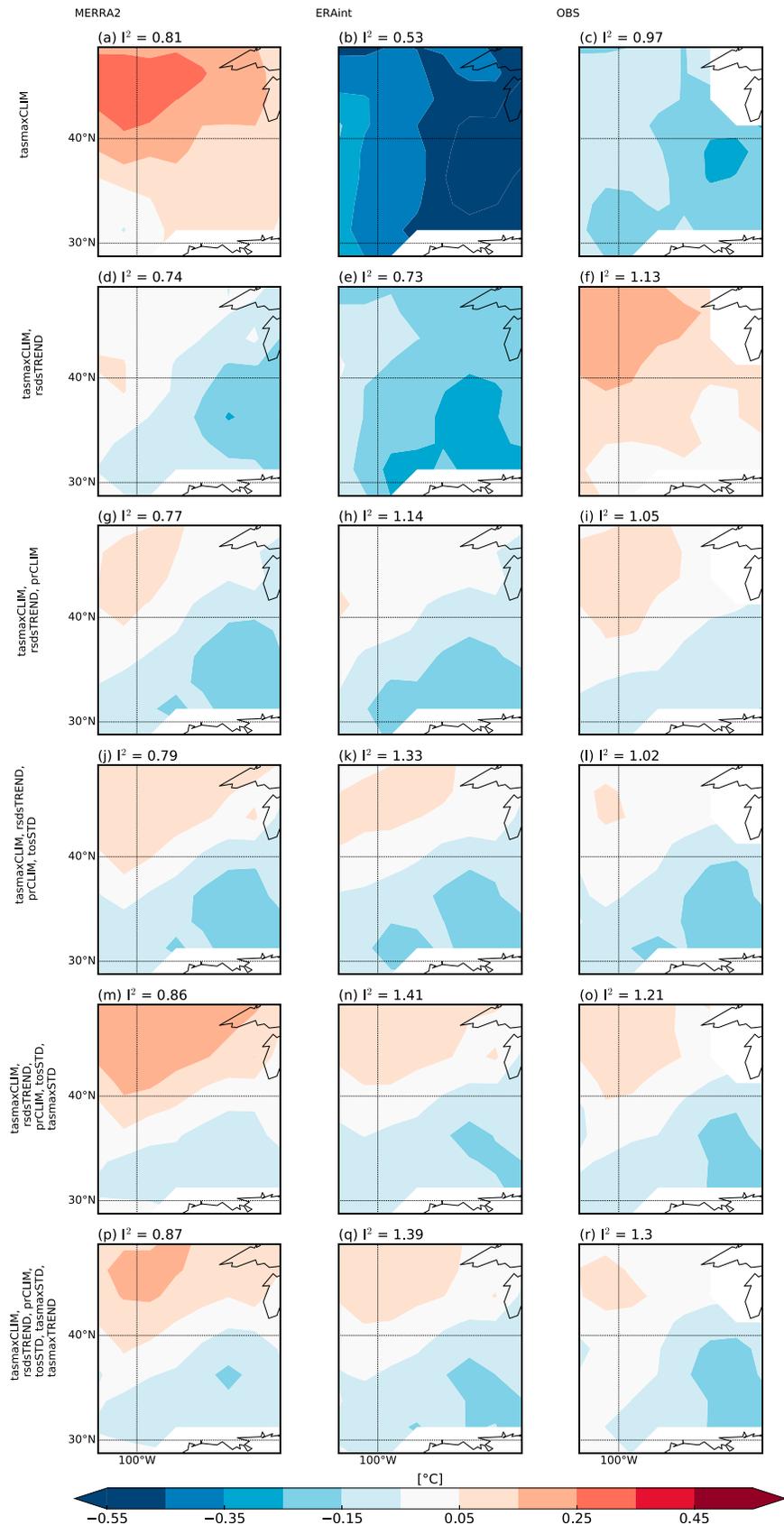


Figure 8. Same as Figure 7 but for Central North America.

and 8e). Including more than two diagnostics generally suggests a larger $\Delta\text{tasmaxCLIM}$ in the northern part of the domain and a smaller $\Delta\text{tasmaxCLIM}$ in the southern part of the domain in the weighted multimodel mean. However, only when using MERRA2 is R^2 smaller than 1 in these cases.

4. Discussion

We test multiple aspects that influence the application of the weighting method proposed by Knutti et al. (2017). The selection of diagnostics used in the weighting process is essential, and many different approaches are possible. While Knutti et al. (2017) base their choice purely on their physical understanding and expert knowledge, we use those diagnostics that are correlated with the projected change in our target diagnostics. While we do not opt for a pure statistical or machine learning method, we use linear regressions, similar to approaches used in the emergent constraints field, to determine which diagnostics are linked with our target diagnostic. In section 2.3 we explained the physical links between these diagnostics and tasmax . In addition, we use multiple statistical algorithms which generally agree with our final choice of diagnostics. While it would be possible to solely base the diagnostic choice on a statistical feature selection method, it has been shown that spurious correlations that occur by chance may result from such exercises (Caldwell et al., 2014; Masson & Knutti, 2013). In this study we only include simple diagnostics based on single variables. The maximum explained variance for $\Delta\text{tasmaxCLIM}$ in North America using these diagnostics is around 60%. Hence, the limited effect of the weighting method could also be influenced by the fact that the predicting diagnostics explain only up to 60% of the variability in $\Delta\text{tasmaxCLIM}$. It is possible that more elaborate diagnostics (such as ENSO indices, aridity index, and correlations between two variables) would lead to better results, and this should be further explored in the future.

Differences between weighted and nonweighted ensemble means when using different data sets become smaller the more diagnostics we take into account. While only using the target diagnostic itself has the largest effect on the spread, this is most likely to be overconfident. Only one variable needs to be modeled accurately in this case, and a single diagnostic is more likely to be tuned to or match by chance the observed historical values. Including more and more diagnostics that are relevant for determining our target diagnostic reduces the risk of spurious correlation causing overconfident results. However, it also reduces the effect of the weighting method. This could be due to multiple reasons such as the spread was not sufficient to begin with, or because the ensemble was already weighted because of good models being replicated more than the bad models, or because the observations are not long enough, not of sufficient quality (too much spread between data sets), or have too much variability to provide a constraint, or because we have already used most of the information from the observations in the model development, evaluation, and calibration. Sanderson et al. (2017) argued that it is less likely that one model performs very well across a large number of variables and diagnostics, therefore, using a larger number of diagnostics will decrease the effect of weighting. However, the interpretation of the spread is different between the unweighted and the weighted multimodel means. The spread of the unweighted multimodel mean is just a spread and is not a measure of uncertainty. It is an ad hoc measure of spread reflecting the ensemble design, or lack thereof, whereas the spread in the weighted multimodel mean can be interpreted as a measure of uncertainty given everything we know. The numbers may be similar, but the interpretation of the spread is very different, and we should have more confidence in the latter.

Borodina et al. (2017) found that aggregating multiple diagnostics, also across seasons, helps to capture the relevant processes to constrain an ensemble. Hence, it is important to use a large enough but not too large number of diagnostics. While it is possible to base this choice on physical understanding only, it can be helpful to investigate a large number of variables and diagnostics to choose the most important ones. A useful test and indication for robustness is to check whether the results change significantly when adding or removing a diagnostic. Adding the fourth and fifth diagnostic when weighting tasmaxCLIM still changes the maps of $\Delta\text{tasmaxCLIM}$ over NAM slightly, but the patterns across the data sets look more and more similar. As mentioned above, these four to five diagnostics explain up to 60% of the variance in the change in our target diagnostic from 1980–2014 to 2065–2099. However, it is still not possible to determine the ideal number of diagnostics to use. For future work, another perfect model test evaluating the future projections of the weighted mean versus the nonweighted mean could provide additional information on how much the weighted mean is improved compared to the nonweighted mean in a perfect model setup, similar to Karpechko et al. (2013).

Observational uncertainty is another important aspect to consider, and it would be dangerous to weight an ensemble too strongly based on performance when observational uncertainty is large. In our case this becomes clear in Figures 7 and 8. Depending on the choice of included diagnostics and observational product, changes in summer maximum temperatures are increased or decreased in the weighted ensemble compared to the nonweighted ensemble (Figures 7 and 8), because the three observational and reanalysis data sets are rather different for climatological JJA maximum temperatures in North America. Depending on which data set is used, one would conclude that the nonweighted CMIP5 ensemble mean is more positively (for ERA-Interim) or more negatively (for MERRA2, HadGHCND) biased (Figure S5). Therefore, weighting based on performance (and independence) will either increase or decrease the absolute weighted multimodel mean. If we trust the station data which HadGHCND is based on more than the reanalysis, which is a valid assumption, the weighted multimodel mean change would be increased with the weighting using tasmaxCLIM only. However, when we include more than one diagnostic the differences between using different observational or reanalysis data sets become smaller. The weighted multimodel mean indicates a small but robust decrease in Δ tasmaxCLIM in the central part of our NAM domain and the southern part of our CNA domain (Figures 7 and 8).

5. Summary and Conclusions

We use the weighting method proposed in Knutti et al. (2017) to constrain summer maximum temperature climatology over North America. Projections for summer maximum temperature in North America range from around 3 to 7 °C increase (\sim 5 °C in the multimodel mean) from the beginning (1980–2014) to the end of the 21st century in the nonweighted CMIP5 multimodel ensemble for the RCP8.5 scenario. For impact research, adaptation strategies, etc., it would be beneficial to decrease this uncertainty as much as possible. Earlier studies suggest the potential to constrain temperature or temperature extremes due to a link between historical model biases in temperature and land surface processes and future temperature increase (Christensen & Boberg, 2012; Sippel et al., 2017).

While model spread is hardly influenced by the weighting, we find the weighted multimodel means to suggest a small reduction of the projected warming in maximum temperature in central North America compared to the nonweighted multimodel mean. This result is in alignment with earlier studies. For instance, Sippel et al. (2017) suggested the CMIP5 multimodel mean to be positively biased due to an overestimation of the influence from the land surface and found projected changes in the warmest day of the year to be reduced in certain regions, among them CNA, when this bias is taken into account. Observational uncertainty is important to consider, especially if different data sets disagree in the region investigated, as they do for summer maximum temperature climatology in North America. The exact number of relevant diagnostics is dependent on the target diagnostic and the region and will need to be investigated on a case by case basis. While there should be an ideal number of diagnostics to include, true out-of-sample tests for the projections are needed to determine these numbers. This is an important avenue to investigate in future studies. In addition, combining the CMIP5 models with a large initial conditions ensemble would allow to investigate the uncertainty from natural variability and to determine if internal variability can be large enough so that model runs will be considered independent even though they come from the same model.

These kind of weighting approaches can help to increase our confidence in future projections from multimodel ensembles. While the spread did not change significantly due to the weighting in our example, the interpretation of the spread is different in the weighted multimodel mean. The spread of the weighted multimodel mean can be interpreted as a measure of uncertainty given everything we know, while in the nonweighted case the spread is not a measure of uncertainty. An important assumption behind this method is that models performing well in historical simulations also have skill in projecting future changes. While we are not able to demonstrate this (yet) using observations, it is difficult to have confidence in models that are not able to reproduce the historical climate reasonably well. The correlation argument in emergent constraints methods is one way of establishing a link between historical and future periods, and it can be tested (to the degree that the models are not all biased in a similar way) in the perfect model approach. To be able to determine which models perform well in the historical climate, we need reliable observations, including the uncertainties associated with these observations. Knowing the uncertainties attached to the observations would help to determine which models are far away from those observations. Additionally, as we have shown,

it is not enough to have good observations for one target variable, but we need reliable observations for all relevant variables. For many climate variables observational records are spatially sparse, time series have gaps and satellite records are still short. While in particular the satellite records are constantly improving, we need to make sure that observations are continued and the data are made available for research.

Acknowledgments

This research was funded by the European Union's Horizon 2020 research and innovation program under grant agreement 641816 (CRESCENDO). Nadja Herger acknowledges the support of the Australian Research Council Centre of Excellence for Climate System Science(CE110001028). We acknowledge the World Climate Research Program's Working Group on Coupled Modeling, which is responsible for CMIP, and we thank the 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 the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. CMIP data can be obtained from <http://cmip-pcmdi.llnl.gov/cmip5/>. MERRA-2 data can be obtained from <https://disc.sci.gsfc.nasa.gov/datasets?page=1&keywords=MERRA-2>. ERA-interim data can be obtained from apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/. HadGHCND data can be obtained from www.metoffice.gov.uk/hadobs/hadghcnd. CERES EBAF data can be obtained from https://ceres.larc.nasa.gov/order_data.php. GPCP data can be obtained from <http://www.esrl.noaa.gov/psd/data/gridded/data.gpcp.html>. HadISST data can be obtained from www.metoffice.gov.uk/hadobs/hadisst. We use python for data analysis and visualization, and functions used in the analysis can be found on github (https://github.com/ruthlorenz/weighting_CMIP).

References

Abramowitz, G., & Bishop, C. H. (2015). Climate model dependence and the ensemble dependence transformation of CMIP projections. *Journal of Climate*, 28(6), 2332–2348. <https://doi.org/10.1175/JCLI-D-14-00364.1>

Abramowitz, G., Leuning, R., Clark, M., & Pitman, A. (2008). Evaluating the performance of land surface models. *Journal of Climate*, 21(21), 5468–5481. <https://doi.org/10.1175/2008JCLI2378.1>

Adler, R., Huffman, G., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., et al. (2003). The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–Present). *Journal of Hydrometeorology*, 4, 1147–1167.

Alexander, L. V., Uotila, P., & Nicholls, N. (2009). Influence of sea surface temperature variability on global temperature and precipitation extremes. *Journal of Geophysical Research*, 114, D18116. <https://doi.org/10.1029/2009JD012301>

Annan, J. D., & Hargreaves, J. C. (2011). Understanding the CMIP3 multimodel ensemble. *Journal of Climate*, 24(16), 4529–4538. <https://doi.org/10.1175/2011JCLI3873.1>

Arblaster, J. M., & Alexander, L. V. (2012). The impact of the El Niño–Southern Oscillation on maximum temperature extremes. *Geophysical Research Letters*, 39, L20702. <https://doi.org/10.1029/2012GL053409>

Baker, N. C., & Taylor, P. C. (2016). A framework for evaluating climate model performance metrics. *Journal of Climate*, 29, 1773–1782. <https://doi.org/10.1175/JCLI-D-15-0114.1>

Borodina, A., Fischer, E. M., & Knutti, R. (2017). Emergent constraints in climate projections: A case study of changes in high-latitude temperature variability. *Journal of Climate*, 30(10), 3655–3670. <https://doi.org/10.1175/JCLI-D-16-0662.1>

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

Caesar, J., Alexander, L., & Vose, R. (2006). Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *Journal of Geophysical Research*, 111, D05101. <https://doi.org/10.1029/2005JD006280>

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., & Sanderson, B. M. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, 41, 1803–1808. <https://doi.org/10.1002/2014GL059205>

Christensen, J. H., & Boberg, F. (2012). Temperature dependent climate projection deficiencies in CMIP5 models. *Geophysical Research Letters*, 39, L24705. <https://doi.org/10.1029/2012GL053650>

Collins, M., Knutti, R., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., et al. (2013). Long-term climate change: Projections, commitments and irreversibility, *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1029–1136).

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553–597. <https://doi.org/10.1002/qj.828>

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, 38(3-4), 527–546. <https://doi.org/10.1007/s00382-010-0977-x>

Dirmeyer, P. A., Jin, Y., Singh, B., & Yan, X. (2013). Evolving land-atmosphere interactions over North America from CMIP5 simulations. *Journal of Climate*, 26(19), 7313–7327. <https://doi.org/10.1175/JCLI-D-12-00454.1>

Donat, M. G., Pitman, A. J., & Seneviratne, S. I. (2017). Regional warming of hot extremes accelerated by surface energy fluxes. *Geophysical Research Letters*, 44, 7011–7019. <https://doi.org/10.1002/2017GL073733>

Fischer, E. M., & Knutti, R. (2014). Detection of spatially aggregated changes in temperature and precipitation extremes. *Geophysical Research Letters*, 41, 547–554. <https://doi.org/10.1002/2013GL058499>

Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., et al. (2016). North American extreme temperature events and related large scale meteorological patterns: A review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, 46(3-4), 1151–1184. <https://doi.org/10.1007/s00382-015-2638-6>

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8), 1095–1107. <https://doi.org/10.1175/2009BAMS2607.1>

Herger, N., Abramowitz, G., Knutti, R., Angéil, O., Lehmann, K., & Sanderson, B. M. (2018). Selecting a climate model subset to optimise key ensemble properties. *Earth System Dynamics*, 9, 135–151. <https://doi.org/10.5194/esd-9-135-2018>

Hirschi, M., & Seneviratne, S. I. (2010). Intra-annual link of spring and autumn precipitation over France. *Climate Dynamics*, 35(7–8), 1207–1218. <https://doi.org/10.1007/s00382-009-0734-1>

Horton, R. M., Mankin, J. S., Lesk, C., Coffel, E., & Raymond, C. (2016). A review of recent advances in research on extreme heat events. *Current Climate Change Reports*, 2(4), 242–259. <https://doi.org/10.1007/s40641-016-0042-x>

Karpechko, A. Y., Maraun, D., & Eyring, V. (2013). Improving antarctic total ozone projections by a process-oriented multiple diagnostic ensemble regression. *Journal of the Atmospheric Sciences*, 70(12), 3959–3976. <https://doi.org/10.1175/JAS-D-13-071.1>

Kato, S., Loeb, N. G., Rose, F. G., Doelling, D. R., Rutan, D. A., Caldwell, T. E., et al. (2013). Surface irradiances consistent with CERES-derived top-of-atmosphere shortwave and longwave irradiances. *Journal of Climate*, 26(9), 2719–2740. <https://doi.org/10.1175/JCLI-D-12-00436.1>

Kharin, V. V., Zwiers, F. W., Zhang, X., & Wehner, M. (2013). Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climate Change*, 119(2), 345–357. <https://doi.org/10.1007/s10584-013-0705-8>

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P., Hewitson, B., & Mearns, L. (2010). Good practice guidance paper on assessing and combining multi model climate projections. In *IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections* (pp. 15).

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10), 2739–2758. <https://doi.org/10.1175/2009JCLI3361.1>

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, 40, 1194–1199. <https://doi.org/10.1002/grl.50256>

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, 44, 1909–1918. <https://doi.org/10.1002/2016GL072012>

- Koster, R. D., Sud, Y. C., Guo, Z., Dirmeyer, P. A., Bonan, G., Oleson, K. W., et al. (2006). GLACE: The global land-atmosphere coupling experiment. Part I: Overview. *Journal of Hydrometeorology*, 7(4), 590–610. <https://doi.org/10.1175/JHM510.1>
- Loikith, P. C., & Broccoli, A. J. (2012). Characteristics of observed atmospheric circulation patterns associated with temperature extremes over North America. *Journal of Climate*, 25(20), 7266–7281. <https://doi.org/10.1175/JCLI-D-11-00709.1>
- Loikith, P. C., & Broccoli, A. J. (2014). The influence of recurrent modes of climate variability on the occurrence of winter and summer extreme temperatures over North America. *Journal of Climate*, 27(4), 1600–1618. <https://doi.org/10.1175/JCLI-D-13-00068.1>
- Loikith, P. C., & Broccoli, A. J. (2015). Comparison between observed and model-simulated atmospheric circulation patterns associated with extreme temperature days over North America using CMIP5 historical simulations. *Journal of Climate*, 28(5), 2063–2079. <https://doi.org/10.1175/JCLI-D-13-00544.1>
- Lorenz, R., Argüeso, D., Donat, M. G., Pitman, A. J., van den Hurk, B., Berg, A., et al. (2016). Influence of land-atmosphere feedbacks on temperature and precipitation extremes in the GLACE-CMIP5 ensemble. *Journal of Geophysical Research: Atmospheres*, 121, 607–623. <https://doi.org/10.1002/2015JD024053>
- Ma, L., Zhang, T., Frauenfeld, O. W., Ye, B., Yang, D., & Qin, D. (2009). Evaluation of precipitation from the ERA-40, NCEP-1, and NCEP-2 reanalyses and CMAP-1, CMAP-2, and GPCP-2 with ground-based measurements in China. *Journal of Geophysical Research*, 114, D09105. <https://doi.org/10.1029/2008JD011178>
- Masson, D., & Knutti, R. (2011). Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *Journal of Climate*, 24(11), 2680–2692. <https://doi.org/10.1175/2011JCLI3513.1>
- Masson, D., & Knutti, R. (2013). Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity. *Journal of Climate*, 26(3), 887–898. <https://doi.org/10.1175/JCLI-D-11-00540.1>
- Meehl, G. A., & Tebaldi, C. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, 305(5686), 994–997. <https://doi.org/10.1126/science.1098704>
- Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J., et al. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, 109, 213–241. <https://doi.org/10.1007/s10584-011-0156-z>
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- Merrifield, A. L., & Xie, S.-P. (2016). Summer U.S. surface air temperature variability: controlling factors and AMIP simulation biases. *Journal of Climate*, 29(14), 5123–5139. <https://doi.org/10.1175/JCLI-D-15-0705.1>
- Mueller, B., & Seneviratne, S. I. (2012). Hot days induced by precipitation deficits at the global scale. *Proceedings of the National Academy of Sciences of the United States of America*, 109(31), 12,398–12,403. <https://doi.org/10.1073/pnas.1204330109>
- Mueller, B., & Seneviratne, S. I. (2014). Systematic land climate and evapotranspiration biases in CMIP5 simulations. *Geophysical Research Letters*, 41, 128–134. <https://doi.org/10.1002/2013GL058055>
- National Center for Atmospheric Research Staff (Eds.) (2017). The climate data guide: SST data: HadISST v1.1.
- Perkins, S. E., Argüeso, D., & White, C. J. (2015). Relationships between climate variability, soil moisture, and Australian heatwaves. *Journal of Geophysical Research: Atmospheres*, 120, 8144–8164. <https://doi.org/10.1002/2015JD023592>
- Perkins, S. E., & Fischer, E. M. (2013). The usefulness of different realizations for the model evaluation of regional trends in heat waves. *Geophysical Research Letters*, 40, 5793–5797. <https://doi.org/10.1002/2013GL057833>
- Perkins, S. E., Pitman, A. J., Holbrook, N. J., & McAneney, J. (2007). Evaluation of the AR4 Climate models simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate*, 20(17), 4356–4376. <https://doi.org/10.1175/JCLI4253.1>
- Pfiffroth, U., Mueller, R., & Ahrens, B. (2013). Evaluation of satellite-based and reanalysis precipitation data in the tropical Pacific. *Journal of Applied Meteorology and Climatology*, 52(3), 634–644. <https://doi.org/10.1175/JAMC-D-12-049.1>
- Rayner, N. A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, 108(D14), 4407. <https://doi.org/10.1029/2002JD002670>
- Sanderson, B., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, 10, 2379–2396. <https://doi.org/10.5194/gmd-10-2379-2017>
- Sanderson, B. M., Knutti, R., & Caldwell, P. (2015a). Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate*, 28(13), 5150–5170. <https://doi.org/10.1175/JCLI-D-14-00361.1>
- Sanderson, B. M., Knutti, R., & Caldwell, P. (2015b). A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, 28(13), 5171–5194. <https://doi.org/10.1175/JCLI-D-14-00362.1>
- Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., et al. (2010). Investigating soil moisture-climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3–4), 125–161. <https://doi.org/10.1016/j.earscirev.2010.02.004>
- Seneviratne, S. I., Wilhelm, M., Stanelle, T., van den Hurk, B., Hagemann, S., Berg, A., et al. (2013). Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. *Geophysical Research Letters*, 40, 5212–5217. <https://doi.org/10.1002/grl.50956>
- Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres*, 118, 2473–2493. <https://doi.org/10.1002/jgrd.50188>
- Sippel, S., Zscheischler, J., Mahecha, M. D., Orth, R., Reichstein, M., Vogel, M., & Seneviratne, S. I. (2017). Refining multi-model projections of temperature extremes by evaluation against land-atmosphere coupling diagnostics. *Earth System Dynamics*, 8(2), 387–403. <https://doi.org/10.5194/esd-8-387-2017>
- Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4), 485–498. <https://doi.org/10.1175/BAMS-D-11-00094.1>
- Tebaldi, C., Hayhoe, K., Arblaster, J. M., & Meehl, G. A. (2006). Going to the extremes. *Climatic Change*, 79(3–4), 185–211. <https://doi.org/10.1007/s10584-006-9051-4>
- Teng, H., Branstator, G., Meehl, G. A., & Washington, W. M. (2016). Projected intensification of subseasonal temperature variability and heat waves in the Great Plains. *Geophysical Research Letters*, 43, 2165–2173. <https://doi.org/10.1002/2015GL067574>
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
- Wang, K., & Dickinson, R. E. (2012). A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Reviews of Geophysics*, 50, RG2005. <https://doi.org/10.1029/2011RG000373.1>
- Waugh, D. W., & Eyring, V. (2008). Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, 8(18), 5699–5713. <https://doi.org/10.5194/acp-8-5699-2008>