Climate Model Evaluation in the Presence of Observational Uncertainty: Precipitation Indices over the Contiguous United States

PETER B. GIBSON, DUANE E. WALISER, HUIKYO LEE, BAIJUN TIAN, AND ELIAS MASSOUD

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California

(Manuscript received 26 October 2018, in final form 27 February 2019)

ABSTRACT

Climate model evaluation is complicated by the presence of observational uncertainty. In this study we analyze daily precipitation indices and compare multiple gridded observational and reanalysis products with regional climate models (RCMs) from the North American component of the Coordinated Regional Climate Downscaling Experiment (NA-CORDEX) multimodel ensemble. In the context of model evaluation, observational product differences across the contiguous United States (CONUS) are also deemed nontrivial for some indices, especially for annual counts of consecutive wet days and for heavy precipitation indices. Multidimensional scaling (MDS) is used to directly include this observational spread into the model evaluation procedure, enabling visualization and interpretation of model differences relative to a “cloud” of observational uncertainty. Applying MDS to the evaluation of NA-CORDEX RCMs reveals situations of added value from dynamical downscaling, situations of degraded performance from dynamical downscaling, and the sensitivity of model performance to model resolution. On precipitation days, higher-resolution RCMs typically simulate higher mean and extreme precipitation rates than their lower-resolution pairs, sometimes improving model fidelity with observations. These results document the model spread and biases in daily precipitation extremes across the full NA-CORDEX model ensemble. The often-large divergence between in situ observations, satellite data, and reanalysis, shown here for CONUS, is especially relevant for data-sparse regions of the globe where satellite and reanalysis products are extensively relied upon. This highlights the need to carefully consider multiple observational products when evaluating climate models.

1. Introduction

Climate model evaluation is useful for characterizing the credibility of future model projections but is challenging when observational products used in the evaluation also have large uncertainties. Recent studies have highlighted substantial differences among several observational precipitation products, particularly in the tails of the distribution (i.e., precipitation extremes). This interproduct spread can be pronounced in both the spatial distribution of precipitation extremes (e.g., Newman et al. 2015; Prein and Gobiet 2017; Angélil et al. 2016; Henn et al. 2018) and in regionally or globally aggregated quantities (e.g., Sillmann et al. 2013; Donat et al. 2014; Kim et al. 2015; Herold et al. 2016).

Various sources of uncertainty exist in observational precipitation products. The development of gridded in situ products involves making a number of decisions when interpolating from point-based gauge measurements to gridded quantities. For example, the choice of interpolation scheme, and how this scheme is applied, is known to be particularly influential in the representation of precipitation extremes (Contractor et al. 2015; Avila et al. 2015). In complex terrain at high elevations, where rain gauge measurements are sparse and do not fully capture the spatial dependence of precipitation on elevation, the way precipitation is corrected for due to orographic enhancement of precipitation is particularly important (Newman et al. 2015; Henn et al. 2018). Precipitation gauges used as input to these products are also well known to suffer from undercatch and evaporation losses in certain conditions (Ciach 2003; Tang et al. 2018).

While satellite products overcome many spatial sampling and coverage issues of in situ observations, their measurements are indirect and inferred from signals

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-18-0230.s1>.

Corresponding author: Peter B. Gibson, peter.gibson@jpl.nasa.gov

emanating from the column interior (e.g., radar returns) or top of the atmosphere (e.g., cold cloud tops). Moreover, limitations for capturing precipitation extremes in satellite products are well documented (e.g., AghaKouchak et al. 2011, 2012; Behrangi et al. 2014). These issues stem from assumptions made in their retrieval algorithms, instrument limitations and inadequate temporal sampling (Tang et al. 2018). Long-term, climatologically relevant, precipitation estimates from satellite are “merged products” as they combine multiple sensors calibrated by precipitation gauge data where possible.

Reanalysis products are historical climate model simulations (from a single consistent model) that are constrained through data assimilation using various historical observations (e.g., satellite, radiosondes, air craft, wind profilers, and in situ observations). Modern reanalysis products differ in terms of the driving dynamical models, model parameterizations, model vertical and horizontal resolutions, model boundary conditions, data sources used in assimilation, data assimilation methodologies, and bias correction schemes among others. Consideration of a diverse range of observational and reanalysis products, each with various and often unique sources of uncertainty, enables model performance to be considered in the context of observational uncertainty (e.g., Herold et al. 2016; Angéilil et al. 2016; Herger et al. 2018). In data-sparse regions of the globe (such as large parts of Africa, South America, and Pacific islands), where reliable in situ observational products are often not available, the use of satellite and reanalysis products are extensively relied upon.

NA-CORDEX (Mearns et al. 2017) is the North American component of the Coordinated Regional Climate Downscaling Experiment (CORDEX; Giorgi and Gutowski 2015) program initiated by the World Climate Research Program (WCRP). A small but growing number of studies have begun analyzing precipitation and precipitation extremes over the contiguous United States (CONUS) in this model ensemble. Diaconescu et al. (2016) evaluated two Canadian RCMs driven by different reanalysis products. Model performance was found to be relatively independent of the particular driving reanalysis and more dependent on the region, season and particular aspect of precipitation. Whan and Zwiers (2017) also analyzed output from the same two Canadian RCMs and found often substantial differences between models in how precipitation variability responds to different modes of variability. Lucas-Picher et al. (2017) analyzed a single RCM at variable horizontal resolution and assessed the sensitivity of different modeled processes (e.g., orographic precipitation and land–sea breezes) to the resolution.

Many of the above studies that detail observational uncertainty in precipitation extremes have had a global focus (e.g., Sillmann et al. 2013; Donat et al. 2014; Herold et al. 2016; Angéilil et al. 2016), often restricting the choice of in situ observations to global but lower-resolution observational products. Other studies have focused on specific regions or catchments within CONUS with emphasis on hydrological applications (e.g., Henn et al. 2018). In terms of model evaluation of precipitation in NA-CORDEX, previous studies have only focused on a few individual RCMs driven by reanalysis (e.g., Diaconescu et al. 2016; Whan and Zwiers 2017).

The overall goal of this study is to provide a more comprehensive evaluation of all models within NA-CORDEX, focusing on CONUS-wide precipitation extremes. The model evaluation procedure is carried out with detailed examination of observational uncertainty derived from multiple products for multiple precipitation indices. To do so, we employ multidimensional scaling (MDS; Borg et al. 2017), a dimension reduction tool, and assess 1) how the spread among observations compares to the spread among models, 2) whether downscaling [RCM simulations driven by either global climate models (GCMs) or reanalysis] noticeably improves fidelity compared to the observational spread, and 3) whether improved resolution in RCMs adds value compared to the observational spread.

2. Methods

a. Observational and reanalysis data

Observational and reanalysis products used in this study are described in Table 1. Gridded in situ products are typically better suited for model evaluation (compared to individual station data) in that they more directly represent the average precipitation over a grid cell as simulated by climate models (Zhang et al. 2011). Gridded in situ products rely on similar rain gauge networks as input data (as described below) but employ substantially different interpolation schemes, elevation corrections, as well as a number of other differences in gridding and processing. Estimating long-term spatially complete precipitation trends from these datasets is challenging due to station location moves, a changing observation network and differing observation times, and other inhomogeneities which can be difficult to control for (<http://www.prism.oregonstate.edu>).

The Oregon State University Parameter-Elevation Regressions on Independent Slopes Model (PRISM) dataset (Daly et al. 2008) uses station data mainly from the Cooperative Observer Program (COOP) and Snowpack Telemetry (SNOTEL) networks, with a vast number

TABLE 1. Gridded observational and reanalysis products used in this study.

Dataset	Reference	Type	Original resolution	Complete temporal length	Temporal length used in this study
CPC Unified (CONUS)	Higgins et al. (2000)	In situ	$0.25^\circ \times 0.25^\circ$	1948–present	1981–2016
Daymet	Thornton et al. (2017)	In situ	~1 km	1980–2017	1981–2016
PRISM	Daly et al. (2008)	In situ	~4 km	1895–present	1981–2016
GPCP-1DD v1.2	Huffman et al. (2001)	Satellite	$1^\circ \times 1^\circ$	1997–2014	1998–2014
PERSIANN-CDR	Ashouri et al. (2015)	Satellite	$0.25^\circ \times 0.25^\circ$	1983–present	1998–2014
TRMM 3b42 v7 (Daily)	Huffman et al. (2007)	Satellite	$0.25^\circ \times 0.25^\circ$	1998–2016	1998–2014
ERA-INT	Dee et al. (2011)	Reanalysis	$0.75^\circ \times 0.75^\circ$	1979–present	1981–2016
MERRA	Rienecker et al. (2011)	Reanalysis	$0.5^\circ \times 0.66^\circ$	1979–2016	1981–2016
MERRA-2	Gelaro et al. (2017)	Reanalysis	$0.5^\circ \times 0.625^\circ$	1980–present	1981–2016
NARR	Mesinger et al. (2006)	Reanalysis	~32 km	1979–present	1981–2016

of smaller networks also used. Over the central and eastern CONUS the daily product also incorporates radar observations at 4-km resolution from the Advanced Hydrometeorological Prediction System (AHPS). PRISM uses a comprehensive linear precipitation–elevation correction scheme that applies weights (separately at each grid point) based on elevation. The station weights are a function of location to nearby stations, proximity to coast, topographic facets, boundary layer conditions, surrounding terrain height and other terrain features. The National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) Unified CONUS dataset (hereafter, CPC) uses station data from the U.S. unified rain gauge dataset, composed of multiple sources ([Higgins et al. 2000](#)). Various quality control measures are used to determine inclusion in the final dataset, including consistency with radar products not used as input data. A simple orographic correction is carried out based on the daily PRISM climatology. Last, Daymet ([Thornton et al. 1997, 2017](#)) uses COOP and SNOTEL station networks as input, correcting for elevation using a simple weighted local linear regression between precipitation and elevation. The weightings are less sophisticated than in PRISM and do not explicitly represent a wide range of nonmonotonic relationships ([Daly et al. 2008](#)).

Satellite products used in this study (Table 1) differ considerably in terms of input data and methodology. The Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks–Climate Data Record (PERSIANN-CDR) product ([Ashouri et al. 2015](#)) primarily relies on 3-hourly gridded satellite infrared measurements as well as monthly data (satellite and gauge) from the Global Precipitation Climatology Project (GPCP) for bias correction. An artificial neural network is applied in PERSIANN-CDR to improve infrared measurements which are longer in record than passive/active microwave sensors relied on heavily in other products. GPCP’s 1° daily product (GPCP 1DD,

[Huffman et al. 2001](#)) combines satellite infrared and microwave measurements scaled by GPCP monthly data (satellite and gauge). Uncertainties are larger in regions poleward of 40° latitude in GPCP 1DD, especially in the daily estimates, due to different input data availability and may bias daily precipitation over northern CONUS. The Tropical Rainfall Measuring Mission (TRMM) 3B42 product ([Huffman et al. 2007](#)) uses calibrated (against monthly gauge) and combined microwave precipitation estimates. Gaps in the microwave estimates are infilled by infrared estimates at a 3-h temporal window. While more recent satellite-based precipitation products exist [e.g., the Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG); [Skofronick-Jackson et al. 2017](#)] the short record length in the current version does not allow characterization of daily precipitation climatology required for the purpose of this study.

Modern global reanalysis products used in this study include ERA-Interim (hereafter, ERA-INT) ([Dee et al. 2011](#)) and MERRA-2 ([Gelaro et al. 2017](#)). The precipitation in these reanalysis products are strongly influenced by their atmospheric models since precipitation measurements are not directly assimilated. We note that MERRA-2 also has an observationally corrected precipitation variable, designed so that simulated precipitation biases do not heavily influence the forcing of the land surface and beyond, however this variable was not analyzed in this study. The earlier generation MERRA reanalysis ([Rienecker et al. 2011](#)) is also included in our analysis to quantify possible improvements in the updated MERRA-2. The final reanalysis product included was the high-resolution North American Regional Reanalysis (NARR; [Mesinger et al. 2006](#)). NARR most notably differs from the other reanalysis products described above in that it does assimilate CONUS rain gauge networks into its latent heating scheme, known to improve precipitation estimates ([Bukovsky and Karoly 2007](#)). The direct assimilation of observed precipitation

TABLE 2. ETCCDI indices used in this study (PR = precipitation)

Index	Long name	Definition	Units
cdd	Consecutive dry days	Maximum annual number of consecutive dry days ($PR < 1$ mm)	days
cwd	Consecutive wet days	Maximum annual number of consecutive wet days ($PR \geq 1$ mm)	days
prcptot	Annual total PR	Annual sum of daily PR	mm
r10mm	Number of heavy PR days	Annual number of days when $PR \geq 10$ mm	days
r20mm	Number of very heavy PR days	Annual number of days when $PR \geq 20$ mm	days
rx1day	Maximum 1-day PR	Annual maximum of 1-day PR totals	mm
rx3day	Maximum 3-day PR	Annual maximum of 3-day PR totals	mm
sdi	Simple daily intensity	Average PR rate on wet days ($PR \geq 1$ mm)	mm day ⁻¹

makes NARR more of a “hybrid” product compared to other traditional reanalysis analyzed in this study.

b. NA-CORDEX model ensemble

The NA-CORDEX model ensemble is composed of 6 RCMs (CRCM5, RCA4, RegCM4, WRF, CanRCM4, and HIRHAM5) driven by either reanalysis (ERA-INT) or one of five GCMs (HadGEM2-ES, CanESM2, MPI-ESM-LR, EC-EARTH, GFDL-ESM2M). The RCMs were run at either 0.44° (~ 50 km) or 0.22° (~ 25 km), with a single higher-resolution run (driven by ERA-INT) at 0.11° (~ 12.5 km). Since the main purpose of this study is model evaluation, all data were taken from the historical experiment (termed “eval” when driven by reanalysis) and not subject to any form of bias correction. Comparing RCMs driven by reanalysis with those driven by GCMs indicates the sensitivity of RCM performance to boundary conditions. RCMs that receive lateral boundary conditions from reanalysis might be expected to be closer to reality than those receiving lateral boundary conditions from GCMs because of extensive data assimilation in the reanalysis. Therefore, errors that develop in the RCM when driven by reanalysis represent errors are likely due to smaller-scale “physics” (including subgrid scale turbulence and other boundary layer processes that are heavily parameterized). In contrast, the RCMs driven by GCMs have propagating errors from both their lateral boundary conditions inherited by the driving GCM and developed by smaller-scale physics in the RCM. Another focus of the NA-CORDEX experimental setup is the sensitivity to RCM’s horizontal spatial resolution. Multiple GCM/RCM pairs (and ERA-INT-RCM pairs) have been run at different spatial resolutions (mainly 50 or 25 km) to allow direct assessment of potential added-value from increased resolution. For a detailed description of the individual RCM configurations within the NA-CORDEX ensemble, the reader is referred to <https://na-cordex.org/rcm-characteristics>.

c. ETCCDI precipitation indices

A number of precipitation indices defined by the World Meteorological Organization (WMO) Expert Team on Climate Change Detection and Indices (ETCCDI;

Zhang et al. 2011) were analyzed (Table 2). These indices span moderate to heavy precipitation extremes including drought-relevant metrics [e.g., consecutive dry days (cdd)]. Heavy precipitation indices from ETCCDI represent more moderate intensity extremes compared to extreme value theory approaches intended to characterize precipitation extremes at intensities greater than 1-yr return periods (e.g., Wehner 2013). ETCCDI indices have become popular in many applications, including comparing consistency between observational products (e.g., Herold et al. 2016), documenting long-term regional and global trends in climate extremes (e.g., Donat et al. 2014), detection and attribution of human influence on climate extremes (e.g., Lu et al. 2018), and climate model evaluation and future projections (e.g., Sillmann et al. 2013; Screen et al. 2015; Diaconescu et al. 2016; Alexander and Arblaster 2017).

Each of the ETCCDI indices (Table 2) was calculated from daily total precipitation resulting in an annual value of each index at each grid cell across CONUS. The “climatology” of each index was then calculated by averaging these annual values over time to give a single long-term mean value at each grid cell. Unless otherwise stated, we are referring to this climatological map of each index when discussing these indices. For both models and observations, ETCCDI indices were calculated on the raw grid then regridded using bilinear interpolation to a 1° CONUS land-only grid (i.e., second-step regridding). The particular “order of operations,” where indices were computed first before remapping to a common grid, has been implemented in other similar studies (e.g., Wehner 2013; Diaconescu et al. 2016; Alexander and Arblaster 2017) with advantages of this second-step regridding approach discussed in Diaconescu et al. (2015). However, to test the sensitivity of results to the particular order of operations chosen, we also analyzed results obtained from first-step regridding, where the original daily precipitation fields were first regridded to a common grid before computing indices (described in section 3). A 1° grid was chosen as this is the coarsest resolution of the observational products (Table 1). The 1° grid was also chosen since

interest was in evaluating the larger-scale (i.e., CONUS-wide) patterns in the climatology of these indices. If instead interest was in evaluating finer-scale precipitation features across individual catchments, remapping all datasets to a finer resolution would be required. However, to verify this, sensitivity tests to target resolution were also carried out, with target grid resolutions of 0.22° and 0.44° .

d. Measure of similarity (T-metric)

ETCCDI indices from observations and models were compared through the T-metric introduced in [Tian et al. \(2017\)](#). This simple metric accounts for differences in both the spatial pattern and amplitude between a pair of observations and models (or more generally between any two products), ranging from 0 (no agreement) and 1 (perfect agreement). [Tian et al. \(2017\)](#) also showed that this metric tended to result in separation between “good” and “bad” degrees of agreement better than other commonly used metrics. The T-metric between two climatologies f and r is calculated as

$$T = [(1 + R)/2] \times [1 - \text{MSE}/(\text{Bias}^2 + \sigma_f^2 + \sigma_r^2)], \quad (1)$$

where R is the linear correlation coefficient; Bias is the difference in area-weighted mean between f and r ; MSE is the area-weighted mean-square error between f and r ; and σ_f^2 and σ_r^2 are the variance of products f and r , respectively.

The T-metric was calculated between all pairs of observational products and models. Since $T(f, r) = T(r, f)$ the resultant matrix (i.e., the matrix generated from all pairwise comparisons using the T-metric, hereafter referred to as the T-matrix) is symmetric and nonnegative. Since the time period between observational products can differ ([Table 1](#)), sensitivity to the time period was also assessed (described below).

e. Multidimensional scaling

MDS is a dimension reduction and visualization algorithm/tool that aims to preserve and map all inter-object distances to a small number of dimensions (usually two-dimensions to aid visualization) ([Borg et al. 2017](#)). While MDS has been commonly used in other fields, very few studies have implemented MDS in the context of climate model evaluation (e.g., [Sanderson et al. 2015](#)). MDS is somewhat similar to principal component analysis (PCA) for dimension reduction, which is commonly implemented in climate science, but the distances input to MDS are not restricted to the correlation or covariance matrix (as in PCA). MDS is arguably more suited for interpreting differences between objects in low dimensional space, whereas PCA gives greater insight into the characteristics of the dimensions themselves (e.g.,

[Hout et al. 2013](#)). For the purpose of this study, the symmetric T-matrix was the distance metric input into MDS, performed separately for each ETCCDI index. After dimension reduction, the resultant distances between objects (i.e., between individual observational products and models) in the first two MDS dimensions enable a visual assessment of similarity (e.g., objects positioned closer together in MDS space share greater similarity compared to objects farther away). We applied MDS to the NA-CORDEX simulations and assessed the spread in observations relative to the spread in models, as well as the potential added value from downscaling and sensitivity to RCM resolution.

The scaling by majorization of a complicated function (SMACOF; [de Leeuw and Mair 2009](#)) solution to the two-dimensional metric MDS problem was implemented in this study using the R package SMACOF. This uses the majorization algorithm, an iterative approach to minimize stress, and is described briefly here. The first step involves defining starting configurations which can be random or predefined; here we use the classical scaling solution (e.g., [Torgerson 1958](#)) as starting configurations. In the second-step stress is computed, defined by the sum of squares. In the third step the Guttman transform ([Guttman 1968](#)) is computed for the configuration, which guarantees a series of nonincreasing stress values with a linear convergence rate. Steps two and three are then iterated until convergence. For a complete mathematical description of each step of the SMACOF approach, the reader is referred to [de Leeuw and Mair \(2009\)](#).

Although MDS visualizes aspects of similarity between all possible pairs of the datasets, a physical interpretation of the MDS dimensions can be somewhat challenging. The importance of this interpretation will depend on the particular questions being probed by the researcher. To improve interpretation and better understand why particular objects may be more/less similar, we have fit and overlaid nonlinear ordination surfaces (i.e., contours) in the MDS space. These ordination surfaces result from fitting generalized additive models (GAM) relating a third variable of interest to the two MDS dimensions. In particular, we assessed how both the area mean and area variance (i.e., the mean and variance calculated from the distribution of gridded values across CONUS) in the climatology of each ETCCDI index, computed for each object, relates to the first two MDS dimensions. Other variables may be chosen to generate ordination surfaces and further aid interpretation, provided that they are one-dimensional n -element vectors (where n is the number of objects/products input into MDS). Various goodness of fit statistics are commonly used to assess how well interproduct distances are

FIG. 1. (a) T-matrix (see section 2) for the climatology of the ETCCDI consecutive wet days index (cwd; days) calculated over P1 (1981–2016, upper triangle of matrix) and P2 (1998–2014, lower triangle of matrix); black shade indicates relatively poor agreement between pairs of observations, and white shade indicates relatively good agreement. The P1 and P2 mean values are shown, calculated as the mean T-metric value of all interproduct comparisons in in periods P1 and P2. (b) Example of a high T-metric value from (a) indicated by blue circle. (c) Example of a low T-metric value from (a) indicated by red circle.

preserved in the dimension reduction performed through MDS. In this study we assess the Kruskal stress (Borg et al. 2017) and R^2 values for each MDS fit to assess the association between the original interproduct distances and those after ordination.

3. Results

a. Observational and reanalysis product comparisons

We begin by calculating the T-metric [Eq. (1)] between each pair of observational and reanalysis products over periods P1 (1981–2016) and P2 (1998–2014) for the climatology of each of the ETCCDI indices (Table 2). As an example, Fig. 1a presents the T-matrix calculated from pairwise comparisons for the consecutive wet days (cwd) index. A large range of T-metric values are shown in Fig. 1a indicating both good agreement (i.e., high T-metric value) between some pairs of observational products but poor agreement (i.e., low T-metric value) between others. For example, the cwd climatology between CPC and PRISM shows good agreement ($T = 0.89$) highlighting agreement in both the spatial pattern and amplitude of the climatology (Fig. 1b). In contrast, the cwd climatology between CPC and MERRA shows poor agreement ($T = 0.13$) related to a substantial excess in the number of

consecutive wet days in MERRA over the south-central and southeast CONUS.

Since satellite-based products used in this study span a shorter time period than in situ and reanalysis products (Table 1), it was important to consider whether the satellite climatology of ETCCDI indices calculated over this shorter period could unfairly bias comparisons. The upper matrix of Fig. 1a shows pairwise T-metric comparisons between observations over period P1, while the lower matrix shows comparisons over period P2. Due to the shorter record length, satellite data products only span 1998–2014 in both upper and lower matrix of Fig. 1a. By comparing P1 and P2 in (upper and lower matrix) for the cwd index, it is shown that the climatology generated by shorter-duration satellite products (1998–2014) retain similar relationships with other observational products compared to the climatology generated over the longer period (1981–2016). This is evident in the mean T-metric distributions (i.e., the mean of all pairwise comparisons) for P1 and P2 ($P1 = 0.48$, $P2 = 0.49$) and with similarly small P1 versus P2 differences found across all other ETCCDI indices analyzed in this study (Fig. 2). The stability of these interproduct relationships across P1 and P2 indicates that the shorter record satellite products used (Table 1) have a suitably stable climatology to warrant intercomparison in this study.

FIG. 2. As in Fig. 1a, but for each ETCCDI index. Figure 1a (cwd) is reproduced here to aid comparisons between indices. As in Fig. 1, the P1 and P2 mean T-metric is shown above, calculated as the mean T-metric value of all interproduct comparisons in periods P1 (1981–2016, upper triangle of matrix) and P2 (1998–2014, lower triangle of matrix). Black shade indicates relatively poor agreement between pairs of observations, and white indicates relatively good agreement.

The T-matrices in Fig. 2 contain a wealth of information regarding observational product agreement and disagreement. Notably the P1 mean T-metric values, which give an overall measure of interproduct agreement for a given ETCCDI index, vary from 0.48 for *cwd* (i.e., relatively poor general agreement between products, as indicated by many darker shaded comparisons) to 0.91 (very good general agreement, as indicated by many lighter shaded comparisons) for total precipitation (*prcptot*). Heavy precipitation indices (i.e., *r20mm* and *rx1day*) also exhibit relatively poor agreement between observational products. It is perhaps not surprising that observational products display better agreement between one another for total precipitation compared to sequential indices (e.g., *cwd*) and other extremes (e.g., *rx1day*, *rx3day*). This may be related to cancellation of errors in total precipitation indices that only become apparent when analyzing sequential or more extreme indices.

In terms of specific observational products that contribute most to the interproduct spread, MERRA (followed by ERA-INT) displays the highest general disagreement with other observational products across most indices. There is also a reasonably large difference

between MERRA and MERRA2, with MERRA2 displaying better agreement with other observational products across all indices. In terms of satellite products, TRMM displays remarkably poor agreement with in situ products for *cwd* yet agrees well with in situ products for other indices including heavy precipitation indices (i.e., *rx3day*). This divergence for TRMM was found to be related to a deficit in the length of wet spells consistent across much of CONUS (see Fig. 4). Last, although in situ products often have relatively good agreement between each other in Fig. 2 (relative to other observational products), for some indices the level of disagreement increases (e.g., *cwd* and *sdii*).

Sensitivity testing was carried out to examine the influence of the chosen order of operations in regridding (Fig. S1 in the online supplemental material). In particular, Fig. 2 shows interproduct spread under second-step regridding whereas Fig. S1 shows interproduct spread under first-step regridding. In comparing the T-metric values between these, it is evident that second-step regridding (compared to first step) produces slightly larger interproduct differences on average (i.e., lower T-metric values). However, the rank order of average

interproduct similarity by ETCCDI index is retained irrespective of the order of operations. Furthermore, the products that stand out as having the highest levels of disagreement with other products across multiple indices (e.g., MERRA and ERA-INT) are also retained irrespective of the order of operations. These sensitivity results suggest that the particular order of operations does not strongly influence the degree of agreement (or lack of) between indices and between observational products described here.

Sensitivity testing was also carried out to examine the influence of the target grid resolution (Fig. S2). Figure 2 shows interproduct differences under second-step regridding to a 1° resolution target grid, whereas Fig. S2 shows interproduct spread under a 0.44° target grid resolution (also under second-step regridding). It is evident from these comparisons that the target grid resolution has very little influence on interproduct spread (average T-metric differences are smaller than 0.02). A similar result was found with a target grid resolution of 0.22° (not shown).

Although focus of this study is on the climatology of ETCCDI indices, we briefly illustrate that gridded annual trends in these observational products (in situ and reanalysis, over a common period 1981–2016) show large disparities across most indices (Figs. S3 and S4). These differences are particularly large for reanalysis products, where ERA-INT displays strong highly significant trends across certain locations not evident in other products. Conversely, the other reanalysis products MERRA and MERRA-2 tend to display abnormally weak and statistically insignificant trends across most of CONUS. In situ products also often disagree concerning the specific location and strength of trends for most indices. CPC and PRISM broadly agree on a general increase in precipitation falling on wet days (sdii) over the eastern half of CONUS but this is not found in Daymet or the reanalysis products. There is also some evidence of consensus across in situ products for a positive trend in very heavy precipitation days (rx1day and rx3day) across the northeastern parts of CONUS. However, the general lack of consensus in trends between the gridded observational products in this study prevents their use for model evaluation of trends at the grid cell scale. The finding that trends in different in situ products can show strong discrepancies is not unexpected (given changes in instrumentation, station location and other inhomogeneities that are not controlled for) and serves to highlight the need for caution when using any trend estimates from these products.

b. Model evaluation with Taylor diagram

Before moving to NA-CORDEX model evaluation through MDS, we briefly show a simple example illustrating how it can be challenging to draw robust

conclusions from a Taylor diagram (e.g., Taylor 2001) in the presence of large observational uncertainty. As discussed in Taylor (2001), in cases where the observational uncertainty is particularly large, the distribution of models on the Taylor diagram might be altered nontrivially if a different observational product were used as reference. To demonstrate this, we show a Taylor diagram approach for probing the potential added value of increasing RCM resolution. To simplify the problem and interpretation, we first focus on a single RCM (driven by ERA-INT) run at three different RCM resolutions (50, 25, 12.5 km) for a single ETCCDI index (cwd, where the observational uncertainty was shown to be relatively large in Fig. 2). As shown in the Taylor diagrams in Figs. 3a–f, the particular choice of reference product can result in significantly rearranged positions of models on the Taylor diagram. The in situ products used as reference (CPC, PRISM, Daymet) give relatively consistent results to suggest that improving model resolution incrementally improves the model relative to observations. However, if any of the satellite products were used as reference, a different conclusion could be reached regarding the sensitivity of model performance to increasing model resolution. Since it would be inefficient to carry out such an analysis with multiple Taylor diagrams in a situation extended to multiple models and multiple ETCCDI indices, we demonstrate how MDS can summarize this information readily into a single diagram (Fig. 3g). In Fig. 3g the spread of observational products is indicated by the blue convex hull with vertices that indicate which products contribute strongly to this spread. By comparing the distance between incremental model resolution increases (red dots) relative to the distance spread of observational products (blue convex hull) in MDS space, it is seen that increasing model resolution increases closeness to in situ products but not necessarily to the satellite products. Additionally, the contribution from each observational product to the observational product spread can also be readily visualized in Fig. 3g. In the following section we extend this simple MDS example to the full NA-CORDEX ensemble, and include interpretation of the MDS dimensions.

c. Model evaluation with MDS

The cwd climatology for all observational products (Table 1) alongside GCMs, RCMs driven by GCMs, and RCMs driven by ERA-INT is shown in Fig. 4. By comparing these climatologies we may wish to ask several questions: How does the spread among observations compare to the spread among models? How does downscaling with RCMs (driven by GCMs or reanalysis) influence model fidelity with observations? What is the sensitivity to RCM resolution? Although information to answer these questions

FIG. 3. (a)–(f) Taylor diagrams for the ETCCDI index cwd with each panel using a different observational reference (REF) product. Smallest to largest circles correspond to ERA-INT and 50-, 25-, and 12.5-km RCM resolutions, respectively. The choice of reference product when assessing the sensitivity to model evaluation is often shown to be important in these diagrams. Note that ERA-INT is not shown in (f) since it was positioned far outside the plot domain. (g) An MDS approach is used to visualize the added value of increased model resolution (open red circles) relative to the observational spread (blue shaded convex hull) in a single diagram. Objects positioned closer in MDS space are more similar to each other based on the T-metric. ERA-INT is indicated by “E.”

is contained in Fig. 4, meaningful interpretation is difficult given the large number of interproduct comparisons to be made, as well as the divergence between observational products. Furthermore, we may wish to ask these questions not just for the cwd index but for all ETCCDI indices in Table 2. For this purpose, we use MDS for dimension reduction and data visualization, allowing us to readily probe

these important questions for model evaluation. Compared to the Taylor diagram, MDS efficiently displays model performance information against the spread in multiple observational products in a single diagram.

The input to MDS is all unique pairwise T-metric comparisons between the 44 products shown in Fig. 4 (i.e., $44 \times 43/2 = 946$ unique comparisons) calculated

FIG. 4. Climatology of cwd (days) for in situ observations (blue, 1981–2016), satellite (red, 1998–2014), reanalysis (green, 1981–2016), ERA-INT forced RCMs (gray, 1989–2009), and GCMs and GCM-forced RCMs (black, 1970–2005). “H” and “L” labels indicate higher (12.5 or 25 km) and lower (50 km) resolution simulations, respectively. This figure is reproduced in Fig. S5 for index sdii.

separately for the climatology of each ETCCDI index. Results for MDS are shown in Fig. 5 for select indices where the average observational interproduct spread was shown to be largest in Fig. 2 (i.e., cww, sdii, rx1day, and r20mm). The first column of Fig. 5 shows the MDS results performed separately for each index and with objects on each MDS map colored/shaped according to product type. Larger distances between objects on the MDS map indicate larger differences between the climatology of the ETCCDI indices, allowing visualization of relative similarities and differences between all products. In this way, the observational spread (which here includes satellite and in situ products, indicated by the red shaded convex hull) can be compared to the overall model spread (indicated by a gray-outlined convex hull). For several indices, the observational spread spans a considerable portion of the model spread. This highlights the importance of using multiple observational products for model evaluation, as the use of any single observational product for benchmarking could lead to a substantially different interpretation of model performance. Although this observational spread is primarily driven by differences in satellite products across most indices, this finding is especially relevant to other data-sparse regions of the globe where satellite products are often exclusively relied upon. It is noteworthy that reanalysis products (green circles in Fig. 5) often lie a relatively large distance (relative to the entire product spread) away from satellite and in situ products in the MDS, which calls for extra caution when using any reanalysis product to evaluate precipitation extremes in climate models.

We further show how the position of objects (i.e., observational products and models) in MDS space can relate to spatial mean and variance differences. The middle and right panels of Fig. 5 (similarly for Figs. 6 and 7) display the MDS space with contours overlaid for the object mean and variance (see section 2 for further details). As an example, for the cww index, in situ and satellite products display a relatively low mean number of consecutive wet days compared to models (as directly evident in Fig. 4 also). Reanalysis products for cww also stand out as having a relatively large area variance in consecutive wet days in MDS space (Fig. 5), related to a large positive bias in the number of consecutive wet days in south and southeast of CONUS as seen in Fig. 4.

To investigate different sources of RCM biases, Fig. 6 shows the same MDS maps as Fig. 5 but with RCMs now colored according to the particular driving GCM. This allows us to examine the extent to which improvements in downscaling relate to RCM choice, while bearing in mind important observational uncertainties. For some indices, especially sdii, downscaling ERA-INT with RCMs leads to robust improvements (gray crosses in

Fig. 6 are now positioned closer to the observational spread) irrespective of the choice of the RCM. However, for other indices, such as cww, particular RCMs can further enhance biases in ERA-INT by moving the RCM farther away from the observational spread. Specific examples of this are evident in Fig. 4, where the CanRCM4 RCM (driven by ERA-INT) introduces a broad positive bias in the number of consecutive wet days across states New Mexico and Colorado not evident in ERA-INT or any of the observational products. For RCMs driven by GCMs the choice of RCM can also either act to amplify existing biases in the GCM or reduce biases (Fig. 6). In general, RCMs downscaled by GCMs also show more consistent improvements for sdii but for most other indices downscaling tends to either show only very slight improvements or degraded performance. For example, some GCMs also show consistently degraded performance after downscaling for the cww index.

A valuable aspect of the NA-CORDEX ensemble is that multiple GCM-RCM pairings and reanalysis-RCM pairings have been run at different RCM resolutions, thereby enabling an assessment of added value in terms of increasing model resolution. For indices related to average and extreme precipitation intensities (i.e., those shown in Fig. 7) increased model resolution consistently increases the mean and spatial variability of precipitation intensities, often (but not always) improving model fidelity with observations. This is evident in Fig. 7, where blue colored objects (higher resolution) tend to shift in the direction of increased mean and variance, determined by the overlaid contours, compared to their lower-resolution pairs (red colored objects). Increased resolution tends to reduce biases in the average precipitation on wet days (sdii) especially in southeastern CONUS, where sdii is consistently too low in the simulations at lower resolution (Fig. S5). In contrast, increasing model resolution has no apparent added value across most RCMs for other indices (e.g., cww and cdd, Fig. S8).

Another important consideration for applying MDS is to check the goodness of fit statistics, describing how well the interobject distances are preserved when mapping from high-dimensional space to two-dimensional MDS space. For all ETCCDI indices, the goodness of fit statistics (Fig. S9) indicate that the MDS projections performed well in preserving interobject distances in the mapping (i.e., R^2 values were between 0.89 and 0.96). Retaining more than two dimensions (i.e., a third dimension) improves the fit somewhat for all indices but presents a clear trade-off with ease of visualization.

d. Relationship to topography

The native resolution of observations and models used in this study (e.g., Table 1) varies greatly. As

FIG. 5. Observation and model comparison through MDS—more similar objects are positioned closer together in each MDS plot. (left) Objects in MDS plot are colored/shaped according to product type, the red shaded convex hull indicates observational spread (in situ and satellite), the black shaded convex hull indicates in situ spread, and the gray convex hull indicates model spread. Also shown are contours overlaid on the same MDS plot in terms of (center) area mean and (right) area variance. For example, in situ products (blue circles) for the *cwd* index are shown to have a smaller area mean and area variance than reanalysis (green circles), which can be verified directly in Fig. 4 also. Figure S6 extends this figure to include all ETCCDI indices in Table 1.

discussed earlier, ETCCDI indices were computed on these native grids before regridding to a common 1° grid. Relatively coarse resolution models (RCMs or GCMs) may not capture important terrain features leading to biases in simulated precipitation at higher elevations. Similarly, for in situ observations, sparseness of rain gauges in high-elevation regions may cause interproduct

differences to become pronounced as sensitivity to the particular choice of interpolation scheme increases. To examine this, we computed interproduct differences [i.e., the coefficient of variation (COV), defined as the interproduct standard deviation divided by the interproduct mean] as a function of elevation in both observations and models (Fig. 8). Notably, in both models and observations,

FIG. 6. As in Fig. 5, but objects are here colored according to the driving GCM. The color of crosses for each RCM relate to which GCM (colored squares) provided lateral boundary conditions. As in Fig. 5, observational spread (in situ and satellite) is indicated by the shaded convex hulls. Figure S7 extends this figure to include all ETCCDI indices in Table 1.

interproduct differences increase (higher COV) considerably in high-elevation regions for the annual counts of heavy precipitation days (i.e., indices r10mm and r20mm). For observations, this suggests that increasing rain gauge coverage in high-elevation regions would likely reduce observational spread for these indices. For models, the spread remains even for high-resolution RCMs (models with 0.22° resolution or higher), suggesting that this resolution is still too coarse to capture important precipitation

features in high elevations, or that other structural model differences (other than resolution) favorably contribute to between model differences here.

4. Discussion

In this study we have implemented multidimensional scaling (MDS) for evaluating NA-CORDEX models in the context of observational uncertainty. We view MDS

FIG. 7. As in Fig. 5, but objects are here colored according to the RCM resolution (red = 50 km, blue = 25 km, purple = 12.5 km). Shapes indicate the particular GCM-RCM pair and reanalysis-RCM pair as shown in the legend (model names have been shortened to the first letter for convenience). As in Fig. 5, observational spread (in situ and satellite) is indicated by shaded convex hulls. Note that a different set of indices is presented here (compared to Fig. 5). Figure S8 extends this figure to include all ETCCDI indices in Table 1.

as a complementary tool to the valuable and well-known Taylor diagram. In the presence of observational uncertainty, one limitation of the Taylor diagram is that models are typically benchmarked against a single reference observational product. Rather than comparing multiple Taylor diagrams for each observational

product used as reference (e.g., as shown in Fig. 3), MDS aims to map all interproduct differences efficiently in a single diagram. As such, the use of MDS should reduce the risk of overinterpreting results when these are heavily dependent on the choice of observation product. MDS is also flexible to using other user-defined distance

FIG. 8. Spatially averaged COV (defined as interproduct standard deviation divided by interproduct mean) for different ETCCDI indices (y axis) across various elevation bins (x axis). (a) Observational products including in situ, satellite, and reanalysis ($n = 10$). (b) Model products including GCMs, GCM-forced RCMs, and ERA-INT-forced RCMs ($n = 34$). (c) Model products are high-resolution RCMs (25 km or greater, $n = 12$). (d) Model products are low-resolution RCMs (50 km, $n = 17$)

measures as input for comparing observations and models; this flexibility should benefit different applications seeking to compare datasets in different ways. One potential drawback of MDS in a practical setting, compared to a Taylor diagram for example, relates to the interpretation of the first two MDS dimensions. However, to circumvent this issue, we have further shown how simple measures (e.g., the area mean and variance) can be mapped to the MDS space to aid physical interpretation of the MDS dimensions (Figs. 5–7).

Applying MDS to the NA-CORDEX multimodel ensemble allowed us to draw a number of conclusions. First, we have shown that observational uncertainty is considerable for some precipitation indices [e.g., consecutive wet days (cww) and average precipitation rate

(sdii)] and that models are better evaluated against this MDS-derived “cloud” of observational uncertainty than against a single reference product (Figs. 2 and 5). Reanalysis products should be used with special caution, given the extent of their interproduct spread, although notable improvements for all precipitation indices were found comparing MERRA to MERRA2. Using an older suite of regional climate models over CONUS, Wehner (2013) also found model performance to depend somewhat on the choice of observational product, though only two different in situ products were compared.

In terms of RCMs in NA-CORDEX, we have shown important sources of uncertainty stem from both the large scale (i.e., choice of GCM) and the small scale (i.e.,

choice of RCM). The added value of dynamical downscaling in NA-CORDEX is apparent only for certain precipitation indices (e.g., *sdi*; Fig. 6). For these indices, dynamical downscaling tends to enhance both the area mean and variance of the precipitation climatology that is otherwise too low in the driving GCM (Fig. 6). This “drizzle problem,” where GCMs simulate too frequent low-intensity precipitation, has been widely reported (e.g., Stephens et al. 2010; Lorenz et al. 2014). Higher-resolution RCMs typically further increase precipitation rates (mean and extremes) compared to their lower-resolution RCMs counterparts (Fig. 7), with increases often in the southeast CONUS (Fig. S5). Wehner et al. (2010) found that increasing model resolution in a GCM improved the representation of very extreme precipitation in this region through cumulus processes that likely better represents the importance of Atlantic tropical cyclones. Bacmeister et al. (2014) suggested that a better representation of summertime disorganized convection in this region could be related to higher resolution, but analyzing subdaily simulated precipitation would be required to further test this in the present study.

However, it was also found that the increased precipitation in simulations at high resolution can often lead to overestimated precipitation during very extreme precipitation events [e.g., in annual 3-day maximum precipitation (*rx3day*); Fig. 7]. A similar finding was reported by Diaconescu et al. (2016) in the higher-resolution version of two Canadian RCMs for the *rx1day* index. In the present study, downscaling for other precipitation indices (e.g., *cwd*) by some RCMs (driven by either reanalysis or GCMs) was also shown to amplify existing biases or create new biases. This is similar to findings reported in Wehner (2013) where an earlier suite of regional models driven by reanalysis were found capable of developing their own errors resulting in substantially altered precipitation climatologies. Apparently, these problems have not been alleviated in some of the RCMs used in the updated NA-CORDEX model ensemble.

Other important questions can also be probed through MDS in the context of regional climate model evaluation. Despite being extremely computationally demanding, recent advances now enable convective parameterizations to be switched off in some model runs (convection permitting simulations; e.g., Prein et al. 2015). Since improvements in these simulations likely manifest in precipitation extremes at high spatial/temporal resolution (Kendon et al. 2017), where observational uncertainty often increases, we suggest MDS would be a particularly useful tool for model evaluation. Internal climate variability in models is also known to have significant influence on historical and future trends in climate extremes at the local–regional scale (e.g.,

Kay et al. 2015; Gibson et al. 2017). Due to computational demand, RCM ensembles like NA-CORDEX do not currently include multiple perturbed initial condition runs; however, should these become available in future ensemble projects, MDS would be well suited for comparing the importance of structural uncertainty in the models with uncertainty imposed by internal variability (e.g., Sanderson et al. 2015). The focus of this study has been on the large-scale CONUS-wide patterns of precipitation indices. It would be interesting to repeat this analysis for key regions at a higher spatial resolution (e.g., Prein et al. 2016) to investigate regional and finer-scale components in the added value of dynamical downscaling in NA-CORDEX. Our results alone do not rule out that added value from dynamical downscaling can manifest at these finer scales.

Last, our assessment of observational uncertainty in this study is based on agreement (or lack of) between a select number of existing observational products. The true uncertainty is likely larger than what was sampled here because the existing products are often not fully independent (i.e., they have similar data inputs despite being generated by different research groups) and do not span the full range of important decisions made when generating these products. None of the observational products used in this study provide direct uncertainty estimates to quantify the sensitivity to various important decisions made during product generation. It would be useful in future research to include the Newman et al. (2015) precipitation dataset (not analyzed in this study), which provides some uncertainty estimates of statistical errors in the dataset through ensemble generation, and map the spread of ensemble members to MDS space. The Newman et al. (2015) dataset was not included in this analysis partly because of its reduced temporal length relative to other in situ products, but future updates to this dataset would be valuable.

5. Conclusions

In this study we have applied MDS to assess the spread in precipitation indices found in the NA-CORDEX multimodel ensemble against the spread in observations and reanalysis. The distance metric input to MDS was the T-metric that depends on both spatial and amplitude agreement between product climatologies. The main conclusions are as follows:

- Observational product spread is sufficiently large for some indices (e.g., *cwd*, *sdi*, *rx1day*, *r20mm*) that model evaluation should explicitly consider this spread (Figs. 2 and 5).
- Modern reanalysis products display especially large spread relative to the model spread, and they should

be used with utmost caution for model evaluation of precipitation extremes (Figs. 2 and 5).

- Downscaling (driven by either ERA-INT or GCMs) leads to reasonably consistent improvements for some indices (e.g., sdii), but improvements are less consistent for some other indices (Fig. 6).
- Downscaling (driven by either ERA-INT or GCMs) can amplify biases for some RCMs and indices (e.g., cwd) (Fig. 6).
- Increased RCM resolution typically increases the average precipitation rate (sdii) and precipitation extremes (rx1day, rx3day, r20mm), sometimes improving model fidelity with observations (Fig. 7).
- The spread among observational products and among model products is found to be considerably larger in high-elevation regions for some precipitation indices (e.g., r10mm and r20mm; Fig. 8).

We suggest MDS can be a useful tool for researchers evaluating large climate model ensembles. MDS is especially useful when evaluating variables that exhibit relatively large observational uncertainties (e.g., climatology of regional precipitation extremes) that can make it challenging to draw robust conclusions from a Taylor diagram alone. We have discussed further avenues for using MDS in this context.

Acknowledgments. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. All data used in this study are publicly available. We thank the modelers and other NA-CORDEX contributors for their efforts and for access to their simulation output. The model evaluation procedure is currently being built into an existing open-source model evaluation toolkit developed by the authors: Regional Climate Model Evaluation System (RCMES; Lee et al. 2018; <https://rcmes.jpl.nasa.gov>).

REFERENCES

- AghaKouchak, A., A. Behrangi, S. Sorooshian, K. Hsu, and E. Amitai, 2011: Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *J. Geophys. Res.*, **116**, D02115, <https://doi.org/10.1029/2010JD014741>.
- , A. Mehran, H. Norouzi, and A. Behrangi, 2012: Systematic and random error components in satellite precipitation data sets. *Geophys. Res. Lett.*, **39**, L09406, <https://doi.org/10.1029/2012GL051592>.
- Alexander, L. V., and J. M. Arblaster, 2017: Historical and projected trends in temperature and precipitation extremes in Australia in observations and CMIP5. *Wea. Climate Extremes*, **15**, 34–56, <https://doi.org/10.1016/j.wace.2017.02.001>.
- Angéilil, O., and Coauthors, 2016: Comparing regional precipitation and temperature extremes in climate model and reanalysis products. *Wea. Climate Extremes*, **13**, 35–43, <https://doi.org/10.1016/j.wace.2016.07.001>.
- Ashouri, H., K.-L. Hsu, S. Sorooshian, D. K. Braithwaite, K. R. Knapp, L. D. Cecil, B. R. Nelson, and O. P. Prat, 2015: PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bull. Amer. Meteor. Soc.*, **96**, 69–83, <https://doi.org/10.1175/BAMS-D-13-00068.1>.
- Avila, F. B., S. Dong, K. P. Menang, J. Rajczak, M. Renom, M. G. Donat, and L. V. Alexander, 2015: Systematic investigation of gridding-related scaling effects on annual statistics of daily temperature and precipitation maxima: A case study for south-east Australia. *Wea. Climate Extremes*, **9**, 6–16, <https://doi.org/10.1016/j.wace.2015.06.003>.
- Bacmeister, J. T., M. F. Wehner, R. B. Neale, A. Gettelman, C. Hannay, P. H. Lauritzen, J. M. Caron, and J. E. Truesdale, 2014: Exploratory high-resolution climate simulations using the Community Atmosphere Model (CAM). *J. Climate*, **27**, 3073–3099, <https://doi.org/10.1175/JCLI-D-13-00387.1>.
- Behrangi, A., K. Andreadis, J. B. Fisher, F. J. Turk, S. Granger, T. Painter, and N. Das, 2014: Satellite-based precipitation estimation and its application for streamflow prediction over mountainous western U.S. basins. *J. Appl. Meteor. Climatol.*, **53**, 2823–2842, <https://doi.org/10.1175/JAMC-D-14-0056.1>.
- Borg, I., P. J. Groenen, and P. Mair, 2017: *Applied Multidimensional Scaling and Unfolding*. Springer, 122 pp.
- Bukovsky, M. S., and D. J. Karoly, 2007: A brief evaluation of precipitation from the North American Regional Reanalysis. *J. Hydrometeorol.*, **8**, 837–846, <https://doi.org/10.1175/JHM595.1>.
- Ciach, G. J., 2003: Local random errors in tipping-bucket rain gauge measurements. *J. Atmos. Oceanic Technol.*, **20**, 752–759, [https://doi.org/10.1175/1520-0426\(2003\)20<752:LREITB>2.0.CO;2](https://doi.org/10.1175/1520-0426(2003)20<752:LREITB>2.0.CO;2).
- Contractor, S., L. V. Alexander, M. G. Donat, and N. Herold, 2015: How well do gridded datasets of observed daily precipitation compare over Australia? *Adv. Meteor.*, **2015**, 325718, <https://doi.org/10.1155/2015/325718>.
- Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, <https://doi.org/10.1002/joc.1688>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- de Leeuw, J., and P. Mair, 2009: Multidimensional scaling using majorization: SMACOF in R. *J. Stat. Software*, **31** (3), <https://doi.org/10.18637/jss.v031.i03>.
- Diaconescu, E. P., P. Gachon, and R. Laprise, 2015: On the re-mapping procedure of daily precipitation statistics and indices used in regional climate model evaluation. *J. Hydrometeorol.*, **16**, 2301–2310, <https://doi.org/10.1175/JHM-D-15-0025.1>.
- , —, —, and J. F. Scinocca, 2016: Evaluation of precipitation indices over North America from various configurations of regional climate models. *Atmos.–Ocean*, **54**, 418–439, <https://doi.org/10.1080/07055900.2016.1185005>.
- Donat, M. G., J. Sillmann, S. Wild, L. V. Alexander, T. Lippmann, and F. W. Zwiers, 2014: Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets. *J. Climate*, **27**, 5019–5035, <https://doi.org/10.1175/JCLI-D-13-00405.1>.
- Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, <https://doi.org/10.1175/JCLI-D-16-0758.1>.

- Gibson, P. B., S. E. Perkins-Kirkpatrick, L. V. Alexander, and E. M. Fischer, 2017: Comparing Australian heat waves in the CMIP5 models through cluster analysis. *J. Geophys. Res. Atmos.*, **122**, 3266–3281, <https://doi.org/10.1002/2016JD025878>.
- Giorgi, F., and W. J. Gutowski Jr., 2015: Regional dynamical downscaling and the CORDEX initiative. *Annu. Rev. Environ. Resour.*, **40**, 467–490, <https://doi.org/10.1146/annurev-environ-102014-021217>.
- Guttman, L., 1968: A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, **33**, 469–506, <https://doi.org/10.1007/BF02290164>.
- Henn, B., A. J. Newman, B. Livneh, C. Daly, and J. D. Lundquist, 2018: An assessment of differences in gridded precipitation datasets in complex terrain. *J. Hydrol.*, **556**, 1205–1219, <https://doi.org/10.1016/j.jhydrol.2017.03.008>.
- Herger, N., G. Abramowitz, R. Knutti, O. Angélim, K. Lehmann, and B. M. Sanderson, 2018: Selecting a climate model subset to optimise key ensemble properties. *Earth Syst. Dyn.*, **9**, 135–151, <https://doi.org/10.5194/esd-9-135-2018>.
- Herold, N., L. V. Alexander, M. G. Donat, S. Contractor, and A. Becker, 2016: How much does it rain over land? *Geophys. Res. Lett.*, **43**, 341–348, <https://doi.org/10.1002/2015GL066615>.
- Higgins, R., W. Shi, E. Yarosh, and R. Joyce, 2000: Improved United States precipitation quality control system and analysis. NCEP/Climate Prediction Center ATLAS 7, https://www.cpc.ncep.noaa.gov/research_papers/ncep_cpc_atlas/7/toc.html.
- Hout, M. C., M. H. Papesh, and S. D. Goldinger, 2013: Multidimensional scaling. *Wiley Interdiscip. Rev.: Cognit. Sci.*, **4**, 93–103, <https://doi.org/10.1002/wcs.1203>.
- Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeorol.*, **2**, 36–50, [https://doi.org/10.1175/1525-7541\(2001\)002<0036:GPAODD>2.0.CO;2](https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2).
- , and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.*, **8**, 38–55, <https://doi.org/10.1175/JHM560.1>.
- Kay, J. E., and Coauthors, 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. Amer. Meteor. Soc.*, **96**, 1333–1349, <https://doi.org/10.1175/BAMS-D-13-00255.1>.
- Kendon, E. J., and Coauthors, 2017: Do convection-permitting regional climate models improve projections of future precipitation change? *Bull. Amer. Meteor. Soc.*, **98**, 79–93, <https://doi.org/10.1175/BAMS-D-15-0004.1>.
- Kim, J., J. Sanjay, C. Matmann, M. Boustani, M. Ramarao, R. Krishnan, and D. Waliser, 2015: Uncertainties in estimating spatial and interannual variations in precipitation climatology in the India–Tibet region from multiple gridded precipitation datasets. *Int. J. Climatol.*, **35**, 4557–4573, <https://doi.org/10.1002/joc.4306>.
- Lee, H., A. Goodman, L. McGibney, D. Waliser, J. Kim, P. Loikith, P. Gibson, and E. Massoud, 2018: Regional Climate Model Evaluation System. *Geosci. Model Dev.*, **11**, 4435–4449, <https://doi.org/10.5194/gmd-11-4435-2018>.
- Lorenz, R., A. Pitman, M. Donat, A. Hirsch, J. Kala, E. Kowalczyk, R. Law, and J. Srinovsky, 2014: Representation of climate extreme indices in the ACCESS1. 3b coupled atmosphere–land surface model. *Geosci. Model Dev.*, **7**, 545–567, <https://doi.org/10.5194/gmd-7-545-2014>.
- Lu, C., Y. Sun, and X. Zhang, 2018: Multimodel detection and attribution of changes in warm and cold spell durations. *Environ. Res. Lett.*, **13**, 074013, <https://doi.org/10.1088/1748-9326/aacb3e>.
- Lucas-Picher, P., R. Laprise, and K. Winger, 2017: Evidence of added value in North American regional climate model hindcast simulations using ever-increasing horizontal resolutions. *Climate Dyn.*, **48**, 2611–2633, <https://doi.org/10.1007/s00382-016-3227-z>.
- Mearns, L. M., and Coauthors, 2017: The NA-CORDEX dataset, version 1.0. NCAR Climate Data Gateway, accessed 18 January 2018, <https://doi.org/10.5065/D6SJ1JCH>.
- Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, <https://doi.org/10.1175/BAMS-87-3-343>.
- Newman, A. J., and Coauthors, 2015: Gridded ensemble precipitation and temperature estimates for the contiguous United States. *J. Hydrometeorol.*, **16**, 2481–2500, <https://doi.org/10.1175/JHM-D-15-0026.1>.
- Prein, A. F., and A. Gobiet, 2017: Impacts of uncertainties in European gridded precipitation observations on regional climate analysis. *Int. J. Climatol.*, **37**, 305–327, <https://doi.org/10.1002/joc.4706>.
- , and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, <https://doi.org/10.1002/2014RG000475>.
- , and Coauthors, 2016: Precipitation in the EURO-CORDEX 0.11° and 0.44° simulations: High resolution, high benefits? *Climate Dyn.*, **46**, 383–412, <https://doi.org/10.1007/s00382-015-2589-y>.
- Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624–3648, <https://doi.org/10.1175/JCLI-D-11-00015.1>.
- Sanderson, B. M., R. Knutti, and P. Caldwell, 2015: Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Climate*, **28**, 5150–5170, <https://doi.org/10.1175/JCLI-D-14-00361.1>.
- Screen, J. A., C. Deser, and L. Sun, 2015: Projected changes in regional climate extremes arising from Arctic sea ice loss. *Environ. Res. Lett.*, **10**, 084006, <https://doi.org/10.1088/1748-9326/10/8/084006>.
- Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh, 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>.
- Skofronick-Jackson, G., and Coauthors, 2017: The Global Precipitation Measurement (GPM) mission for science and society. *Bull. Amer. Meteor. Soc.*, **98**, 1679–1695, <https://doi.org/10.1175/BAMS-D-15-00306.1>.
- Stephens, G. L., and Coauthors, 2010: Dreary state of precipitation in global models. *J. Geophys. Res.*, **115**, D24211, <https://doi.org/10.1029/2010JD014532>.
- Tang, G., A. Behrangi, D. Long, C. Li, and Y. Hong, 2018: Accounting for spatiotemporal errors of gauges: A critical step to evaluate gridded precipitation products. *J. Hydrol.*, **559**, 294–306, <https://doi.org/10.1016/j.jhydrol.2018.02.057>.
- Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, <https://doi.org/10.1029/2000JD900719>.
- Thornton, P. E., S. W. Running, and M. A. White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.*, **190**, 214–251, [https://doi.org/10.1016/S0022-1694\(96\)03128-9](https://doi.org/10.1016/S0022-1694(96)03128-9).

- , M. M. Thornton, B. W. Mayer, Y. Wei, R. Devarakonda, R. S. Vose, and R. B. Cook, 2017: Daymet: Daily surface weather data on a 1-km grid for North America, version 3. ORNL Distributed Active Archive Center, accessed 1 February 2018, <https://doi.org/10.3334/ORNLDAAAC/1328>.
- Tian, B., and Coauthors, 2017: Development of a model performance metric and its application to assess summer precipitation over the U.S. Great Plains in downscaled climate simulations. *J. Hydrometeor.*, **18**, 2781–2799, <https://doi.org/10.1175/JHM-D-17-0045.1>.
- Torgerson, W. S., 1958: *Theory and Methods of Scaling*. Wiley, 460 pp.
- Wehner, M. F., 2013: Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. *Climate Dyn.*, **40**, 59–80, <https://doi.org/10.1007/s00382-012-1393-1>.
- , R. L. Smith, G. Bala, and P. Duffy, 2010: The effect of horizontal resolution on simulation of very extreme US precipitation events in a global atmosphere model. *Climate Dyn.*, **34**, 241–247, <https://doi.org/10.1007/s00382-009-0656-y>.
- Whan, K., and F. Zwiers, 2017: The impact of ENSO and the NAO on extreme winter precipitation in North America in observations and regional climate models. *Climate Dyn.*, **48**, 1401–1411, <https://doi.org/10.1007/s00382-016-3148-x>.
- Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers, 2011: Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 851–870, <https://doi.org/10.1002/wcc.147>.