Vol. 77: 23–43, 2019
https://doi.org/10.3354/cr01541

CLIMATE RESEARCH
Clim Res

Published online January 8

Weighting a regional climate model ensemble: Does it make a difference? Can it make a difference?

Melissa S. Bukovsky[1,*], Joshua A. Thompson[2], Linda O. Mearns[1]

[1]National Center for Atmospheric Research, Boulder, CO 80305, USA
[2]Global Weather Corporation, Boulder, CO 80301, USA

ABSTRACT: We explore the effect of weighting using the performance metrics developed for weighting in the ENSEMBLES program on the regional climate model (RCM) ensemble produced as a part of the North American Regional Climate Change Assessment Program (NARCCAP). We consider weighting a reanalysis-driven ensemble, as well as the effect on baseline GCM-driven ensemble mean bias, and mean climate change projections. This work evaluates when, where, and how weighting affects the ensemble mean results. We conclude that, in most cases, the metrics and resulting weights do not substantially differentiate the simulations. Also, the metrics do not always produce the same quality ranking results as an in-depth process-level analysis, implying that important, region-specific processes may be missing from this more universally applicable set of metrics. Moreover, it is found that when the metrics are used as weights, they do not consistently improve ensemble mean bias, as was found in the ENSEMBLES program. Furthermore, this analysis notably finds that weighting does not substantially change mean climate change projections unless all weight is applied to one or a few of the simulations that sit towards the extremes of the ensemble distribution, nor can it. We demonstrate why this is so.

KEY WORDS: Weighting · Metrics · Regional climate · NARCCAP

1. INTRODUCTION

When dealing with ensembles of global or regional climate models (GCMs or RCMs), the question of whether to weight or not to weight the ensemble often arises. Weighting better-performing members of an ensemble is assumed to create better ensemble means with less inherent uncertainty (e.g. Giorgi & Mearns 2002, Tebaldi et al. 2005). However, there is no consensus on how to weight ensembles (Knutti 2010), and while an unweighted ensemble mean often outperforms any given ensemble member (e.g. Knutti et al. 2010, Wehner 2013), a weighted ensemble mean may not produce a more robust ensemble mean or a substantially different result (e.g. Weigel et al. 2010, Chen et al. 2017).

In the European ENSEMBLES regional climate downscaling project, one objective was to explore the use of performance-based RCM weights for eventual use in creating probability distributions of regional projections, as weighting based on performance has been suggested as a method for possible uncertainty reduction (e.g. Giorgi & Mearns 2002, Knutti et al. 2010). Thus, 6 performance metrics were adopted to evaluate and weight the ENSEMBLES RCM simulations (Christensen et al. 2010). Five of the metrics were developed to measure the performance of the RCMs in simulating sub-GCM scale climate features, specifically emphasizing where the RCMs should be capable of adding value over GCM performance. The sixth metric was designed to test the performance of the RCMs in reproducing large-scale weather regimes, as they contribute to regional climate characteristics. Thus, multiple metrics were used to ensure a simulation could perform well by a range of measures, verifying reliability, and minimizing the possible effects of compensating systematic bias.

*Corresponding author: bukovsky@ucar.edu

In this work, we explore the effect of the ENSEMBLES metrics on the RCM simulations produced as a part of the North American Regional Climate Change Assessment Program (NARCCAP; Mearns et al. 2012, 2013). In ENSEMBLES, the goal was to investigate the consequence of weighting an RCM ensemble with weights specifically designed for RCMs. Our goal is the same, but we expand on the ENSEMBLES effort by applying the metrics to the NARCCAP simulations, thus allowing the metrics to be tested over a different region of the world, and over a greater number of diverse sub-continental scale climates. In Christensen et al. (2010), the European continent was broken down into 8 sub-regions. Here, we divide the USA into 16 sub-regions based on eco-climatic zone. Other efforts to examine the effect of similar performance-based metrics have usually employed fewer metrics — thus not effectively sampling measures of model performance and reliability — and/or have been limited to fewer regions, which limits general applicability (e.g. Holtanová et al. 2012, Foley et al. 2013, Eum et al. 2014, Gillet 2015, Chen et al. 2017, Ring et al. 2018). Furthermore, in the ENSEMBLES work, the metrics were applied only to the reanalysis-forced simulations. We expand on their method by applying the metrics, where appropriate, to the NARCCAP GCM-driven simulations. Given the results of this experiment using the ENSEMBLES metrics for weighting, we also briefly question when, where, and how any type of generic weighting scheme would ever significantly change ensemble mean results. Finally, we discuss the results of the metrics and weighting in comparison to a couple of in-depth analyses of the NARCCAP simulations. Therefore, our focus herein is not on providing an assessment of the differential credibility of the NARCCAP simulations, though this may be implied. We instead focus on the process, actual effect, and value and limitations of weighting such an ensemble via a metrics system that is designed to assess reliability and perhaps decrease uncertainty in future projections.

2. METHODS

2.1. NARCCAP

The ENSEMBLES performance metrics are calculated for the 6 NARCCAP RCMs forced with the National Centers for Environmental Prediction (NCEP)/Department of Energy (DOE) Reanalysis II (hereafter NCEP; Kanamitsu et al. 2002) and the 12 simulations that result from forcing the 6 RCMs with 4 different Coupled Model Intercomparison Program 3 (CMIP3) era GCM simulations (Mearns et al. 2007). Future simulations utilize the Special Report on Emissions Scenarios (SRES; Nakićenović et al. 2000) A2 emissions scenario; the twentieth-century (20c3m) emission representation is used for the baseline period. All simulations were completed with a 50 km horizontal resolution. A thorough description of these simulations is available in Mearns et al. (2012) or at www.narccap.ucar.edu. Table 1 provides an overview of the RCMs and GCMs; Table 2 presents the

Table 1. Regional and global climate models (RCMs and GCMs) used in the North American Regional Climate Change Assessment Program (NARCCAP), their identifying acronyms (RCM acronyms are as used in the NARCCAP model archive), and relevant references

Acronym	Details	References
RCMs		
CRCM	Canadian RCM	Caya & Laprise (1999)
ECP2	Experimental Climate Prediction Center's version of the Regional Spectral Model	Juang et al. (1997)
HRM3	Third-generation Hadley Centre RCM	Jones et al. (2003)
MM5I	Fifth-generation Pennsylvania State University – National Center for Atmospheric Research (NCAR) Mesoscale Model	Grell et al. (1993)
RCM3	International Centre for Theoretical Physics RCM version 3	Giorgi et al. (1993a,b), Pal et al. (2007)
WRFG	Weather Research and Forecasting model	Skamarock et al. (2005)
GCMs		
CCSM	NCAR CCSM version 3.0, run 5	Collins et al. (2006)
CGCM3	Canadian Global Climate Model version 3, run 4	Flato et al. (2000)
GFDL	GFDL climate model version 2.1, runs 1 and 2	GFDL GAMDT (2004)
HADCM3	Hadley Centre Climate Model version 3, this run is not part of the CMIP3 archive	Gordon et al. (2000), Pope et al. (2000)

Table 2. NARCCAP GCM-driven simulations. All combinations are marked with an X

	CCSM	CGCM3	GFDL	HADCM3
CRCM	X	X		
ECP2			X	X
HRM3			X	X
MM5I	X			X
RCM3		X	X	
WRFG	X	X		

RCM-GCM simulation combinations. When referring to an RCM and its driving GCM, the forcing simulation follows the RCM in lower case, e.g. WRFG-ccsm; otherwise, all acronyms are in upper case. When referring to an RCM forced by NCEP, in sections where only the NCEP-driven simulations are discussed, the RCM is referred to without the name of the driver attached.

The focus of this manuscript is on winter (December–January, DJF), summer (June–August, JJA), and annual (ANN) results, although spring (March–May, MAM) and autumn (September–November, SON) are included. For the NCEP-driven simulations, the NARCCAP ensemble spans 1980–2004, and for the GCM-driven simulations, 1971–1999. The future simulation period, used in the analysis of projections, spans 2041–2069. The metrics, due to limitations in the observationally based comparison datasets, are only calculated for 1980–1999 for the GCM-driven simulations.

Two of the RCMs (the CRCM and ECP2) use spectral nudging, which regularly forces the simulations back toward the large-scale driving conditions in the interior of the domain (instead of just at the boundaries). Therefore, these models are more constrained to follow the parent reanalysis or GCM, particularly at large-scales. This trait is relevant to the results herein.

2.2. Verification datasets

Several observationally based datasets are used when calculating the metrics. These include: 500 hPa geopotential height from the NCEP reanalysis; 1/8° spatial resolution, gridded precipitation and temperature from Maurer et al. (2002) (hereafter M-OBS; available at www.engr.scu.edu/~emaurer/gridded_obs/index_gridded_obs.html), bilinearly interpolated to a 1/2° grid; and monthly, 1/2° gridded precipitation and temperature from the Climate Research Unit (CRU) TS version 2.10 dataset (hereafter CRU; see Mitchell & Jones 2005; available at https://crudata.uea.ac.uk/cru/data/hrg/).

2.3. Regions

The ENSEMBLES weighting metrics were calculated over all of the regions shown in Fig. 1. Region abbreviations are given in Table 3. For more information on these regions, please see Bukovsky (2011).

2.4. ENSEMBLES weighting metrics

Six performance-based metrics were developed within the European ENSEMBLES project, which were then combined into individual RCM weights for use in multi-RCM analyses. An overview of these metrics and their resulting weights as applied to the

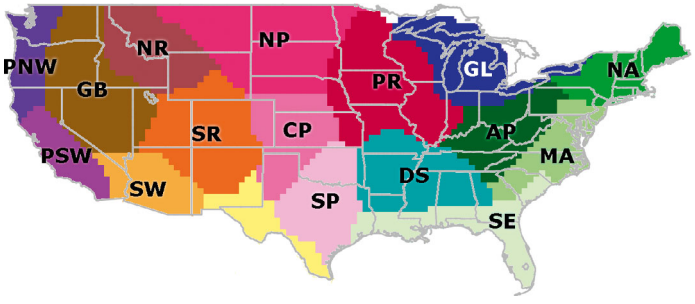

Fig. 1. Map of regions used in this study. Each regional abbreviation is defined in Table 3. The 1 region without an abbreviation (in yellow) was not used in this study as much of it lies outside of the USA

Table 3. Names of regions illustrated in Fig. 1

Abbreviation	Name
AP	Appalachia
CP	Central Plains
DS	Deep South
GB	Great Basin
GL	Great Lakes
MA	Mid-Atlantic
NA	North Atlantic
NP	Northern Plains
NR	Northern Rockies
PNW	Pacific Northwest
PSW	Pacific Southwest
PR	Prairie
SE	Southeast
SW	Southwest
SP	Southern Plains
SR	Southern Rockies

ENSEMBLES RCMs is given in Christensen et al. (2010). For further information on the metrics described below, including the motivation behind each metric, please see the included references. Every attempt was made to reproduce these metrics exactly; however, small changes were necessary to accommodate dataset and domain differences. Notation for each metric is also retained from Christensen et al. (2010) and the other related publications in the same special issue of Climate Research at www.int-res.com/abstracts/cr/v44/n2-3. A list of the metrics, including their identification tags, is given in Table 4 for reference.

2.4.1. Large-scale circulation metric (f1)

This metric tests a simulation's ability to reproduce weather regimes (WR) (Sanchez-Gomez et al. 2008, 2009). WR are identified in NCEP using a *k*-means clustering algorithm on 500 hPa daily geopotential height anomalies. A principle component (PC) analysis is used on the anomalies to reduce the number of degrees of freedom first. Enough PCs are kept to explain 90% of the variance. Clustering of the PC anomalies is done over an expanded North American/North Pacific region covering 10°–80° N and 170°–320° E for 1980–2004. Twelve WR clusters are calculated, following Riddle et al. (2013) who found that for annual analyses 12 clusters are optimal for North America. Europe, on the other hand, required 4 WR (Sanchez-Gomez et al. 2009). A map of the 12 cluster centroids is provided in Fig. 2.

Once the cluster centroids are identified, over a common NARCCAP domain covering 15.25°–75.25° N and 159.75°–29.75° W, daily 500 hPa geopotential height anomalies from the RCMs and NCEP are matched to a given WR cluster centroid. Each day is matched to the regime that has the highest spatial correlation over the common domain.

Five skill scores are then calculated from this chronology of WR for each RCM versus NCEP for June–September (JJAS), December–March (DJFM), and the year (taken as the JJAS and DJFM average), as done in ENSEMBLES. They include: a spatial correlation between NCEP and the RCM composite for each WR (f111), the difference between NCEP and an RCM in terms of the mean frequency of occurrence for each WR (f112), the difference in the mean persistence time for each WR (f113), an RCM/NCEP ratio of the variance of the timeseries of

Table 4. Reference list of metrics

Identifier	Metric
f1	Large-scale circulation
f2	Mesoscale metric
f3	Probability density distribution
f4	Extremes
f5	Temperature trends
f6	Annual cycle

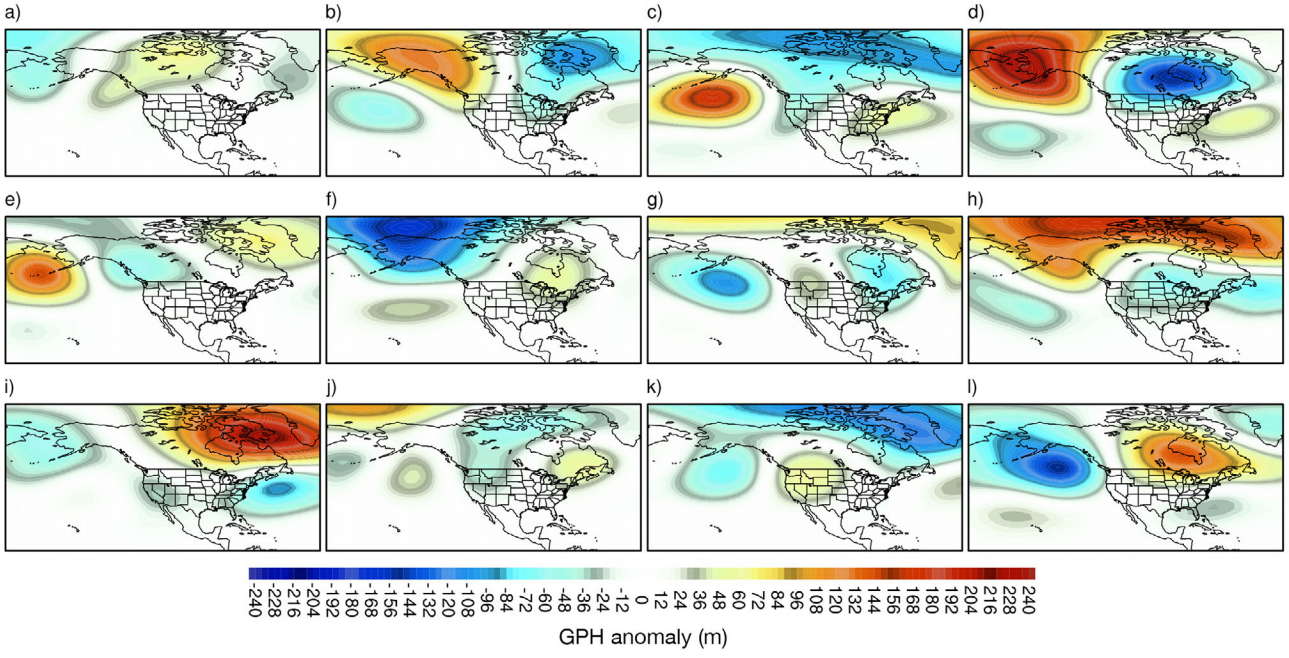

Fig. 2. 500 hPa geopotential height (GPH) anomalies associated with each cluster centroid

the frequency of each WR in each year (f121), and the temporal correlation of the NCEP and RCM timeseries of the annual frequency of each WR (f122). For application to the GCM-driven simulations, f122 was not included, as the GCMs are not expected to reproduce the weather regimes with the same temporal phasing.

The 5 skill scores are then normalized to be between 0 and 1, multiplied together, and normalized again to sum to 1 across the models (for application to the ensemble mean) to obtain the final values for this metric for each season. A final annual value for all of North America is obtained by averaging between the 2 seasons and normalizing. This is the only metric calculated across the North American NARCCAP domain and not individual subregions. For final application, each region receives the same value.

It is important to note at this point that we partly change the meaning (i.e. the intent, not the calculation, except as noted above) of this metric between its application to the NCEP-driven and GCM-driven ensembles. This metric was originally meant to both measure a model's performance in reproducing the observed large-scale circulations that drive regional climates and measure how well an RCM *follows its driver* (Sanchez-Gomez et al. 2009, Christensen et al. 2010), both of which are sensible for an RCM performance metric, particularly when only applying the metric to reanalysis driven simulations. We maintain this definition for the NCEP-driven runs, but cannot maintain both parts of the original intent for the GCM-driven simulations. We compare the simulations to NCEP, instead of their parent GCMs, turning this into a pure verification metric, instead of measuring how well the RCMs follow their drivers. The reason for this is 2-fold. (1) We believe it is more important to know how well a simulation produces a climate that is most like reality. (2) The NARCCAP ensemble contains nudged simulations, and as intended, they follow their parent simulations' large-scale best. In ENSEMBLES, the 1 simulation that utilized nudging was not included in the final results, as the relatively superior performance in f1 would have resulted in the highest weight, by far, based on only 1 measure (ENSEMBLES 2009). There were still 15 simulations in their final weighted ensemble. We do not discard the nudged simulations here, as it would mean throwing out 1/3 of the simulations, leaving only 8 produced by 4 RCMs. Instead, we present results of the weighting with and without the inclusion of metric f1 in the final weight, as the results vary substantially depending on whether or not it is used, particularly in the NCEP-driven simulations, where nudging matters most. While we could just discard this metric — as it disproportionately highlights nudged models in the reanalysis-driven simulations and is more of a GCM performance metric when used with the GCM-driven simulations — its use allows us to better examine the effect of different weight distributions.

2.4.2. Mesoscale metric (f2)

This metric tests the information the RCM adds at the mesoscale in seasonal average precipitation and 2 m temperature (Coppola et al. 2010). The mesoscale component is found by subtracting a 250 km, 5×5 grid cell running average from each grid cell. In ENSEMBLES, a 9×9 grid box running average was subtracted instead, as the RCM resolution was 25 km. Five skill scores comparing each RCM mesoscale field to that from CRU are calculated for each season. These skill scores measure the spatial correlations of each field individually (mainly measuring the ability to capture the spatial patterns near complex topography and coasts), the interannual variability and RMSE of the mesoscale signals for each variable (to measure the skill in reproducing the magnitude and sign), and a spatial correlation between the temperature and precipitation fields (measuring their spatial interconnection). As in f1, these skill scores are normalized, multiplied together, and normalized again to obtain a final value for this metric for each region and season.

2.4.3. Probability density distribution metric (f3)

This method examines the statistical properties of the empirical probability distribution functions (PDFs) of daily and monthly precipitation and daily maximum and minimum temperature (Kjellström et al. 2010). A skill score is generated from daily precipitation and daily minimum and maximum temperatures based on the overlap of M-OBS and each RCM's PDFs. A skill score for monthly precipitation is also calculated, but consists of 5 parts that measure an RCMs ability to capture different aspects of the generated PDF. The final value for this metric is an average of the skill score from daily precipitation, the average of the maximum and minimum daily temperature skill scores, and the square root of the monthly precipitation skill score. These are then normalized, as above, to give a final value for each season and each region.

2.4.4. Extremes metric (f4)

The extremes metric tests the ability of an RCM to reproduce 99th, 99.9th, and 99.99th percentile daily precipitation across each given region (Lenderink 2010) and it also uses generalized extreme value (GEV) theory to assess 5 yr return periods in daily precipitation and maximum and minimum temperature (ENSEMBLES 2009). A basic transformation of percent bias versus M-OBS for each of the 3 extreme daily precipitation percentiles is calculated giving values between 0 and 1 for each percentile. These are then averaged to obtain the first skill score for each region and season. The second skill score is derived from 5 yr return periods of daily precipitation and minimum and maximum temperature. A GEV distribution is fit to a timeseries of seasonal maximums using L-moments. The final score for this second part is based on the difference in the upper and lower confidence bounds (at a 0.2 confidence level) for the RCM 5-year return period and the bias in the 5-year return value versus that from M-OBS. These 2 skill scores are multiplied together and normalized to obtain the final values for this metric for each region and season.

2.4.5. Temperature trends metric (f5)

The ability of an RCM to reproduce seasonal and annual mean 2 m temperature trends is tested in this metric. The value for the metric in each region is a basic skill score with values ranging from 0 to 1 comparing the linear slope/trend for a given region from the RCM to that in CRU, as described in Lorenz & Jacob (2010). This metric is only used with the NCEP-driven simulations and not the GCM-driven simulations, as the GCMs are not expected to contain the signals forcing the historical trends with the same temporal phasing. Multiple GCM realizations would likely be needed to characterize the internal variability and encompass the observed trend.

2.4.6. Annual cycle metric (f6)

The annual cycle metric examines bias in the amplitude and period in the annual cycle of monthly mean 2 m temperature and precipitation against CRU (ENSEMBLES 2009, Christensen et al. 2010). The index used is S, where:

$$S = \frac{4(1+R)^4}{(\sigma + 1/\sigma)^2 (1+R0)^4} \quad (1)$$

and R is the correlation coefficient between the RCM and observations, $R0$ is the maximum attainable correlation ($R0 = 1$), and σ is the SD of the RCM normalized by the SD of the observations ($\sigma = \sigma_{RCM}/\sigma_{OBS}$). S is calculated separately for 2 m temperature and precipitation, the 2 values are averaged together and then renormalized. For this metric only, the same value is used in the final weight for the annual weight as well as the individual season weights, as it is computed on a full annual cycle only, and not subperiods within that cycle.

2.4.7 Final weights (W)

Following the baseline approach in Christensen et al. (2010), the final weight is obtained by multiplying the metrics together. Using this method, as opposed to an additive one, a model must perform well in all metrics to obtain a high weight and the importance of each metric is retained in the final result. This also assures no compensation for bad performance in one metric given good performance in another. Once the product is taken, the values are renormalized to sum to 1 for the final weight. Christensen et al. (2010) only address annual weights and not seasonal ones. Here, however, we discuss weights for individual seasons as well. As in ENSEMBLES, the summer final weight is the JJAS seasonal value from the large-scale metric (f1) and the JJA seasonal value from the other metrics multiplied together, and for winter, the DJFM seasonal value from f1 and DJF from the others (except for f6, in which the same value is used for all seasons).

3. RESULTS AND DISCUSSION

3.1. NCEP-driven simulation metrics

Winter, summer, and annual results from metrics f2–f6 are displayed in Fig. 3. Results for spring and fall are given in Fig. S1 in the Supplement at www.int-res.com/articles/suppl/c077p023_supp.pdf. Across regions, the metrics are very evenly distributed for f3 (PDF metric), f4 (extremes metric), and f6 (annual cycle metric). That is, the RCMs perform almost equally, and consistently across regions and seasons in these metrics. These measures do not express that the RCMs performed well or poorly, only that they performed similarly, better, or worse relative to the other RCMs. However, there is a little more variability in performance in the extremes met-

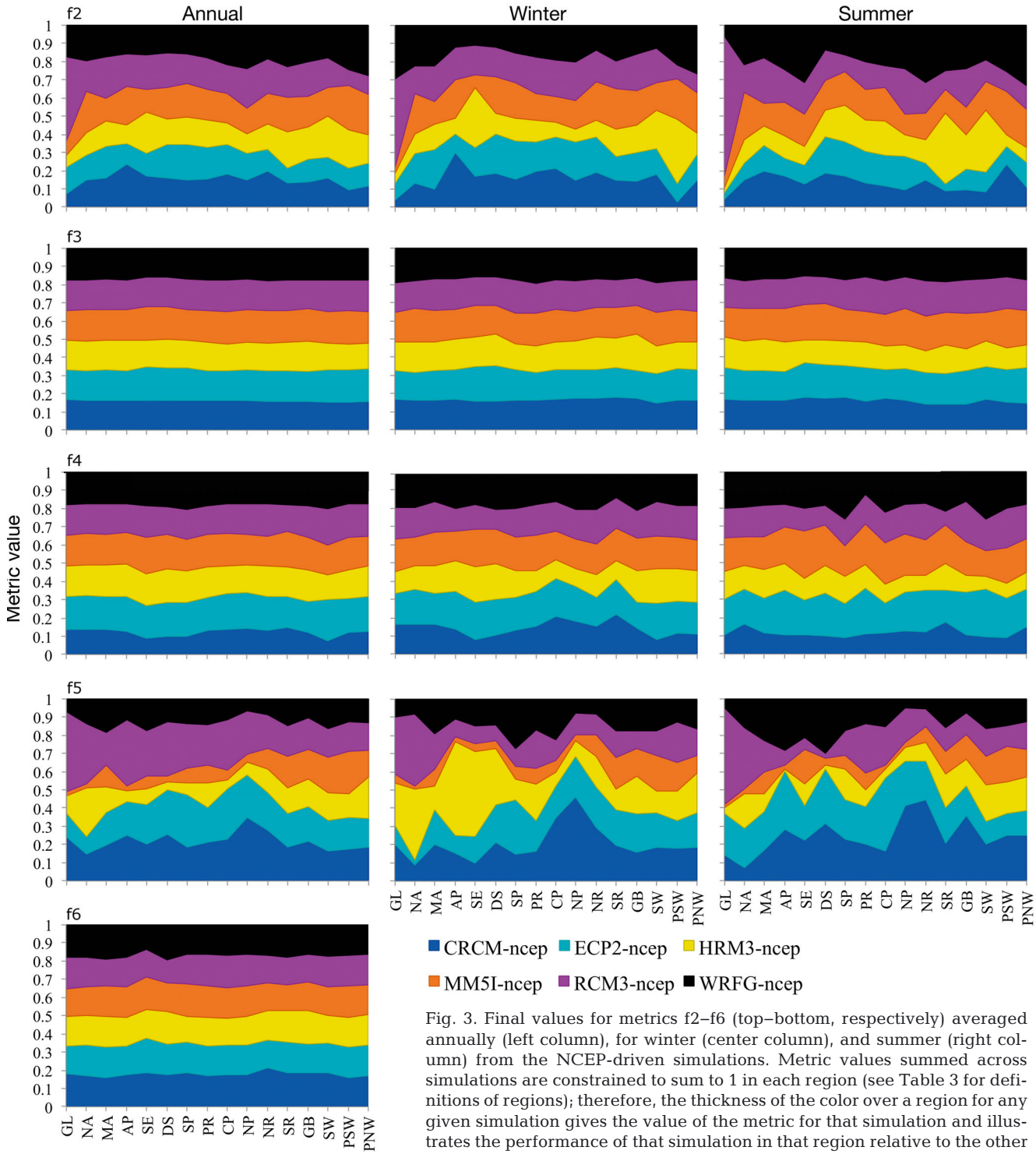

Fig. 3. Final values for metrics f2–f6 (top–bottom, respectively) averaged annually (left column), for winter (center column), and summer (right column) from the NCEP-driven simulations. Metric values summed across simulations are constrained to sum to 1 in each region (see Table 3 for definitions of regions); therefore, the thickness of the color over a region for any given simulation gives the value of the metric for that simulation and illustrates the performance of that simulation in that region relative to the other simulations (thicker is better)

ric, especially in summer, originating from regions where warm-season convection is important (e.g. the Southwest, Plains, and Southeast regions). The CRCM, for instance, scores lower than other RCMs in the extremes metric in summer in many regions, due to its handling of precipitation extremes (not shown). This is not surprising, however, given that this RCM does produce extremes that are damped, likely because of its use of nudging (Alexandru et al. 2009). Likewise, the HRM3 contains a relatively low score in many regions, originating

from the temperature component of the extremes metric (not shown), and possibly related to its known warm bias (Mearns et al. 2012). This same bias likely also contributes to the HRM3's low performance in f3 via the temperature half of the metric (where it is ranked lowest in the majority of regions). Metrics f2 (mesoscale metric) and f5 (trends metric) produce greater variability in performance between models and regions than the other 3 metrics across seasons (Fig. 3). This is somewhat different than what was found in ENSEMBLES, where f2 and f4 produced the greatest performance variability. The most obvious departure in score is in f2 from the RCM3 in the Great Lakes region, where it receives a very high score. The RCM3 also has a high score in f5, though not to the same extent. Another noteworthy outlier in f2 is the HRM3 in summer in the Southwest and Southern Rockies (f2 scores here are the highest, second to the RCM3 in the Great Lakes region only), where it is capturing the orographically forced spatial distribution of precipitation and temperature and their interannual variability in a manner that is superior to the other RCMs. This result is not surprising, given its known skillful performance in simulating various aspects of the North American monsoon (Bukovsky et al. 2013, 2015). Aside from a few other more minor outliers, the RCMs perform fairly equally in f2 relative to the results of f5. High variability in f5 is expected, however, given previous examination of the temperature trends produced by these models in Bukovsky (2012). Given Bukovsky (2012), for example, it is not surprising the MM5I performs poorly across all seasons and regions in this metric, as it has a very strong, very widespread, progressive warming bias. Much of the variations otherwise come from the RCMs' differing ability to capture features such as the summer and fall 'warming hole' that is present across the central part of the continent during this period, the strength and pattern of winter warming, and the extent of the cooling in spring across the northern USA during this period. Overall, the 2 nudged RCMs, the CRCM and ECP2, as well as the RCM3 often outperform the HRM3 and MM5I in this measure, particularly outside of the western USA.

Scores from f1, the large-scale metric, are provided in Fig. 4. Whereas scores from the other metrics are often relatively uniform, and between 0.1 and 0.2, with the exception of some outliers, scores from f1 do not indicate relatively consistent performance across the RCMs. As illustrated in Fig. 4a, the CRCM scores substantially higher than the other RCMs in this metric in winter (0.54), summer (0.70), and in the annual average (0.62). ECP2 is always the second highest,

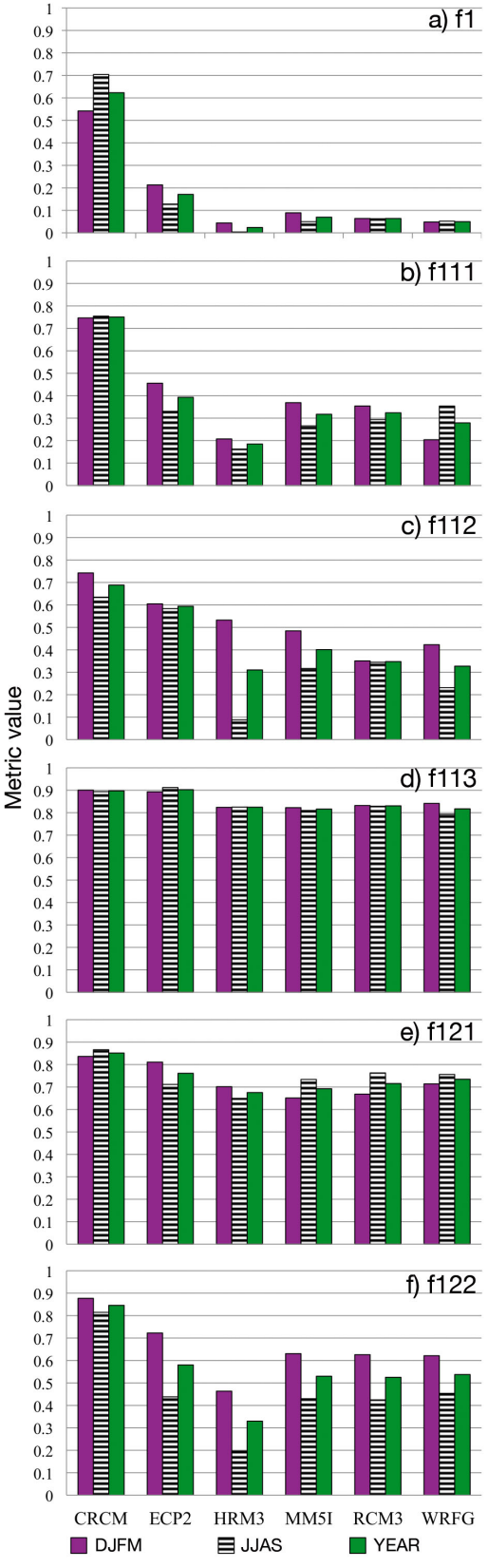

Fig. 4. Values for (a) metric f1 and (b–f) its sub-components for the NCEP-driven simulations

but not by nearly as much (0.21–0.13). Both the CRCM and ECP2 use nudging. As in the ENSEMBLES simulations, this yields a large advantage (Sanchez-Gomez et al. 2009), as they are being nudged to one of the same fields used to compute the WR. However, this metric is in part meant to measure how well an RCM follows its driver. The HRM3, on the other hand, always scores lowest in this metric. As illustrated in Fig. 4b-e, these differences do not originate in all of the sub-components of f1. The performance of the RCMs in f113 and f121 are more uniform than in the other sub-metrics. That they are uniform in f113 is not surprising, as this is a measure of average persistence time in a given WR. With 12 regimes, this value is almost always between 1 and 3 d in summer and 1 and 4 d in winter in the RCMs and NCEP, with an average of 2.5 d in any weather regime in NCEP; thus, there is very little spread. There would be more spread if the maximum amount of time spent in each WR was used instead of the average, as on average across the WR, the maximum time spent in any WR in the series is 10.5 d in NCEP. The greatest spread in performance is in f111, the spatial correlation of the 500 hPa geopotential height anomaly composited for days assigned to a given WR. The smaller bias in the large-scale in the nudged models is clearly an advantage here. As the sub-metrics are multiplied together to come up with the final value for f1, any low or high score in a component will be reflected. As the CRCM does not perform poorly in any sub-metric (and is often the best), the performance in f111 highly dictates its final high score. As HRM3 performs poorly in f111, but also has the lowest score in other sub-metrics where there is less spread, it obtains the lowest score. The HRM3's low score in f1 in summer is further influenced by low values in f112 and f1222.

When f1 is not included in a region's final weight, the RCM that carries the most weight varies by region (Fig. 5). When f1 is included, because there is a much greater spread in the values than the other metrics, the CRCM carries the greatest weight in almost all regions at all times, and the HRM3 has almost no weight in many regions, particularly in summer. This is also true of the MM5I in many regions. The overall influence of f1 on the final weights is made more obvious by averaging across regions for each RCM, as done in Fig. 6. While the average weight without f1 is more uniform across RCMs (about 0.1–0.3), including f1 the weight varies between near 0 and 0.6.

3.2. GCM-driven simulation metrics

In ENSEMBLES, the metrics and final weights were calculated using only reanalysis-driven RCM simulations, not GCM-driven simulations. Other work with the NARCCAP ensemble suggests that the performance of the RCMs when driven with the GCMs can be

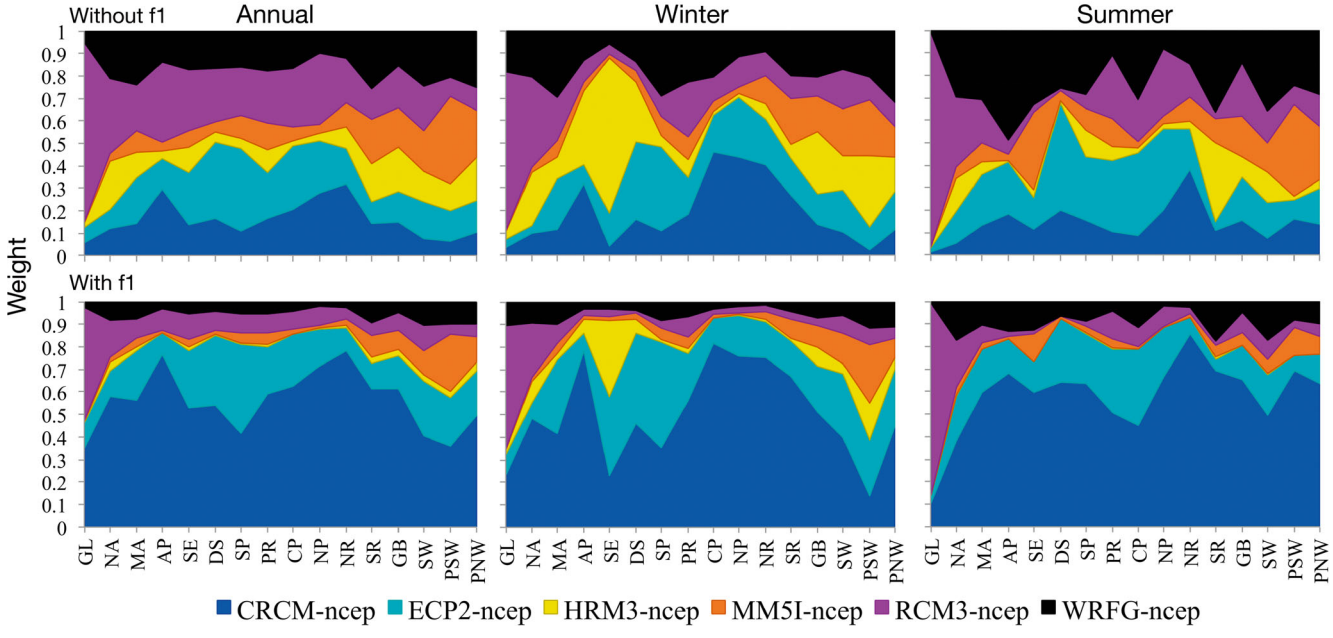

Fig. 5. Final weights for the NCEP-driven simulations with (bottom) and without (top) the large-scale metric (f1) included in the calculation

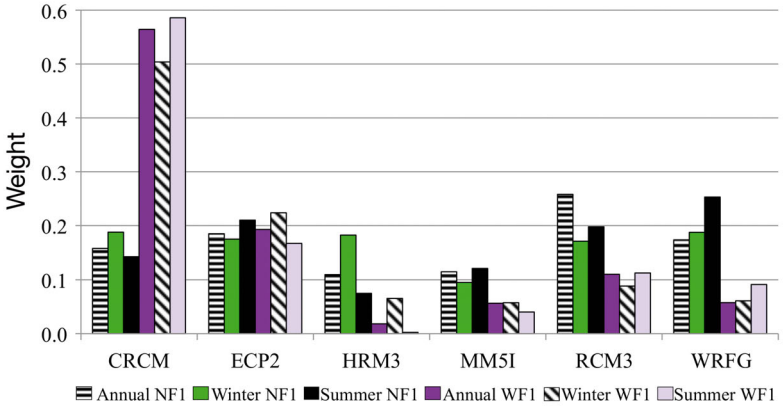

Fig. 6. Final weights for the NCEP-driven simulations averaged across all regions with (WF1) and without (NF1) the large-scale metric (f1) included in the calculation. If all of the simulations received equal weight, the weights would be 0.167

very different from their corresponding reanalysis-driven simulations (e.g. Bukovsky et al. 2013, Wehner 2013, cf. McCrary & Mearns 2017, R. R. McCrary & L. O. Mearns unpubl.); therefore, we have calculated the metrics and final weights for each GCM-driven simulation as well. In doing so, we removed metrics and metric components that would be most affected by internal GCM variability: f4, the trend metric; and part f122 of the large-scale metric. We also changed the overall intent of f1 (but not the methodology), as discussed in Section 2.4.1. In this application, f1 is now purely a verification metric, comparing the simulations to NCEP, and not also assessing how well the simulations follow their driver. Since the GCM simulation of large-scale WR is not necessarily very accurate (compared to NCEP), the nudged simulations lose the very clear advantage they demonstrated when forced by NCEP.

The annual average final metric values for the GCM-driven simulations are shown in Fig. 7. Overall, the annual average results are similar to the seasonal results (Fig. S2 in the Supplement). Similar to the NCEP-driven results, metrics f3, f4, and f6 show relatively uniform performance between the simulations with greater variability between simulations in metric f2, the mesoscale metric. The performance of some of the

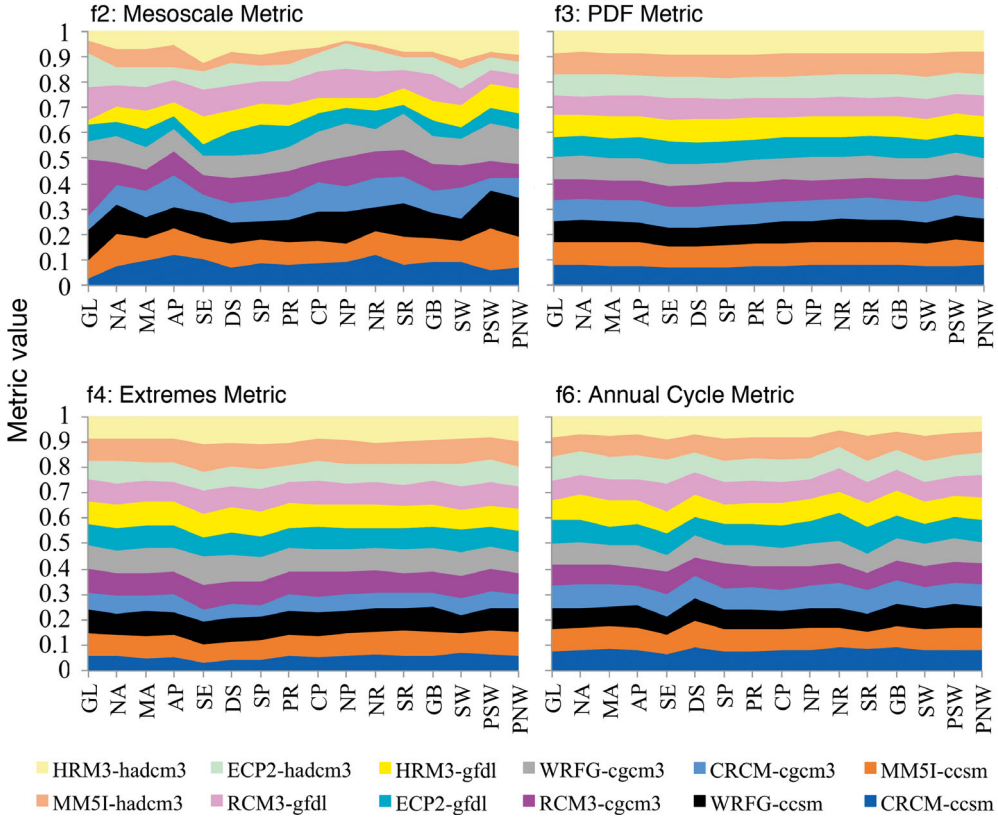

Fig. 7. Annually averaged final values for metrics f2, f3, f4, and f6 from the GCM-driven simulations. As in Fig. 3, the thickness of a color above a given region indicates the value of the metric for the simulation assigned to that color, and metric values must sum to 1 in each region

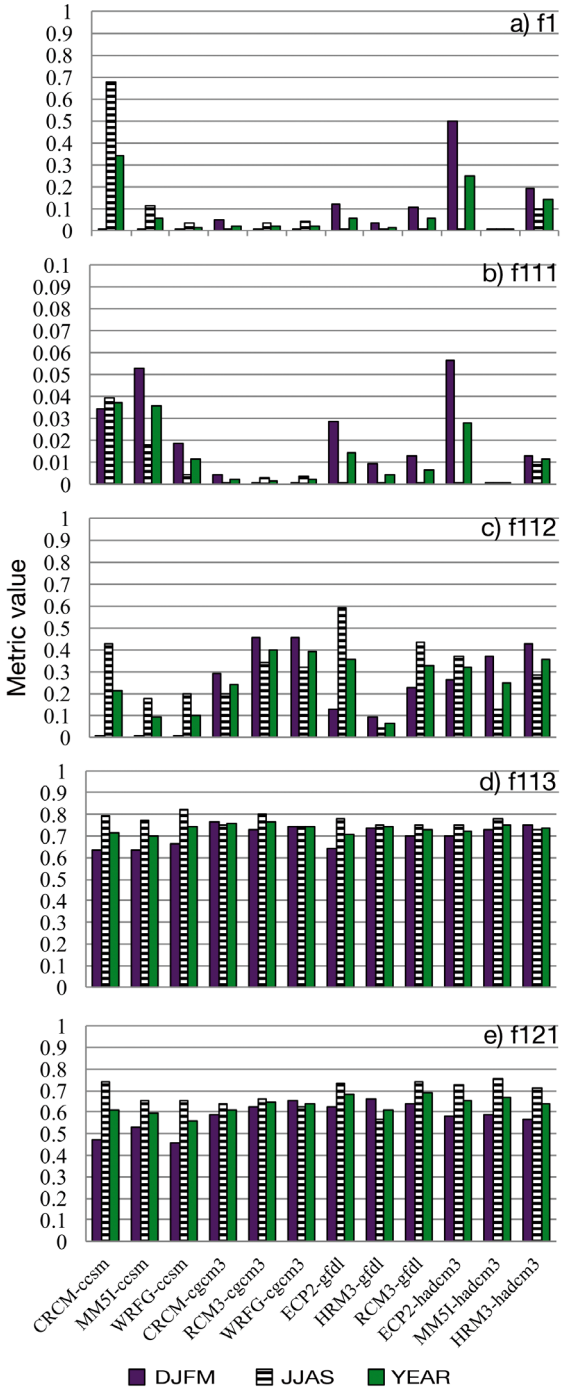

Fig. 8. Values for metric f1 and its sub-components for the GCM-driven simulations. Note that the scale on (b) is now different from the other panels by 1 order of magnitude

RCMs when forced with a GCM is also comparable to their NCEP-forced simulation. For example, the RCM3 performs well in the mesoscale metric here (Fig. 7) as well as when forced by NCEP (Fig. 3) in the Great Lakes region (GL). Also, the lack of heavy precipitation extremes in the CRCM leads to a relatively poorer performance compared to the other simulations in the extremes metric regardless of driver. Relative performance in these metrics does not cluster by RCM or GCM in all cases, however. For instance, the MM5I-hadcm3 and ECP2-gfdl both score worse in the mesoscale metric in regions of complex terrain relative to their counterparts.

The GCM-driven simulations perform considerably worse than the NCEP-driven simulations in metric f1, the large-scale metric. The relative performance can be seen by comparing the scores from the subcomponents of f1 in Figs. 4 & 8. In this case, the nudged models do not have an advantage as performance is closely tied to how well the driving GCMs reproduce the large-scale circulation and the various WR. However, when a nudged RCM is paired with a GCM that better captures North American weather regimes, it does score higher than the other RCMs paired with that GCM. This is the case for the CRCM-ccsm in summer and the ECP2-hadcm3 in winter (Fig. 8a).

Scores in part f111 (the spatial correlation of the height pattern for all days identified as a given weather regime) are particularly poor, with values that are lower by about an order of magnitude compared to the NCEP-driven simulations (cf. Figs. 4b & 8b). Furthermore, due to poor performance in winter in f112 (percent of days spent in each weather regime), the CCSM-driven simulations virtually drop out of metric f1 completely in winter (due to an inability to capture regimes a, f, and l; see Fig. 2).

Performance in f1 suggests that if we were to weight the ensemble using it, the results may not be very meaningful. This would give us little confidence in a weighted ensemble mean; therefore, in presenting results, we will continue to separate the results of the final weighting with and without f1. While we could discard f1, it demonstrates the effects of a diverse weight distribution well.

As in the NCEP-driven simulations, inclusion versus exclusion of the large-scale metric in the final weight creates a much greater separation in weight between simulations (Fig. 9; and for weights averaged across all regions, see Fig. S3 in the Supplement), although for a different reason. Weights excluding f1 are relatively more uniform across the simulations and regions. Without f1, the WRFG-cgcm3 simulations stands out as performing better than most simulations in most regions in both seasons. The HRM3-gfdl also performs best in summer in most regions according to these 4 metrics, and it does particularly well in the Southwest (SW). Without

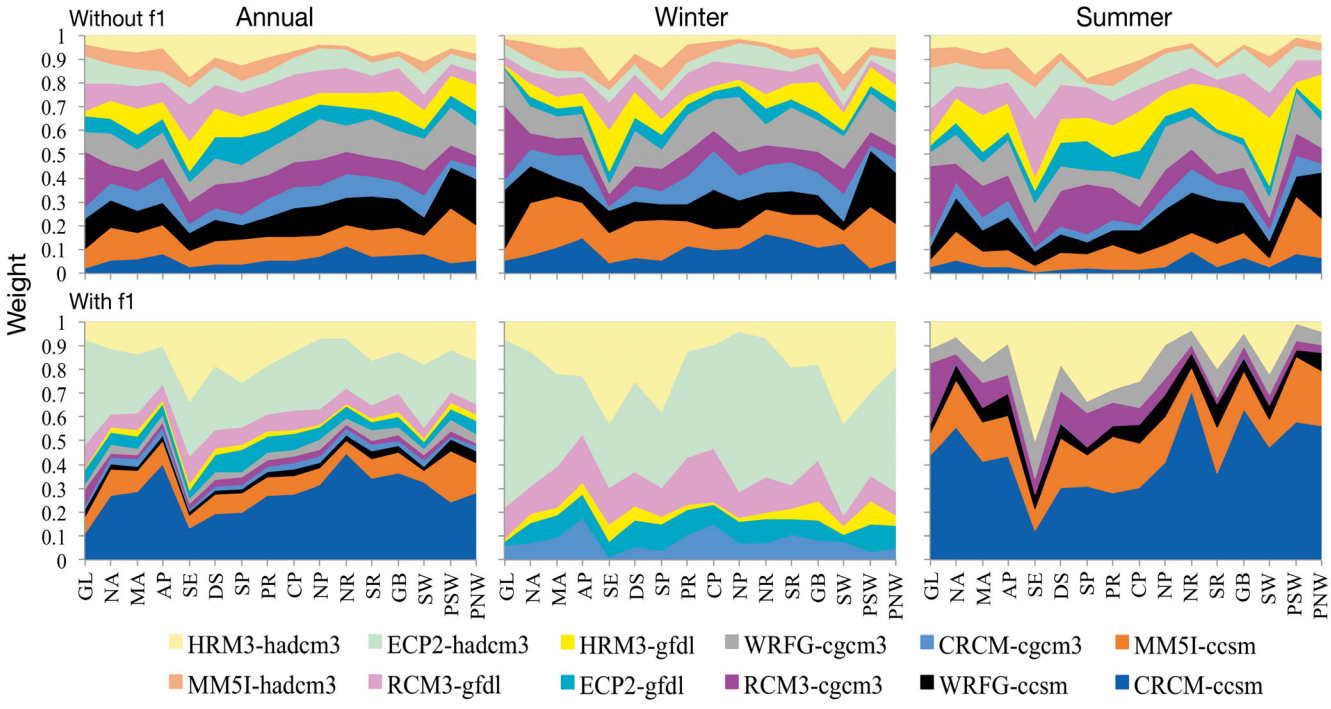

Fig. 9. Final weights for the GCM-driven simulations with (bottom) and without (top) the large-scale metric (f1) included in the calculation. In a 'model democracy', the weights would equal 0.083

f1, the MM5I-hadcm3 simulations performs most poorly in many regions in winter and summer, along with the ECP2-gfdl in winter and the CRCM-ccsm in summer. When f1 is included in the final weights, the CRCM-ccsm moves from the lowest weighted model in summer (and second lowest annually), to the highest weighted model in summer (and highest annually). The HRM3-gfdl, because of its very low score in f1, switches from one of the highest weighted models in summer to a weight that is virtually 0 when f1 is used. Similarly, the HRM3-hadcm3 simulation final weight increases such that it moves from the poorer-weighted half of the simulations, to the highest weight in winter and one of the higher weights in

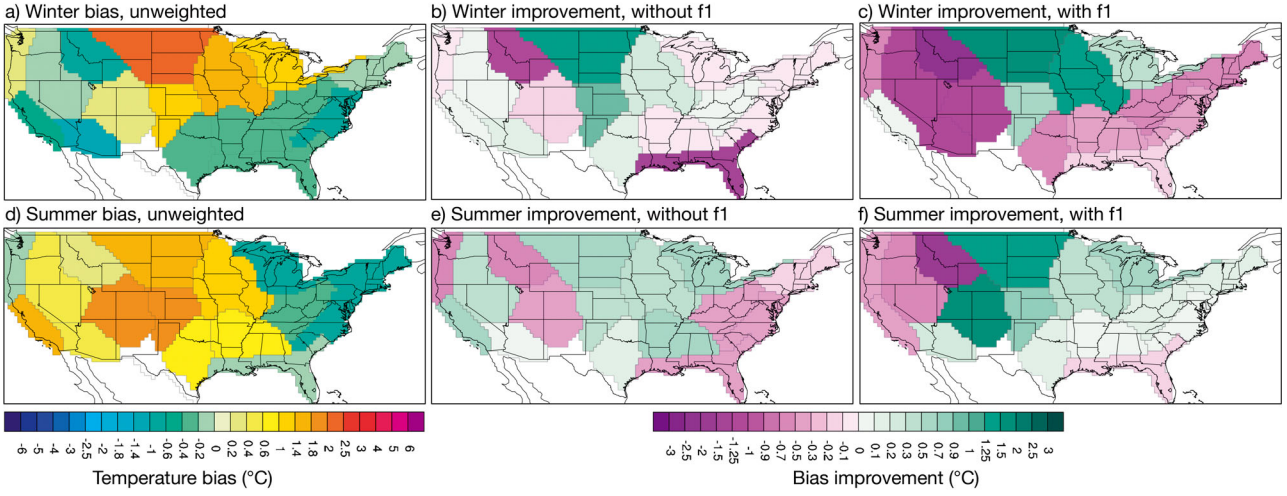

Fig. 10. Unweighted NCEP-driven ensemble mean temperature bias for (a) winter, (d) summer. Bias is defined as the ensemble mean minus CRU dataset difference. (b,c,e,f) Improvement with weighting on the ensemble mean temperature bias. Improvement is the difference in bias between the weighted and unweighted ensemble mean, but where positive (negative) improvement indicates that the bias has become closer to (farther from) 0 with weighting by the indicated amount. Weighted ensemble improvement is shown excluding f1 from the final weight (b,e) and including it (c,f)

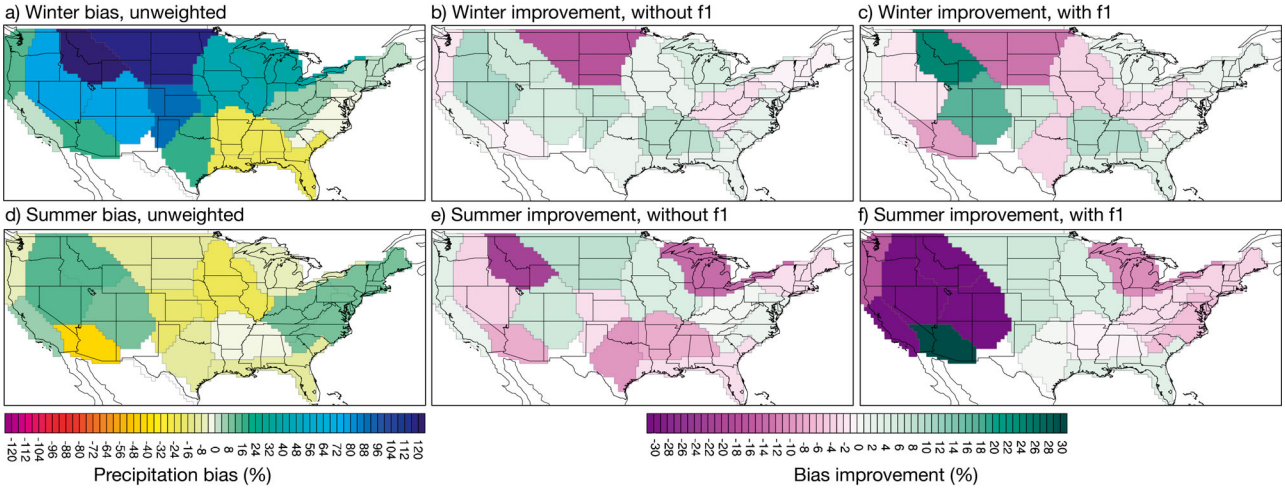

Fig. 11. As in Fig. 10, but for ensemble mean precipitation percent bias

summer in most regions with the addition of f1. Interestingly, the set of simulations with a weight that is not effectively 0 in summer is almost exactly opposite that of the set that is not effectively 0 in winter (see Fig. 9, bottom panels). Only the HRM3-hadcm3 has any significant amount of weight in both seasons in most regions when f1 is included, and the MM5I-hadcm3 has a near-0 weight in all regions in both seasons. Considering they have the same driver, this result is surprising, and also not comparable to those RCMs' performances when driven with NCEP. Furthermore, it is only comparable in rank to their weights without f1 when driven by the HADCM3, in that the HRM3 has a somewhat higher weight than the MM5I in most regions.

3.3. Effect of weighting on ensemble mean bias: Does it make a difference?

The effect of weighting on the performance of the ensemble mean for the NCEP-driven simulations relative to the CRU dataset for mean temperature and precipitation is illustrated in Figs. 10 & 11. The change in the bias with weighting is given as 'improvement', where a positive improvement indicates that the bias has decreased (moved closer to 0) by the indicated value. Negative 'improvement' indicates that the bias has increased (moved farther from 0). Figs. 10 & 11 show the effect of weighting excluding and including f1 from the final weights for each region. Each region is given the weight calculated from the performance of each metric in that individual region (not a uniform weight averaged across all regions), i.e. the weights given in Fig. 5 for the NCEP-driven simulations. The effect of weighting on the bias is only shown and discussed here for the NCEP-driven simulations versus CRU as similar conclusions can be drawn from the GCM-driven simulations. Results from the GCM-driven simulations versus CRU are available in Figs. S4 & S5 in the Supplement. Similarly, since the same conclusions can also be drawn using M-OBS for comparison instead of CRU, we will only show and discuss results using CRU (results using M-OBS can be found in Figs. S6–S9 in the Supplement). This is not to say that the comparison dataset does not change the bias, however, it is just that we end up with the same general conclusions using either.

Overall, improvement in bias with weighting for both mean temperature and precipitation is mixed, with the weights increasing or decreasing the ensemble mean bias depending on the region and season. Positive or negative improvement is often small relative to the ensemble mean bias for weighting without f1. In some cases, the change in bias with weighting is quite large relative to the unweighted bias though, for good or bad, and this most often occurs in the western half of the USA. The effect across variables is not always consistent either. For example, in the Northern Plains (NP) and Central Plains (CP) regions for winter temperature, the positive improvement with weighting notably decreases the bias (Fig. 10). However, at the same time, bias in DJF precipitation substantially increases with weighting in NP relative to other regions (Fig. 11), but the bias is already so high that this change is not large relative to that region's unweighted bias (Fig. 11). In some regions, weighting flips the sign of the bias as well (not shown), improving (e.g. GL DJF temperature, NP JJA

precipitation) but more often exacerbating (e.g. Great Basin [GB], Southeast [SE] DJF temperature, GL JJA precipitation) the bias.

Furthermore, including f1 in the final weight often increases the magnitude of the 'improvement' for better or worse. Generally, including f1 illustrates the mixed effect of putting most of the weight on the CRCM and ECP2 simulations versus spreading the weight more evenly across the set of 6 NCEP-driven simulations. The effect of including f1 is much amplified in the GCM-driven ensemble, where weights are also skewed towards only a few simulations as well, with more large increases in bias across the USA in winter (see Figs. S4 & S5)

By and large, there is no clear signal that weighting improves the ensemble mean bias. That the weights sometimes lead to a more biased ensemble mean is not unexpected given that the weights are not based on metrics that are explicitly related to mean bias. This was also the case in Christensen et al. (2010) with these metrics. It is in contrast to weighting schemes like the one applied in Haughton et al. (2015) though, where the mean bias in global mean temperature improved when a measure of mean performance was used to weight a GCM ensemble. Therefore, including a measure of mean performance here may improve these results, but given the influence of the other metrics, this cannot be assumed, since the relative performance in all metrics influences the final weight. Different methods for equally combining the metrics could also be tested in place of the multiplicative approach, but as in Christensen et al. (2010) when this was tested, we suspect we would also end up with the same rank of simulations in any given region but with damped spread and, therefore, similar but damped results for bias improvement. We could also pick and choose which metrics to use, but this would introduce more subjectivity into the approach as we would need to determine what aspects of performance are most important, and this would likely vary by region, decreasing the usefulness of an approach that is meant for universal application. It also may not decrease bias in the ensemble mean.

3.4. Effect of weighting on ensemble mean projections: Can it make a difference?

Given the results of Section 3.3, we hypothesize that unless most of the weight is applied to 1 or 2 simulations that have changes toward the extremes of the ensemble distribution, the effect of the weights on the ensemble mean change would not be significant. We also question whether weights that were based on multiple metrics of performance could ever create weights that differentiated the models in a meaningful way, even if the metrics were reasonably justified, as the weights might become more diluted. The greater the number of metrics and sub-metrics involved, the more uniform the weights seem likely to become, and this would also be true as the number of models increased. We see another issue here as well, related to the necessary limitation that the weights sum to 1 for application to the ensemble. The results before the normalization process (not shown for most metrics) are often more revealing regarding model performance than they are afterward. With a 0–1 normalization limitation, it seems very likely that any weighting scheme would not substantially differentiate between simulations unless it involved few performance metrics and/or in a location where performance between models is *very* heterogeneous. Therefore, in this section we also apply a generic set of 'all possible weights' to test these hypotheses when applying the metric-based weights to the projections.

In Fig. 12 we demonstrate the effect of the metric-based weights on the ensemble mean projections of precipitation and temperature, but also the potential effect of *any* combination of weights on the ensemble. In Fig. 12 we focus on 1 season and 3 regions that broadly represent the possible results across all of the regions and seasons (the full set of results is shown in Figs. S10–S15 in the Supplement). For Fig. 12 (and Figs. S10–S15), 1 000 000 random, uniformly distributed weight *combinations* (sets of 12) that sum to 1 were created and applied to the NARCCAP ensemble of simulation projections over each region to create a PDF of possible weighted climate changes. We consider any ensemble mean change that is outside of these bounds to be significantly different from the unweighted ensemble mean. In this way, we are commenting on statistical significance and not practical significance. Some of the differences discussed below may matter from a climate change impacts point of view even if they are not statistically significant.

For the North Atlantic region (Fig. 12a,b), the weighted ensemble mean with or without f1 is not significantly different than the unweighted mean for precipitation or temperature. For temperature, weighting tends to pull the ensemble mean to a slightly lower value of change, as most of the weight is still on a model or models that project changes near the ensemble mean. This includes the ensemble mean with f1, where approximately 57% of the weight is placed on the ECP2-hadgem (Fig. 9), which projects a change of

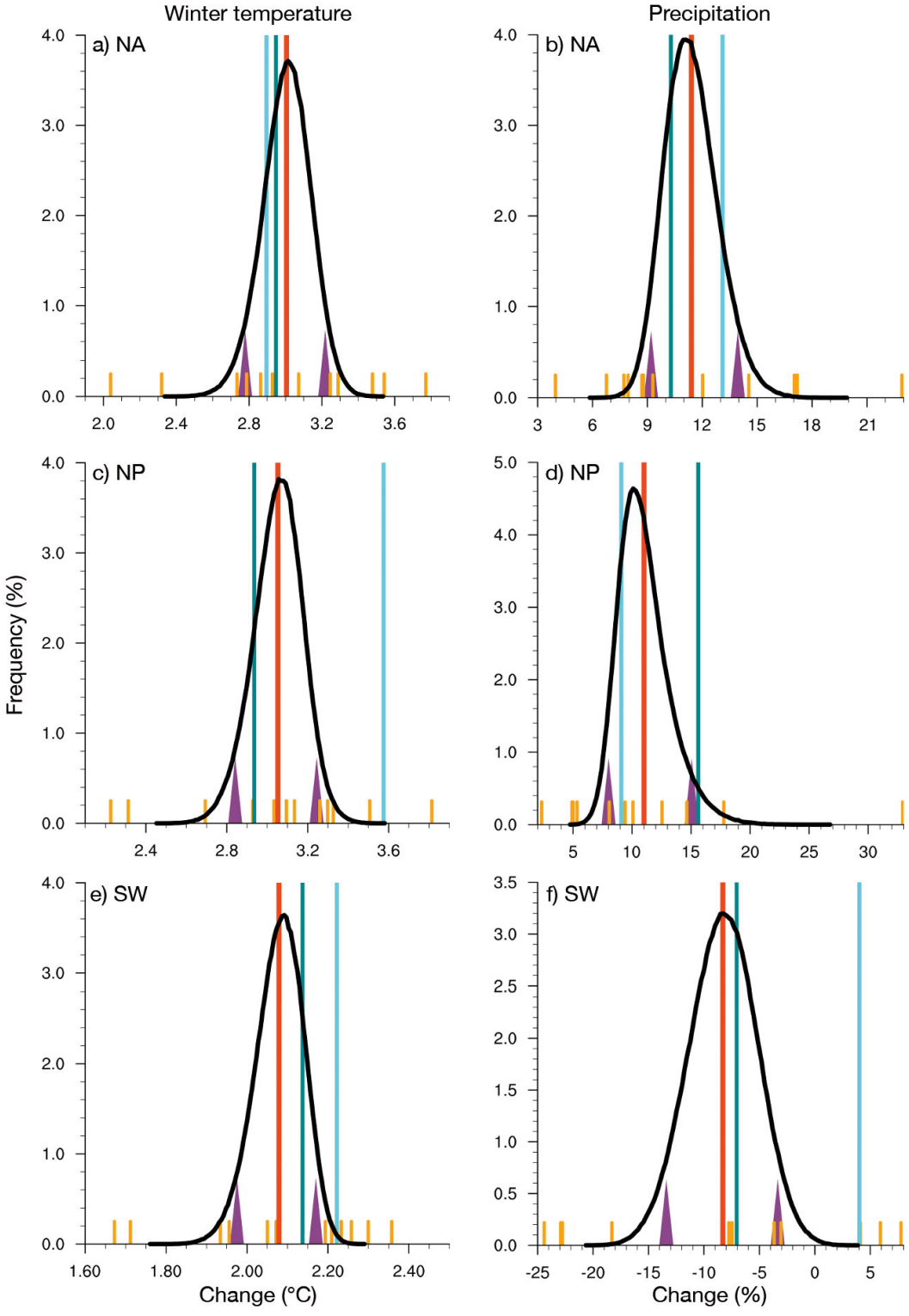

Fig. 12. Projections of (a,c,e) winter temperature and (b,d,f) precipitation with and without weights from 3 regions (a,b: NA, North Atlantic; c,d: NP, Northern Plains; e,f: SW, Southwest). Black curve represents the frequency of a given ensemble mean change using 1 000 000 different sets of random, uniformly distributed weights. The purple triangles indicate the 5th and 95th percentile values of change from this distribution. The dark orange line indicates the change from the unweighted ensemble mean; the light turquoise line that from the ensemble mean weighted using the metrics without f1; and the dark turquoise line that from the ensemble mean weighted using the metrics with f1. The short orange lines indicate the individual simulation projections

+2.9°C, just below the unweighted ensemble mean. Thus, the weighted means are not significantly different. Results for precipitation in this region are similar.

For the Northern Plains region (Fig. 12c,d), one weighted mean for both temperature and precipitation is significantly different than the unweighted mean. For temperature, the weighted mean including f1 is at the high extreme of what is possible given the 'all possible weights' distribution, mostly because 66% of the weight is on the one simulation (ECP2-hadgem) with the greatest temperature increase. For precipitation, this is one of only two regions and seasons where the weighted ensemble mean without f1 is significantly different than the unweighted mean (the other is in annual mean temperature in the PSW). The weights without f1 in this region are fairly uniform, but there is an emphasis on the simulations with the greatest projected increase in precipitation, the WRFG-cgcm, which receives 23% of the weight (Fig. 9). Also, in the NP, weights without f1 favor the CCSM- and CGCM-driven simulations, which also tend to project a greater increase in precipitation. Combined, this leads to the very rare occurrence of the weighted ensemble without f1 being significantly different than the unweighted ensemble.

In the SW region (Fig. 12e,f), the weighted mean using f1 demonstrates with both precipitation and temperature the effect of putting most of the weight on models with larger changes. For temperature, about 41% of the weight is put on the simulation with the greatest temperature change, pushing the ensemble mean weighted with f1 outside of the significance bounds. With precipitation, 81% of the weight ends up on the 2 simulations with the greatest positive change, leading the ensemble mean weighted with f1 to be at the very edge of what is possible within the 'all possible weights' distribution.

When f1 is included in the final weight in particular, it is not uncommon for a metric-weighted mean to be significantly different than the unweighted mean, but this is because most of the weight is put on a simulation or two with more extreme changes relative to the rest of the simulations. This is particularly common in winter temperature changes, where this happens in 14 of the 16 regions (88%), and 7 of the 16 regions (44%) for winter precipitation. It is otherwise much less common when the weights are more evenly distributed (when f2–f6 are used only) or weighting favors simulations with changes that are near the mean. For summer precipitation, this occurs in 2 regions, and for temperature in 6, while for the annual mean change, it occurs in 3 regions for precipitation and 2 for temperature.

Overall, the results of this experiment show that a large percentage of weight must be put on a model or models with values towards the extremes of the ensemble of possible changes to obtain an ensemble mean that is significantly different than the unweighted ensemble mean. It also illustrates that this is an unlikely outcome in any weighting system where the weights must sum to 1, unless there is one factor that dominates over all the rest, justified or not, as f1 does here.

3.5. Metrics vs. in-depth analysis

The ENSEMBLES metrics were developed to be generally useful in measuring RCM performance over any region. Therefore, regionally specific processes that may be of great importance to the perceived quality of an RCM simulation are not included in the metrics or final weights. It may be possible, as a result, for a model to 'score' well and have problems that are not clearly captured by the metrics. To illustrate this point, we will briefly compare and contrast the results of this weighting effort with the results of in-depth NARCCAP analyses that have been completed over 2 very different regions in summer with a focus on precipitation.

3.5.1. North American Southwest

Summer precipitation in the North American SW region is produced as a part of the North American monsoon system (NAMS). There are a number of well-defined processes that encompass NAMS that affect when and where precipitation falls. Bukovsky et al. (2013, 2015) (hereafter BUK1315) provide an indepth examination of the NARCCAP simulations with the explicit goal of establishing their differential credibility in simulating NAMS. The overall results of BUK1315 for the NCEP-driven simulations do not deviate substantially from those of the combined metrics, excluding f1, in that the simulations are of a similar quality, as each simulation has variables/processes that it reproduces slightly better or worse than the others. Here, if the final weight excluding f1 is used to rank the NCEP-driven simulations, the WRFG is weighted highest, followed by the ECP2, RCM3, HRM3, MM5I, and finally the CRCM. In BUK1315, the WRFG, HRM3, and CRCM are found to be the better performing simulations mostly because they reproduce the Gulf of California low-level jet best, which is important for moisture flux into the

SW, and therefore precipitation production. The MM5I, for example, in BUK1315 is the poorest performing RCM, as it has the weakest onshore flux of moisture (no low-level jet), and a very strong dry-bias in SW-area monsoon precipitation as a result, which is consistent with its lower weight. The metrics agree with BUK1315 when weighting the WRFG highly, but the HRM3 always contains a weight that falls in the bottom half of the results, contrary to the process-level assessment.

The differential performance of the GCM-driven simulations using the metrics versus the results of the in-depth analysis in BUK1315 is much more complex, and there are some striking differences. Using the final summer weights to rank the simulations, the HRM3-gfdl is the best performer in the SW excluding f1, followed by the RCM3-gfdl, ECP2-hadcm3, and HRM3-hadcm3. The CRCM-ccsm receives the lowest weight without f1. When f1 is included, the CRCM-ccsm receives the most weight, followed by the HRM3-hadcm3, MM5I-ccsm, WRFG-cgcm3, WRFG-ccsm, and RCM3-cgcm3. The others receive weights that are virtually 0. Specifically, in Bukovsky et al. (2015), the model that evaluated as most credible was the HRM3-hadcm3 for the SW. This is partly because this RCM captures regional processes well, but also because the HADCM3 GCM provided the least biased boundary conditions for this region. That this simulation is ranked in the top half when f1 is excluded and is second overall when f1 is included is consistent with the assessment of BUK1315. We see the greatest disagreement in the other highly weighted simulations.

BUK1315 show that the other GCM-driven simulations all have significant problems, most of which are inherited from the GCMs, and some of which are quite detrimental to the simulations, but few of these are picked up on using the metrics. For example, the GFDL-driven simulations, 2 of which receive the most weight when f1 is excluded, inherit a large-scale forcing problem from the GFDL that causes an extraordinarily high precipitation bias from mid-summer into winter over the SW. This does present itself in the raw results of f6 in the annual cycle of precipitation to some extent (not shown), but once averaged with the temperature sub-component of f6 and normalized, this problem is washed out in the results (see Fig. 7). Additionally, according to metric f1, the CCSM and its child RCMs reproduce North American large-scale weather regimes in a way that is relatively better than the others. Therefore, they receive much of the weight when f1 is included. However, the CCSM-driven simulations in BUK1315 are found to be the least credible. This is due to their inability to well simulate many important monsoon-related processes (e.g. the monsoon anticyclone, tropical easterly waves, the El Niño–Southern Oscillation, atmospheric moisture content, etc.). This combination of problems results in these simulations containing almost no precipitation in the SW in summer, which affects metrics f2–f6 and the final weight when f1 is excluded. When the weights are used on the ensemble mean, these differences are observable. When the set of weights excluding f1 is used (Fig. S5e), the bias in SW summer precipitation decreases, as putting more weight on the wetter simulations compensates for the propensity of dry biased simulations in the ensemble. When the set of weights with f1 is applied to the ensemble mean (Fig. S5f), the dry bias is exacerbated, as most of the weight is then from simulations that have very insufficient precipitation during monsoon season. This again suggests that f1 is not an appropriate metric for performance in this region.

3.5.2. North Atlantic

Thibeault & Seth (2015) (hereafter TS15) evaluated the credibility of the NARCCAP GCM-driven simulations for summer in the North Atlantic region (NA). They focused on factors that are important in distinguishing wet summers versus dry summers and whether or not the models capture the relevant processes. This included anomalies of 500 hPa heights, 850 hPa winds, and moisture convergence. In the end, no model performed well in all measures, but TS15 did identify 4 'better' models, the MM5I-ccsm, WRFG-ccsm, HRM3-gfdl, and RCM3-gfdl, in no particular order. In the NA region, 2 of these are consistent with the metric weighting, the WRFG-ccsm and MM5I-ccsm are weighted highest when f1 is excluded, and because the CCSM-driven runs do well with metric f1 in summer, they still receive high weights when f1 in included (Fig. 9). The 2 GFDL-driven simulations may do well using the region-specific measures in TS15, but they do not both perform well in all of the metrics. While no model receives a particularly low weight in the NA when f1 is excluded, the HRM3-gfdl does receive one of the highest weights and the RCM3-gfdl one of the lowest, and none of the GFDL-driven simulations receive a non-negligible weight when f1 is included. In TS15, the GFDL-driven simulations best reproduce the observed large-scale circulation anomalies associated with wet summers. Clearly there is a difference, then, in the ability to reproduce the observed large-scale

weather regimes over North America in metric f1, and the ability to reproduce the large-scale drivers of wet and dry climates in the NA. This is similar to what is seen in the Southwest in Bukovsky et al. (2015), with the CCSM-driven simulations having the greatest error in the placement and magnitude of the monsoon anticyclone, a regionally important large-scale feature, and the GFDL- and HADCM-driven simulations having the least error. Therefore, we would again argue that f1, as it is currently formulated, may not be the best measure of large-scale circulation performance in the RCMs, particularly in summer for regions in North America.

4. CONCLUSIONS

Performance metrics and weights initially created for use with the ENSEMBLES simulations were applied here to the NARCCAP simulations. While we can obtain a model with the highest 'score' within the NCEP-driven simulations, the final weights do not substantially differentiate performance between the models in many regions, across seasons, or across the whole USA combined unless the large-scale metric, f1, is included (Fig. 6). This is because, as expected, the models that are nudged to the large-scale, perform better in the large-scale metric. That the metrics did not substantially differentiate between the models when only f2–f6 are used is similar to the result obtained in Christensen et al. (2010); however, one of the ENSEMBLES reanalysis-driven simulations did emerge as a 'winner' over Europe with a substantially higher weight, unlike here. In Christensen et al. (2010), nudged models were not included in the analysis, so the beneficial effect of nudging did not cause interpretation problems in f1 results, or a uniformly strong differentiation of weights in all regions, as it does here. In metrics f2–f6, though, the skill of an RCM usually depends on what metric, season, or region is examined. Generally, any one simulation does not perform uniformly well everywhere in all metrics at all times of the year, and all RCMs have similar performance everywhere in several metrics.

Similar results are found for the GCM-driven simulations, with relatively little differentiation when only metrics f2–f6 are used. Again, the skill of a simulation depends heavily on which metric or region is considered. The RCMs perform similarly to one another in metrics f3–f6, with some slight differentiation via f2, the mesoscale metric. As with the NCEP-driven simulations, while we could identify a 'best' simulation given the metrics, the final weights (without f1) are fairly uniform (with weights ranging from 0.04–0.11, where 0.083 would be the weight in a 'model democracy'; see Fig. S3). Overall, however, for both the GCM-driven and NCEP-driven results, RCM3 and WRFG simulations always rank in the top, higher-weighted half of simulations, with CRCM in the bottom-half, and HRM3, ECP2, and MM5I simulations mixed within the ensembles (regardless of driver), if rank is determined by the average annual weight across all regions excluding f1 (cf. Fig. 6, Fig. S3). Within the GCM-driven ensemble only, there is no clear delineation in performance/rank by parent GCM using the same measure. This suggests that the purposeful design of metrics f2–f6 for RCM-specific performance evaluation may in fact provide some delineation of performance by RCM, with less regard for the quality of the boundary conditions, at least in this case. As the ENSEMBLES program did not apply the metrics to GCM-driven simulations, it is unknown if the relative performance by total weight for their RCMs would carry through to the GCM-driven simulations. This could also be difficult to assess without an ensemble like NARCCAP where each RCM systematically downscaled multiple GCMs.

However, when the metric-based weights are used on the ensemble mean, we show that they do not consistently improve the bias over the unweighted ensemble mean for precipitation or temperature in the NCEP- or GCM-driven simulations (see Section 3.3.). This result is similar to the findings in other studies (e.g. Fowler and Ekström 2009, Christensen et al. 2010, Knutti et al. 2010, Weigel et al. 2010). The increased bias in some regions with weighting is likely the result of a combination of factors. (1) There is no direct measure of mean bias in the metrics; (2) the metrics do not capture all of the possible errors or processes that could feed into model mean skill; and (3) relative performance can still give high weight to models that are not performing well (as in Klocke et al. 2011).

Comparing weights in 2 regions to in-depth analyses in Section 3.5 highlights the fact that metrics that are expected to be meaningful across a wide variety of regions clearly fail to take account of the effects of important region-specific processes. This is evidenced when the universal metrics do not identify the same well or poor performing simulations as the in-depth analyses. Creating weights that are specifically tailored to a region's processes based on in-depth analysis may be a worthwhile exercise, but may also be subject to similar complications. For example, in Bukovsky et al. (2015), a binary quasi-metric based on process-level analysis was explored

in the discussion, and it successfully differentiated the quality of the simulations, and may have produced a wide range of weights. However, a similar approach would likely not work to substantially differentiate credibility or produce diverse weights in TS15, as no simulation performed similarly well across all process-based measures.

Given the experiment in Section 3.3 using baseline climate simulations, the number of metrics and sub-metrics included in the final weights, and the 0–1 normalization required for application of the weights, we believed that using the metric-based weights on our ensemble means for climate projections would not produce substantial differences in the regional mean climate projections. Thus, when applying the metric-based weights to the projections in Section 3.4., we also applied a large, generic set of 'all possible weights' to create a PDF of possible climate changes that could result from any similar weighting scheme.

Overall, the results of this latter experiment illustrate that it is unlikely that a weighted ensemble mean change will be significantly different than an unweighted ensemble mean change in any weighting system where the weights must sum to 1 unless, unsurprisingly, there is one factor that dominates over the rest. If a large percentage of weight is put on a simulation or the simulations that are correspondingly more extreme than the rest, then significant differences in ensemble means are possible, but one would need to closely examine whether or not that outcome was credible.

Consequently, we find that weighting 'does make a difference', in that it can change the ensemble mean. But, the difference after weighting is not always better, meaningful, or significant. We also find that weighting 'can make a difference', but this depends heavily on the distribution of weight, and usually requires that most of the weight go to few of the more extreme simulations.

Overall, we do not judge these metrics as useless. On their own, the sub-metrics are useful for highlighting potential model problems. They become diluted in combination though, and do not substantially differentiate the simulations. We also do not see these outcomes regarding the weighting as an argument for 'model democracy' in an ensemble. It is clear from in-depth analyses that some simulations are more credible than others, and should be given more weight. However, given this work, weighting an ensemble in any systematic, meaningful way across many regions using performance metrics is likely to bear limited fruit.

Acknowledgements. We thank Samuel Somot, Emilia Sanchez Gomez, Erasmo Buonomo, Tomas Halenka, and Erika Coppola for their correspondence and assistance in reproducing the ENSEMBLES metrics. We also thank Steve Sain and Tammy Greasby for useful discussions and calculations in the early stages of this work. The authors also acknowledge the support of the NOAA Climate Program Office Modeling, Analysis, Predictions and Projections (MAPP) Program. This work was supported under Grant NA11AOR4310111. This work was also supported by the NCAR Weather and Climate Impacts Assessment Science Program funded by the National Science Foundation (NSF) under the NCAR cooperative agreement managed by Dr. Linda O. Mearns, by the Strategic Environmental Research and Development Program (SERDP) under Contract 2516, and by DOE RGCM grant DE-SC0016438. We also acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) for the use of the Geyser analysis cluster provided by NCAR's Computational and Information Systems Laboratory, sponsored by the NSF. We also thank NARCCAP for providing the model data used in this paper. NARCCAP was funded by the NSF, the U.S. Department of Energy (DoE), the National Oceanic and Atmospheric Administration (NOAA), and the U.S. Environmental Protection Agency Office of Research and Development (EPA). NCEP Reanalysis 2 data were provided by the NOAA/OAR/ESRL PSD, Boulder, CO, USA at www.esrl.noaa.gov/psd/.

LITERATURE CITED

Alexandru A, de Elia R, Laprise R, Separovic L, Biner S (2009) Sensitivity study of regional climate model simulations to large-scale nudging parameters. Mon Weather Rev 137:1666–1686

Bukovsky MS (2011) Masks for the Bukovsky regionalization of North America. www.narccap.ucar.edu/contrib/bukovsky/

Bukovsky MS (2012) Temperature trends in the NARCCAP regional climate models. J Clim 25:3985–3991

Bukovsky MS, Gochis DJ, Mearns LO (2013) Towards assessing NARCCAP regional climate model credibility for the North American monsoon: current climate simulations. J Clim 26:8802–8826

Bukovsky MS, Carrillo CM, Gochis DJ, Hammerling DM, McCrary RR, Mearns LO (2015) Towards assessing NARCCAP regional climate model credibility for the North American monsoon: future climate simulations. J Clim 28:6707–6728

Caya D, Laprise R (1999) A semi-implicit semi-Lagrangian regional climate model: The Canadian RCM. Mon Weather Rev 127:341–362

Chen J, Brissette FP, Lucas-Picher P, Caya D (2017) Impacts of weighting climate models for hydro-meteorological climate change studies. J Hydrol (Amst) 549:534–546

Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M (2010) Weight assessment in regional climate models. Clim Res 44:179–194

Collins WD, Bitz CM, Blackmon ML, Bonan GB and others (2006) The Community Climate System Model: CCSM3. J Clim 19:2122–2143

Coppola E, Giorgi F, Rauscher SA, Piani C (2010) Model weighting based on mesoscale structures in precipitation and temperature in an ensemble of regional climate models. Clim Res 44:121–134

ENSEMBLES (2009) ENSEMBLES deliverable D3.2.2. RCM-specific weights based on their ability to simulate the present climate, calibrated for the ERA40-based simulations. http://ensembles-eu.metoffice.com/deliverables.html

Eum HI, Gachon P, Laprise R (2014) Developing a likely climate scenario from multiple regional climate model simulations with an optimal weighting factor. Clim Dyn 43: 11–35

Flato GM, Boer GJ, Lee WG, McFarlane NA, Ramsden D, Reader MC, Weaver AJ (2000) The Canadian Centre for Climate Modeling and Analysis global coupled model and its climate. Clim Dyn 16:451–467

Foley A, Fealy R, Sweeney J (2013) Model skill measures in probabilistic regional climate projections for Ireland. Clim Res 56:33–49

Fowler HJ, Ekström M (2009) Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. Int J Climatol 29:385–416

GFDL GAMDT (The GFDL Global Atmospheric Model Development Team) (2004) The new GFDL global atmosphere and land model AM2-LM2: evaluation with prescribed SST simulations. J Clim 17:4641–4673

Gillett NP (2015) Weighting climate model projections using observational constraints. Philos Trans R Soc A 373: 20140425

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method. J Clim 15:1141–1158

Giorgi F, Marinucci MR, Bates GT (1993a) Development of a second-generation regional climate model (RegCM2). I. Boundary-layer and radiative transfer processes. Mon Weather Rev 121:2794–2813

Giorgi F, Marinucci MR, De Canio G, Bates GT (1993b) Development of a second-generation regional climate model (RegCM2). II. Convective processes and assimilation of lateral boundary conditions. Mon Weather Rev 121:2814–2832

Gordon C, Cooper C, Senior CA, Banks H and others (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16:147–168

Grell GA, Dudhia J, Stauffer DR (1993) A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech Note NCAR/TN-398+1A. NCAR, Boulder, CA

Haughton N, Abramowitz G, Pitman A, Phipps SJ (2015) Weighting climate model ensembles for mean and variance estimates. Clim Dyn 45:3169–3181

Holtanová E, Mikšovský J, Kalvová J, Pišoft P, Motl M (2012) Performance of ENSEMBLES regional climate models over Central Europe using various metrics. Theor Appl Climatol 108:463–470

Jones RG, Hassell DC, Hudson D, Wilson SS, Jenkins GJ, Mitchell JFB (2003) Workbook on generating high-resolution climate change scenarios using PRECIS. Hadley Centre for Climate Prediction and Research, Met Office, Bracknell. www.uncclearn.org/sites/default/files/inventory/undp17.pdf

Juang HMH, Hong SY, Kanamitsu M (1997) The NCEP regional spectral model: an update. Bull Am Meteorol Soc 78:2125–2143

Kanamitsu M, Ebisuzaki W, Woollen J, Yang SK, Hnilo JJ, Fiorino M, Potter GL (2002) NCEP-DOE AMIP-II Reanalysis (R-2). Bull Am Meteorol Soc 83:1631–1643

Kjellström K, Boberg F, Castro M, Christensen JH, Nikulin G, Sánchez E (2010) Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. Clim Res 44:135–150

Klocke D, Pincus R, Quaas J (2011) On constraining estimates of climate sensitivity with present-day observations through model weighting. J Clim 24:6092–6099

Knutti R (2010) The end of model democracy? Clim Change 102:395–404

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758

Lenderink G (2010) Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations. Clim Res 44:151–166

Lorenz P, Jacob D (2010) Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40. Clim Res 44:167–177

Maurer EP, Wood AW, Adam JC, Lettenmaier DP, Nijssen B (2002) A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States. J Clim 15:3237–3251

McCrary RR, McGinnis S, Mearns LO (2017) Evaluation of snow water equivalent in NARCCAP simulations, including measures of observational uncertainty. J Hydrometeorol 18:2425–2452

Mearns LO, McGinnis S, Arritt R, Biner S, and others (2007) The North American Regional Climate Change Assessment Program dataset. National Center for Atmospheric Research Earth System Grid data portal, Boulder, CO (updated 2014)

Mearns LO, Arritt R, Biner S, Bukovsky MS and others (2012) The North American Regional Climate Change Assessment Program: overview of phase I results. Bull Am Meteorol Soc 93:1337–1362

Mearns LO, Sain S, Leung LR, Bukovsky MS and others (2013) Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP). Clim Change 120:965–975

Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. Int J Climatol 25: 693–712

Nakićenović N, Alamco J, Davis G, de Vries B, Fenhann J, Gaffin S and others (2000) Special report on emissions scenarios. Cambridge University Press, Cambridge. http://pure.iiasa.ac.at/id/eprint/6101/1/emissions_scenarios.pdf

Pal JS, Giorgi F, Bi X, Elguindi N and others (2007) Regional climate modeling for the developing world: the ICTP RegCM3 and RegCNET. Bull Am Meteorol Soc 88: 1395–1409

Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3. Clim Dyn 16: 123–146

Riddle EE, Stoner MB, Johnson NC, L'Heureux ML, Collins DC, Feldstein SB (2013) The impact of the MJO on clusters of wintertime circulation anomalies over the North American region. Clim Dyn 40:1749–1766

Ring C, Pollinger F, Kaspar-Ott I, Hertig E, Jakobeit J, Paeth H (2018) A comparison of metrics for assessing state-of-the-art climate models and implications for probabilistic projections of climate change. Clim Dyn 50:2087–2106

Sanchez-Gomez E, Cassou C, Hodson DLR, Keenlyside N, Okumura Y, Zhou T (2008) North Atlantic weather

regimes response to Indian-western Pacific Ocean warming: a multi-model study. Geophys Res Lett 35:L15706

Sanchez-Gomez E, Somot S, Déqué M (2009) Ability of an ensemble of regional climate models to reproduce weather regimes over Europe-Atlantic during the period 1961–2000. Clim Dyn 33:723–736

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, Powers JG (2005) A description of the Advanced Research WRF version 2. NCAR Tech Note NCAR/TN-468+STR. NCAR, Boulder, CO

Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainties in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. J Clim 18:1524–1540

Thibeault JM, Seth A (2015) Toward the credibility of Northeast United States summer precipitation projections in CMIP5 and NARCCAP simulations. J Geophys Res Atmos 120:10050–10073

Wehner MF (2013) Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. Clim Dyn 40:59–80

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. J Clim 23:4175–4191

Editorial responsibility: Filippo Giorgi, Trieste, Italy

Submitted: January 15, 2018; Accepted: October 2, 2018
Proofs received from author(s): November 29, 2018