Clim Dyn
DOI 10.1007/s00382-015-2537-x

Evaluation of large-scale meteorological patterns associated with temperature extremes in the NARCCAP regional climate model simulations

Paul C. Loikith · Duane E. Waliser · Huikyo Lee · J. David Neelin · Benjamin R. Lintner · Seth McGinnis · Linda O. Mearns · Jinwon Kim

Received: 5 November 2014 / Accepted: 20 February 2015

Abstract Large-scale meteorological patterns (LSMPs) associated with temperature extremes are evaluated in a suite of regional climate model (RCM) simulations contributing to the North American Regional Climate Change Assessment Program. LSMPs are characterized through composites of surface air temperature, sea level pressure, and 500 hPa geopotential height anomalies concurrent with extreme temperature days. Six of the seventeen RCM simulations are driven by boundary conditions from reanalysis while the other eleven are driven by one of four global climate models (GCMs). Four illustrative case studies are analyzed in detail. Model fidelity in LSMP spatial representation is high for cold winter extremes near Chicago. Winter warm extremes are captured by most RCMs in northern California, with some notable exceptions. Model fidelity is lower for cool summer days near Houston and extreme summer heat events in the Ohio Valley. Physical interpretation of these patterns and identification of well-simulated cases, such as for Chicago, boosts confidence in the ability of these models to simulate days in the tails of the temperature distribution. Results appear consistent with the expectation that the ability of an RCM to reproduce a realistically shaped frequency distribution for temperature, especially at the tails, is related to its fidelity in simulating LMSPs. Each ensemble member is ranked for its ability to reproduce LSMPs associated with observed warm and cold extremes, identifying systematically high performing RCMs and the GCMs that provide superior boundary forcing. The methodology developed here provides a framework for identifying regions where further process-based evaluation would improve the understanding of simulation error and help guide future model improvement and downscaling efforts.

Keywords Temperature extremes · Regional climate modeling · Large-scale meteorological patterns · North America · Model evaluation

P. C. Loikith (✉) · D. E. Waliser · H. Lee
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91101, USA
e-mail: paul.c.loikith@jpl.nasa.gov

D. E. Waliser · J. Kim
Joint Institute for Regional Earth System Science and Engineering, University of California Los Angeles, Los Angeles, CA, USA

J. D. Neelin
Department of Atmospheric and Oceanic Science, University of California, Los Angeles, CA, USA

B. R. Lintner
Department of Environmental Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

S. McGinnis · L. O. Mearns
Institute for Mathematical Applications to the Geosciences, National Center for Atmospheric Research, Boulder, CO, USA

1 Introduction

Temperature extremes are associated with severe impacts across a range of societal sectors including human health, agriculture, and energy production. Furthermore, anticipated changes in temperature extremes resulting from anthropogenic global warming are expected to have an increasingly severe impact on society (Seneviratne et al. 2012). Several recent studies provide evidence that externally forced changes are already observable over many parts of the world (Coumou et al. 2013; Donat et al. 2013; Min et al. 2013; Morak et al. 2013; Peterson et al. 2013;

Published online: 04 March 2015

Zwiers et al. 2011) with more substantial changes anticipated for the future (Bindoff et al. 2013; Coumou and Robinson 2013; Meehl and Tebaldi 2004; Sillmann et al. 2013). In light of this, it is crucial to carefully assess the ability of current-generation climate models to capture extreme events. Towards this goal, the present study evaluates the ability of a suite of dynamically downscaled regional climate models (RCMs) participating in the North American Regional Climate Change Assessment Program (NARCCAP; Mearns et al. 2012) to simulate key large-scale meteorological patterns (LSMPs) associated with extreme temperature days over North America.

LSMPs associated with extreme warm temperatures have been characterized for individual events (Beniston and Diaz 2004; Dole et al. 2011; Meehl and Tebaldi 2004). Using cluster analysis, Stefanon et al. (2012) show anticyclonic anomalies at 500 hPa associated with heatwave patterns over Europe. Similarly, using empirical orthogonal function analysis, Lau and Nath (2012, 2014) isolated strong associations between anticyclonic anomalies and extreme heat for several regions of North America and Europe. Loikith and Broccoli (2012) developed several metrics using gridded observations to identify and describe LSMPs associated with daily warm and cold temperature extremes systematically over the North American domain, finding that most temperature extremes are associated with synoptic scale forcing. Loikith and Broccoli (2015) further evaluated the LSMPs in a suite of global climate models (GCMs) contributing to the Coupled Model Intercomparison Project Phase 5 and found that the ensemble generally reproduces the observed spatial patterns of LSMPs associated with temperature extremes.

While the focus of regional downscaling is often on improving the representation of small-scale features and extremes (e.g. heat waves in Vautard et al. (2013)), large-scale patterns and their relation to temperature and temperature extremes have been analyzed in RCMs. Bowden et al. (2012) investigated the representation of weather regimes and their impacts on temperature in RCMs over North America and evaluated the benefits of interior nudging on the simulation of temperature and precipitation. Sanchez-Gomez et al. (2009) evaluated the ability of a suite of RCMs to simulate large-scale weather regimes over Europe and found that in some cases the RCMs degrade the representation of the driving large-scale patterns. Linking dynamics and extreme events, Clark and Brown (2013) evaluated the influence of LSMPs on European heat extremes using an ensemble of regionally downscaled model simulations.

At larger scales, recurrent modes of low frequency weather and climate variability, e.g. the Northern Annular Mode (NAM), El Niño Southern Oscillation (ENSO), Pacific North America Pattern (PNA), have been associated with extreme temperatures over North America (Gershunov and Barnett 1998; Griffiths and Bradley 2007; Kenyon and Hegerl 2008; Wettstein and Mearns 2002), often in conjunction with synoptic scale LSMPs (Loikith and Broccoli 2014). Westby et al. (2013) demonstrated the influence that these modes have on unusually cold and warm temperature events and suggest that current generation GCMs often have difficulty in reproducing these associations.

Because temperature extremes occur in the tails of the temperature probability distribution function (PDF), characteristics of the PDF tails at a given location provide insight into the physical processes controlling or influencing extremes (Loikith et al. 2013). Moreover, the shape of the PDF has important implications for the vulnerability of a particular location to future changes in extremes (Ruff and Neelin 2012) making temperature PDFs an important target for model evaluation.

While the spatial scale of typical LSMPs associated with temperature extremes is large enough (~10^3 km) that model grid resolution is not expected to have much influence on model fidelity, the occurrence of extreme temperature events at particular times or locations may depend on local scale effects that are sensitive to resolution, such as topography or coastlines. Furthermore, the mechanisms associated with extreme daily temperatures for one location may differ substantially from nearby locations, especially in regions of complex terrain or along coastlines. For example, the synoptic conditions associated with extreme warm temperatures in coastal Southern California promote local downslope winds commonly referred to as "Santa Ana's" (Hughes and Hall 2009) while different conditions are associated with extreme warmth tens of km inland. In this sense, LSMPs should be viewed as proxies for processes such as temperature or moisture advection or more local scale processes that influence local temperature. Therefore, data with high spatial resolution is desirable for both understanding the mechanisms associated with and predicting changes in extreme temperatures on scales that are relevant to society. To avoid the computation expense required to run GCMs at high resolution, RCMs are commonly used to downscale GCM output over a target region. Although downscaling does not guarantee improved model performance over GCMs, several studies (Di Luca et al. 2011; Feser et al. 2011; Paeth and Mannig 2012) have demonstrated added value for extreme events.

Biases in NARCCAP surface temperature have been previously documented (Kim et al. 2013; Mearns et al. 2012; Rangwala et al. 2012). Motivating the current study (Loikith et al. 2015) comprehensively evaluated PDF morphology in the NARCCAP hindcast RCM suite. While the NARCCAP RCMs reproduce temperature skewness with reasonable fidelity in the winter, larger model-observation disagreement is evident in summer. This illustrates the difficulty in simulating the tails of the temperature PDF and

Table 1 Full names of RCMs and GCMs contributing to NARCCAP with associated references

RCMs	Model name	References
CRCM	Canadian regional climate model	Caya and Laprise (1999)
ECP2	NCEP regional spectral model	Juang et al. (1997)
HRM3	Hadley regional model 3	Jones et al. (2004)
MM5I	PSU/NCAR mesoscale model	Grell et al. (1993)
RCM3	Regional climate model version 3	Pal et al. (2007)
WRFG	Weather research and forecasting	Shamarock et al. (2005)
GCMs		
CCSM	NCAR community climate system model, version 3	Collins et al. (2006)
GFDL	GFDL climate model, version 2.1	Anderson et al. (2004)
HADCM3	Hadley centre climate model, version 3Q0	Gordon et al. (2000) and Pope et al. (2000)
CGCM3	Canadian global climate model, version 3	Flato et al. (2000)

consequently temperature extremes, at least in some seasons. The present study builds mechanistically on (Loikith et al. 2015), with the overarching goal of identifying where and with which RCM–GCM configurations temperature extremes are simulated with the highest fidelity by plausible physical processes.

The rest of this paper is organized as follows. Section 2 discusses the model and reference data, and Sect. 3 outlines the methodology. Section 4 focuses on four selected cases, with a domain-wide analysis described in Sect. 5. RCM simulation rankings and concluding remarks are presented in Sect. 6.

2 Data

2.1 Model data

This work evaluates a total of 17 simulations produced using six RCMs contributing to NARCCAP (Mearns et al. 2009, 2012), (http://www.narccap.ucar.edu). Data are provided every 3 h on a 50 km horizontal grid. The National Center for Environmental Prediction (NCEP) Reanalysis II (Kanamitsu et al. 2002) provides the boundary conditions for the hindcast experiment (six simulations) and four GCMs provide boundary conditions for the remaining eleven historical simulations (Table 1). The NARCCAP domain covers all of the conterminous United States (US) and much of Canada and Northern Mexico. As model development progresses, higher resolution RCMs are becoming available over North America e.g. Wang and Kotamarthi (2013), however the coordinated, multi-RCM/multi-GCM framework for NARCCAP allows for systematic evaluation of multiple RCMs and the influence from choice of boundary forcing.

The NCEP-driven simulations officially cover the years 1980–2004, however data for all variables was only available for the 23-year period of 1980–2002. The GCM-driven simulations cover the years 1971–2000, however all datasets were not complete after 1998. Therefore, the years 1976–1998 are used to evaluate the GCM-driven simulations in an effort to have the same sample size as the hindcasts while maximizing temporal overlap. Surface air temperature (TAS), sea level pressure (SLP), and 500 hPa geopotential height (Z500) are used to compute LSMPs.

2.2 Reference data

Two reference datasets are employed. The Wang and Zeng (2014) 2-m temperature dataset, based on the National Aeronautics and Space Administration (NASA) Modern Era-Retrospective Analysis for Research and Applications (MERRA) reanalysis, is used to identify extreme temperature days and to compute TAS LSMPs. This dataset, introduced in Wang and Zeng (2013) is produced by bias correcting MERRA (Rienecker et al. 2011) reanalysis hourly 2-m air temperature with monthly gridded in situ maximum and minimum 2-m air temperature from the Climate Research Unit Time Series version 3.10 (CRU 3.10; Mitchell and Jones (2005)). This dataset (MERRA–CRU from now on) is a global, land only, hourly TAS dataset on a 0.5° latitude/longitude grid mesh with substantially reduced uncertainty compared with standard MERRA reanalysis. The NCEP North American Regional Reanalysis (NARR; Mesinger et al. (2006)) is used as reference for SLP and Z500. NARR is originally provided on a 32 km grid. NARR is not used to define TAS extremes because NARR does not assimilate TAS, introducing bias (Loikith et al. 2015; Mesinger et al. 2006).

2.3 Data processing

Daily means were computed from the NARCCAP 3-hourly and MERRA–CRU 1-hourly output. All NARCCAP TAS

data are regridded to a common 0.5° latitude/longitude grid mesh, the same as the MERRA–CRU grid, using a kriging algorithm implemented with a thin plate spline (TPS) routine (Fields 2006). Surface elevation is provided as a covariate and interpolation is performed only over land grid points. NARR SLP and Z500 data are interpolated to the same grid using a more computationally efficient linear method based on Delaunay triangulation (Lee and Schachter 1980; Youn et al. 2006). Because it was not as crucial to reduce smoothing and preserve extremes for SLP and Z500, it was decided to use this more efficient interpolation method over kriging. All reference data are subset to match the NCEP-driven time period of 1980–2003.

3 Methodology

All data are de-seasonalized by subtracting the daily climatological mean from each day. Evaluation was performed for the seasons of summer (June, July, August; JJA) and winter (December, January, February; DJF) and warm and cold temperature extreme days are defined as those days falling within the lower (cold) and upper (warm) 5 % of the temperature anomaly distribution. For DJF (JJA) there are 90 (92) days per season for 23 years resulting in a total sample size of 2070 (2116) days. This results in about 104 (106) extreme temperature days for each type of extreme (warm and cold). The exception is for the HRM3–HadCM3 and MM5I–HadCM3 runs, which have a 360-day calendar resulting in 90-day JJA seasons.

LSMPs are constructed by computing the composite mean of the anomaly fields for each variable (TAS, SLP, and Z500) for all extreme warm or cold temperature days at a given grid point (see Figs. 2, 3, 4, 5, 6 for examples). TAS, SLP, and Z500 are chosen to represent the spatial extent of the anomalously warm/cold airmass, the near-surface circulation and thermal advection, and the mid-tropospheric circulation respectively. Ensemble-mean LSMPs are computed by averaging all six (eleven) LSMPs for the NCEP- (GCM-) driven suites of simulations.

There are other valid methods to defining extreme temperature days, each with benefits and limitations. Here, the percentile threshold definition was chosen so that all locations had an equal frequency of extreme days, even in the case of a highly skewed frequency distribution. Additionally, the choice of 5 % over a more lenient or stringent threshold limits the analysis to days that are relatively infrequent yet results in a relatively large sample size for computing LSMPs. One limitation to this choice is that each extreme event is considered independent, even if it is part of a multi-day outbreak. This has the benefit of constructing an LSMP that is necessarily associated with the most extreme days, but could impact the statistical robustness of the composite anomalies in the case of small number of independent samples. In the four cases discussed in Sect. 4, most of the days included are temporally separated from other extreme days (80 % of days for Chicago DJF, 81 % of days for California DJF, 66 % of days for Ohio JJA, and 64 % of days for Houston JJA). The slower progression of summer synoptic systems likely results in a lower percentage of independent days in the JJA examples compared with DJF.

The comparison metrics are based on the root mean square error, normalized by the spatial standard deviation of the reference pattern (RMSE hereafter), computed between the model and reference LSMPs. The data are area weighted by multiplying all grid cells by the square root of the cosine of latitude before summing the difference. Only data within a 4500 (4000) km radius of a target grid cell for which the LSMP is computed are included in the metric comparisons for DJF (JJA). Although the precise value of this radius is arbitrary, we suggest a value on the order of several 1000 km is reasonable for including large-scale structure while at the same time excluding areas too far away to be relevant to the extreme temperature occurrence. A larger threshold is used for DJF than JJA because of the typically larger spatial scale of winter compared with summer LSMPs (Loikith and Broccoli 2012).

4 Individual cases

Four individual cases are selected to evaluate and analyze the LSMPs associated with extreme temperature days in detail. The four cases were chosen to exemplify a range of model behaviors, dynamical conditions, and societal impacts. Temperature anomaly distributions are presented in Fig. 1 for each case to provide qualitative comparison of the distribution tails. Figures 2, 3, 4, 5 and 6 present the LSMPs while Fig. 7 summarizes the results of Figs. 2, 3, 4, 5 and 6 in the form of a portrait diagram. Throughout this section, both Figs. 1 and 7 are referenced for each individual case.

4.1 Chicago DJF cold extremes

Winter cold extremes near Chicago, as demonstrated during the winter of 2013–2014, can have severe impacts on society including disruptions to transportation, increases in energy demand, and threats to human health and safety. The temperature distribution for Chicago (Fig. 1a, e) is characterized by a modest long cold tail in MERRA–CRU, consistent with the longer-than-Gaussian cold tail found in station data in Ruff and Neelin (2012). Most of the NCEP-driven runs capture the overall distribution shape, as they

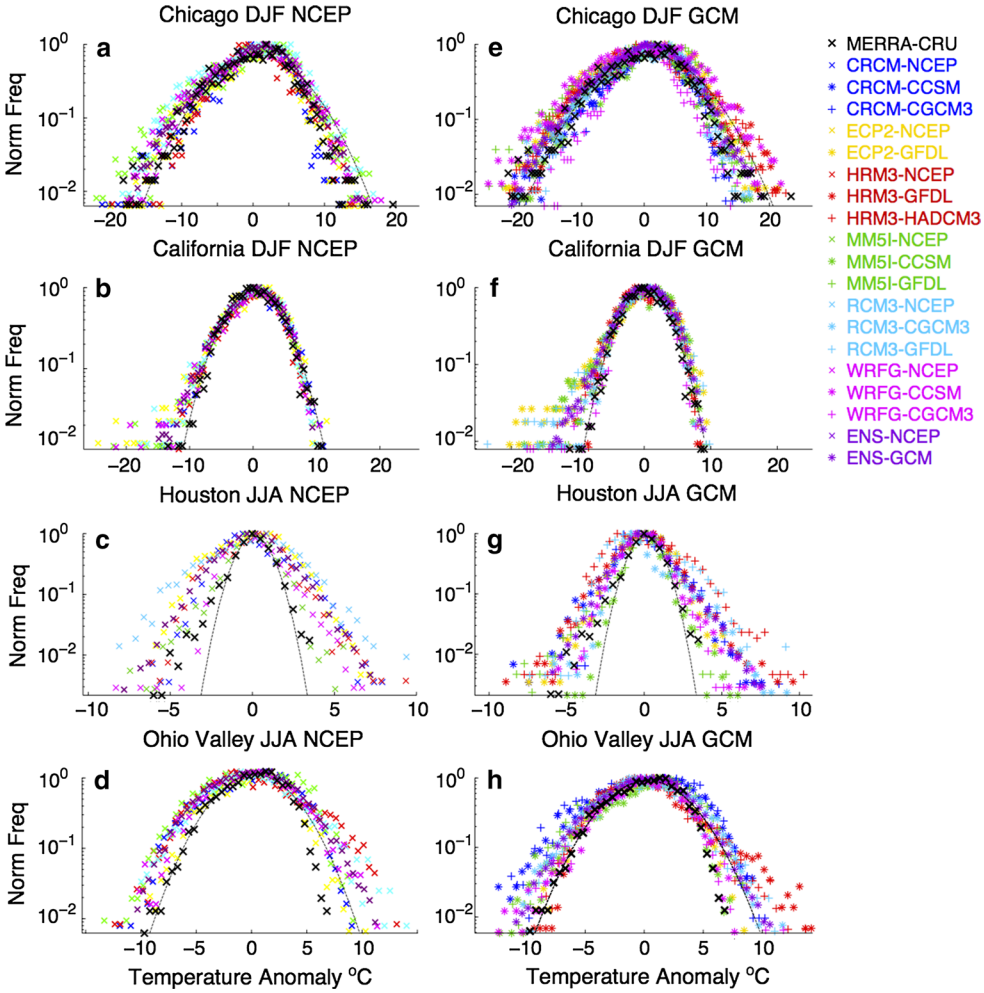

Fig. 1 Daily temperature anomaly distributions for **a**–**d** NCEP- and **e**–**h** GCM-driven simulations at the four individual grid cells as discussed in Sect. 4 and indicated on the maps in Figs. 2, 3, 4, 5 and 6. Bin widths are 0.5 °C and frequencies are normalized by the maximum bin count and plotted on a log scale. *Black X's* are MERRA–CRU and the *dotted curves* are the Gaussian fit to the core of the MERRA–CRU distribution. Bin counts are plotted for each simulation based on the *color* and *symbol* in the legend on the *right*

exhibit weak asymmetry with a slightly long cold tail and slightly short warm tail, while there is notably larger spread among the GCM-driven (Fig. 1e) simulations. Days in the cold tail occur as part of a large cold airmass covering much of North America (Fig. 2a). Strong positive SLP anomalies (Fig. 2d) associated with the cold airmass of Arctic origin are present locally, to the west and south, and extend upstream to the northwest while negative SLP anomalies associated with a cyclone along the strong baroclinic zone on the leading edge of the cold airmass are present downstream. A deep Z500 trough (~200 m) is located overhead and slightly downstream (Fig. 2g).

Both the NCEP- and GCM-driven runs capture these features with the most notable difference being the weaker positive SLP anomalies to the west and southwest of Chicago. The summary panel in Fig. 7a shows that the ensemble mean reflects the performance of most of the individual ensemble members with RMSE values generally near or below 0.5, indicating error substantially lower than the spatial variability of the reference pattern. The largest disagreement is for SLP, consistent with Fig. 2d–f. LSMPs associated with extreme warm days (LSMPs not shown) are also well simulated albeit with slightly poorer agreement for SLP. In most cases, the NCEP-driven simulations exhibit similarly high fidelity to their GCM-driven counterparts, with the ensemble mean patterns showing nearly identical comparison metric values.

The dominant influence of synoptic-scale atmospheric circulation along with the lack of locally influential geographic or topographic features render this region and season one for which climate models may be expected to perform well in simulating extreme winter cold, regardless of horizontal resolution, which these results support. While the proximity of Lake Michigan to the east may affect local climate, the dominance of advection from the north and west makes it unlikely that the lake, and the way it is represented in the RCMs, would have a substantial influence on extreme cold winter temperatures. The strong performance at reproducing the conditions associated with extreme cold events here boosts confidence in the ability of these models to simulate temperature extremes in such regions.

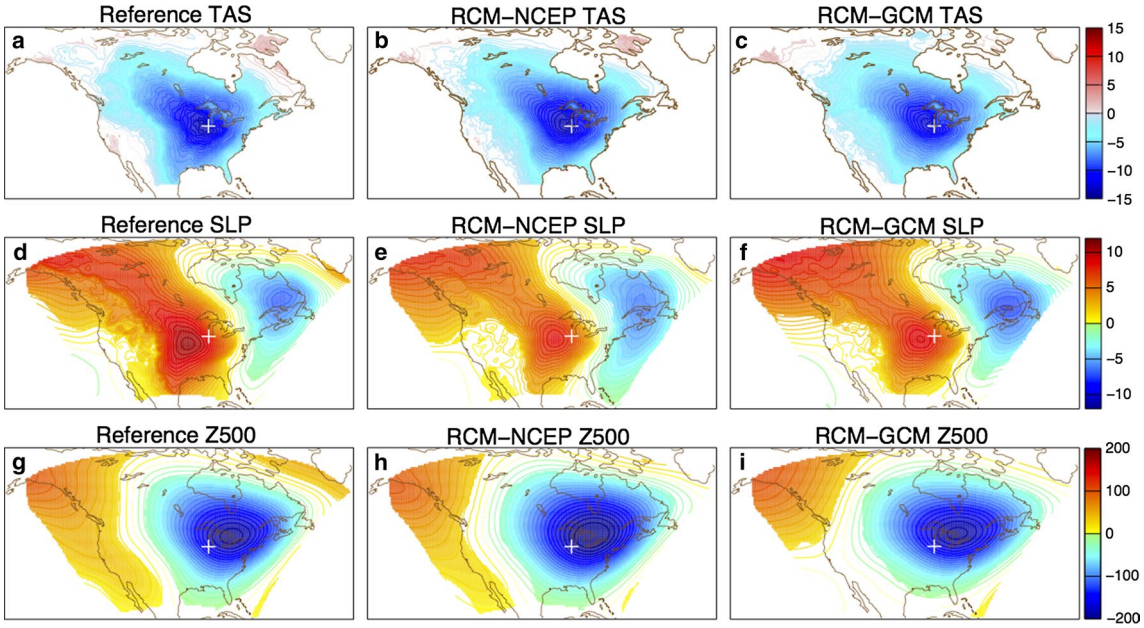

Fig. 2 LSMPs for **a–c** TAS (°C), **d–f** SLP (hPa), and **g–i** Z500 (m) anomalies at the Chicago grid cell indicated by the *white plus symbol*. All results are for extreme cold DJF days. The *left column* is for the reference data, the *middle* for the NCEP-driven RCMs, and the *right column* for the GCM-driven RCMs. All model LSMPs are composite means of the individual ensemble member LSMPs. For the reference panels, only grid cells with anomalies significantly different from zero at the 5 % significance level according to a *t* test are *shaded*. For the model panels, only grid cells where at least half of the models contributing to the ensemble mean show statistical significance and have the same sign anomaly as the reference pattern are shaded. See Sect. 4.1 for discussion

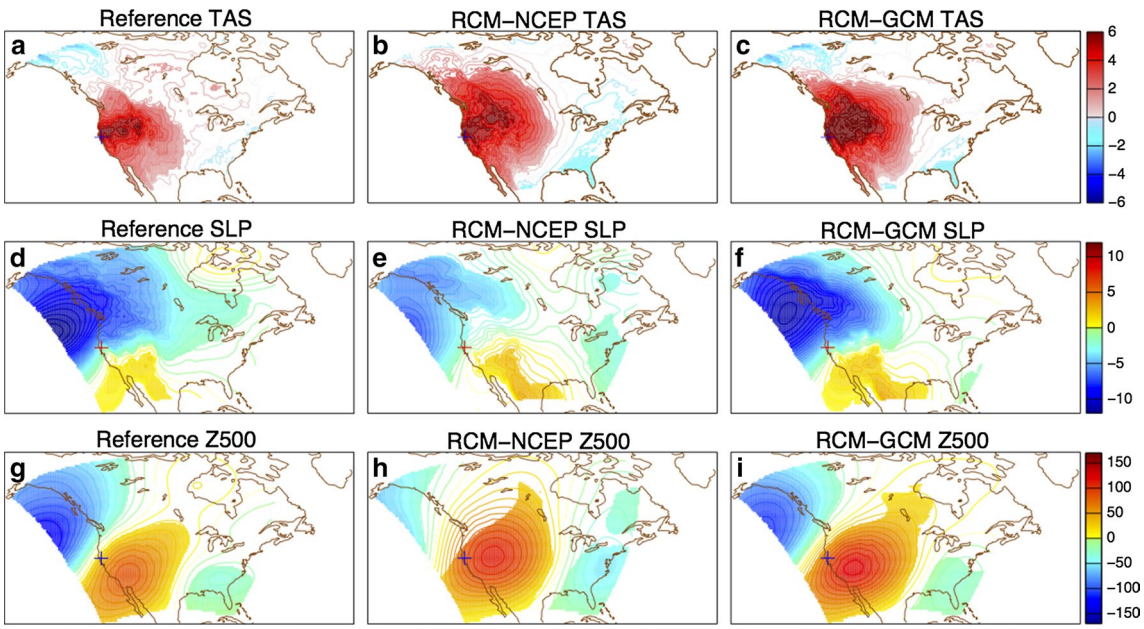

Fig. 3 Same as in Fig. 2 except for extreme warm DJF days at the California example indicated by the *plus symbol*. See Sect. 4.2 for discussion

4.2 Northern California DJF warm extremes

The complex orography of California leads to multiple climate zones across a relatively small horizontal distance making this region a model resolution challenge. MERRA–CRU shows a nearly Gaussian frequency distribution in Fig. 1b, f; on the other hand, many of the models show long cold tails, especially for the ECP2, RCM3,

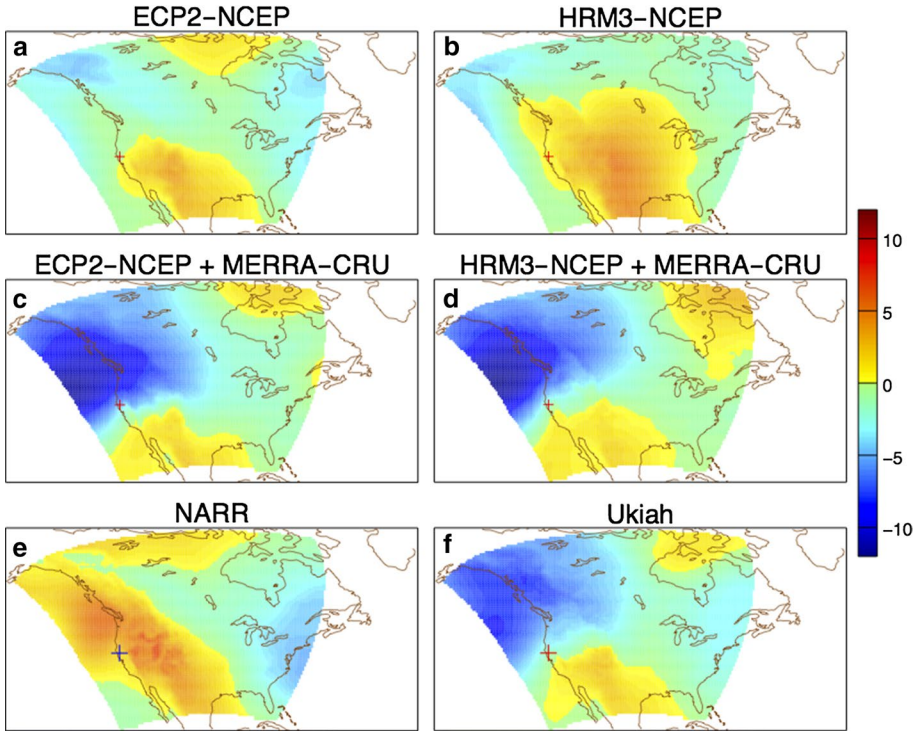

Fig. 4 SLP composite LSMPs (hPa) for the California grid cell for **a** the ECP2–NCEP and **b** HRM3–NCEP model runs, with extreme days defined based on the model climatology. **c**, **d** Same as **a**, **b** except extreme days are defined in the MERRA–CRU reference dataset. **e** SLP LSMP for extreme warm days in NARR and **f** for extreme warm days in Ukiah station data. All SLP values are from NARR, while the source of the surface temperature anomalies used to identify extreme days differs. See Sect. 4.2 for discussion

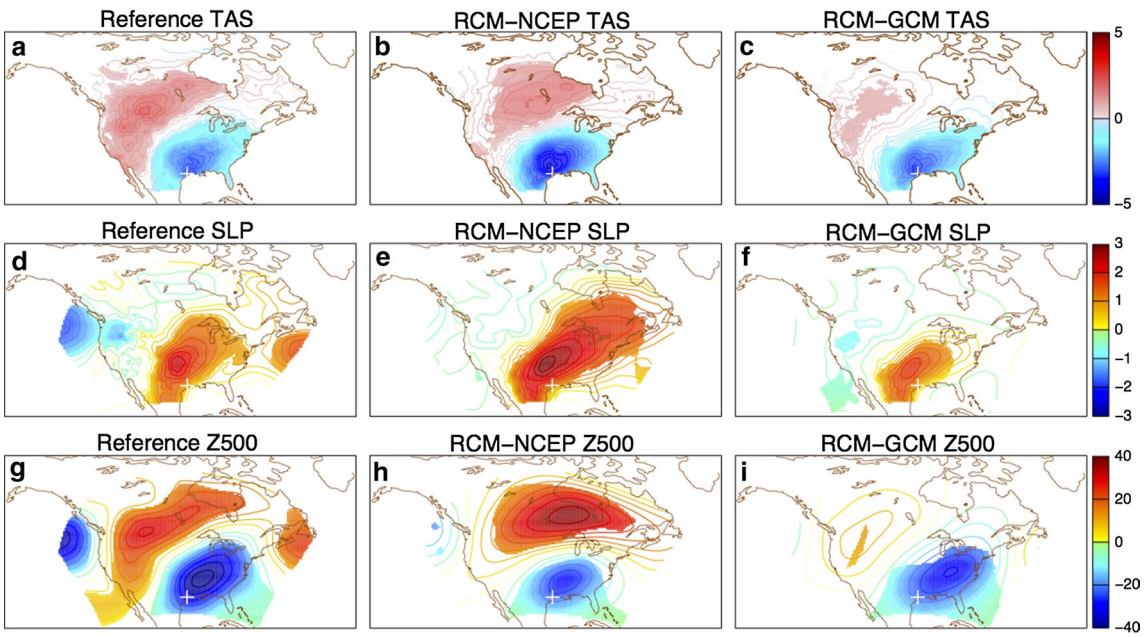

Fig. 5 Same as in Fig. 2 except for extreme cold JJA days at the Houston grid cell indicated by the *white plus symbol*. See Sect. 4.3 for discussion

and WRFG simulations, though the warm tails agree well.

The reference pattern in Fig. 3d shows that at this grid cell, the warmest days occur when SLP anomalies promote a strong southerly component to the low-level winds. The grid cell is located in the middle of a strong SLP gradient with large negative anomalies to the northwest, indicative of an upstream cyclone, and weak positive anomalies to the south and east. Strong warm anomalies encompass much of the domain (Fig. 3a), but the warmest anomalies appear to

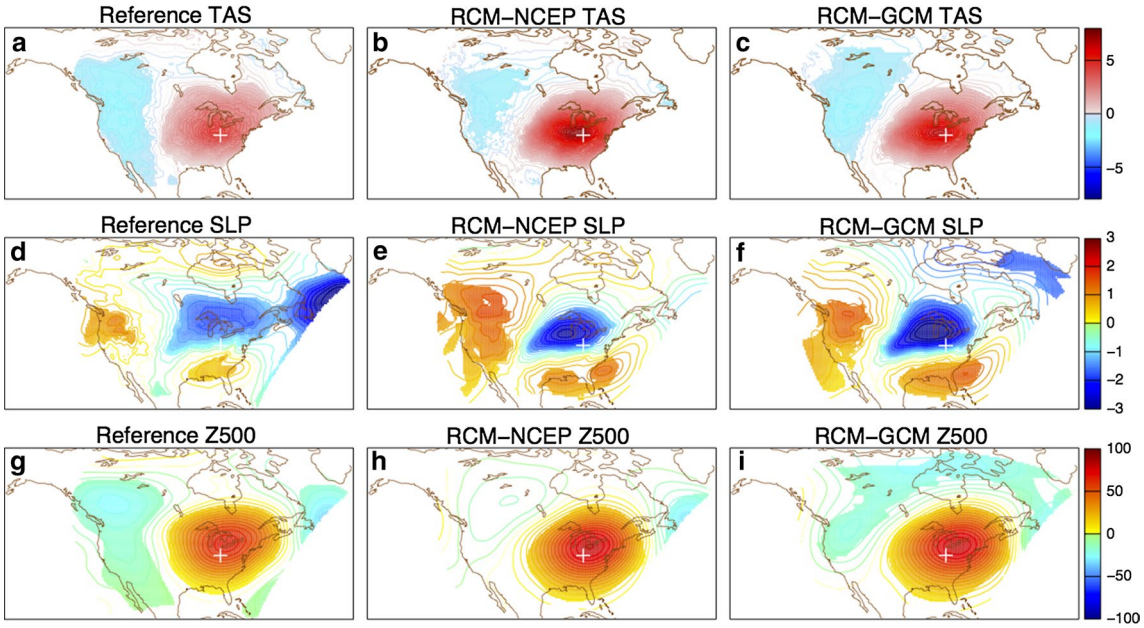

Fig. 6 Same as in Fig. 2 except for extreme warm JJA days at the Ohio Valley grid cell indicated by the *white plus symbol*. See Sect. 4.4 for discussion

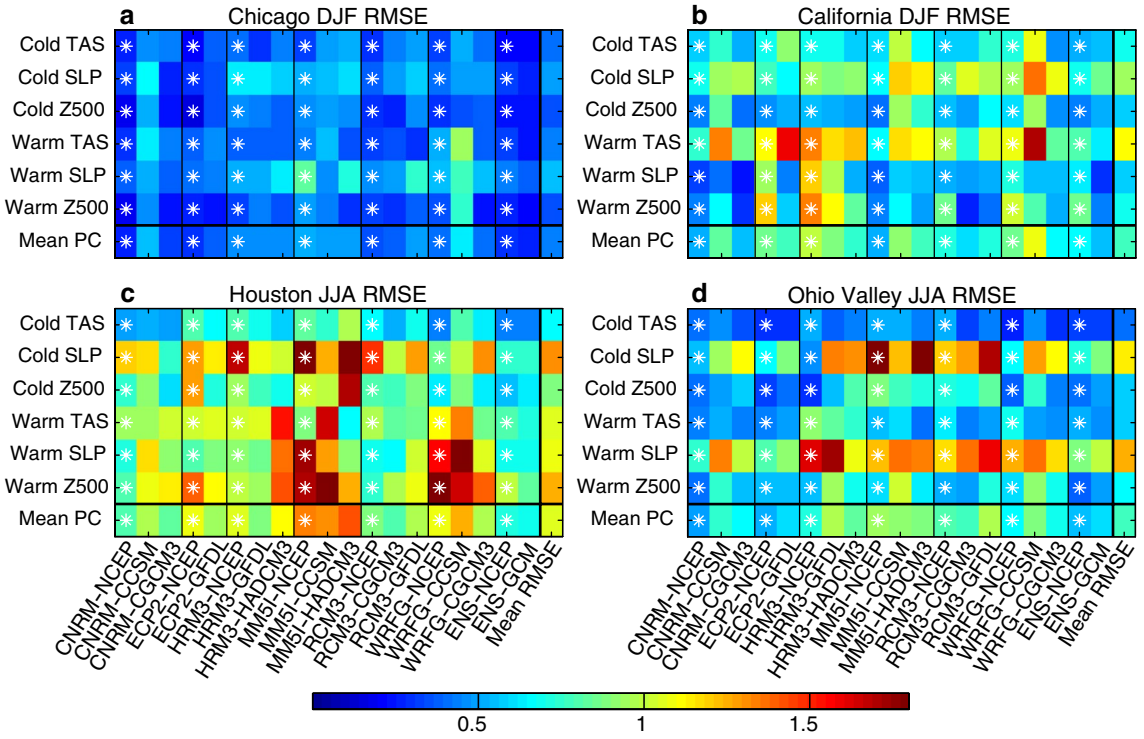

Fig. 7 The RMSE between the reference and simulated LSMPs, normalized by the spatial standard deviation of the reference LSMP for all NARCCAP runs and all variables for cold and warm temperature extremes. Results for **a** are discussed in Sect. 4.1, **b** Sect. 4.2, **c** Sect. 4.3, and **d** Sect. 4.4. Results for the NCEP-driven runs are represented with a white star. Model configurations using different RCMs are separated by *vertical black lines*. The *bottom row* is the mean value for the *corresponding column* and the *column on the far right* is the mean value for the *corresponding row*, excluding the ENS–NCEP and ENS–GCM values

align with the maximum SLP gradient. At Z500, an amplified wave train pattern is clear with a downstream ridge and upstream trough associated with the upstream surface cyclone.

Overall, the NCEP- and GCM-driven TAS patterns show more widespread warm anomalies than observed, while the GCM-driven TAS pattern also shows widespread very warm anomalies (>6 °C). The synoptic setup of the SLP patterns is well reproduced by the models with the GCM-driven ensemble showing better agreement than the NCEP-driven simulations. The negative SLP anomalies to the north and west are substantially weaker than reference for the NCEP-driven simulations. Z500 shows similar behavior with the amplitude and orientation of the wave train better realized in the GCM-driven ensemble.

It is somewhat counterintuitive that simulations driven by reanalysis would resemble the reference pattern with less fidelity than simulations driven by GCMs. It is apparent in Fig. 7b that ECP2–NCEP and HRM3–NCEP are the leading contributors to the larger error for the NCEP-driven ensemble mean LSMPs. To investigate further, Fig. 4a, b shows the SLP LSMPs for ECP2–NCEP and HRM3–NCEP respectively. The most striking difference between these patterns and the reference pattern in Fig. 3d is the lack of strong negative SLP anomalies to the north and west of the grid cell. In ECP2–NCEP, the field is largely dominated by positive anomalies to the south with weak anomalies to the north and west while HRM3–NCEP shows a large area of strong positive SLP anomalies over much of the US. The HRM3–NCEP pattern is suggestive of an offshore component to the low level winds. These patterns are robust at more restrictive extremes thresholds (not shown). Along coastal California, such conditions promote surface warming primarily by inhibiting the moderating influences of the predominant onshore flow. While these patterns are completely different than the reference pattern, it is not unreasonable to expect anomalously warm temperatures under these synoptic conditions.

To investigate the role of boundary forcing, Fig. 4c, d depicts the SLP composites for ECP2–NCEP and HRM3–NCEP produced using the same days contributing to the reference LSMP. Because these are hindcasts driven by reanalysis, it is expected that the RCMs strongly resemble observed conditions for a given day. For both RCMs, the SLP pattern is very similar to the reference pattern suggesting the errors in Fig. 4a, b are not being introduced by the boundary forcing. The key difference is that the local surface temperature anomalies for these days, while anomalously warm, are not above the 95th percentile of the distribution (not shown). This indicates that the RCMs are able to produce large-scale dynamics given realistic boundary forcing, but unable to realize extreme warm temperatures resulting from these dynamics. This may result in part from problems related to the simulation of the boundary layer or surface processes including the local influence of topographical features. If model evaluation was based solely on temperature in this case, ECP2 and HRM3 may be misleadingly chosen as being well suited for making future projections of extremes, which highlights the value of analyzing LSMPs.

To rule out uncertainty in the reference dataset, the LSMPs are recomputed using days that are extremely warm in NARR TAS (processed the same way as the RCMs as described in Sect. 2.3). As in the other reference LSMPs, NARR SLP is used to compute the composites, but here NARR TAS rather than MERRA–CRU TAS determines which days contribute to the composite. The NARR SLP pattern shown in Fig. 4e is fundamentally different from the reference pattern in Fig. 3d but shows some resemblance to the ECP2–NCEP and HRM3–NCEP patterns. The NARR SLP gradient is highly suggestive of warming due to inhibition of onshore winds. Because NARR does not assimilate surface temperature (Mesinger et al. 2006), it is reasonable to suspect biases in the NARR TAS.

To reconcile the differences between the MERRA–CRU and NARR-based results, station data from nearby Ukiah, California (39.1°N × 123.2°W), obtained from the National Climate Data Center's Global Surface Summary of the day product, are used to identify extreme temperature days. The SLP composite pattern in Fig. 4d is computed using NARR SLP in the same manner as the reference pattern in Fig. 3d, but extreme days are defined with the Ukiah station data. The resulting LSMP strongly resembles the reference LSMP in Fig. 3d, obtained using the MERRA–CRU temperature climatology. Together this supports MERRA–CRU as a reliable TAS observational dataset for this grid cell, and suggests that NARR TAS is biased, and that despite being forced by reanalysis, ECP2–NCEP and HRM3–NCEP reproduce the dynamics associated with extreme warm DJF temperatures with low fidelity at this grid cell.

While it is hard to definitively explain why GCM-driven simulations using these same RCMs perform better, it is possible that additional biases introduced by the GCM boundary conditions compensate for the inherent bias in the RCMs. While the LSMPs are not shown, it is interesting to note in Fig. 7b that the RMSE is also higher for the NCEP-driven ECP2 and HRM3 for the SLP patterns associated with extreme cold days than the corresponding GCM-driven simulations suggesting problems with the simulation of the dynamics associated days in both tails of the temperature distribution. Furthermore, previous work has identified HRM3 as possessing an outstanding warm bias across much of the domain in winter and summer (Kim et al. 2013), potentially indicative of issues with the simulation of temperature in this run.

4.3 Houston JJA cold extremes

Anomalously cold summertime temperatures in the Houston region are not associated with widespread climate impacts, however, Ruff and Neelin (2012) identified Houston as having a non-Gaussian long cold tail and Loikith et al. (2015) documented large uncertainty in temperature PDF tails in NARCCAP hindcasts. This suggests that capturing realistic characteristics of extremes is challenging here with important implications for future projections. Figure 1c, g shows a long cold tail for MERRA–CRU, consistent with Ruff and Neelin (2012). Overall, the RCMs show difficulty in reproducing this feature, with several members showing pronounced long warm tails, especially in the HRM3 and RCM3 configurations. The spread is also greater for the GCM-driven simulations than those forced by reanalysis.

Cold extremes in the Houston region occur as part of an area of negative temperature anomalies that radiate outward from the Gulf of Mexico (Fig. 5a). Warm anomalies are present across the Rocky Mountains of the US and central and western Canada. Significant positive SLP anomalies are largely coincident with the region of negative TAS anomalies (Fig. 5d), suggestive of a continental airmass of high latitude origin with the SLP gradient suggestive of northeasterly anomalies in the surface wind. This pattern has commonalities with the synoptic mechanisms for Chicago DJF extreme cold in Sect. 4.1, albeit with weaker and smaller scale anomalies. A negative Z500 anomaly center is located to the north of Houston with positive anomalies over central Canada (Fig. 5g). While Houston is far removed from an active storm track and regions of large horizontal temperature gradients in the summer, the overall synoptic pattern is suggestive of a cool airmass originating from higher latitudes.

The TAS pattern for the NCEP-driven runs resembles the reference pattern (Fig. 5b). The GCM-driven simulations also reproduce the TAS pattern well locally. The NCEP-driven SLP pattern shows stronger positive anomalies than the reference (Fig. 5e) while the GCM-driven positive anomaly is spatially smaller (Fig. 5f). At Z500, both the NCEP- and GCM-driven RCMs capture the negative anomaly center near Houston, but the GCM-driven pattern resembles a more progressive wave-train than is apparent in the reference albeit without a statistically robust upstream ridge (Fig. 5h, i).

Figure 7c shows relatively strong model performance for TAS patterns with weaker performance for Z500 and even weaker performance for SLP patterns for extreme cold days near Houston. Similar to the California case, several RCMs manifest larger errors in SLP when forced by NCEP reanalysis than a GCM (CRCM, ECP2, HRM3, RCM3) even though the resulting ensemble mean pattern shows similar RMSE for NCEP- and GCM-driven simulations. Interestingly, TAS LSMPs associated with cold extremes are reproduced better than warm extremes at Houston. Warm extremes may be more difficult to capture in these LSMPs because of the likely influence from small-scale features or processes that influence the surface energy budget, such as anomalous soil moisture (Berg et al. 2014; Fischer et al. 2007).

4.4 Ohio Valley JJA extreme warm days

Summer heatwaves are often associated with the most severe societal climate impacts and as such are arguably the most prominent research focus for future changes in temperature extremes. Extreme summer heat in the Ohio Valley region of the US often occurs in conjunction with high humidity levels, further exacerbating the health impacts of such conditions. The temperature distribution for the Ohio Valley grid cell in Fig. 1d, h shows a slightly short warm tail for MERRA–CRU with most simulations exhibiting a wider distribution at both tails. This points to issues in simulating the magnitude of extreme warm events. The short warm tail at this location would result in a relatively large increase in the number of extreme warm exceedances due to a simple shift in the temperature PDF (Ruff and Neelin 2012), which underscores the importance of proper simulation of warm extremes here.

Heatwaves in the Ohio Valley are part of an anomalously warm airmass encompassing much of the eastern half of North America (Fig. 6a) with the grid cell on the eastern edge of the warmest temperatures. The negative SLP anomalies to the north of the grid cell (Fig. 6d) promote southwesterly flow, suggesting ongoing neutral or warm air advection. The extent of the warm anomalies is captured well by the models (Fig. 6b, c) with some indication of a warm bias, consistent with the wider tails of the anomaly distribution in Fig. 1. The broad characteristics of observed SLP anomalies are evident in the models; however, both the NCEP- and GCM-driven runs show a more amplified SLP pattern (Fig. 6e, f). At Z500, both the position and extent of the positive anomalies centered over the Ohio Valley and Great Lakes are captured well by the models (Fig. 6g–i), except the anomalies are more positive in the simulations than in the reference, consistent with the warm bias.

Despite the qualitatively reasonable pattern agreement, the models typically exhibit high RMSE values, especially for SLP (Fig. 7d). Values greater than one indicate that the difference between the reference and model patterns is larger than the spatial variability of the reference pattern. One contributing factor to this is the difference in SLP anomaly sign and strength in the northern portion of

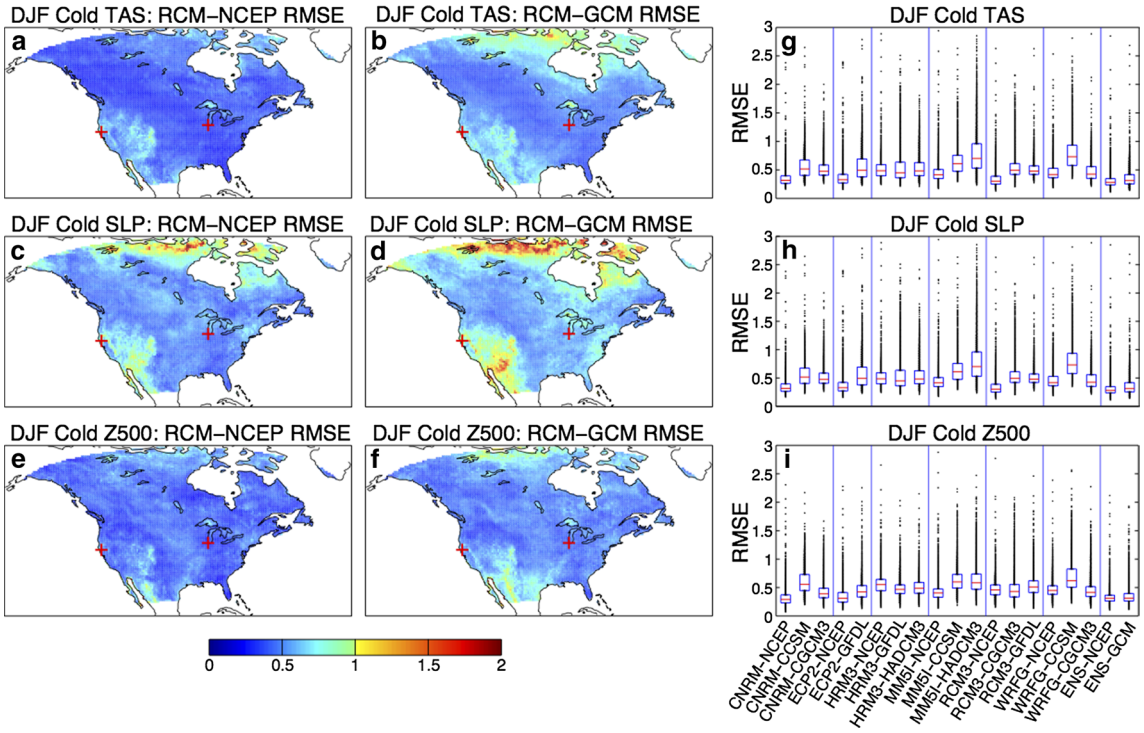

Fig. 8 **a**–**f** Median RMSE computed between the reference LSMP and each of the (*left*) six NCEP- and (*right*) eleven GCM-driven simulated LSMPs for DJF cold extremes. All RMSE values are normalized by the spatial standard deviation of the reference pattern before the median is computed. Maps are for **a**, **b** TAS, **c**, **d** SLP, and **e**, **f** Z500 LSMPs. *Red plus symbol*'s indicate the California and Chicago grid cells. **g**–**i** Box-and-whisker plots showing the RMSE value at every grid cell in the domain for all NARCCAP runs. The *horizontal red lines* indicates the median values while the blue boxes outline the 25th and 75th percentiles and *black dots* are outliers. The RCM and driving boundary forcing is labeled along the *x*-axis and organized such that the NCEP-driven runs are always to the left of the GCM-driven runs for that same RCM. The *vertical blue lines* delineate each of the RCMs, with the ensemble mean on the right. See Sect. 5 for discussion

the domain. While largely not statistically significant, the GCM-driven runs show negative anomalies over northern Canada while the NCEP RCMs and the reference show positive anomalies. The NCEP-driven runs generally perform better than the GCM-driven runs in this case. The LSMPs for days in the cold tail of the distribution show better agreement, possibly because there is a stronger synoptic component to unusually cool days while extreme heat is influenced by more local-scale processes such as land–atmosphere coupling that may not be captured in these LSMPs.

The stronger agreement for the TAS and Z500 patterns and weaker agreement for SLP suggests that temperature extremes at this grid cell may be more influenced by anomalies at Z500 than the near-surface circulation. Because of the relatively weak horizontal temperature gradients present during the summer, it is reasonable that near-surface circulation is not as influential as subsidence under a large Z500 ridge allowing the models to capture the extent and magnitude of the warm airmass with reasonable fidelity while exhibiting poor representation of the SLP pattern.

5 Evaluation of LSMPs over North America

Expanding on the cases presented in Sect. 4, RMSE is computed for each LSMP at each continental grid cell. Each value plotted on the maps in Figs. 8, 9, 10 and 11 is the median RMSE value from the six (eleven) NCEP- (GCM-) driven ensembles for that grid cell. The median RMSE is chosen over the mean to reduce the influence of outliers.

5.1 DJF cold extremes

Figure 8a–f shows the RMSE across the entire domain for LSMPs associated with extreme cold DJF days. Box-and-whisker plots (Fig. 8g–i) show the RMSE value at every grid cell for each ensemble member grouped by RCM. In the box-and-whisker plots, the simulations are aligned so that all simulations with the same RCM are grouped together, with the NCEP-driven run on the left. TAS LSMPs for extreme cold DJF days (Fig. 8a, b) show low error over much of the domain for both the NCEP- and GCM-driven simulations with the only areas of elevated RMSE over the mountains of the western US and far northern Canada.

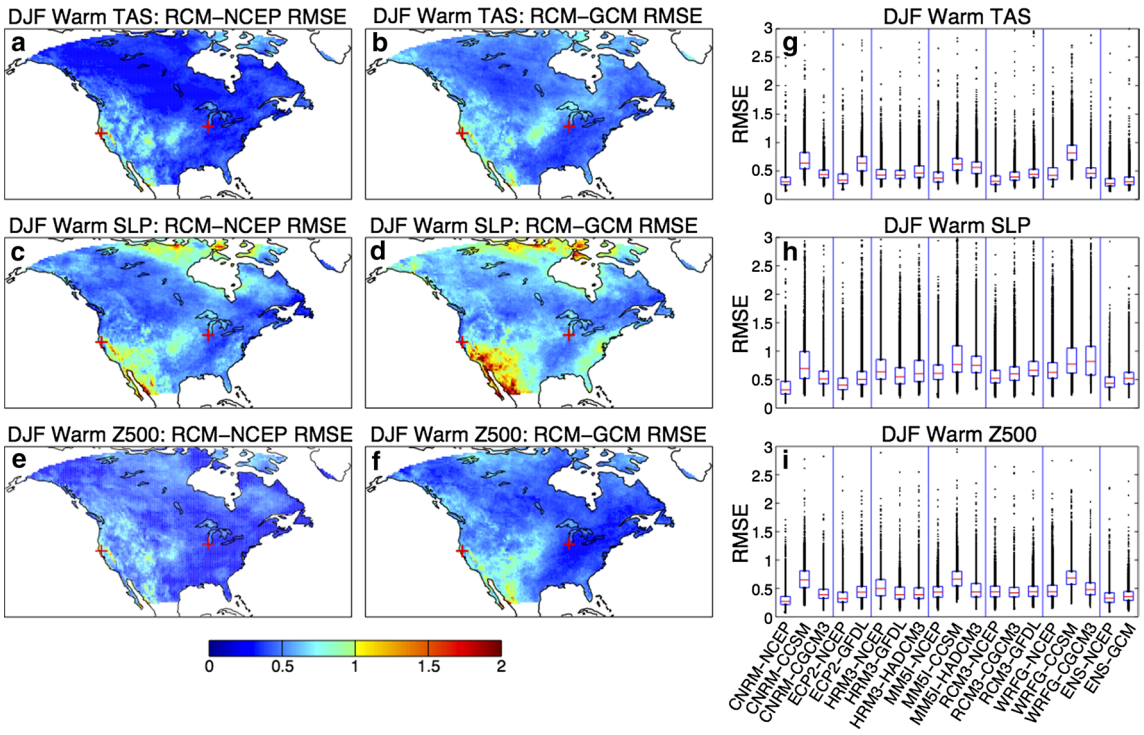

Fig. 9 Same as in Fig. 8, except for DJF warm extremes

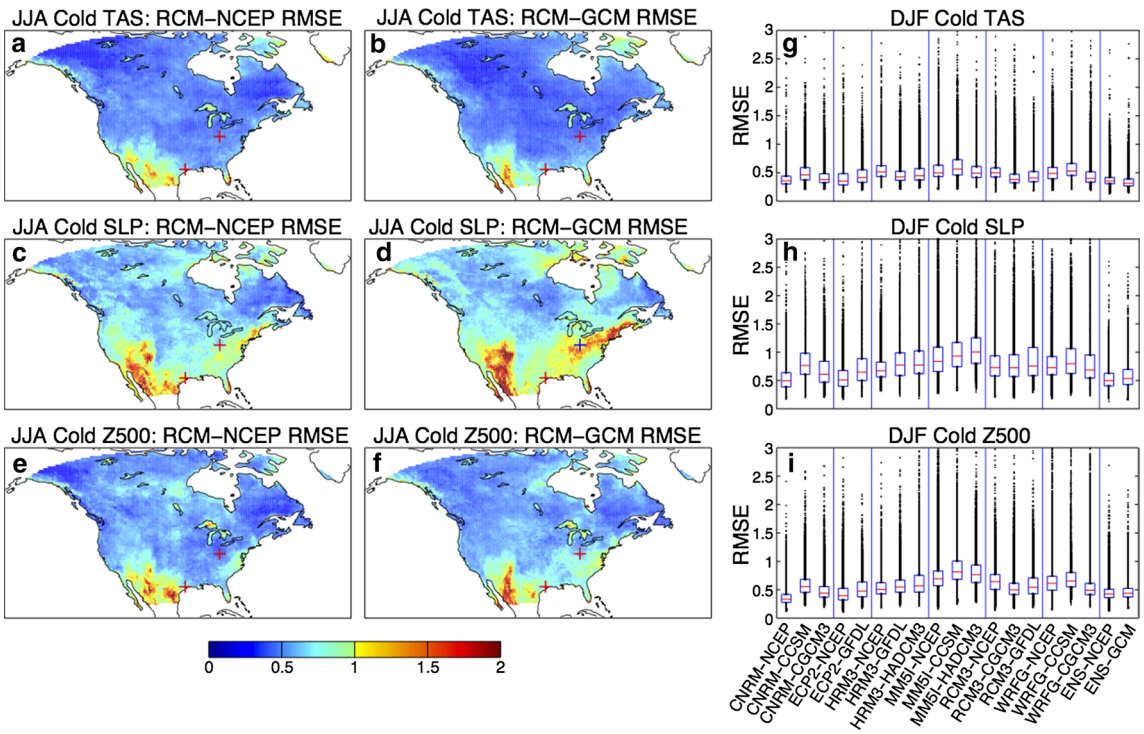

Fig. 10 Same as in Fig. 8 except for JJA cold extremes, with the *plus symbol*'s representing the Ohio Valley and Houston grid cells

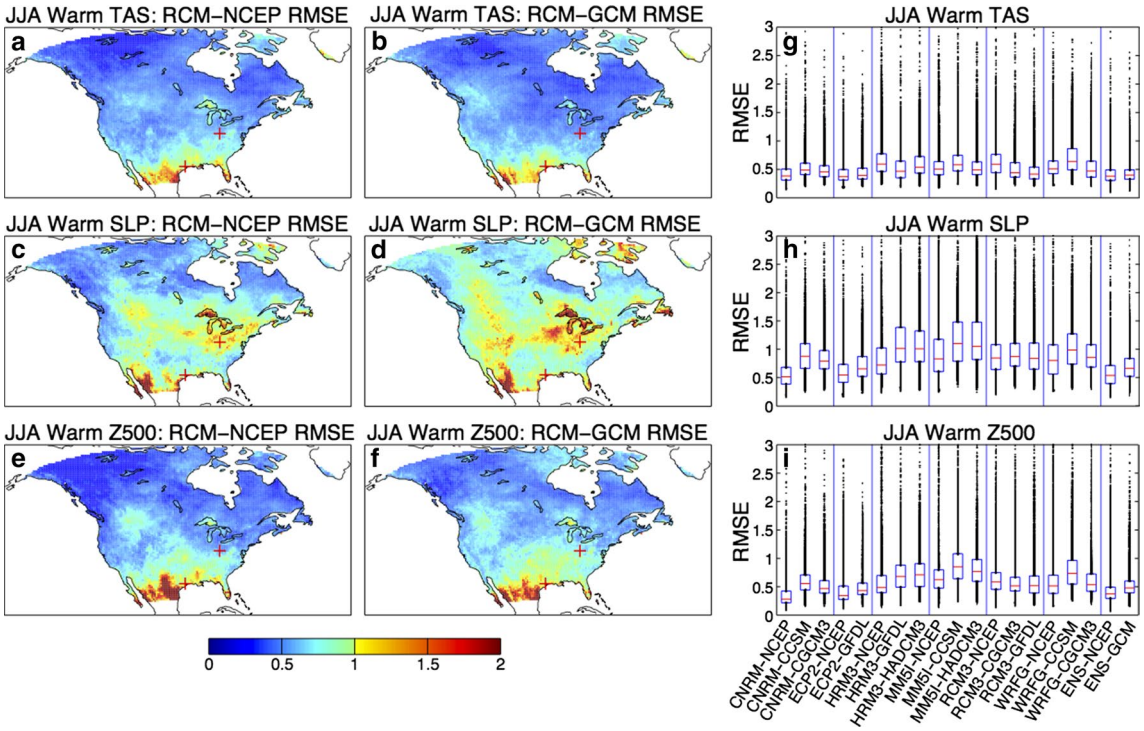

Fig. 11 Same as in Fig. 8, except for JJA warm extremes, with the *plus symbol's* representing the Ohio Valley and Houston grid cells

Overall, RMSE is higher for GCM-driven simulations compared with the NCEP-driven runs. Individual models tend to reflect similar behavior, as the median RMSE values are higher for the GCM-driven runs except for HRM3.

RMSE is higher for the SLP patterns (Fig. 8c, d), with high error over northern Canada and the southwestern portion of the domain, amplified in the GCM–RCM simulations. CRCM–NCEP, ECP2–NCEP and the ensemble means show the lowest RMSE overall, with relatively high values for many of the GCM-driven runs including WRFG–CCSM, RCM3–GFDL, CRCM–CCSM, and MM5I–CCSM (Fig. 8h). In some cases there are very large differences between the NCEP- and GCM-driven simulations. For example, CRCM–NCEP has notably low RMSE while CRCM–CCSM has relatively high values. MM5I–NCEP compared with MM5I–CCSM shows similar behavior. This suggests that substantial error is being introduced by the driving GCM, especially CCSM.

Z500 patterns (Fig. 8e, f) largely match the spatial patterns of SLP RMSE but with lower error. The box plot in Fig. 8i also resembles the boxplot for SLP with CCSM configurations standing out as having relatively high error. Because TAS shows elevated RMSE in many of the same regions as SLP and Z500, it follows that proper simulation of the dynamics is important for reproducing the extent and strength of the anomalous airmasses associated with extreme cold DJF temperature events. In all cases, the Chicago example presented in Sect. 4 (Fig. 2) is regionally representative with low RMSE over much of the central and eastern portion of the domain.

5.2 DJF warm extremes

The maps of RMSE for extreme warm DJF days (Fig. 9a–f) show elevated error in many of the same regions as for extreme cold DJF days (Fig. 8). TAS pattern agreement (Fig. 9a, b) is relatively weak over the western and central US and stronger over the eastern and northern portions of the domain. These values, as for DJF cold extreme LSMPs, are mostly less than one, indicating that error is smaller than the spatial variability of the reference pattern. Individual ensemble members vary with all NCEP-driven runs except HRM3 showing lower RMSE values than their GCM-driven counterparts (Fig. 9g). CRCM, ECP2, and RCM3 stand out as strong performers when driven by NCEP, while RCM3–CGCM3, CRCM–CGCM3, and HRM3–GFDL stand out as superior GCM-driven simulations. The CCSM-driven simulations show elevated RMSE similar to the DJF extreme cold results.

SLP patterns (Fig. 9c, d) exhibit the highest error, especially in the GCM-driven runs. The southwestern US and northern Mexico along with far northern Canada exhibit the highest RMSE, with GCM-driven values substantially higher than the NCEP simulations. There are a number of

likely sources contributing to the elevated error at high latitudes. Observational uncertainty may be larger here, since in situ observations are sparser than at lower latitudes. Additionally, the RCMs may have difficulty in simulating the physics and dynamics of the very stable boundary layer. Lastly, the proximity of the high latitudes to the domain boundary may interfere with the spatial extent of the LSMP that is being evaluated, with the potential for important features existing beyond the RCM boundary. The individual ensemble members (Fig. 9h) show similar relative behavior as in TAS and for DJF extreme cold SLP (Fig. 8h) in most cases with CRCM–NCEP and ECP2-NCEP showing the lowest and the three CCSM configurations showing the highest RMSE overall. The California grid cell is within a local area of moderate RMSE, however as discussed in Sect. 3, error is slightly larger for the NCEP- compared with the GCM-driven simulations. This is not the case domain-wide.

At Z500 (Fig. 9e, f), RMSE is lower than for SLP although regions of elevated RMSE are generally coincident with SLP and TAS. In general, the simulations that perform well (poorly) for one variable also perform well (poorly) for the other variables. For example, CRCM–NCEP and ECP2–NCEP exhibit notably low error for TAS, SLP, and Z500, while the opposite is true for the three CCSM-driven runs. This multi-variate consistency highlights the important role of large-scale dynamics on the occurrence of temperature extremes. This also suggests that there is considerable error being inherited from CCSM as the NCEP-driven runs of CRCM, MM5I, and WRFG have considerably lower RMSE than their CCSM-driven counterparts.

5.3 JJA cold extremes

For JJA cold extremes, the TAS RMSE (Fig. 10a, b) is low over much of the domain with elevated values from Mexico northward into the southwest US. Boundary forcing does not have a strong influence on the RMSE values domain wide, as RMSE is only slightly higher for RCMs driven by NCEP compared to GCMs. Similar to the DJF examples, RCMs forced with CCSM show elevated error relative to the corresponding hindcast results.

SLP RMSE (Fig. 10c, d) shows large values in some of the same regions as in TAS. RCM–NCEP configurations show somewhat elevated RMSE along the Appalachian Mountains through the northeastern US with this area highly amplified and expanded in the GCM-driven simulations. Mirroring the DJF behavior, the best RCM configurations are CRCM–NCEP and ECP2–NCEP (Fig. 10h). MM5I–HADCM3 stands out as having the highest RMSE for SLP. Z500 shows relatively high RMSE (Fig. 10e, f) in the same areas as SLP. The area along the southwestern Great Plains and eastern Rocky Mountains shows consistently high RMSE for all three variables, suggesting a substantial dynamical contribution to the error in simulating extreme temperatures there. CRCM–NCEP and ECP2–NCEP also show the lowest RMSE for Z500 (Fig. 10i), while the MM5I configurations stand out as having the highest error overall, as for SLP. The Houston example is generally within a coherent region of low to moderate RMSE for all three variables, suggesting the results in Sect. 4.3 are somewhat representative of the northwestern Gulf of Mexico coast.

5.4 JJA warm extremes

Of the four types of temperature extremes analyzed here, warm summertime extremes are associated with the most severe impacts and are often associated with other extreme conditions such as drought (Fischer et al. 2007) and air pollution (Jacob and Winner 2009). Furthermore, it is anticipated that warm extremes will become more intense and persistent due to anthropogenic climate warming (Seneviratne et al. 2012). The TAS patterns for warm summertime extremes (Fig. 11a, b) show low RMSE over the northern 2/3 of the domain with elevated values over the southern third. Intra-ensemble variability is relatively low (Fig. 11g), however CRCM–NCEP and ECP2–NCEP stand out as having relatively low error while WRFG–CCSM shows the highest RMSE. In some cases the NCEP-driven error is larger than the GCM-driven simulations, but the ensemble means reflect nearly identical error for either boundary forcing.

SLP pattern agreement shows relatively high RMSE domain wide with higher error in the GCM-driven runs. On the other hand, the Pacific coast of the US and Canada and portions of Northern Canada show relatively low error. Individual model error is often very large with median RMSE values near 1.0 for the MM5I–GCM simulations. Consistent with other cases, CRCM and ECP2 have superior skill when forced with NCEP although when forced by GCMs they produce substantially larger errors.

Circulation anomalies aloft are associated with lower RMSE than SLP, especially over the central US and the mountains of the western US and Canada (Fig. 11e, f). The geographic distribution of RMSE is qualitatively similar between TAS and Z500, while SLP is distinct. This suggests that circulation at Z500 may be more important for the occurrence and magnitude of temperature extremes than SLP. The behavior described for the Ohio Valley case in Sect. 4.4 is consistent with this hypothesis, as it is located in a coherent region of low RMSE for TAS and Z500 and high RMSE for SLP. These results suggest the need for further analysis of other key processes, such as the influence of low soil moisture on extreme heat, that may influence or

control NARCCAP RCM fidelity in simulation of extreme warm summer temperatures.

6 Discussion and conclusions

6.1 RCM–GCM configuration ranks

To provide an assessment of the relative merit of the RCMs and the driving GCMs, the simulations are ranked with respect to a single measure of RMSE (Fig. 12). This measure is derived by computing the spatial mean of the RMSE values for each model simulation and for each variable (TAS, SLP, Z500). Then, the average of those values is computed to get the overall RMSE-based performance metric plotted in Fig. 12. This is performed separately for DJF and JJA.

CRCM and ECP2 show the lowest RMSE when driven by NCEP of any individual simulation configuration in both seasons. These are the only two RCMs that use spectral nudging in the domain interior, likely contributing to the superior performance when driven with reanalysis (Mearns et al. 2012). The NCEP-driven multi-model ensemble mean also has low RMSE, although higher than CRCM–NCEP, while the GCM-driven ensemble mean has lower RMSE than any individual GCM-driven run. The superior performance of the GCM-driven ensemble mean indicates that intra-ensemble bias is not systematic, as averaging all eleven LSMPs reduces error relative to any individual run. Consistent with expectations, RCMs driven by NCEP show lower RMSE than when the same RCM is driven by a GCM, in most cases. Three exceptions are for HRM3 in DJF and WRFG and RCM3 in JJA. In these cases, it is possible that error introduced by the driving GCM compensates for inherent biases in the RCM, resulting in an overall lower RMSE compared with the NCEP-driven hindcast.

Consistent with the findings in Sect. 5, CCSM boundary forcing tends to introduce larger error than other GCMs. For example, in DJF three of the four lowest ranked simulations are driven by CCSM. The apparent effect of CCSM is particularly notable for CRCM, which when driven by NCEP ranks the highest and when driven by CCSM ranks the fourth lowest. MM5I falls from 7th with NCEP to 18th with CCSM and WRFG from 10th to 19th (last). Similar results are found for JJA where CRCM falls from 1st place when forced by NCEP to 9th when forced by CCSM. WRFG–CCSM also ranks 17th and MM5I–CCSM 19th. Simulations produced using CGCM3 and GFDL as boundary forcing are generally ranked the highest both overall and compared with other boundary conditions for the same RCM. For DJF and JJA, GFDL and CGCM3 account for six out of the top 7 RCM–GCM. The two top performing

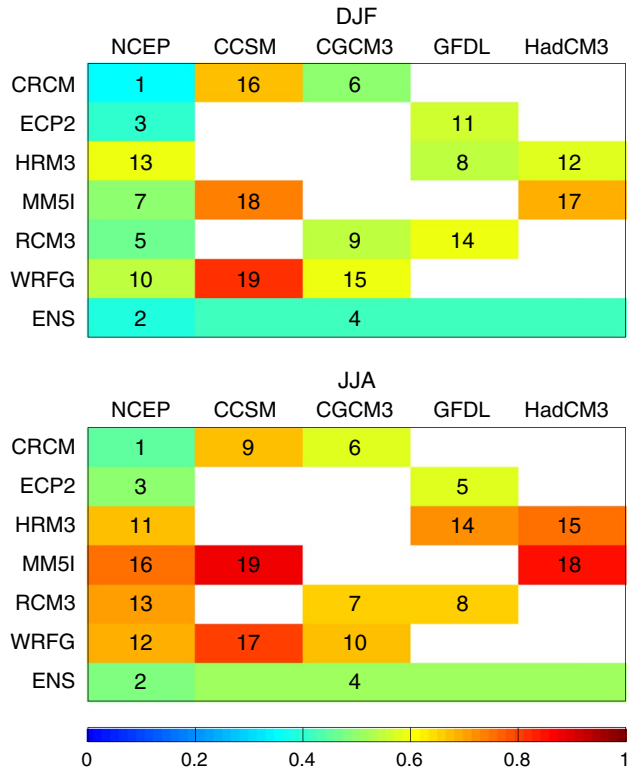

Fig. 12 Simulation rankings for (*top*) DJF and (*bottom*) JJA for all simulations with *each row* representing an RCM and *each column* the driving boundary forcing. *Colors* indicate total mean RMSE (described in Sect. 6.1) and the numbers indicate the simulation ranking, out of 19, relative to the other simulations based on the total RMSE. *White areas* are where no simulation is available with the indicated configuration

RCMs when forced by NCEP, CRCM and ECP2, show relatively small changes when forced by CGCM3 and GFDL respectively, suggesting these combinations are strong candidates for projecting future changes in temperature extremes in both seasons.

One caveat in using this overall RMSE score as a performance metric is that some RCM configurations may perform well in some locations while performing poorly in others. For example, ECP2–NCEP ranks third highest overall in both seasons but performs with relatively low skill for the California example (Sect. 4.2; Fig. 4a). Thus, while Fig. 12 distills evaluation of these models into a single generalized performance metric, if the goal is to project future changes in temperature extremes for a specific location of interest, it is important to consider the RCM–GCM configuration rankings at that location. It is also important to note that this generalized metric only reflects model performance for LSMPs associated with extreme temperature days, and does not provide a definitive ranking for overall simulation of the climate.

6.2 Conclusions and future direction

This study comprehensively evaluates the LSMPs associated with extreme winter and summer temperature days in the NARCCAP hindcast and historical GCM-driven experiments. Regional downscaling aims to improve over coarser resolution models in areas with complex topography or other features that necessitate finer resolution. In terms of extreme events, while the LSMPs themselves do not occur on such small scales, the temperature extremes may, and in areas of complex terrain or sharp climate zone gradients, starkly different LSMPs can be associated with extremes for neighboring locations. LSMPs can also be associated with processes, such as adiabatic warming due to downslope winds or convective precipitation that can influence temperature extremes at more local scales. Therefore, the information provided by physically interpreting and comparing model simulated LSMPs to those derived from observations, can both boost confidence in the ability of a model to simulate temperature extremes and identify areas where models may be challenged. Furthermore, analysis and physical interpretation of LSMPs helps to identify areas for which models may simulate realistic extremes but with incorrect underlying mechanisms. The ECP2–NCEP and HRM3–NCEP hindcasts exemplify this for the California example presented in Fig. 4. Here, evaluating model performance based solely on reproduction of temperature would yield misleading results.

Overall, LSMPs associated with temperature extremes are simulated with highest fidelity away from complex topography, where synoptic-scale dynamics are highly influential on local temperature, and during winter as compared with summer. In addition to LSMPs, summer extremes may be associated with other processes such as smaller scale and weaker circulation and land–atmosphere coupling related to soil moisture, resulting in lower model skill scores compared with winter. Focusing on four individual cases, LSMPs were represented well for Chicago winter cold days, well for northern California winter warm days with some exceptions, and with less skill for Houston summer cold days and Ohio Valley summer hot days. In many cases examined, especially in DJF, areas of elevated error for TAS are coincident with areas of elevated error for SLP and Z500, indicative of the dynamical interplay among all three variables. This underscores the importance of evaluating LSMPs in relation to temperature extremes. The ability of the models to reproduce realistic temperature distributions for the four individual cases in Sect. 4 appears to have some relation to the simulation of LSMPs.

Based on the results of this analysis, CGCM3 and GFDL appear to be best suited in terms of boundary forcing GCMs for future projections of temperature extremes. When driven by reanalysis, CRCM and ECP2 demonstrate superior skill at reproducing LSMPs, suggesting that these RCMs may be the best suited for making future projections of temperature extremes, especially when used in conjunction with CGCM3 and GFDL boundary forcing. CRCM and ECP2 are also the only two RCMs to use interior nudging, likely contributing to the superior performance when driven by reanalysis. CCSM consistently proves to be an inferior GCM for boundary forcing, with CCSM-forced simulations showing substantially higher error than the NCEP-driven hindcasts using the same RCM.

Future efforts should focus on process-based evaluation of the NARCCAP suite, as in Bukovsky et al. (2013) for the North American Monsoon, focusing on cases identified as having low model fidelity. Such improved understanding of model error would provide a baseline for evaluating the efficacy of dynamical downscaling at higher resolutions versus other modeling efforts such as performing statistical downscaling, producing high resolution global simulations, or improving model physics. For example, the California case in Sect. 4 suggests that two of the RCMs analyzed may simulate local temperature extremes of proper amplitude but in relationship to physical processes inconsistent with observations. Here, high resolution is expected to be crucial given the influence of topographical features. Investigation of the potential causes of errors in key processes may also improve understanding of why some RCMs show better skill at reproducing the LSMPs when forced by GCMs compared with NCEP.

The Houston case is also an illustrative example of a strong candidate for further process-oriented evaluation. Both warm and cold extremes show substantial differences from reference (Fig. 7c). For warm extremes, the effects of land–atmosphere coupling through soil moisture feedback or improper representation of the sea breeze front and convective precipitation could contribute to the generally large model error. Furthermore, while high resolution is necessary to capture many of these processes, this may be a situation for which improved model physics would stimulate the greatest improvement in model fidelity.

Results for the heat wave case in the Ohio Valley show that TAS and Z500 patterns are reasonably simulated, while SLP patterns have large errors. This is likely indicative of the relative importance of a Z500 ridge for the occurrence of extreme warm temperatures compared with near-surface circulation features. However, further analysis of the surface energy budget as it relates to anomalous soil moisture and synoptic-scale subsidence under the Z500 ridge are reasonable targets for future process-based evaluation for this case.

The results of this work also identify places where future efforts would not be productively spent, as exemplified by Chicago in the winter. It is somewhat expected, but nonetheless encouraging, that the NARCCAP RCMs

realistically simulate the LSMPs associated with winter temperature extremes here. This indicates that the key processes are well simulated in this region and further increases in resolution are unlikely to foster major improvements in model fidelity.

Beyond these four examples, the domain-wide focus of this evaluation framework allows for systematic identification of all regions where further analysis and simulation improvement efforts may be productively focused and where strong model performance lends confidence to future climate simulations. Additionally, the methodology employed here could extend to other extreme phenomena that are associated with characteristic LSMPS, such as precipitation extremes.

Ultimately it is the future aim for this work to develop a framework for generalized and systematic evaluation of the ability of RCMs to simulate temperature extremes based on several diagnostics and metrics. Combined with the evaluation of temperature distributions in Loikith et al. (2015) and planned future process-based evaluation, we aim to develop a suite of generalized performance metrics, similar to that presented in Fig. 12, that can be used to rank RCM–GCM configurations. While this work has focused on NARCCAP, such metrics can be applied to any suite of simulations, providing readily interpretable information based on statistics, dynamics, and processes that are all key to extreme temperature simulation.

Acknowledgments Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Part of this research was funded by NASA National Climate Assessment 11-NCA11-0028 (P.C.L., J.K., H.L., D.E.W) and NSF AGS-1102838 (J.D.N.). We also thank the NARCCAP team for production of the model simulations and archiving of the data. We thank Joyce Meyerson for her valuable contributions to figure production.

References

Anderson JL et al (2004) The new GFDL global atmosphere and land model AM2–LM2: evaluation with prescribed SST simulations. J Clim 17:4641–4673

Beniston M, Diaz HF (2004) The 2003 heat wave as an example of summers in a greenhouse climate? Observations and climate model simulations for Basel, Switzerland. Glob Planet Change 44:73–81

Berg A, Lintner BR, Findell KL, Malyshev S, Loikith PC, Gentine P (2014) Impact of soil moisture–atmosphere interactions on surface temperature distribution. J Clim 27:7976–7993

Bindoff NL et al (2013) Detection and attribution of climate change: from global to regional. In: Stocker TF, Qin D, Lattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge and New York

Bowden JH, Nolte CG, Otte TL (2012) Simulating the impact of the large-scale circulation on the 2-m temperature and precipitation climatology. Clim Dyn 40:1903–1920

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

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

Clark RT, Brown SJ (2013) Influences of circulation and climate change on European summer heat extremes. J Clim 26:9621–9632

Collins WD et al (2006) The community climate system model: CCSM3. J Clim 26:9621–9632

Coumou D, Robinson A (2013) Historic and future increase in the global land area affected by monthly heat extremes. Environ Res Lett. doi:10.1088/1748-9326/1088/1083/034018

Coumou D, Robinson A, Rahmstorf S (2013) Global increase in record-breaking monthly-mean temperatures. Clim Change 118:771–782

Di Luca A, de Elía R, Laprise R (2011) Potential for added value in precipitation simulated by high-resolution nested regional climate models and observations. Clim Dyn 38:1229–1247

Dole R et al (2011) Was there a basis for anticipating the 2010 Russian heat wave? Geophys Res Lett. doi:10.1029/2010gl046582

Donat MG et al (2013) Updated analysis of temperature and precipitation extreme indices since the beginning of the twentieth century: the HadEX2 dataset. J Geophys Res 118:1–16

Feser F, Rockel B, von Storch H, Winterfeldt J, Zahn M (2011) Regional climate models add value to global model data: a review and selected examples. Bull Am Meteorol Soc 92:1181–1192

Fields DT (2006) Tools for spatial data. National Center for Atmospheric Research, Boulder. http://www.cgd.ucar.edu/Software/Fields

Fischer EM, Seneviratne SI, Vidale PL, Lüthi D, Schär C (2007) Soil moisture–atmosphere interactions during the 2003 European summer heat wave. J Clim 20:5081–5099

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

Gershunov A, Barnett TP (1998) ENSO Influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: observations and model results. J Clim 11:1575–1586

Gordon C et al (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 G, Dudhia J, Stauffer DR (1993) A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398-STR

Griffiths ML, Bradley RS (2007) Variations of twentieth-century temperature and precipitation extreme indicators in the Northeast United States. J Clim 20:5401–5417

Hughes M, Hall A (2009) Local and synoptic mechanisms causing Southern California's Santa Ana winds. Clim Dyn 34:847–857

Jacob DJ, Winner DA (2009) Effect of climate change on air quality. Atmos Environ 43:51–63

Jones RG, Hassell DC, Hudson D, Wilson SS, Jenkins GJ, Mitchess JFB (2004) Workbook on generating high resolution climate change scenarios using PRECIS. Hadley Centre for Climate Prediction and Research

Juang H-MH, Hong S-Y, Kanamitsu M (1997) The NCEP regional spectral model: an update. Bull Am Meteorol Soc 78:2125–2143

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

Kenyon J, Hegerl GC (2008) Influence of modes of climate variability on global temperature extremes. J Clim 21:3872–3889

Kim J et al (2013) Evaluation of the surface climatology over the conterminous United States in the North American regional climate change assessment program hindcast experiment using a regional climate model evaluation system. J Clim 26:5698–5715

Lau N-C, Nath MJ (2012) A model study of heat waves over North America: meteorological aspects and projections for the twenty-first century. J Clim 25(14):4761–4784

Lau N-C, Nath MJ (2014) Model simulation and projection of European heat waves in present-day and future climates. J Clim 27(10):3713–3730

Lee DT, Schachter BJ (1980) Two algorithms for constructing a Delaunay triangulation. Int J Comput Inf Sci 9:219–242

Loikith PC, Broccoli AJ (2012) Characteristics of observed atmospheric circulation patterns associated with temperature extremes over North America. J Clim 25:7266–7281

Loikith PC, Broccoli AJ (2014) The influence of recurrent modes of climate variability on the occurrence of winter and summer extreme temperatures over North America. J Clim 27:1600–1618

Loikith PC, Broccoli AJ (2015) Comparison between observed and model simulated atmospheric circulation patterns associated with extreme temperature days over North America using CMIP5 historical simulations. J Clim 28:2063–2079

Loikith PC, Lintner BR, Kim J, Lee H, Neelin JD, Waliser DE (2013) Classifying reanalysis surface temperature probability density functions (PDFs) over North America with cluster analysis. Geophys Res Lett 40:3710–3714

Loikith PC et al (2015) Surface temperature probability distributions in the NARCCAP hindcase experiments: evaluation methodology, metrics and results. J Clim 28:978–997

Mearns LO, Gutowski W, Jones R, Leung R, McGinnis S, Nunes A, Qian Y (2009) A regional climate change assessment program for North America. EOS 36:311–312

Mearns LO et al (2012) The North American regional climate change assessment program: overview of phase I results. Bull Am Meteorol Soc 93:1337–1362

Meehl GA, Tebaldi C (2004) More intense, more frequent, and longer lasting heat waves in the 21st century. Science 305:994–997

Mesinger F et al (2006) North American regional reanalysis. Bull Am Meteorol Soc 87:343–360

Min S-K, Zhang X, Zwiers F, Shiogama H, Tung Y-S, Wehner M (2013) Multimodel detection and attribution of extreme temperature changes. J Clim 26:7430–7451

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

Morak S, Hegerl GC, Christidis N (2013) Detectable changes in the frequency of temperature extremes. J Clim 26:1561–1574

Paeth H, Mannig B (2012) On the added value of regional climate modeling in climate change assessment. Clim Dyn 41:1057–1066

Pal JS et al (2007) Regional climate modeling for the developing world: the ICTP RegCM3 and RegCNET. Bull Am Meteorol Soc 88:1395–1409

Peterson TC et al (2013) Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United States: state of knowledge. Bull Am Meteorol Soc 94:821–834

Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parameterizations in the Hadley Centre climate model: HadCM3. Clim Dyn 16:123–146

Rangwala I, Barsugli J, Cozzetto K, Neff J, Prairie J (2012) Mid-21st century projections in temperature extremes in the southern Colorado Rocky Mountains from regional climate models. Clim Dyn 39:1823–1840

Rienecker MM et al (2011) MERRA: NASA's modern-era retrospective analysis for research and applications. J Clim 24:3624–3648

Ruff TW, Neelin JD (2012) Long tails in regional surface temperature probability distributions with implications for extremes under global warming. Geophys Res Lett. doi:10.1029/2011gl050610

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

Seneviratne SI, et al (2012) Changes in climate extremes and their impacts on the natural physical environment. In: Field CB, Barros V, Stocker TF, Qin D, Dokken DJ, Ebi KL, Mastrandrea MD, Mach KJ, Plattner G-K, Allen SK, Tignor M, Midgley PM (eds) Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of working Groups I and II of the intergovernmental panel on climate change (IPCC). Cambridge University Press, Cambrige and New York, pp 109–230

Shamarock 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

Sillmann J, Kharin VV, Zhang X, Zwiers FW, Bronaugh D (2013) Climate extremes indices in the CMIP5 multimodel ensemble: part 1. Model evaluation in the present climate. J Geophys Res 118:1716–1733

Stefanon M, D'Andrea F, Drobinski P (2012) Heatwave classification over Europe and the Mediterranean region. Environ Res Lett. doi:10.1088/1748-9326/1087/1081/014023

Vautard R et al (2013) The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. Clim Dyn 41:2555–2575

Wang J, Kotamarthi VR (2013) Assessment of dynamical downscaling in near-surface fields with different spectral nudging approaches using the nested regional climate model (NRCM). J Appl Meteorol Climatol 52:1576–1591

Wang A, Zeng X (2013) Development of global hourly 0.5° land surface air temperature datasets. J Clim 26:7676–7691

Wang A, Zeng X (2014) Global hourly 0.5-degree land surface air temperature datasets. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder. doi:10.5065/D6PR7SZF. Accessed 24 June 2014

Westby RM, Lee Y-Y, Black RX (2013) Anomalous temperature regimes during the cool season: long-term trends, low-frequency mode modulation, and representation in CMIP5 simulations. J Clim 26:9061–9076

Wettstein JJ, Mearns LO (2002) The influence of the North Atlantic–Arctic oscillation on mean, variance, and extremes of temperature in the Northeastern United States and Canada. J Clim 15:3586–3600

Youn D, Choi W, Lee H, Wuebbles DJ (2006) Interhemispheric differences in changes of long-lived tracers in the middle stratosphere over the last decade. Geophys Res Lett. doi:10.1029/2005gl024274

Zwiers FW, Zhang X, Feng Y (2011) Anthropogenic influence on long return period daily temperature extremes at regional scales. J Clim 24:881–892