Future Projections of the Large Scale Meteorology Associated with California Heat Waves in CMIP5 Models

Erool Palipane[1] and Richard Grotjahn[1*]
[1]*Department of Land, Air and Water Resources, University of California, Davis, CA, 95616, USA*

January 2018 (Revised: May 2018; Re-revised July 2018)
Submission to Journal of Geophysical Research: Atmospheres

* *Corresponding author*: Richard Grotjahn (grotjahn@ucdavis.edu)

Corresponding author address:
Atmospheric Science Program, One Shields Ave., Dept. of L.A.W.R. University of California Davis, Davis CA USA 95616

Key Points (140 characters)

- The California Central Valley heat waves synoptic pattern does not change in frequency or intensity from 1961-2000 to 2061-2100 in models.
- Heat waves are much more frequent and predominantly of one type when using historical thresholds due to the change in the climate 'mean'.
- A multi-model average has 4x as many heat waves, lasting 2x as long, with 1.5x the 20-year return value relative to historical values.

Abstract

Previous work showed that climate models capture historical large-scale meteorological patterns (LSMPs) associated with California Central Valley (CCV) heat waves including both ways these heat waves form. This work examines what models predict under the RCP4.5 and RCP8.5 scenarios. Model performance varies, so a multi-model average weights each model based on its historical performance in four parameters. An LSMP index (LSMPi) defined using upper atmosphere variables captures dates of past extreme surface temperature maxima. LSMPi correlates well with all values of CCV-average surface maximum temperature. LSMPi distributions in future simulations shift ~0.6 standard deviations higher between 1961-2000 and 2061-2100 for RCP 8.5 data. Based on the *historical* climatology, future scenarios show a large increase in the frequency and duration of heat waves in every model. Four times as many heat waves occur and their median duration doubles, using historical thresholds. Of the two ways heat waves form, Type 1 has similar frequency in the future. But, Type 2 becomes much more common because Type 2 has a preexisting hot anomaly in Southwestern Canada, much like the historical to future climatological change in that region (a "global warming" signal). The 20-year return value anomaly increases by 30-40%. The average of the 50 hottest temperatures increases 3.5-6K depending on the scenario. When extreme values are defined using the *future* climatology, the models and their average have no consistent increase or decrease of distribution properties such as: shape, scale, and return values of the extremes compared to historical values.

Plain Language Summary (200 words limit)

The hottest heat waves in California occur during specific weather patterns. Computer models that simulate global climate can include such patterns because the patterns have large horizontal size. We average results from 13 climate models. More weight is given those models that are better than other models at generating these weather patterns and other properties of California heat waves during 1961-2000. . A heat wave is when the daily maximum temperature at each Central Valley location is among the warmest 5% of June-September days during the years 1971-2010. We examine two scenarios for the future, one with continued increase in 'greenhouse' gases and another that has less increase. Our averages find several things for 2061-2100. In the future, depending on which scenario, daily maximum temperatures will exceed those highest 5% values six to eight times as often as now, about one fifth to one quarter of summer. The hottest days will be 3.5-6 degrees Celsius (about 7-12 degrees Fahrenheit) hotter than the hotter days in recent memory. Those heat wave weather patterns are neither stronger nor occur more often in the future, instead many more heat waves occur in future simulations from a general warming of western North America.

Index Terms (five or less)

3337 Global climate models (1626, 4928) - Primary

3305 Climate change and variability

4313 Extreme events (1817, 3235)

3364 Synoptic-scale meteorology

0429 Climate Dynamics.

Keywords (six or less): future heat waves, California heat waves, large-scale meteorological patterns simulation, future summer variability, climate model simulations of heat waves

1 Introduction

Extreme heat (>40C) occurs during June through September over the California Central Valley (CCV). Our prior work discusses how large-scale meteorological patterns (LSMPs) are associated with CCV heat waves (Grotjahn & Faure, 2008; Grotjahn, 2011, 2013). Lee and Grotjahn (2016; hereafter LG2016) show that CCV heat waves (HWs) can form by two ways discussed below. Grotjahn and Lee (2016; hereafter GL2016) show climate models' varying ability to create simulated heat waves in historical conditions. This paper extends GL2016 to future climate scenarios by applying the LSMP context to identify and understand possible future changes in CCV extreme heat events during summer. Our specific questions include: will events occur more often than in the past? Will events become more severe? Will events last longer? Will changes from historical to future climate be due mainly to a shift in the climatological conditions (a 'global warming signal') or in the LSMP properties?

Coumou et al. (2013), Perkins et al. (2012), Russo et al. (2014) and others have discussed the effects global warming may have on local heat event characteristics (frequency, duration, intensity) in the mid-latitudes. Other studies explore possible physical mechanisms behind the changes in these heat wave events. These mechanisms include: changes in sub-seasonal atmospheric variability (Teng et al., 2013), variations in the quasi-stationary waves (Screen & Simmonds, 2014; Petoukhov et al., 2013) and weakening of the boreal storm tracks in the summer (Lehmann et al., 2014).

The synoptic situation during CCV heat waves includes these factors. Subsidence causes warming of the air from adiabatic compression, clear skies support radiant heating, and warm advection all heat the CCV air (Grotjahn 2011; Horton et al. 2016). Also, offshore winds (Grotjahn, 2011; Lau & Nath 2012; Grotjahn et al. 2016) block a cooling sea breeze. To have offshore winds, Grotjahn (2011) finds the largest warm temperature anomaly in the lower troposphere lies just offshore, helping to set up the low level pressure gradient force that opposes a cooling sea breeze. Grotjahn (2011) also notes that the subsidence inversion becomes shallower and more rapidly warmed by solar radiation.

Other studies of California HWs include historical (Gershunov et al., 2009; Clemesha et al., 2017), future (Gershunov and Guirguis, 2012; Pierce et al., 2013), and both (Mastrandera et al., 2011) time periods.

Large scale features associated with California HWs (i.e. LSMPs) are resolved by climate models and provide a context to examine different models' predictions of future CCV heat waves. Grotjahn (2011) developed a LSMP index based on upper air data that correlates highly with surface maximum temperatures over the CCV and captures most extreme events; it works well in part because soil moisture over the region is similar each year due to drought every summer and copious irrigation. Grotjahn (2013) compares the CCSM4 simulated LSMPs and other properties of heat waves to corresponding reanalysis data. LG2016 and GL2016 use a cluster analysis to sort CCV heat waves into two types based on LSMPs leading up to heat events. One cluster ("Type 1") has

cold anomalies prevailing over the NW US and western Canada several days before CCV heat event onset and the CCV heat wave develops quickly in the day before onset. The other cluster ("Type 2") has a preexisting hot anomaly over SW Canada for several days prior to CCV heat onset, then a southwestward extension of the hot anomaly initiates the CCV heat wave. LG2016 find roughly equal numbers of Type 1 and Type 2 events plus a smaller number of events that mix both types.

This work builds on our previous work to answer those questions above. We consider 13CMIP5 models' simulations of RCP4.5 and RCP8.5 scenarios. We develop a simple multi-model average based on each model's historical performance. We improve the diagnostics of the LSMPs from our prior work and apply those diagnostics to estimate how the extremes may change in the future, including the two cluster types.

The next section describes the data and methods developed to understand the future changes in CCV heat waves. The third section describes results and section four has conclusions.

2 Data and Methodology

2.1 Data Used for the Analysis

NOAA NCDC daily surface maximum temperatures at 15 stations in the CCV are used to identify historical extreme HWs. (Figure 1 in LG2016, also reproduced in the Supporting Materials, shows the station locations.)

The NCEP-NCAR reanalysis (Kalnay et al., 1996, hereafter NNRA1) data are used for: the formulation of the LSMPs via composites, the development of an improved LSMP index, distinguishing the two cluster types of HWs, and verification and comparison with corresponding quantities in the model data. The NNRA1 data are from 1971-2010. ERA-interim (Dee et al., 2011) data (1979-2010) are used to cross-check the NNRA1 results; as found by LG2016, the results during the overlapping period are essentially the same. Hence the NNRA1 reanalysis is used because its longer record includes more extreme heat wave events. As in Grotjahn (2011), the event onset for upper air data is always at 12 GMT.

CMIP5 model data from historical, RCP4.5, and RCP8.5 simulations are studied. Model historical data are from simulated years 1961-2000; the RCP simulations are for 2061-2100. Climate model simulations are not weather forecasts, so a specific date has only accidental similarity between models and the reanalysis. The model and reanalysis periods are offset to take advantage of better upper air observations at later times while the historical simulations end before 2010. Some models have parallel simulations (ensemble runs) with the sub-daily upper air data we need archived; as available, data from ensemble runs are included. Table S1 and the GL2016 supplementary

information give a description of the model data used and how many grid points for a specific model are considered to lie within the CCV.

The zonal wind anomaly (Ua), the meridional wind anomaly (Va) and the temperature anomaly (Ta) were examined at: 250hPa, 500hPa and 850hPa at every 6hr snapshot time (i.e. at 0, 6, 12, 18GMT). The anomalies are with respect to corresponding long term daily mean (LTDM) values. The LTDM values for each variable are found by the methodology described in LG2016. To summarize: corresponding days of the year are averaged to create an initial LTDM at each location; since the initial LTDM has sizeable day to day variation on a 40 year average, the initial LTDM is Fourier transformed and only the first five harmonics used to reconstruct the smoother, final LTDM. Daily anomalies at each location are constructed by subtracting the corresponding daily final LTDM values.

LTDM values for surface maximum temperatures are found for each station. Those LTDM values are subtracted from the measured value to find temperature anomalies. The daily anomalies are normalized by the long-term mean seasonal average standard deviation at each location. The resulting normalized anomalies make values at different stations inter-comparable. These normalized anomalies are labeled 'Tnamax' here. The daily average of the 15 Tnamax values over the CCV is labeled '**avTnamax'.** Our prior work uses this methodology.

2.2 Definitions of a Heat Event

The methodology to identify CCV heat waves in the CMIP5 models is similar to that used by GL2016. Daily maximum surface temperatures at the CCV grid points are examined for each model. Analogous to station data, Tnamax values are calculated for each model grid point in the CCV. At least half of the CCV grid points (stations) must reach or exceed the 95^{th} percentile threshold for at least three consecutive days for the date to qualify as a HW for that model (observations).. The normalizations used to calculate Tnamax make CCV grid points intercomparable over the summer.

Future scenarios use data from 2061-2100. HWs in RCP4.5 and RCP8.5 data are defined two different ways. One way, labelled Fh, uses the same threshold values of surface avTnamax as calculated in historical simulations to define a future heat wave. The second way, labelled Ff, uses the 95th percentile of avTnamax values during the future time period of each respective RCP scenario.

The following label conventions designate how the threshold is defined when choosing candidate heat waves. There are five combinations of time period (F or H), threshold (f or h), and RCP (4.5 or 8.5).CMIP5_Hh: HWs from CMIP5 historical runs using thresholds based on the model's historical data - HWs from CMIP5 future runs using thresholds based on the model's historical data are RCP4.5_Fh and RCP8.5_Fh for the RCP 4.5 and 8.5 scenarios respectively. HWs from CMIP5 future runs using thresholds based on the model's future data are RCP4.5_Ff and RCP8.5_Ff for the

RCP 4.5 and 8.5 scenarios respectively. Comparing 'Ff' to 'Hh' results emphasizes changes in LSMP properties. The 'Fh' to 'Hh' comparison includes both the general trend as well as LSMP changes, so how properties differ between 'Fh' and 'Ff' emphasizes the general trend.

2.3 Clustering Methodology

We use separate calculations to determine which cluster an event belongs to and to assess the strength of each cluster type present in every event.

Previous work, reported in LG2016, showed two groupings of the air parcel trajectories that arrive near the northwest California coast. That location is the center of the hot anomaly at 850 hPa that is fundamental to CCV HWs. GL2016 choose the hottest 28 heat wave events (from 1977-2010) to form the composite cluster patterns. They show that ERA-interim (from 1979) and NNRA1 cluster patterns are essentially the same. The criteria find 32 HWs (from 1971-2010); these form the composite cluster patterns.

To assign each event to a cluster type in model data, projection coefficients are calculated in a 'domain' well above the Earth's surface where large and consistent differences occur between the cluster composites in NNRA1 data. After some testing, the domain bounded by135-120W and 40-55N was chosen to determine to which cluster an event belongs. (Figures S1-S4 in the Supporting Materials show the regions used and how they have opposite anomaly patterns, analogous to GL2016.) In this domain target values at -2 days lag (i.e. before onset) are used of: temperature anomaly at 850hPa (Ta850) and 500hPa (Ta500) plus zonal wind anomaly at 500hPa (Ua500).

Projection coefficients ($P_{k,n}$) are calculated for the domain and variables stated above.

$$P_{k,n} = \frac{\sum_i \sum_j (q_{i,j,n} . Q_{k,i,j})}{\sum_i \sum_j (Q_{k,i,j})^2}$$

Here k indicates cluster type 1 or 2; n indicates a date (i.e. during an event) and i,j is a grid point in the longitude, latitude domain. The summations are over all grid points in the domain. q is the variable during an individual event while Q is the corresponding variable in the cluster mean field calculated from the NNRA1 data. There are three combinations of variable, level and time before onset, hence three projection coefficients for each event and cluster type. The three projections are averaged to obtain an average projection for each event and cluster type. Each pair of average projections for each event is shown later on scatter plots. The larger, positive projection determines the cluster type for most events. However, if the average projection onto one cluster differs by less than 0.3 from the average projection on the other cluster or if both projection coefficients are negative, the event is labelled a 'mixed' type. The method discussed thus far is used only to determine the cluster *type* of an event.

2.4 Updating a Large Scale Meteorological Pattern Index (LSMPi)

The strength of each event is measured by the LSMPi value. Grotjahn (2011, 2013, 2016) introduced a "circulation index" (Ci) that measures how similar a pattern on a given day is to the heat wave composite pattern in corresponding variables. The Ci in Grotjahn (2011) uses the temperature anomaly at 850 hPa (Ta850) and meridional wind anomaly at 700 hPa (Va700) values averaged over the event onset dates (labeled 'target composites'). Corresponding daily fields are projected (un-normalized and separately) onto the target composites of Ta850 and Va700 in regions that are highly consistent between ensemble members. The Ci was an optimal weighted combination of these two projections each day. 'Extreme' dates were the hottest 1% of the Tnamax values during the entire data record. The levels and variables were chosen to match the daily climate model data available to the author at that time. Later work, such as GL2016, used different levels, variables, and regions for the projections and also use more stations in the CCV surface maximum temperature average; again, the choices were dictated by available data and optimized matching.

This study improves upon this Ci definition. To distinguish this new index from the earlier one, it is labeled the LSMP index, or LSMPi. The following approaches are used:
(i). Use only data on the heat wave onset date
(ii). Focus on regions with high consistency (measured by the 'sign count[1]'; see Grotjahn, 2011)
(iii). Focus on rectangular regions with anomaly extrema (relative maxima and minima) that are also common to both cluster types
(iv). Test spatially-varying weighting proportional to the sign count.

The LSMPi variables and the regions used are these: Ta850 in region 128-119W, 29-46N; meridional wind component anomaly, Va500 in region 142-132W, 37-51N; zonal wind component anomaly, Ua500 in region 128-111W, 28-37N.

The LSMPi is a simple projection of three daily observed fields onto the corresponding target composite fields over the indicated regions. The match between LSMPi and avTnamax values on dates of extreme avTnamax was improved by including weights in the projection calculation, where the weights, $w_{i,j}$ are proportional to the sign count at each location. Thus, grid points in the region where the anomaly signs are more consistent between past events are given more weight. Grid points with smaller sign count have less weight in the projection calculation. The following equation calculates the LSMP index for 850hPa temperature.

$$I_{w,n}(Ta850) = \frac{\sum_i \sum_j w_{i,j}\, \overline{T}_{\mathrm{a}}(\mathrm{i,j})\, \mathrm{T_a(i,j,n)}}{\sum_i \sum_j w_{i,j}\, \overline{T}_{\mathrm{a}}(\mathrm{i,j})^2}$$

Here: $I_{w,n}(T850)$ is a weighted, normalized projection for a specific day n based on the temperature anomalies at 850hPa level; i and j are the longitude and latitude pointers respectively. The summations are over the ranges of i and j for the specified region over which the projection is made. $T_a(i, j, n)$ is the anomaly value of the temperature for that specific day n and grid point (i, j).

[1] The sign count is calculated for each variable at each grid point. At HW onset, if the anomaly value is positive at a grid point the sign value is +1 and if negative the sign value is -1. The sign values at onset of all the HWs are added together at the grid point, then divided by the number of HWs. The sign count thus lies between -1 and +1.

$\bar{T}_a(i, j)$ is the corresponding target composite at that particular grid point calculated from the onset dates of the 32 events. The weight $w_{i,j}$ is the same as the sign count at that grid point calculated from the 32 onset events. Analogous indices using each velocity component were also calculated from projections over their respective regions defined above.

The weights were adjusted to optimize the LSMPi match for extreme avTnamax values. The circulation index is defined as $LSMPi = w1*I_{w,n}(Ta850)+w2*I_{w,n}(Va500)+w3*I_{w,n}(Ua500)$. Here the w1, w2, and w3 weights are constrained such that w1+w2+w3=1. To optimize the weighting, the root mean square difference between avTnamax and LSMPi for each combination of w1, w2, and w3 was calculated. All possible combinations (in 0.01 increments) were tested. An optimal combination (w1=0.68, w2=0.02, w3=0.30) minimized the root mean square difference between the LSMPi value and the avTnamax value over the summers.

These LSMPi values are compared against avTnamax values using scatter plots (shown later). In addition, the distribution of LSMPi values for all days are binned then fit with a curve using the Johnson system (Johnson, 1949) for all days in every group of 40 summers. Estimation of the Johnson parameters is done from quantiles. The procedure of Wheeler (1980) is used. From these fitted curves, we show how the distributions of LSMPi values change between the Hh, Ff, and Fh cases.

2.5 Determining Extreme Event Skill

This work focuses on extreme events. Hence, some metrics from matching event avTnamax with LSMPi extreme values are calculated: The avTnamax that corresponds to the 95th percentile is called Ts95. Then a cubic polynomial regression line fits dates when the CCV stations mean (avTnamax) is ≥ Ts95. That regression line defines the LSMPi-Ts95 value corresponding to Ts95. LSMPi-Ts95 varies for different combinations of Ta850, Va500, and Ua500.

Some standard metrics are based on these contingency table quantities:

N_all: number of points where either LSMPi ≥ LSMPi-Ts95 *or* the avTnamax is ≥ Ts95.

N_s: Number of points where LSMPi ≥ LSMPi-Ts95 *and* avTnamax is ≥ Ts95 (these are successful matches).

N_u: Number of points where avTnamax is ≥ Ts95 *and* LSMPi < LSMPi-Ts95 (LSMPi is unsuccessful because an event is occurring by this measure but the LSMPi value is below the threshold to signal an event).

N_o: Number of points where LSMPi ≥ LSMPi-Ts95 and avTnamax is < Ts95 (LSMPi is unsuccessful because it exceeds the threshold to signal an event but the avTnamax values are not high enough to indicate an event).

The contingency table provides FAR (false alarm ratio, =N_o/(N_o+N_s) and POD (probability of detection, =N_s/(N_u+N_s). These indices assess skill detecting rare events (Stephenson et al. 2008, Marzban 1998). FAR is the number of 'false alarms' divided by the sum of the 'hits' plus false alarms. A 'hit' is when the avTnamax value is above the 95% threshold and the LSMPi is

above the regression curve value for that avTnamax threshold, i.e. both quantities indicate a heat wave. A 'false alarm' is where the LSMPi value is above its threshold but the avTnamax is not. The POD is the number of hits divided by the sum of hits plus misses. A miss is where the avTnamax is above its threshold but the LSMPi is not. A better match between LSMPi and avTnamax is when FAR is smaller and POD is larger. FAR and POD both range from 0 to 1.

2.6 Determining Multi-model mean Weights

The models are not equally adept at capturing the number and intensity of heat wave events in the historical period (e.g. GL2016). So, a multi-model mean should not weight each model simulation equally. If differences in model skill are large, unequal weighting improves the projection (Weigel et al. 2010). Various methods were tested to devise an objective weight for each model's contribution to the weighted model-mean. The Kolmogorov-Smirnov test (on cumulative distribution functions found by the Johnson method) proved unsatisfactory; the test rated some models worse than others even when those others matched NNRA1 properties better. Several measures of error in Wehner (2013) were tested (with the weight proportional to the inverse of the error) but the weights were similarly unsatisfactory. Since the multi-model average is used to estimate some basic properties of extreme events, such as their intensity, frequency, and distribution of high values, then metrics of those properties are used. How well one model creates some historical HW property versus another model depends on the property. So, the weights are based on comparisons between model and reanalysis for multiple historical properties. The reanalysis and historical multi-model means compare well in tables and figures shown later. It is assumed that the relative model skill in simulating historical HWs is our best guess of future relative skill. The weighting scheme selected uses four, squared, inverse, normalized, model-relative, differences. The difference in variable 'v' for model 'm', $d_{v,m}$, is the model value minus NNRA1 value divided by the NNRA1 value of the variable. The inverse of $d_{v,m}$ is used but normalized by the sum of the inverse $d_{v,m}$ values from all models, hence a model's weight depends upon that model's performance relative to corresponding values of other models. The inverse is defined as $b_{v,m} = (1/d_{v,m}) / \left\{ \sum \left(1/d_{v,l} \right) \right\}$ where the summation is over all the models 'l', including model 'm'.

The four variables for each model m are: 1) LSMPi mean divided by its standard deviation; 2) the number of days with LSMPi >1 divided by the total number of days; 3) the value of the shape parameter from a generalized Pareto (GP) fit; 4) the value of the scale parameter from the GP fit. These variables are from the 40 year historical period and the weights are assumed to hold for all future periods. The $b_{v,m}$ values for each of the variables are combined to get a root mean squared total, S_m as: $S_m = \sqrt{b_{1,m}^2 + b_{2,m}^2 + b_{3,m}^2 + b_{4,m}^2}$. The final model weight W_m is defined relative to other models by dividing by the sum of the corresponding 'S' from every model 'l': $W_m = S_m / \sum S_l$. Therefore, all the W_m values sum to one.

2.7 LSMP pattern metrics

Four metrics are calculated to assess how similar each model's LSMP is to the corresponding reanalysis LSMP. The LSMP is the ensemble mean of an anomaly field at the onset of all HWs in the reanalysis and model 40-year historical periods. Bias ($B_{v,m}$) and percent error ($PE_{v,m}$) when variable 'v' is Ta850 and model 'm' are:

$$B_{Ta850,m} = \frac{\sum_{i}^{N}\sum_{j}^{M}\left|w_{i,j}\,C_j\right|\left(MT_{i,j,m} - RT_{i,j}\right)}{\sum_{i}^{N}\sum_{j}^{M}\left|w_{i,j}\,C_j\right|} \quad , \quad PE_{Ta850,m} = 100.\,\frac{\sum_{i}^{N}\sum_{j}^{M}\left|w_{i,j}\,C_j\right|\,\left|MT_{i,j,m} - RT_{i,j}\right|}{\sum_{i}^{N}\sum_{j}^{M}\left|w_{i,j}\,C_j\,RT_{i,j}\right|}$$

where $1 \le i \le N$ is the range in longitude, $1 \le j \le M$ is the range in latitude for the domains defined in §2.4, $C_j = \cos(\varphi_j)$ where φ_j is the latitude of each grid point, $W_{i,j}$ equals the sign count for the reanalysis ensemble, $RT_{i,j}$ is the value of the *reanalysis* ensemble mean at the point i,j, and $MT_{i,j,m}$ is the value of the *model 'm'* ensemble mean at the point i,j (an average of however many events model 'm' created). The units of $B_{Ta850,m}$ are K. These quantities are used to assess the hot anomaly centered quite close to the area of interest.

Two measures of the larger structure of the LSMP are the pattern correlation ($Cor_{v,m}$) and reanalysis projection ($Prj_{v,m}$). These quantities are defined for Ta850as:

$$Cor_{T850,m} = \frac{\sum_{i=istrt}^{iend}\sum_{j=jstrt}^{jend}\left\{\left(MT_{i,j,m} - \overline{MT}_{i,j,m}\right)\left(RT_{i,j} - \overline{RT}_{i,j}\right)\right\}}{\left\{\sum_{i=istrt}^{iend}\sum_{j=jstrt}^{jend}\left(\left(MT_{i,j,m} - \overline{MT}_{i,j,m}\right)^2\left(RT_{i,j} - \overline{RT}_{i,j}\right)^2\right)\right\}^{\frac{1}{2}}} \;,\; \mathrm{Pr}_{T850,m} = \frac{\sum_{i=istrt}^{iend}\sum_{j=jstrt}^{jend}\left\{MT_{i,j,m}\,RT_{i,j}\right\}}{\sum_{i=istrt}^{iend}\sum_{j=jstrt}^{jend}\left\{RT_{i,j}\right\}^2}$$

The overbar indicates the average value for all the points in the domain. The domain used for these variables is much broader and captures more of the LSMP. For Ta850, the domain encompasses the large hot anomaly (centered off the northern California coast) and the cold anomalies flanking it to the west and east. Since the domain includes hot and cold anomalies, the overbar terms tend to be small.

3 Results

3.1 Model Representation of the Primary LSMP

The LSMP contributeing most strongly to the LSMPi is in the temperature anomaly at 850 hPa. Accordingly, how well models capture this pattern at heat wave onset is a primary indicator of how

well models simulate California heat waves (Grotjahn, 2011, 2013). Table 1 lists bias, percent error, pattern correlation, and pattern projection of each model's ensemble mean relative to the ensemble mean of the reanalysis as described in section 2.7.

Bias and percent error over the small region (128W-119W by 29N-46N) are designed to capture the stronger and more consistent (as measured by sign count, Grotjahn, 2011) hot anomaly. As discussed in Grotjahn (2011) this anomaly sets up pressure and wind fields to oppose penetration inland of a cooling sea breeze. Many models have a negative bias meaning their temperature anomaly is not hot enough, though the bcc models have a positive bias. The percent error varies from ~10% to ~40%. Higher resolution does not guarantee lower bias or percent error.

The pattern correlation and projections extend over a large region (175W-95W by 20N-60N) that captures the stronger pattern of cold-hot-cold anomalies that extends from near the date line to the middle of North America. The pattern correlations range from 0.93 to 0.72 with 8 models having $Cor_{Ta850,m} \geq 0.9$. Hence, the models are capturing not just the hot anomaly but the cold anomalies upstream and downstream. (Figure S12 in the Supporting Materials shows this LSMP for several models.) While the correlation describes the pattern, the projection includes the magnitude of the anomaly in the model. The projections have a broader range than the correlations. Most models have projection less than one, consistent with their cold bias. Models with larger negative (cold) biases have projections notably less than their correlations. Models with positive (warm) biases have projections that exceed their correlations. Higher resolution only partly yields better pattern match. For example, the bcc models have quite different resolutions; both models have positive bias; the higher resolution model has larger pattern correlation, but the bias pushes the lower resolution model to a better projection.

3.2 Past and Future Event Number and Duration

This section discusses how the climate model HWs change between historical and future climate simulations. For this analysis, Hh (1961-200) data are compared with (2061-2100) Fh and Ff data.

The HW definition has a minimum duration of three extremely hot days. Figure 1 is a histogram of consecutive days above the threshold (specified in §2.2). Longer durations are less common than shorter durations above the threshold. For Hh panels, almost all the higher resolution models (Figure 1a) do a reasonable job simulating the distribution found in reanalysis data. Most of the coarser resolution models tend to overestimate the duration of events (Figure 1b).

Figure 1 shows that heat wave durations increase in the future simulations when using each model's historical threshold (Fh cases), and more so for RCP8.5 data. For example, in the HADGEM2-CC model RCP8.5_Fh scenario, heat wave events that last 5 days are more common than heat wave events lasting 3 or 4 days and there are three times as many events as in the model's Hh data. Inmcm4 and NorESM1-M have large numbers of events in Fh cases, but these models also have a many more events in their Hh data than occur in the reanalysis. Other models have between three

and four times as many extreme heat wave events in Fh versus Hh data. Table 2 lists the total number of events for RCP8.5 by each model as well as the weighted model mean. In the CCSM4 model, the RCP8.5_Fh events are, on average, 2.6 days longer than for Hh simulations while the RCP4.5_Fh events are 1.3 days longer; these averages are over 6 ensemble runs in each case. For the bcc-csm1-1-m model, the increase of average duration is 0.5 days in the RCP4.5_Fh case, but 2.5 days in the RCP8.5_Fh case. HadGEM2-CC has a larger change in average duration: 2.3 days for RCP4.5 and 5.9 days for RCP8.5. A few models (notably the MIROC models) show much longer increases in average event duration. The multi-model mean is 1.4 (3.7) days longer for RCP4.5_Fh (RCP8.5_Fh).

Comparing Hh and Ff cases, finds generally little change in the average duration or general shape of the histograms, especially for models having more than one ensemble member (CCSM4 has 6 members; HadGEM2-CC has 3). Hence, the frequency and duration of the weather patterns, i.e. the LSMPs producing the HWs are likely little-changed from their historical values. This point is developed further below.

The longest events generally last between 7-10 days in the higher resolution models in Hh simulations. For all models the longest event becomes longer in each future simulation, typically doubling (or more) in length for RCP4.5_Fh cases and tripling (or more) for RCP8.5_Fh cases. The longest duration increases from 8 days in Hh to 28 days in RCP8.5_Fh in CCSM4 and from 8 to 53 days in HadGEM2-CC. (Some longest events exceed the ranges plotted in Figure 1.) Comparing Hh to RCP8.5_Ff, the longest duration increases in 8 out of the 13 models and is more than three times longer in CCSM4 and more than six times longer in the HadGEM2-CC for the CMIP5_Fh RCP8.5 scenario. Comparing Hh to CMIP5_Fh for the RCP4.5 scenario is about two times for CCSM4 and nearly three times longer for HadGEM2-CC. Specificaly, in HadGEM2-CC simulations, the number of Fh events in RCP4.5 is greater than in RCP8.5, the longest event is more than twice as long (53 vs 22 days). However, comparing Hh and Ff cases finds little difference in the length of the longest events (similar to the average duration results).

Grotjahn (2016) has similar histograms using durations above one standard deviation for Hh and Fh simulations by CCSM4. He found RCP8.5 durations above one standard deviation to be most common at four and five days, a histogram structure different than found for CCSM4 here, but similar to the result for HadGEM2-CC. He also found the number of events declines more slowly for longer durations than shown here. His results are consistent with the general warming comparable to one standard deviation, but much less than the 95^{th} percentile used here. Mastrandrea et al. (2011) define HW duration as at least 5 consecutive days with maximum temperature 5C above the LTDM; they find duration little changed on average at CCV locations from 1950-2000. They use two downscaling methods to find HW durations increase on average roughly 10 or 15 days from the historical period to 2100 for three to six models' simulations of the 'A2' (Nakicenovic et al. 2000) scenario. Historical HWs as they define them last about twice as long as ours so the average duration increase they find seems consistent with ours.

Zonal wind and temperature anomalies increase in the projection domain between RCP4.5 and RCP8.5 simulations. But, there is not a clear increase in the events from the RCP4.5 to RCP8.5 simulations. Some RCP8.5_Fh simulations (CCSM4, bcc-csm1-1-m, CNRM-CM5, and inmcm4, GFDL-ESM2G and GFDL-ESM2M) do increase the number of events from the RCP4.5 to the RCP8.5 simulations. However, the other models (including coarser resolution models MIROC-ESM, MIROC-ESM-CHEM and FGOALS-g2) have fewer heat events in RCP8.5 than RCP4.5. Though there are fewer events they last longer. For all but one model (MIROC-ESM) the average number of days each year that are during a HW is greater for RCP8.5 than RCP4.5. (MIROC-ESM has essentially the same number of HW days, <1% difference, in both Fh scenarios.) The Supporting Materials show percent of HW days for all models and scenarios. Oleson et al. (2018) find total HW days/year increasing above historical values for both RCP scenarios (RCP8.5 roughly double the RCP4.5 increase) at CCV locations in CESM large ensemble simulations. Schoetter et al. (2015) find a similar result for European HWs.

The multi-model weighted average numbers of events (Table 2) are essentially the same between Hh and Ff simulations (35.6 and 36.3 respectively) and are similar to the reanalysis number of 32. However, the number of events using historical thresholds in the future (Fh data) is four times as large for RCP8.5 simulations. So, the multi-model average has a HW on 25% of the days in the RCP8.5_Fh (18% of the RCP4.5_Fh days) compared with 3% of the Hh and Fh days.

3.3 Past and Future Number of Events by Cluster Type

Most heat wave events have LSMPs that cluster into one of two types using the projection methodology described in §2.3. Average projection values for each pair of cluster types for each event are plotted in Figure 2. The projection method was developed for the NNRA1 data (which match corresponding values for ERA-Interim data as a check). The NNRA1 data in Figure 2 nicely separate events along a line between the two clusters, with one mixed type. The NNRA1 data show that if an event projects strongly on one cluster type, then that event often projects weakly or negatively the other cluster type. Although simulated historical heat waves in the models are not so neatly along a line, most model events separate into one of the two types in a way that is similar to the reanalysis result. As noted in Table 2, the models vary a bit in terms of their relative fractions of type 1, 2, or mixed. Like GL2016, the models have more mixed events than the reanalysis, but the proportion of events in each type is not much different than the reanalysis for most models. Our updated cluster projection scheme clearly improves the separation between the clusters than shown by similar figures in: LG2016 for the reanalysis and GL2016 for these models. GL2016 shows the distance between the centroids for clusters one and two; repeating the calculation here finds that distance increases on average by 67% for these models and the multi-model weighted mean distance increases by 73% to become within 8% of the NNRA1 value. (Individual model details are in the Supporting Materials).

The projection procedure was applied to the RCP4.5 and RCP8.5 simulations using historical thresholds (Fh). These data are not plotted but the numbers of events of each type are included in

Table 2 for the RCP8.5 simulations. The greater number of events in the future using historical thresholds is not evenly split between the two cluster types but is disproportionately found in Type 2. Cluster Type 2 is characterized by a preexisting hot anomaly in southwestern Canada, but the future climatology in the models is several degrees warmer than historically, especially over the continents and extending over the adjacent oceanic areas. (The CCSM4 future climatology and how that climate change maps onto projection areas used to distinguish the two clusters are in the Supporting Materials.) The domain used for the cluster type designation has a cool anomaly for Type 1 and a warm anomaly for Type 2 at 850 and 500 hPa. Hence, the future climatology alone favors the Type 2 projection.

As noted, the future climatology (Ff) has a similar number of events as in the historical period. The split between the two types changes between historical (Hh) and future (Ff) simulations in the models. In CCSM4 and MIROC-ESM-CHEM, Type 2 events double and Type 1 are fewer. In contrast, CNRM-CM5 has half as many more Type 1 but fewer Type 2 events. Other models change the balance between event types between these extremes. The balance between the two event types in RCP4.5 simulations is similar though some models have opposite changes compared to RCP8.5 results. The models do not show a systematic change. Thus, the multi-model average in the future (Ff) is very similar to the recent past (Hh). In short, neither cluster type LSMP is more common in the future. Sillmann et al. (2013a) assess CMIP5 model performance on ETCCDI temperature extremes (http://etccdi.pacificclimate.org/list_27_indices.shtml) and single out the MIROC models for criticism. Cheng et al. (2015) find FGOALs-g2 and these MIROC models to have a larger mean error than other models we examine.

The Hh, Fh, and Ff results taken together indicate that the frequency and magnitude of the LSMP are not changed noticeably but that the increase in HWs based on historical thresholds (Fh) is due primarily to a change in the climatology, i.e. to the ‘global warming signal’. This result seems consistent with Brewer and Mass (2016) who show: 1) a general increase in the geopotential heights at 500 hPa especially near the Northwestern US and 2) a similar number of troughs and ridges centered near longitude range 123W during 1970-1999 and 2071-2100. 123W is near where our anomaly LSMP peaks. Similarly, Horton et al. (2015) do not find a robust change in this region during the recent past 25 years.

3.4 Past and Future Cluster Strength

The strength of each event is measured by the largest avTnamax that occurs during the event. These largest avTnamax values can be further stratified by the cluster type. Figure 3 shows the evolution of event strength by cluster type over each 40-year period. Figure 3 does not have large trends or significantly more events in the latter part of the reanalysis or most historical simulations. The results use anomalies defined from averages over the period of each dataset, so the use of offset historical periods is justified.

The large future increase of Type 2 events in Fh results is immediately obvious in the preponderance of blue symbols. The increased strength of events is also easily seen. In general,

most models tend to have similar distributions in the Hh and Ff panels. But, within the Ff panels, the number of events per decade increases towards the end of the period for most models, especially for RCP8.5.The main exceptions are the MIROC models; their disparity from other models has been noted before (Sillmann et al., 2013a; Cheng et al., 2015).

Since avTnamax values are normalized by the standard deviation, the peak values of those future temperatures in some models are quite high. For RCP8.5_Fh, CCSM4 has a half-dozen events exceeding four standard deviations above the historical mean. Similar results are found for other models, including the other four highest resolution models, plus NorESM1-M and GFDL-CM3. The other two GFDL models and FGOALS-g2 do not have quite as strong events. The remaining models, especially the MIROC models have stunningly high peak average temperatures as numerous events exceed 5 standard deviations and in the MIROC-ESM model two events exceed eight standard deviations above the historical mean. The MIROC models and to a lesser degree the bcc-csm1-1 results are consistently different from the other models in having larger scatter and extreme avTnamax values in historical as well as future climatological situations.

3.5 Past and Future LSMP Index Distributions

Scatter plots (Figure 4) compare the LSMPi values with CCV-average avTnamax values on all 4880 days of summer (1971-2010) from the NNRA1 and 1961-2000 from the models. Similar plots are in Grotjahn (2013) and Katz and Grotjahn (2014). Each panel in Figure 4 shows just the extreme values avTnamax. Contingency table scores: FAR and POD (§2.5) are included in each panel. It is best if FAR has low value and POD has high value in their 0 to 1 ranges. All models have POD > FAR.

The LSMPi was developed to best fit avTnamax on the few onset dates of HWs using the NNRA1 data. The climate models also have a strong correspondence between high LSMPi and high avTnamax. Nearly all models outperform the reanalysis judging from the FAR and POD values. Collapsing the relationship to a regression curve (Figure 4) shows that the relationship between LSMPi and avTnamax varies between models. Most models have a nearly linear regression curve meaning the match between LSMPi and avTnamax extends from moderate to high values of avTnamax. Such models that show a consistent LSMPi to avTnamx relationship for very high temperatures reinforce applying LSMPi to future climate simulations. However, MIROC-ESM models have a large spread of low LSMPi values during high avTnamax dates while the inmcm4 model has a large range of avTnamax values for high LSMPi dates, both situations reduce the match between the two quantities; but since both situations do not occur together in these models, their FAR and POD scores are better than for the reanalysis.

Figure 5 shows the historical and future distributions of LSMPi>1 values. This figure is similar to Figure 7 in GL 2016, but the figure here shows all the extreme values not just LSMPi values on the onset days. The reanalysis distribution is plotted in every panel as a blue dotted curve. The Hh simulations (dotted red curves) seem to underestimate the standard deviation of the LSMPi

distribution in several models, especially CCSM4, NorESM1-M, the MIROC models, and FGOALS-g2.. However, the bcc models, CNRM-CM5, and HadGEM2-CC values match the reanalysis well over the distribution range shown. Grotjahn (2016) noted the CCSM4 distribution being narrower than the reanalysis. Cheng et al. (2015) find FGOALS-g2 colder, in climatology and return values near the CCV, than the other models common to our study and theirs.

Figure 5 shows future scenarios using both historical (Fh) and future (Ff) climatologies to define anomalies. The number of events and relative strength of the events are very similar between Ff and Hh results. Ff and Hh distributions in Figure 5 are also have highly similar high tails, though some models differ from this general conclusion. HadGEM2-CC and two GFDL models have lower probability density values in Ff than in Hh results for both RCP scenarios. Model inmcm4 has lower values for RCP8.5 than either historical or RCP4.5 results. Since there are many more heat waves that last longer in the future when using historical thresholds, the Fh curves in Figure 5 are systematically shifted to higher LSMPi values relative to the Hh and Ff curves. The amount of shift varies between models; the multi-model average shift ~0.25. Brewer and Mass (2016) show 700 hPa summer temperature changes (1970-99 versus RCP8.5 2071-2100) that vary widely between models (e.g. larger for GFDL-CM3 than for GFDL-ESM2M) consistent with Figure 5. Grotjahn (2016) noted CCSM4 increases negative skew in the future. Lau and Nath (2012) find a similar skew change for GFDL simulations of future climate (A1B scenario) for the Pacific Northwest US. Here, some models have more negative but others have less negative skew change from RCP4.5Fh to RCP8.5Fh. So, the multi-model mean has little change in skewness.

Some qualitative impressions from Figure 5 can be made quantitative by calculating scale and shape parameters from a Generalized Pareto distribution (GP) fit. The GP scale parameter (Figure 6a) varies by ~0.1 between models relative to the multi-model mean and the reanalysis value (0.32). The direction of the change in GP scale between cases is generally consistent. Except for the CNRM-CM5 and inmcm4 models, the scale increases for RCP4.5_Fh and even more for RCP8.5_Fh. The amount of increase varies greatly between models. However, the multi-model average scale is a third larger for RCP4.5 and more than half again larger for RCP8.5_Fh. The GP shape parameter is negative for the reanalysis and nearly all cases by all models. Negative shape means the tail is unbounded. The models are not consistent about the change of GP shape between the cases. Because the shape results are so equivocal shape is shown in the Supplemental Materials.

Return value also provides information on a distribution's high tail and is shown in Figure 6b. The 20-year return value may be interpreted as that value having a 5 % chance of being exceeded in any particular year. The return values in Ff cases are generally very close to Hh values for each model (LSMPi= 1.3-2). The differences between Ff and Hh values are smaller than the range among the models. So again, the large scale pattern for the heat wave is not occurring more intensely in the future if one uses the future climatology to define the anomalies. The return values for Fh cases (LSMPi= 2-2.8) are systematically >50% larger than the historical values. The multi-model averages are 1.76 for Hh, 2.14 and 2.24 for RCP4.5 and RCP8.5, respectively. As Figure 4 shows, different models have a different relation between LSMPi value and corresponding near surface

664 temperature. For many models LSMPi increases more slowly than temperature; so, LSMPi 20-year
665 return values >2 imply very high if not unprecedented surface temperatures.
666
667 A broad estimate of the hotter surface temperatures is shown in Figure 6c. The estimate is
668 calculated from the average of the 50 highest avTnamax values for each case and model. Each
669 average is multiplied by delT = 3.97K. This delT is the value used to normalize the temperature
670 anomalies on average for the CCV stations during summer. The difference between the future
671 climate value and the historical value for the model is plotted in Figure 6c. Relative to future
672 climatology (blue and red dots), the models vary about zero, consistent with other Ff results shown
673 above. The future simulations relative to historical values finds a consistent increase that is larger
674 for RCP8.5_Fh. The amount of increase ranges from 2 to 8K for RCP4.5_Fh and 4 to 11K for
675 RCP8.5. The multi-model averages are: 3.3 and 6K for RCP4.5 and RCP8.5, respectively.
676
677 A simple quantitative metric for a trend (Figure 6d) is to subtract the average of the 30 highest
678 avTnamax values in the first 20 years from the corresponding average over the last 20 years of each
679 period. While the reanalysis does not have a trend, there is a trend in the multi-model Hh mean
680 though the trend varies a lot between models. Inspection of Figure 3 shows most models trend
681 towards more events later in the RCP8.5_Ff simulations and that is consistent with Figure 6d. There
682 is no clear trend in the RCP4.5 data over the 40-year period. Grotjahn (2016) showed similar results
683 for CCSM4 data and slightly different comparison periods. The multi-model average trends in C/20
684 years are 1 for Hh, and 1.4 (1.5) for RCP8.5_Ff (Fh).
685
686 Radiative forcing increases until mid-century in RCP4.5 (e.g. Russo et al. 2014). The minimal trend
687 in RCP4.5 multi-model mean reflects max temperatures over land areas that asymptote to a nearly
688 constant value by ~2070 (Sillmann et al., 2013b) in multi-model averages. When decadal
689 differences compare a historical period to late in the Twenty-first Century, then RCP4.5 has a trend
690 (Karin et al., 2013; Sillmann et al., 2013b; Pierce et al., 2013). However, Gershunov and Guirguis
691 (2012) find no obvious trend in CCV 'daytime relative' HWs in CNRM simulations (A2 scenario)
692 of 1950-2100.
693
694
695 **4 Summary**
696
697 How general properties of CCV heat waves (HWs) change for RCP4.5 and RCP8.5 scenarios are
698 studied using 13 climate models. Future results use anomalies defined relative to either historical
699 climatology ('Fh' data) or the climatology of the future period ('Ff' data). Forty-year simulations
700 by each model, both historical ('Hh') and future, are compared to detect relative changes. Two
701 cluster types of patterns lead up to CCV HW onset. These climate models develop both types.
702 There are thus five groupings of model output: one Hh, and two scenarios each of Fh and Ff data.
703 Each of these five categories is split into the two cluster types. Forty years of NNRA1 reanalysis
704 data of are used for comparison.
705

HW type and intensity can be related to the upper air large scale meteorological patterns (LSMPs) via indices. This work improves upon the LSMP-based index of HW intensity (GL2016) and assignment of HW type (LG2016). The index LSMPi provides a compact and accurate way of characterizing the LSMP developed by a model during a HW. Station surface maximum temperature anomalies, normalized by the local standard deviation, then averaged over the CCV define 'avTnamax'. Daily LSMPi is strongly related to avTnamax in scatter plots and a corresponding regression curve is calculated for each model. The link between LSMPi and avTnamax is *stronger* in models than in the reanalysis. Since LSMPi properties characterize HWs in the models, it is useful to examine the LSMPi statistics. How well each model's historical simulations match four statistical properties of the reanalysis LSMPi defines weights used to calculate multi-model means. The models that match the reanalysis better are given more weight in the multi-model mean.

Like GL2016, most models capture the frequency of Hh HWs, though some models develop twice as many heat waves. Model distributions of duration are comparable to that from the reanalysis, though the models developing many more HWs have a larger fraction of short (3-day) events. Like GL2016, the split between event types varies between models. In Fh scenarios, there are 4x as many events and their average durations are twice as long versus historical simulations. The increase is mainly in cluster Type 2 events; that cluster has a pre-existing HW over Canada not present in cluster Type 1. (The Supporting Materials show the Type 2 pattern along with the change in future climatology for CCSM4.) In future scenarios, these models have higher average temperatures over the continents, thereby explaining the asymmetric preference for Type 2 HWs. However, when HWs are defined as extremes relative to the future climatology, then the number of events and the proportion of each cluster type are both very similar to the corresponding historical values. These results are consistent with Brewer and Mass (2016) and Horton et al. (2015). Therefore the large scale patterns that create the HWs are not occurring more or less frequently in the future. To the extent that the LSMPi represents the variability of the summer temperatures (shown in Grotjahn, 2011) then the future variability is the same as in the historical simulations. That result means that the increases in heat waves, their intensity, and other metrics shown are primarily due to a warming of the average conditions..

Examining avTnamax values in models, Hh and Ff data are very similar, while Fh data have many more days with higher values. How high the values reach varies greatly between models. Most models have peak avTnamax values between 2-3 standard deviations for Hh and Ff calculations while most Fh values range between 2-4 (2-5) standard deviations above the mean in RCP4.5 (RCP8.5) data. However, a few models have Fh values up to 8 standard deviations.

The number of events is larger for RCP4.5_Fh than RCP8.5_Fh in six models and vice versa for the other seven. Those six models have events with much longer durations than historically. Fewer events occur when they last longer. Hence, the fraction of the summer experiencing a HW using historical thresholds increases for 12 of the 13 models. The multi-model average finds ~3% of summer days are HWs for the Hh and both Ff cases; ~18% of RCP4.5_Fh days, and 25% of the

RCP8.5_Fh days. Oleson et al. (2018) find similar fractions for CCV locations. The multi-model mean duration increases from 4.21 to 5.56 to 7.87 days in Hh, RCP4.5_Fh, and RCP8.5_Fh data. Again, the patterns are not lasting longer than corresponding historical patterns when the future climatology is used to define them.

Multi-model averages find for Fh LSMPi: the Generalized Pareto scale parameter increases (by more than 50%) and the 20-year return period value increases by almost 30% in the RCP8.5 data; both are consistent with the LSMPi distribution shifting to higher values. However, the sign of the RCP4.5 temperature trend *within* 2061-2100 is not consistent between the models. Extreme temperatures also increase. The extreme values in RCP scenarios are consistently larger than historical values in all models, though the amount of increase from the 1961-2000 values and those a century later varies widely, by a factor of three. These increases are comparable to those found by other multi-model studies (e.g. Kharin et al. 2013). An estimate based on historical scaling finds the multi-model average is >3C warmer for RCP4.5 and 6C hotter for RCP8.5 scenarios compared to historical conditions.

Acknowledgements

This research was funded in part by NSF grant 1236681 and also supported by the USDA National Institute of Food and Agriculture, Hatch project Accession #1010971. Coding assistance was provided by Dr. Yun-Young Lee. Additional support provided by NASA grant NNX16AG62G and Department of Energy Office of Science award number DE-SC0016605, ``An Integrated Evaluation of the Simulated Hydroclimate System of the Continental US.". NOAA station data retrieved from http://ipm.ucanr.edu/WEATHER/wxactstnames.html. NCEP Reanalysis data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html. ECMWF ERA-Interim data used in this project are from the ECMWF data server: http://apps.ecmwf.int/datasets/data/interim-full-daily/ . We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. These model data are available from the Earth System Grid Federation at https://esgf-data.dkrz.de/search/cmip5-dkrz/ . Programs used here are available from http://grotjahn.ucdavis.edu/EWEs .

787 References
788
789 Brewer, M. C., & Mass, C. F. (2016) Projected changes in western U.S. large-scale summer
790 synoptic circulations and variability in CMIP5 models. *Journal of Climate, 29,* 5965-5978.
791 https://doi.org/10.1175/JCLI-D-15-0598.1
792 Cheng, L., Phillips, T. J., & Kouchak, A. (2015) Non-stationary return levels of CMIP5 multi-
793 model temperature extremes. *Climate Dynamics. 44* , 2947-2963.
794 https://doi.org/10.1007/s00382-015-2625-y
795 Clemesha, R. E. S., Guirguis, K., Gershunov, A., Small, I.J., & Tardy, A. (2017) California heat
796 waves: their spatial evolutions, variation, and coastal modulation by low clouds. *Climate*
797 *Dynamics*, https://doi.org/10.1007/s00382-017-3875-7
798 Coumou, D., Robinson, A., & Rahmstorf, S. (2013). Global increase in record-breaking monthly-
799 mean temperatures. *Climatic Change*. *118* 771–82. https://doi.org/10.1007/s10584-012-
800 0668-1
801 Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA -
802 Interim reanalysis: Configuration and performance of the data assimilation system,
803 *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553-597.
804 https://doi.org/10.1002/qj.828
805 Gershunov, A., Cayan, D. R., & Iaccobellis, S. F. (2009) The great 2006 heat wave over California
806 and Nevada: Signal of an increasing trend. *Journal of Climate, 22,* 6181-6203
807 https://doi.org/10.1175/2009JCLI2465.1
808 Gershunov, A., & Guirguis, K. (2012) California heat waves in the present and future. *Geophysical*
809 *Research Letters*, *39*, L18710, https://doi.org/10.1029/2012GL052979
810 Grotjahn, R. (2011). Identifying extreme hottest days from large scale upper air data: a pilot scheme
811 to find California Central Valley summertime maximum surface temperatures, *Climate*
812 *Dynamics*, *37*(3-4), 587-604. https://doi.org/10.1007_s00382-011-0999-z
813 Grotjahn, R. (2013). Ability of CCSM4 to simulate California extreme heat conditions from
814 evaluating simulations of the associated large scale upper air pattern, *Climate Dynamics*,
815 *41*(5-6), 1187-1197. https://doi.org/10.1007/s00382-013-1668-1
816 Grotjahn, R. (2016). Western North American extreme heat, associated large scale synoptic-
817 dynamics, and performance by a climate model, in J. Li, R. Swinbank, R. Grotjahn, H.
818 Volkert (Eds.), *Dynamics and Predictability of Large-scale, High-Impact Weather and*
819 *Climate Events*, (pp. 198–209). Cambridge, England: Cambridge University Press, ISBN
820 978-1-107-07142-1

821 Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., et al. (2016). North
822 American extreme temperature events and related large scale meteorological patterns: a
823 review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics, 46*,
824 1151–1184. https://doi.org/10.1007/s00382-015-2638-6
825 Grotjahn, R. & Lee, Y.-Y. (2016). On climate model simulations of the large-scale meteorology
826 associated with California heat waves, *Journal of Geophysical Research:*
827 *Atmospheres*, *121*, 18–32. https://doi.org/10.1002/2015JD024191

828 Horton, D. E., Johnson, N. C., Singh, D., Swain, D. L., Rajaratnam, B., Diffenbaugh, N. S. (2015)
829 Contribution of changes in atmospheric circulation patterns to extreme temperature trends.
830 *Nature* 522, 465–469. https://doi.org/10.1038/nature1550
831 Horton, R. M., Mankin, J. S., Lesk, C., Coffel, E., & Raymond, C. (2016), A Review of Recent
832 Advances in Research on Extreme Heat Events. *Current Climate Change Reports*, *2*, 242-
833 259. https://doi.org/10.1007/s40641-016-0042-x.
834 Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation.
835 *Biometrika*, *36*. 149-176.
836 Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven D., Gandin, L., et al. (1996). The
837 NCEP/NCAR 40-year reanalysis project, *Bulletin of the American Meteorological Society*,
838 *77*(3), 437-471.
839 Katz, R. W., & Grotjahn, R. (2014) Statistical methods for relating temperature extremes to Large-
840 Scale Meteorological Patterns. US CLIVAR *Variations*, *12* (1), 4-7.
841 Kharin, V. V., Zwiers, F. W., Zhang, X., & Wehner, M. 2013: Changes in temperature and
842 precipitation extremes in the CMIP5 ensemble. *Climatic Change*, 119, 345-357.
843 https://doi.org/10.1007/s10584-013-0705-8
844 Lau, N.-C., & Nath, M. J. (2012) A model study of heat waves over North America: Meteorological
845 aspects and projections for the Twenty-First Century. *Journal of Climate*, *25*, 4761-4784
846 https://doi.org/10.1175/JCLI-D-11-00575.1
847 Lee, Y.-Y. & Grotjahn, R. (2016). California Central Valley summer heat waves form two ways.
848 *Journal of Climate*, *29*, 1201-1217. https://doi.org/10.1175/JCLI-D-15-0270.1.
849 Lehmann, J., Coumou, D., Frieler, K., Eliseev, A. V., & Levermann, A. (2014) Future changes in
850 extratropical storm tracks and baroclinicity under climate change. *Environmental Research*
851 *Letters*, *9*(8)084002. https://doi.org/10.1088/1748-9326/9/8/084002
852 Marzban, C. (1998) Scalar measures of performance in rare-event situations. *Weather and*
853 *Forecasting*, *13*, 753–763.
854 Mastrandera, M. D., Tebaldi, C., Snyder, C. W., & Schneider, S. H. (2011) Current and future
855 impacts of extreme events in California. *Climatic Change*, *109*, S43-S70.
856 https://doi.org/10.1007/s10584-011-0311-6
857 Nakicenovic, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., Gaffin, S., et al. (2000)
858 *Intergovernmental Panel on Climate Change Special report on emissions scenarios*.
859 Cambridge University Press, Cambridge ISBN 0521804930, 612 pp
860 Oleson, K. W., Anderson, G. B., & Jones, B. et al. (2018) Avoided climate impacts of urban and
861 rural heat and cold waves over the U.S. using large climate model ensembles for RCP8.5
862 and RCP4.5 *Climatic Change*, *146*, 377-392. https://doi.org/10.1007/s10584-015-1504-1
863 Perkins, S. E, Alexander, L. V., & Nairn, J. R. (2012) Increasing frequency, intensity and duration
864 of observed global heatwaves and warm spells. *Geophysical Research Letters*, *39* (20),
865 L20714, http://dx.doi.org/10.1029/2012GL053361
866 Petoukhov ,V., Rahmstorf, S., Petri, S., & Schellnhuber, H. J. (2013) Quasiresonant amplification
867 of planetary waves and recent Northern Hemisphere weather extremes. *Proceedings of the*
868 *National Academy of Sciences of the United States of America*, *110*, 5336–5341.
869 https://doi.org/10.1073/pnas.1222000110.

Pierce, D. W., Das, T., Cayan, D. R., Maurer, E. P., Miller, N. L., Bao, Y., et al. (2013) Probabilistic estimates of future changes in California temperature and precipitation using statistical and dynamical downscaling. *Climate Dynamics, 40*, 839-856. https://doi.org/10.1007/s00382-012-1337-9

Russo, S., Dosio, A., Graversen, R.G., Sillmann, J., Carrao, H., & Dunbar, M.B (2014) Magnitude of extreme heat waves in present climate and their projection in a warming world. *Journal of Geophysical Research: Atmospheres, 119*, 500–512. https://doi.org/10.1002/2014JD022098

Schoetter, R., Cattiaux, J. & Douville, H. (2015) Changes of western European heat wave characteristics projected by the CMIP5 ensemble. *Climate Dynamics,* 45, 1601. https://doi.org/10.1007/s00382-014-2434-8

Screen, J. A, & Simmonds, I. (2014). Amplified mid-latitude planetary waves favour particular regional weather extremes. *Nature Climate Change, 4*, 704–709. https://doi.org/10.1038/NCLIMATE227

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013a) Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres, 118*, 1716–1733. https://doi.org/10.1002/jgrd.50203

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013b) Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres, 118*, 2473-2493. https://doi.org/10.1002/jgrd.50188

Stephenson D. B, Casati, B., Ferro, C. A. T., & Wilson, C. A. (2008) The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications, 15*(1), 41–50, https://doi.org/10.1002/met.53

Teng, H., Branstator, G., Wang, H., Meehl, G. A., & Washington, W. M. (2013) Probability of US heat waves affected by a subseasonal planetary wave pattern. *Nature Geoscience, 6*, 1–6. https://doi.org/10.1038/ngeo1988

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

Weigel, A. P., Knutti, R., Liniger, M. A., & Appenzeller, C. (2010) Risks of model weighting in multimodel climate projections. *Journal of Climate* 23, 4175-4191. https://doi.org/10.1175/2010JCLI3594.1

Wheeler, R. E. (1980) Quantile estimators of Johnson curve parameters. *Biometrika*, *67*(3) 725-728 https://doi.org/10.2307/2335153

Figure Captions

Figure 1. Histogram of heat waves duration (in consecutive days) for CMIP5 models for each of the groupings: Hh, Ff and Fh (both RCP4.5 and RCP8.5 scenarios). The historical period is 1961-2000 while the future period is 2061-2100. Included in the figure are the length of the longest event and the average duration. For models with more than one ensemble member, each bin is divided by the ensemble size. The longest event in each ensemble member was found, added together, and then divided by the number of ensembles for that model to produce the number shown. a) Six CMIP5 models with corresponding NCEP-NCAR reanalysis values for 1971-2010 shown for comparison. b) seven more CMIP5 models, c) the multi-model mean duration histograms.

Figure 2. Projection coefficients onto each cluster type for all heat waves in the reanalysis and the models. The projections are onto upper air variables in a specific region as detailed in the text. Red dots are events that are primarily type 1 while blue dots are primarily type 2; green dots are mixed type events. These data are for 40-year historical periods. Events in all ensemble members are shown; CNRM-CM5, NorESM1-M, MIROC-ESM, both bcc, and all three GFDL models have three ensemble members; HadGEM2-CC and FGOALS-g2 have two members, and the remainder one member.

Figure 3. Maximum avTnamax temperature during each event as a function of time in each 40-year period. The peak value of each event is color-coded such that red circles are cluster type 1, blue circles designate type 2, and green circles are the mixed type. The layout of the reanalysis and model groupings matches figure 1: a) reanalysis and six models; b) seven more models. To make the results in different models and groupings comparable, only one ensemble member is used for each grouping.

Figure 4. Scatterplots of daily avTnamax (abscissa) and corresponding LSMP index (ordinate) for every day of the CMIP5_Hh simulations. The best fit curve uses the points where avTnamax is >1. Also included are the FAR (False alarm ratio) and the POD (probability of detection).

Figure 5. Distribution functions of LSMPi >1 for all historical (Hh) summer days (June-September). The NCEP-NCAR reanalysis (1971-2010) (blue dotted) curve is on all panels for reference. Model data are shown in a format similar to Figure 2._Red dotted curves are model Hh (1961-2000) data. Future scenarios (2061-2100) use green curves for RCP 4.5 and purple curves for RCP 8.5 data, with solid lines for Ff data and dashed lines for Fh data.

Figure 6. Distribution properties for the models. The black dots are Hh data, the red dots are RCP4.5_Ff data, the blue dots are RCP8.5_Ff data, the green dots are RCP4.5_Fh, and the purple

945 dots are RCP8.5_Fh data. Corresponding values for the multi-model weighted average and the
946 NCEP-NCAR reanalysis is also shown. a) Generalized Pareto (GP) scale parameter for the
947 extremes in the models examined for the five groupings. The threshold for the extremes is
948 LSMPi>1 (The LSMPi values >1 were all declustered to make the data independent prior to the
949 calculation as recommended for GP calculations. To calculate the GP function, we have used all the
950 ensembles available for each model. b) 20-year return values of LSMPi in the models and
951 reanalysis. c) Temperature anomaly difference from the Hh data of the 4 groups of future scenarios.
952 The anomaly in each group is the mean of the 50 largest avTnamax values for each group
953 multiplied by the delT value, where detT is the magnitude of the temperature normalization
954 averaged over the summer and all CCV stations. Here the delT value equals 3.97C. d) The trend
955 within each grouping, calculated as the average of the 30 largest avTnamax values during the last
956 20 years minus the corresponding values for the first 20 years. These values are also multiplied by
957 the delT value, so these trends have units of C/20 years.
958

959

Figure 1.

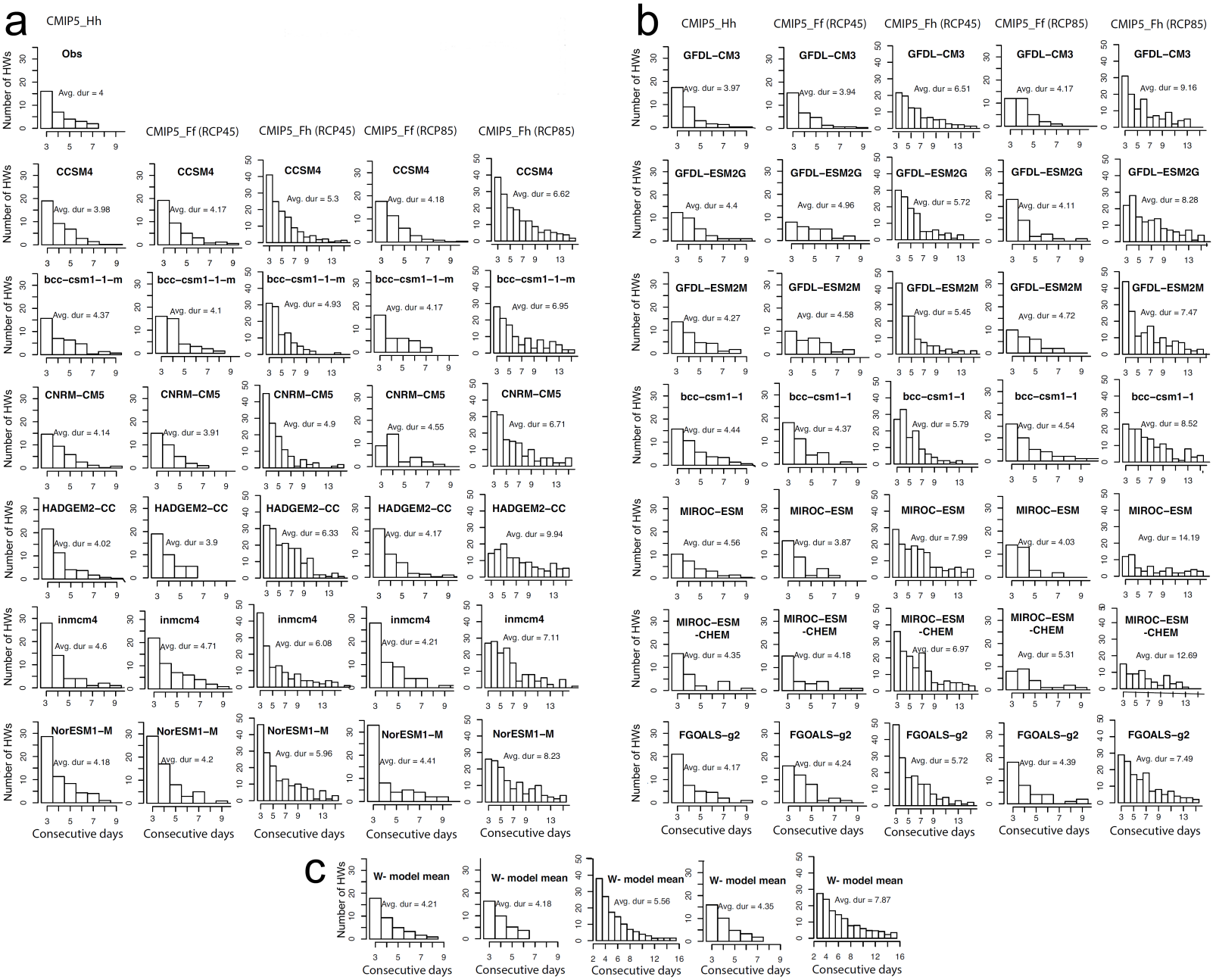
a
CMIP5_Hh
CMIP5_Ff (RCP45)
CMIP5_Fh (RCP45)
CMIP5_Ff (RCP85)
CMIP5_Fh (RCP85)
Obs
Avg. dur = 4
CCSM4
Avg. dur = 3.98
Avg. dur = 4.17
Avg. dur = 5.3
Avg. dur = 4.18
Avg. dur = 6.62
bcc−csm1−1−m
Avg. dur = 4.37
Avg. dur = 4.1
Avg. dur = 4.93
Avg. dur = 4.17
Avg. dur = 6.95
CNRM−CM5
Avg. dur = 4.14
Avg. dur = 3.91
Avg. dur = 4.9
Avg. dur = 4.55
Avg. dur = 6.71
HADGEM2−CC
Avg. dur = 4.02
Avg. dur = 3.9
Avg. dur = 6.33
Avg. dur = 4.17
Avg. dur = 9.94
inmcm4
Avg. dur = 4.6
Avg. dur = 4.71
Avg. dur = 6.08
Avg. dur = 4.21
Avg. dur = 7.11
NorESM1−M
Avg. dur = 4.18
Avg. dur = 4.2
Avg. dur = 5.96
Avg. dur = 4.41
Avg. dur = 8.23
Number of HWs
Consecutive days
b
CMIP5_Hh
CMIP5_Ff (RCP45)
CMIP5_Fh (RCP45)
CMIP5_Ff (RCP85)
CMIP5_Fh (RCP85)
GFDL−CM3
Avg. dur = 3.97
Avg. dur = 3.94
Avg. dur = 6.51
Avg. dur = 4.17
Avg. dur = 9.16
GFDL−ESM2G
Avg. dur = 4.4
Avg. dur = 4.96
Avg. dur = 5.72
Avg. dur = 4.11
Avg. dur = 8.28
GFDL−ESM2M
Avg. dur = 4.27
Avg. dur = 4.58
Avg. dur = 5.45
Avg. dur = 4.72
Avg. dur = 7.47
bcc−csm1−1
Avg. dur = 4.44
Avg. dur = 4.37
Avg. dur = 5.79
Avg. dur = 4.54
Avg. dur = 8.52
MIROC−ESM
Avg. dur = 4.56
Avg. dur = 3.87
Avg. dur = 7.99
Avg. dur = 4.03
Avg. dur = 14.19
MIROC−ESM −CHEM
Avg. dur = 4.35
Avg. dur = 4.18
Avg. dur = 6.97
Avg. dur = 5.31
Avg. dur = 12.69
FGOALS−g2
Avg. dur = 4.17
Avg. dur = 4.24
Avg. dur = 5.72
Avg. dur = 4.39
Avg. dur = 7.49
Number of HWs
Consecutive days
c
W- model mean
Avg. dur = 4.21
Avg. dur = 4.18
Avg. dur = 5.56
Avg. dur = 4.35
Avg. dur = 7.87
Number of HWs
Consecutive days

Figure 2.

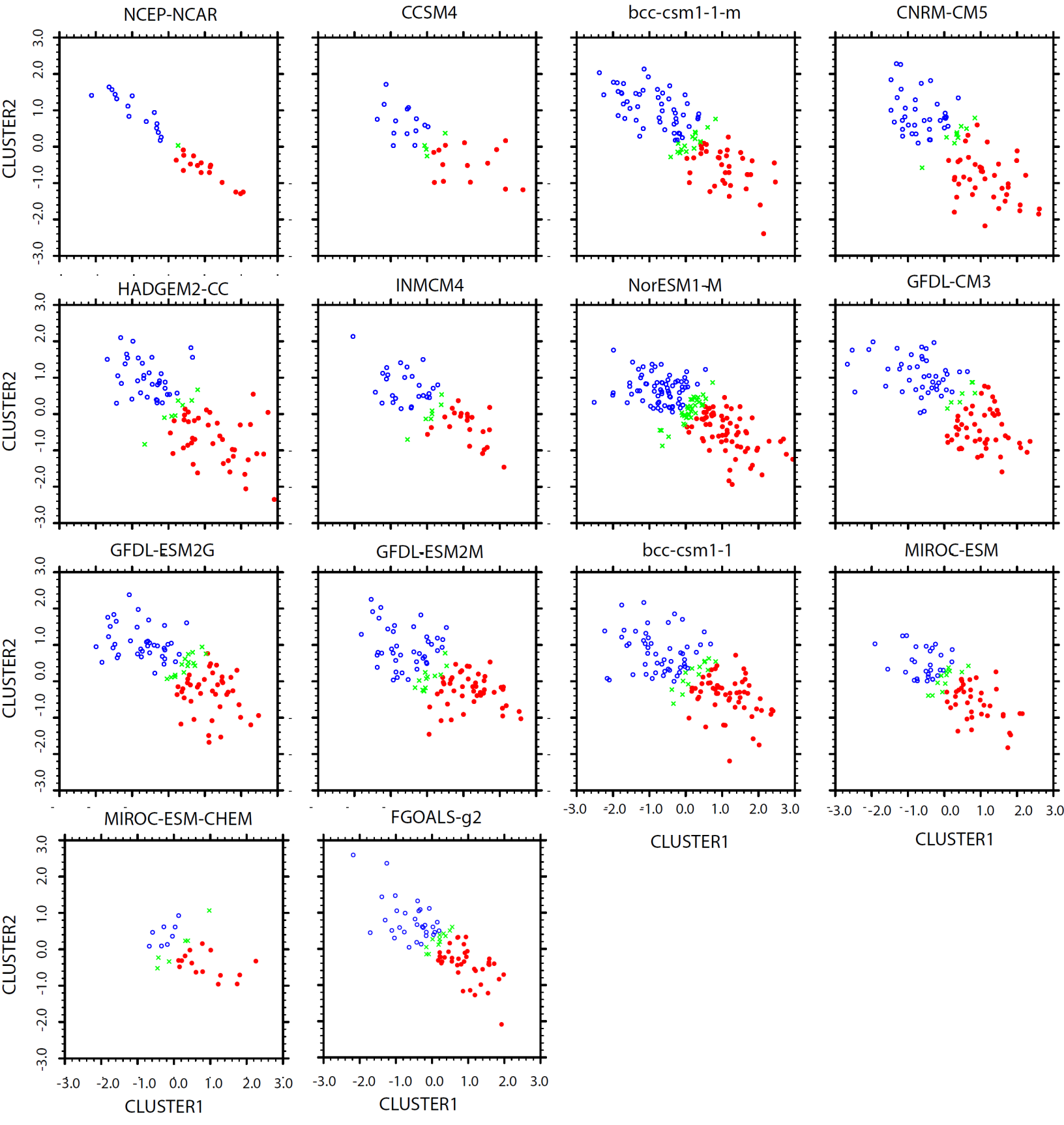
NCEP-NCAR
CCSM4
bcc-csm1-1-m
CNRM-CM5
HADGEM2-CC
INMCM4
NorESM1-M
GFDL-CM3
GFDL-ESM2G
GFDL-ESM2M
bcc-csm1-1
MIROC-ESM
MIROC-ESM-CHEM
FGOALS-g2
CLUSTER2
CLUSTER1
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

Figure 3.

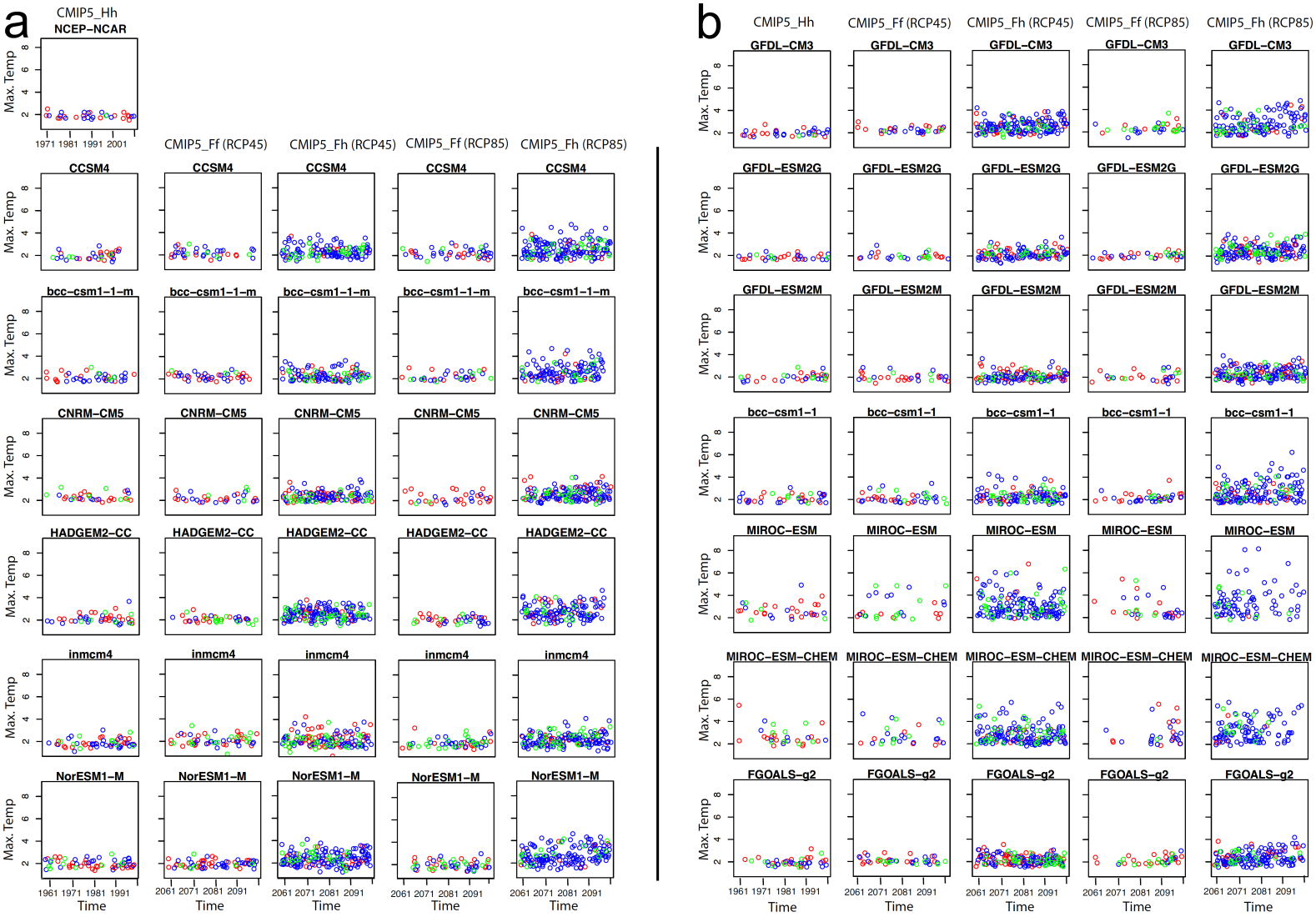
a
CMIP5_Hh
NCEP–NCAR
CMIP5_Ff (RCP45)
CMIP5_Fh (RCP45)
CMIP5_Ff (RCP85)
CMIP5_Fh (RCP85)
CCSM4
bcc–csm1–1–m
CNRM–CM5
HADGEM2–CC
inmcm4
NorESM1–M
Max. Temp
Time
b
CMIP5_Hh
CMIP5_Ff (RCP45)
CMIP5_Fh (RCP45)
CMIP5_Ff (RCP85)
CMIP5_Fh (RCP85)
GFDL–CM3
GFDL–ESM2G
GFDL–ESM2M
bcc–csm1–1
MIROC–ESM
MIROC–ESM–CHEM
FGOALS–g2
Max. Temp
Time

Figure 4.

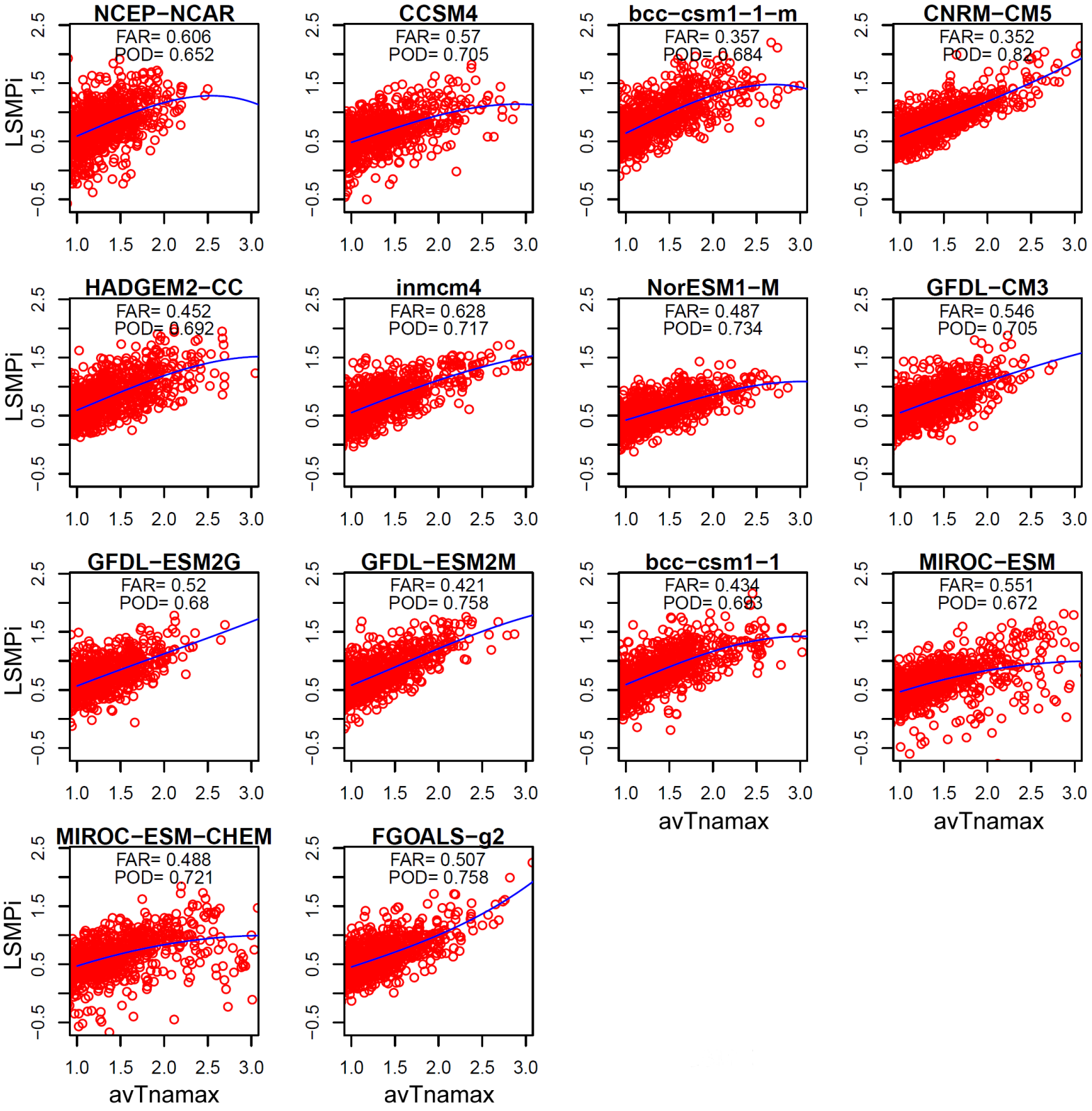
NCEP−NCAR
FAR= 0.606
POD= 0.652
CCSM4
FAR= 0.57
POD= 0.705
bcc−csm1−1−m
FAR= 0.357
POD= 0.684
CNRM−CM5
FAR= 0.352
POD= 0.82
HADGEM2−CC
FAR= 0.452
POD= 0.692
inmcm4
FAR= 0.628
POD= 0.717
NorESM1−M
FAR= 0.487
POD= 0.734
GFDL−CM3
FAR= 0.546
POD= 0.705
GFDL−ESM2G
FAR= 0.52
POD= 0.68
GFDL−ESM2M
FAR= 0.421
POD= 0.758
bcc−csm1−1
FAR= 0.434
POD= 0.693
MIROC−ESM
FAR= 0.551
POD= 0.672
MIROC−ESM−CHEM
FAR= 0.488
POD= 0.721
FGOALS−g2
FAR= 0.507
POD= 0.758
LSMPi
avTnamax
−0.5
0.5
1.5
2.5
1.0
1.5
2.0
2.5
3.0

Figure 5.

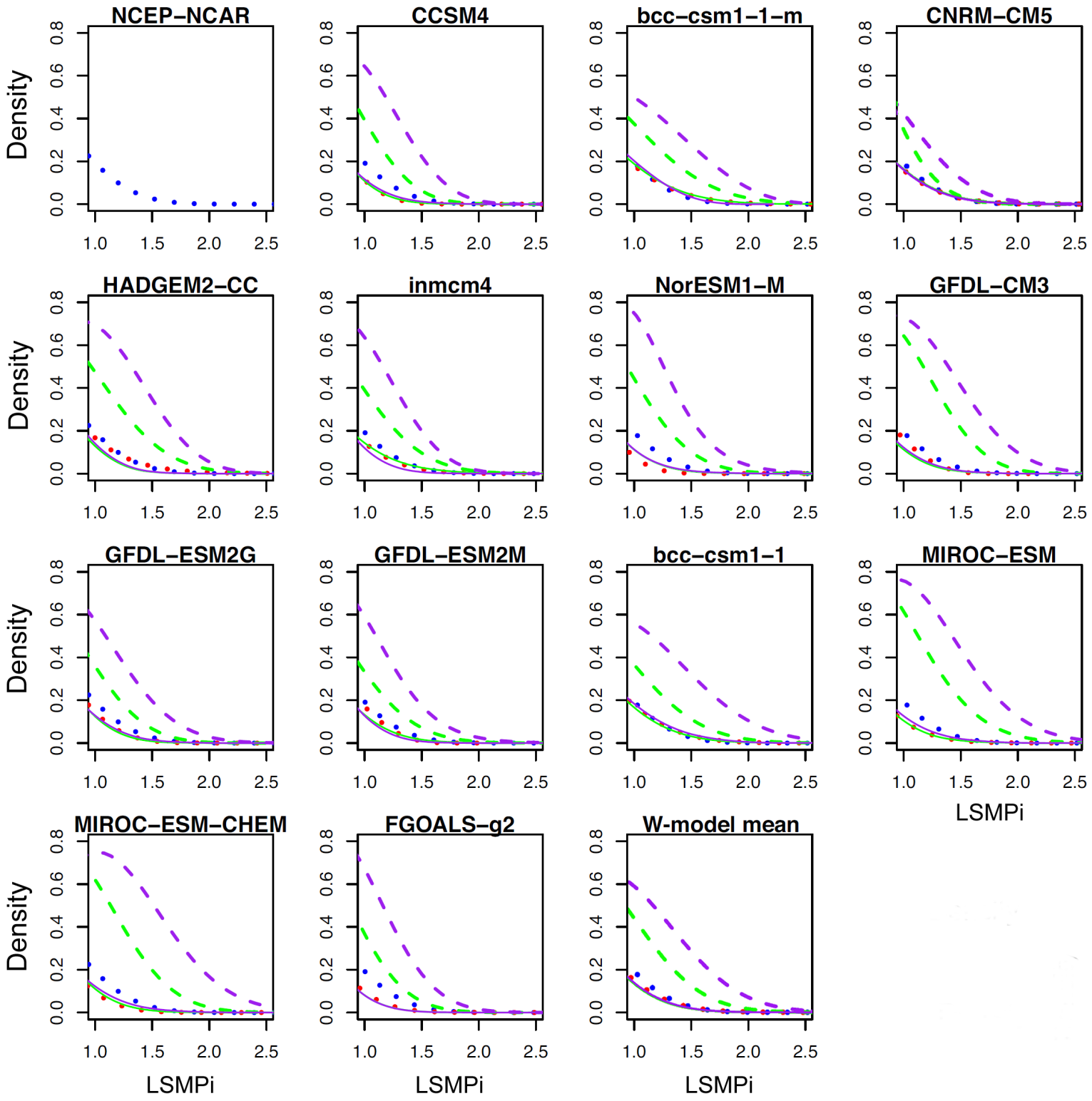
NCEP–NCAR
CCSM4
bcc–csm1–1–m
CNRM–CM5
HADGEM2–CC
inmcm4
NorESM1–M
GFDL–CM3
GFDL–ESM2G
GFDL–ESM2M
bcc–csm1–1
MIROC–ESM
MIROC–ESM–CHEM
FGOALS–g2
W-model mean
Density
LSMPi
0.0
0.2
0.4
0.6
0.8
1.0
1.5
2.0
2.5

Figure 6.

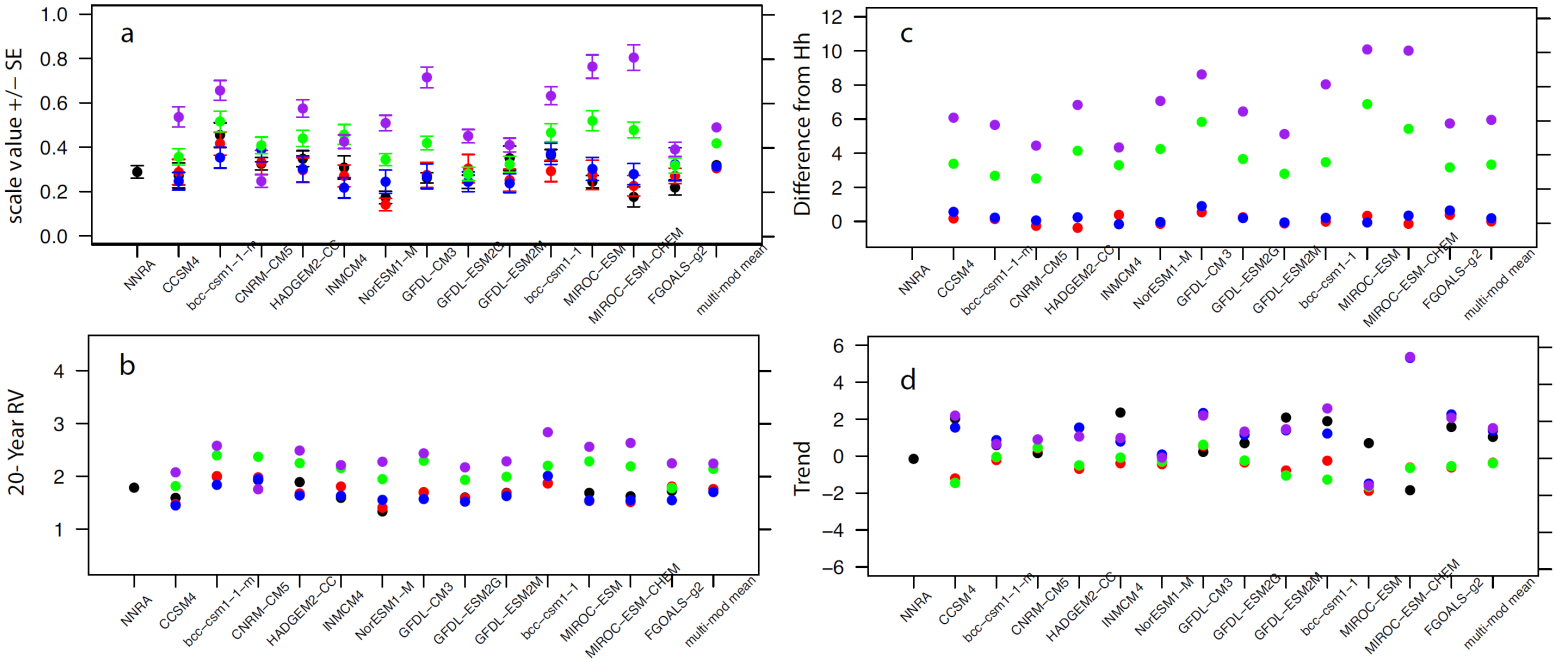
a
scale value +/− SE
1.0
0.8
0.6
0.4
0.2
0.0
NNRA
CCSM4
bcc−csm1−1−m
CNRM−CM5
HADGEM2−CC
INMCM4
NorESM1−M
GFDL−CM3
GFDL−ESM2G
GFDL−ESM2M
bcc−csm1−1
MIROC−ESM
MIROC−ESM−CHEM
FGOALS−g2
multi−mod mean
c
Difference from Hh
12
10
8
6
4
2
0
b
20- Year RV
4
3
2
1
d
Trend
6
4
2
0
−2
−4
−6

Table 1. *Metrics of model ability to capture the LSMP anomaly temperature at 850 hPa. Highest and lowest magnitudes in each column are in bold. Models listed from higher to lower horizontal resolution.*

Model	Ta_{850} Bias (K)	Ta_{850} Error (%)	Pattern correlation	Projection (95-175W; 20-60N)	Horizontal Resolution (lon x lat)
CCSM4	-0.20	**9.8**	**0.93**	0.91	288x192
Bcc-csm1-1-m	0.92	19.2	0.92	**1.21**	320x160
CNRM-CM5	-0.43	15.6	0.83	0.81	256x128
HADGEM2-CC	-0.24	10.3	0.90	0.85	192x144
INMCM4	-1.11	23.1	0.90	0.70	180x120
NORESM1-M	**-1.92**	**38.5**	0.84	0.60	144x96
GFDL-CM3	-0.37	14.5	0.91	0.90	144x90
GFDL-ESM2G	-0.52	16.6	0.92	0.95	144x90
GFDL-ESM2M	**-0.15**	11.8	0.89	0.83	144x90
BCC-CSM1-1	0.19	17.9	0.91	0.98	128x64
MIROC-ESM	-1.58	34.7	0.85	0.56	128x64
MIROC-ESM-CHEM	-1.35	33.3	**0.72**	**0.54**	128x64
FGOALS-G2	-1.05	25.3	0.90	0.71	128x64

Table 2: *Number of events occurring in models during 40 year periods for historical (Hh; 1961-2000) and future climate (Fh or Ff; 2061-2100) scenarios and the multi-model weights and resultant means. Reanalysis data from 1971-2010 included for comparison*.

Model	Hh			RCP8.5_Fh			RCP8.5_Ff			W_m
Event types	# event	Type 1	Type 2	# event	Type 1	Type 2	# event	Type 1	Type 2	
NNRA1	**32**	**16**	**15**							
CCSM4	34	15	14	168	17	128	**44**	**12**	**27**	.1109
bcc-csm1-1-m	36.67	13.33	17	126	16	97	41	17	19	.0534
CNRM-CM5	**33.33**	**13.67**	**12.67**	154	33	98	**33**	**21**	**10**	.0935
HadGEM2-CC	44	20.5	17.5	136	22	99	41	17	17	.0947
inmcm4	58	23	26	166	28	107	58	14	23	.0168
NorESM1-M	58.67	23	23	162	14	131	59	19	24	.0125
GFDL-CM3	33.33	16	14.33	143	18	103	33	8	12	.2076
GFDL-ESM2G	33.67	14	13	167	35	100	35	17	13	.1059
GFDL-ESM2M	34.33	15.33	13.33	171	43	106	29	14	10	.1047
bcc-csm1-1	41.33	18.67	17.33	159	21	121	41	17	19	.0754
MIROC-ESM	28	12.67	9.67	92	2	81	33	13	15	.0578
MIROC-ESM-CHEM	31	15	8	110	6	92	29	10	18	.0595
FGOALS-g2	41.5	19.5	15.5	161	28	115	38	19	8	.0072
Multi-model weighted average	**35.6**	**15.8**	**14.2**	**147.7**	**22.8**	**104.5**	**36.3**	**14.0**	**15.6**	