On Climate Models Simulations of the Large-scale Meteorology Associated with California Heat Waves

Richard Grotjahn[1] and Yun-Young Lee[1]

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

June 2015

Revised September 2015

Re-revised November 2015

Published in Journal of Geophysical Research, Atmospheres

DOI: http://dx.doi.org/10.1002/2015JD024191

How to cite:

Grotjahn, R., and Y.-Y. Lee (2016), On climate model simulations of the large-scale meteorology associated with California heat waves, J. Geophys. Res. Atmos., **121**, 18–32, doi:10.1002/2015JD024191.

* *Corresponding author address*:

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

Corresponding author email: grotjahn@ucdavis.edu

Key Points:

California Central Valley heat waves are simulated by 14 climate models

All climate models capture both ways the heat waves develop

Models with higher horizontal resolution but lower stratospheric top tend to perform better

Abstract

Previous work considered how well the large scale meteorological patterns (LSMPs) associated with California Central Valley (CCV) heat waves are captured by a climate model. Recent work found two distinct types of LSMPs and key parcel trajectories occurring prior to heat wave onset. This study searches for those two types of heat waves in two additional reanalyses and in historical simulations by 14 climate models. The reanalyses develop both types with similar properties; their differences are used as a conservative estimate of acceptable differences between datasets. All the models develop heat waves of both types, but the models vary quite a bit in: the separation between the two types, the magnitudes of the two types, and the frequency of occurrence of the two types. The best models match a third to a half of the properties found in reanalyses. Models tend to have lower-valued projections onto the two types than reanalyses, consistent with a systematic tendency to center the hottest 850 hPa temperatures onshore instead of just offshore. Models of higher horizontal resolution tend to simulate better the two types. There is some evidence that models with a low top (with relatively poorly resolved stratosphere) also simulate the clusters better.

Index Terms (five or less)

3337 Global climate models (1626, 4928)

4313 Extreme events (1817, 3235)

3364 Synoptic-scale meteorology

4317 Precursors

Keywords (six or less): heat waves large-scale meteorological patterns simulation,

1. Introduction

This article assesses how CMIP5 (Coupled Model Intercomparison Project Phase 5) models simulate the large scale meteorological patterns (LSMPs) associated with heat waves affecting the California Central Valley (CCV). A recent review [*Grotjahn et al.*, 2015] discusses extreme statistics, dynamics, model simulations, and trends of LSMPs associated with North American heat events. That review did not find a systematic study of the LSMPs generated by climate models affecting the CCV, thereby motivating this work reported here.

Grotjahn and Faure [2008]; *Grotjahn* [2011, 2013, 2015] found that regional scale extreme heat in the CCV is linked to LSMPs that are an equivalent barotropic, nearly-stationary wave train (ridge-trough-ridge) across the North Pacific and western North America. A ridge over the region of extreme heat is expected from simple thermodynamics and is evident in various studies. The ridge being part of a larger pattern is evident in several studies, such as: *Bumbaco et al.* [2013] for Washington and Oregon heat waves; *Loikith and Broccoli* [2012], *Lau and Nath* [2012] and other studies of Midwest US events; *Chen and Konrad* [2006] and other studies of eastern US events.

Lee and Grotjahn [2015; hereafter LG2015] found that while the LSMP at the heat wave onset was similar for all extreme events, the LSMP structures leading up to onset tended to develop in two different ways. They group most events into two such clusters, with a few events in a 'mixed' group sharing traits of both clusters. In all cases, anomalously hot lower tropospheric temperatures over a region centered offshore near the NW coast of California and SW coast of Oregon is key to extreme CCV heat waves. Heat in that region is key because it migrates the 'thermal trough' in sea level pressure (SLP) to the coast and the corresponding SLP gradient opposes a cooling sea breeze [*Grotjahn*, 2011]. One cluster of events tends to form from a strong lower tropospheric hot temperature anomaly that forms in the key region only immediately before onset, with cold anomalies prevailing off the NW US coast several days before. In this cluster, air parcels tend to travel across the Pacific before sinking at the key region just prior to onset. The second cluster develops hot temperatures in the key region as a southwestward extension of a hot anomaly in southwestern Canada that exists several days prior. In this second cluster, air parcels tend to have small horizontal motion with some tending to migrate from the northeast, east, or southeast while sinking over at the key region while parcels on the north side of the key region travel from the southwest. Neither formation process excludes the other process; so a small fraction of events, the mixed type, appears to be a mixture of the two cluster types.

It should be noted that the two types here differ from the two types ('daytime' versus 'nighttime') identified by *Gershunov et al*. [2009] whose simulation by four climate models over the CCV is discussed in *Gershunov and Guirguis* [2012]. In this study we emphasize simulation of the two ways a 'daytime' CCV heat wave forms in that our events are defined based upon the daytime maximum temperatures. Another notable difference is Gershunov and Guirguis emphasize sea level pressure whereas upper air variables associated with LSMPs are emphasized here.

Land conditions at the surface or below influence regional scale heat events like the CCV heat waves studied here. Land use and land cover change (e.g. from irrigated farm to urban area) can reduce or amplify the area experiencing extreme heat [*Grossman-Clarke et al.* 2010; *Wang et al.* 2013]. Surface energy budget analysis finds that low soil moisture strongly contributes to hot extremes in some regions, such as much of Europe [*Fischer et al.* 2007; *Hirschi et al.* 2011]. However, soil moisture content is not a major factor for the CCV because most farmlands in the CCV are heavily irrigated and the surrounding region receives little or no precipitation every summer.

The CCV is geographically complex (figure 1) so local thermally-driven circulations caused by terrain slopes (mountain-valley winds) are mixed with land-sea breezes. Because hot spells are associated with easterly flows [*Grotjahn*, 2011] air moving in that direction sinks down into the Central valley, warms adiabatically, and opposes a cooling sea breeze while also lowering the subsidence inversion, these conditions all favor the formation of extreme hot spells. This complex topography is not resolved by climate models however the larger scale pattern (i.e. the LSMPs) always associated with CCV heat waves is resolved and that is why this research focuses upon those patterns.

Grotjahn [2013, 2015] examined all CCV hot spells that develop in the National Center for Atmospheric Research (NCAR) Community Climate System Model version 4 (CCSM4). The CCSM4 hot spells were compared to corresponding fields in the National Climatic Data Center (NCDC) station data as well as the NCAR/NCEP reanalysis 1 dataset [*Kalnay et al.* 1996]. (NCEP is the National Centers for Environmental Prediction.) This NCAR/NCEP reanalysis dataset will be called NNRA1 hereafter. LSMPs, surface temperatures, and a 'circulation index' that measured how similar the pattern on a given day was to the ensemble mean of extreme hot spells in two reanalysis fields were tracked in historical and future climate simulations. The LSMPs have such large scale structure (spanning the Pacific and much of North America) that they are resolved directly by the model, and while some details of the LSMP are sensitive to local processes (such as: the CCV, which is missing in the CCSM4 topography and the model resolution being too coarse to resolve sea breezes) the LSMPs in the model-generated heat waves are shown in *Grotjahn* [2013] to be very similar to the corresponding LSMPs in NNRA1 data during observed CCV heat waves. Hence, our focus on the LSMP avoids dependence upon the precise surface temperature values which are strongly affected by local processes that may be missing or poorly simulated by a model.

This work builds upon that earlier work by expanding the analysis to include 13 more models and by specific focus upon how well the models represent the LSMPs of these two clusters of heat wave events. Including more models than *Grotjahn* [2013] has obvious benefits such as broader comparison of different model formulations present in the CMIP5 datasets and models used by many more studies. Separating some of the analysis on the basis of the two cluster types is done because the dynamical development immediately prior to the heat wave onset is very different for the two clusters. Hence, it is possible (and it will be seen) that individual models capture the dynamics of each cluster in different ways. The differing model performance can provide clues to model improvements and output adjustments.

The next section describes the data and the methods for obtaining and comparing the LSMPs for the two types of clusters both in the reanalysis and the model data. The third section presents the main results. Finally there is a brief summary.

2. Data and Methods

Heat wave events considered in this study are exactly same with the events in our prior paper (LG2015). Daily maximum near-surface temperature time series were used to isolate the CCV heat wave events at 15 National Oceanic and Atmospheric Administration (NOAA) Cooperative Observer Program (COOP) stations covering the whole CCV area (Figure 1). These data are post-processed for quality control and archived at the NCDC as part of the U.S. Historical Climatology Network (for details, see Menne et al. [2009] and references therein). The methodology to isolate the events is discussed below where the extension of the methods to climate model output is described.

This study mainly uses the NNRA1 [*Kalnay et al.* 1996] for verification and comparison with model simulations of the upper-air LSMPs for each of the two heat wave clusters. NNRA1 has adequate time and space resolution: 6 hours, 2.5 degrees longitude, and 2.5 degrees latitude resolution. LSMPs evolve over several days and have half-wavelength scale greater than 20 degrees in both latitude and longitude. Boreal summer season for the CCV used extends from June through September (JJAS, 122 days). The 34 years of 1977 to 2010 are considered. NNRA1 data are available at http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html.

Two other reanalysis datasets, the NCEP/DOE AMIP-II Reanalysis (NDRA2; see *Kanamitsu et al.* [2002]) and ERA-Interim produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) described by *Dee et al.* [2011] are also analyzed. These other datasets do not extend as far back in time, hence more events are captured by the NNRA1 data emphasized. These other two reanalyses are presented for two similar reasons. First, they confirm the fidelity of two types of heat waves grouping. Second, they establish a range of variation that could be expected with different models representing the same events. All three reanalyses have the same 6 hourly temporal resolution. The spatial resolutions differ and are 2.5 degrees longitude by 2.5 degrees latitude for NDRA2 and 1 degree longitude by 1 degree latitude for ERA-interim. There are 32 summer seasons from 1979 to 2010 considered in these two reanalyses. Availability of data online is found at http://apps.ecmwf.int/datasets/data/interim-full-daily/ for ERA-Interim and http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html for NCEP-DOE.

Phase 5 of the Coupled Model Intercomparison Project (CMIP5; *Taylor et al*. [2012]) provides an ideal opportunity to assess how well state-of-the-science Atmosphere-Ocean General Circulation Models (AOGCMs) represent the two types of CCV heat waves. This study analyzes historical simulations by 14 models (listed in Table 1) which are available through the portal, the Earth System Grid - Center for Enabling Technologies (ESG-CET; at http://pcmdi9.llnl.gov/). Since the common running period of CMIP5 historical simulations is 1950 to 2005, we choose the 34 summer seasons from 1972 to 2005 for the comparison with NNRA1. Standard CMIP5 output of 6 hourly data only has 3 pressure levels (250, 500, and 850 hPa). LG2015 uses data at 700hPa and 600 hPa so an adjustment is made here to replace data at those levels with 850hPa and 500hPa data. These substitutions make the cluster mean fields different from those of LG2015 but still similar in major features. The mean fields used in defining the clusters are discussed below as are representative time sequences so that the reader need not consult LG2015 to visualize the differences in the primary features of the LSMPs for the two cluster types.

The CMIP5 data archived vary from model to model. The focus here upon the LSMPs prior to the heat wave onset requires historical simulations of upper air temperature and zonal wind to be available at 6 hourly intervals. These data are available to us from 14 models at their own horizontal resolution. The models are listed in Table 1 and ordered from higher to lower resolution. The models also have differing numbers of vertical levels. CMIP5 models are classified into either high-top (HT) or low-top (LT) based on their representation of the stratosphere and the threshold between HT and LT is at 1 hPa of their lid height [*Charlton-Perez et al.*, 2013; *Lee and Black*, 2015]. This study examines the dependency of the heat wave classification in CMIP5 simulations on both horizontal and vertical top.

From observations, 28 heat wave events were identified from 15 NCDC COOP stations (figure 1) using the criteria that at least 6 stations must surpass the 95% level for at least 3 days using normalized daily maximum surface temperature anomalies. The anomalies are formed by subtracting the long term daily mean for the calendar date. The normalization is by the station's corresponding standard deviation. Further details are in LG2015. Local surface maximum temperatures corresponding to the NCDC station dataset are not available for model simulations. Instead, daily maximum near surface temperature time series at grid points around the CCV area are used for individual model's simulations.

The grid points chosen were a partially subjective decision based on these considerations: the 'box' surrounding the grid point (and represented by it) must be mainly over the CCV, the grid point must have minimal or no oceanic influence, and sufficient points should be chosen to have a balance of information from the southern and northern parts of the CCV. The 'box' represented by each grid point used is indicated by shading in the supporting document figure. Several topographic heights are also indicated. Obviously, none of the models can resolve the low-elevation, nearly flat CCV, but instead there is a broad slope from the sea to higher elevations to the east. It is this lack of a CCV that motivates our focus on LSMP properties, as mentioned above.

Since grid resolution and origin vary between models, the number of points within the CCV for each model was manually determined and is indicated in Table1. The number of grid points within the CCV area is proportional to the horizontal resolution as expected. Hence, the criterion for minimum number of grid points exceeding the 95% threshold was adjusted for each model based on the number of CCV grid points in the model (also in Table 1). The duration criterion was unchanged: each of a minimum number of grid points must exceed its own 95% threshold for at least 3 consecutive days. Average CCV heat wave duration in most models is similar to that observed. The number of events is also similar to observed, though two models had notably more frequent hot spells; the models whose number of events are greater or less than observed by 15% are marked with asterisks in Table 1. The average duration of events, ~4 days, was very similar across the models, but four days is only one day more than the minimum number specified in the event identification criteria.

In LG2015, 28 heat wave events are assigned to groups based on the pattern dissimilarity of three target fields using the K-means clustering technique. Although model running is executed at multiple levels in sigma coordinates, the CMIP5 outputs of 6-hourly data are available at only three standard pressure levels as mentioned before. To maximize the number of models surveyed, we adjusted our cluster criteria to match the paucity of levels available for some models. This study employs three anomaly 'target fields' into the cluster analysis: -2 days zonal wind at 500 hPa, -2 days temperature at 850 hPa, and -1 day temperature at 850 hPa over the domain of 140°W-100°W, 25°N-60°N. The negative sign indicates time before event onset; hence -2 days is two days before event onset. As discussed in LG2015, the clusters differ not in the pattern at onset, but in how the pattern at onset is achieved. As stated above, the 'target fields' used by LG2015 are different, the corresponding fields used by LG2015 are: -2 days 700 hPa zonal wind, -2 days 600 hPa temperature, and -1 day 700 hPa temperature over 150°W-100°W, 20°N-60°N domain. This change was necessitated by the lack of CMIP5 data for some models at 700 and 600 hPa. As will be seen, the clustering is not sensitive to this change of fields and domain, providing some proof of the robustness of the two clusters. The change applied to NNRA1 data had negligible impact on projections of individual events onto each cluster mean and no event switched to a different cluster by using these different levels. When applying the same clustering procedure, the clustering membership of events is exactly the same as LG2015, which supports the robustness of our heat wave events classification. Once cluster membership is defined, two cluster composite means are calculated from the NNRA1 data.

In a study of regional climate model simulations of hot days during winter, Loikith et al. [2015] aggregate results from events together and then use normalized root mean squared error (RMSE) as the comparison metric. RMSE is a gross measure of difference which does not fit the purposes of this study. Here, the emphasis is upon comparing the patterns (models versus reanalyses LSMPs) while showing the range of simulated events relative to the two types of heat waves observed, so a metric suited to that comparison is used.

Spatial projection analysis is applied to sort into the two clusters the individual events in each of the two additional reanalyses and 14 CMIP5 simulations. Projection coefficients ($p_{k,j}$) are calculated for the same domain of the 'target fields' above.

$$p_{k,j} = \frac{\sum_{i=1}^{N}\left(x_{i,j}y_{i,k}\right)}{\sum_{i=1}^{N}\left(y_{i,k}\right)^2} \qquad for\ k=1,2\ and\ \ j=n$$

where k indicates a cluster; j indicates an event; i is a grid point in the domain; n is the total number of events; N is the total number of grid points in the domain; x is the variable of an individual event (j) from a dataset to be projected onto the corresponding variable y of the cluster (k) mean field calculated from NNRA1 data. Since the individual simulations have their own grid structure, the composite mean field of a variable (y) is interpolated to the model grid resolutions in advance for the projection analysis. There are 3 combinations of variable, level, and time before onset. The resultant three projection coefficients are averaged and plotted. The main difference from a pattern correlation is the equation above does not have the x squared summation in the denominator, therefore the spatial projection includes amplitude information in addition to pattern similarity.

In a scatter plot of projection coefficients, events are plotted based upon their projection coefficient for each cluster's mean. Many events have much larger coefficient onto one cluster mean than the other and are designated as belonging to the cluster for which that event has larger cluster projection. A few events are assigned as 'mixed' events when the difference of two correlation coefficients is less than 0.30461. This threshold is deduced from the scatter plot of NNRA1. Events having both negative coefficients are sorted into the 'mixed' as well since they do not correspond to either cluster. (No observed event has this property, but some model events do.) Except for these mixed events, individual events fall into the same clusters for all three reanalyses, though the event projections do change. This change between renanalyses is useful to gauge the spread of events in scatter plots of CMIP5 data.

This study examines the composite maps for the two different clusters. These maps are used both for the projection coefficient calculation and also to show synoptically (if subjectively) how the heat waves develop in each model. These maps clearly show differences of corresponding LSMPs and related dynamics. Differences are seen between the two clusters and also between reanalyses and models.

3. Heat Wave Types in Reanalyses and Model Simulations

Figure 2 shows the LSMPs for three combinations of variable, level, and time before onset using NNRA1 data. Only those grid points of each field that are shared by at least 2/3 of the members in each cluster are shaded in the plotting. The box shows the smaller region used by the clustering algorithm. The shading is based on sign counts; which examine the sign of each member of an ensemble and adds the signs of all members at each grid point. If the anomaly for a particular event is positive, the value is +1, if negative it is -1. Hence if a cluster has six members and four members have positive sign and two members have negative sign at a grid point, the sign count is +2. To facilitate comparison with sign counts for other reanalyses and models having different numbers of events, the sign count is divided by the total number of events to obtain the sign count fraction. In this example the sign count fraction is 1/3 which corresponds to 2/3 of the members having a positive sign. Figure 2 emphasizes parts of the LSMPs that are consistent among at least 2/3 of the events comprising that cluster (equivalent to a sign count fraction greater or equal to 1/3 of the events in the composite). The clusters have clearly different evolutions towards a similar pattern at onset.

In cluster one (left column of figure 2) there is often a pre-existing thermal trough in the Gulf of Alaska (see top row of figure 3) and so the zonal wind has strong westerlies to the south of strong easterlies. (LG2015 show a corresponding geopotential height trough in their figure 9.) The trough has cooler temperatures across the Gulf of Alaska into western Canada one and two days prior to onset.

In cluster two (right column of figure 2) there is a strong positive temperature anomaly over SW Canada and NW US with an associated geopotential height ridge (e.g. figure 9 in LG2015) for several days prior to CCV heat wave onset, hence strong westerlies are north of easterlies (in the anomaly fields) prior to onset. The temperature remains hot over Canada and NW US in contrast with the cool temperatures that prevail there in cluster one.

The region outlined by the boxes in figure 2 is thus chosen to highlight differences prior to the onset of the CCV heat wave. This box location is selected to emphasize where the members of one cluster are consistent and different from the members of the other cluster. Other domains were tested and the cluster membership is not sensitive to the choice of domain.

Figure 3 shows the time evolution of the ensemble mean for cluster one in NNRA1 data as well as 3 models. The NCAR CCSM4 model is shown to connect this work with *Grotjahn* [2013] and mainly because it is a popular model (a special issue of Journal of Climate is devoted to this model, described by *Gent et al.* [2011]). HadGEM2-CC is also based on a popular model frequently used in model-intercomparison studies in diverse research subjects [e.g. *Ahlström et al.*, 2012; *Bellouin et al.*, 2011; *Charlton‐Perez et al.*, 2013; *Dai*, 2013; *Gillett and Fyfe*, 2013; *Kawatani and Hamilton*, 2013; *Kawazoe and Gutowski Jr*, 2013; *Kim and Yu*, 2012; *J-Y Lee and Wang*, 2014; *Manzini et al.*, 2014; *Mishra et al.*, 2014; *Purich et al.*, 2014; *Woollings et al.*, 2014] and this version includes Troposphere, Land Surface and Hydrology, Aerosols, Ocean and Sea-ice, Terrestrial Carbon Cycle, and Ocean Biogeochemistry models. These two models have some of the better simulations of the LSMP evolutions based on qualitative comparison as well as our cluster projection results discussed below. A third model is chosen to bracket the LSMP simulation discussion since this model has the lowest resolution and consequently is one of the lower-performing models *based on the cluster projection measure* results discussed below.

The NNRA1 data in figure 3 show a pre-existing cold anomaly in the Gulf of Alaska in most members of cluster one that splits with a portion remaining over the north Pacific (amplifying a trough there, not shown) and the remaining portion migrating eastward across Canada. The temperature anomaly in the key region just offshore of northern California develops only within the last 2 days. All the models have this prior cold anomaly over western Canada, extending into the north Pacific (as selected by the clustering algorithm) but there is more consistency in this feature by the CCSM4 and HadGEM2-CC models than for FGOALS-g2. Compared to the reanalysis, the models are more consistent with this feature. All models develop the warm anomaly just prior to onset, again as captured by the cluster algorithm. A difference from NNRA1 that is common to all models is the temperature anomaly is centered more over the land in the models than it is in reanalysis data, a problem noted for CCSM4 by *Grotjahn* [2013, 2015].

In cluster two the pre-existing strong temperature anomaly in SW Canada and NW US is present in NNRA1 and all models (as expected from the projection coefficient used to populate the ensemble means shown next) (Figure 4). The north Pacific cold anomaly -- outside the domain used for the clustering -- is somewhat better captured by FGOALS-g2 than by the CCSM4 and HadGEM2-CC models at times prior to onset; while the cold anomaly is a bit too far south in FGOALS-g2 it is more consistently present in combination with the warm anomaly, suggesting a more consistent wave train in geopotential than the other two models. All models develop the warm anomaly on the southern side of the pre-existing anomaly and over the key region near the northern California coast just prior to onset. Interestingly, the anomaly for CCSM4 has a lobe *offshore* not present in the other two models (and not in the other cluster for this model) but somewhat closer to reanalyses positioning.

Figure 5 shows how each event projects onto the two NNRA1 cluster means. The panels are scatter plots of the two projection coefficients for the three reanalyses and for the 14 models examined here. Other information plotted includes: dashed lines that separate members of each cluster from each other and identify ‘mixed’ events within the dashed lines (or having negative projection onto both NNRA1 cluster means), grey triangles for the average projections for each cluster (labeled ‘centroids’), and a line showing the distance between these cluster mean projections (between centroids). Text information identifies the fraction of events in that dataset that belong to each cluster.

The reanalyses panels in figure 5 are intended to illustrate the differences that may arise from different models treating the same events. There is a tendency for ERA-interim to have larger magnitudes for the dominant projection for each member in each cluster; while the amplification is less consistent for NCEP-DOE data, the cluster means are also larger for these data. All reanalyses have about half the events classified as cluster one and about 3/8 of the events as cluster two. The 15-30% increased magnitude on average is apparent in the table 2 which has the locations of the centroids for each reanalysis. Also apparent in table 2 is that *the two clusters have little projection upon each other*, indeed the projection of one upon the other has negative sign. Hence the two clusters are independent groups. Table 2 shows that on average, mixed events in the NCEP-NCAR and ERA-interim data have comparable projection onto the two cluster means and the projection is positive, and in the case of NCEP-DOE data it is not small. The mixed projection coefficients reinforce our designation of the mixed events as hybrids combining properties of both clusters. The larger amplitude of the projections in ERA-interim and NCEP-DOE data also increase the scatter of events about the cluster mean (or centroid) which tends to be slightly larger about cluster two. The rightmost column in table 2 shows a good separation between the cluster means; the centroid separation being about three or more times the average spread of events about their respective centroids.

The model panels in figure 5 are ordered beginning with the highest resolution models (top row) progressing to the lowest resolution model at the right end of the bottom row. All the models have members within each of the two clusters. There is some tendency for the clusters to be better separated for the models with higher resolution. This trend with declining resolution causes two models of lowest resolution to have the most members in the mixed category. In addition, the lower resolution models tend to have smaller separation between the centroids than higher resolution models, though this trend is less consistent. Except for CCSM4 and bcc-csm1-1-m, all the models on the top two rows have most of their members in cluster one, similar to the reanalyses; those two exceptions have more members of cluster two. Of the lowest-resolution models on the bottom row, three have most of their members in the mixed category. Those lower-resolution models are having difficulty separating the two cluster types, though they and all other models do produce events that are clearly members of each cluster.

The models in table 2 have clusters with similar centroids as the reanalyses that are well separated, though both properties are reproduced least well by the lowest resolution model. Like the reanalyses and as might be expected from the criteria for populating this category, the average of the mixed events projects similarly onto both cluster means. The distance between centroids tends to decline with lower resolution as does the scatter about each event, so that spread around each centroid remains roughly a third (or less) the size of the distance between centroids (GFDL-CM3 is an exception, being only about half). All the models have less separation between the centroids than any of the reanalyses. Further, visual inspection of figure 5 creates the impression that for most models the separation between the clusters of the points for each of the models is less distinct than for the reanalyses. Models having smaller separation between centroids also tend to have smaller values for the location of the centroids, meaning that *those models not only have trouble separating the clusters, but tend to have weaker amplitudes* of the LSMPs than in reanalyses. Several models seem to prefer one cluster over another; MRI-ESM1, MIROC-ESM, inmcm4, FGOALS-g2 have trouble producing cluster two with the pre-existing SW Canada hot anomaly; whereas bcc-csm1-1-m, bcc-csm1-1, maybe CCSM4 and GFDL-ESM2M have some trouble producing a distinct cluster one.

In a prior study, *Lee and Black* [2013] found two primary modes of boreal winter extratropical low-frequency variability: the North Atlantic Oscillation (NAO) and Pacific-North American (PNA) patterns were unexpectedly better simulated in models with a low top (LT in Table 1) than those with higher stratospheric vertical resolution (HT in Table 1). Any improvement by LT is hard to see in Table 2. MRI-ESM1 (an HT model) and bcc-csm1-1-m (a LT model) have comparable resolution while the centroid locations, spread about them, and distance between centroids are slightly better (more similar to all 3 reanalyses) in the LT model; on the other hand the membership of the clusters is perhaps a little worse in the LT model. Similarly, HadGEM2-CC (a HT model) has a cluster membership and most properties in Table 2 that are generally similar to the reanalyses. GFDL-CM3 (a HT model) does seem to perform less well in Table 2 measures than the other GFDL models (all LT) with comparable resolution, though the qualitative impression from Figure 5 is that the two LT GFDL models seem better than the HT model.

In short, lowered resolution has a tendency to degrade the simulation of distinct heat wave clusters and possibly HT has a degrading effect as well, but there are likely other factors causing the differences seen.

Figure 6 is produced in an effort to summarize key information from both Table 2 and Figure 5. The abscissa is the ratio of events in cluster one divided by the events in cluster two. The ordinate is the distance between the two centroids. Labels for the three reanalyses (with NNRA1 marked by a multiplication symbol) and for the 14 models include other information such as resolution (lower number is higher resolution), whether it is a LT or HT model, and whether the model (grey dot) has more mixed cluster members than one or more of the clusters. The three reanalyses are grouped near the top around ratio 1.3, reflecting the excellent separation between clusters and similar membership size in each cluster type. None of the models has as great a separation between average cluster projections, though bcc-csm1-1-m, the second highest resolution model, comes closest and yet it has more mixed events than cluster one events. An arbitrary rectangle centered on the NNRA1 data is drawn to separate the half of the models that seem closer to the reanalyses from the half that are further from the reanalyses. Standard deviations of two metrics (0.86 for the ratio of events and 0.21 for the distance between two centroids) within 14 models are referenced for the rectangle range. The horizontal rectangle range is set to two values with the distance of one standard deviation (0.86) from NNRA1. The vertical range is set to two values with the distance of two standard deviations (0.43) from NNRA1. A rectangle is arbitrary and not ideal, but the intent (in combination with the dot darkness) is to isolate models within that have both a good ratio between the cluster membership and good separation between them. This rectangle shows a tendency to select models with higher resolution but there are exceptions (lower resolution models: bcc-csm1-1, GFDL-ESM2G, and GFDL-ESM2M are within the rectangle; higher resolution MRI-ESM1 is outside). There is some tendency for HT models to be outside the rectangle (the one exception being HadGEM2-CC) but there are LT models outside as well.

Comparing figure 6 with Table 2, six of the seven models in the rectangle have the seven highest fractions of their nine properties in Table 2 within the range of the reanalyses. (The exception is bcc-csm1-1, the lowest resolution model in the figure 6 rectangle. The model not in the rectangle, but in the top seven matching reanalyses ranges is MRI-ESM1, the second highest resolution model.) As above, our use of ranges of reanalyses values is intended as a rough indicator of ranges that have acceptable deviation from NNRA1 data. If anything, this acceptable range is overly conservative, since the reanalyses are covering the same events (with a few extra in NNRA1 data due to its longer duration). The bcc-csm1-1-m has seven of nine values within reanalyses ranges while GFDL-ESM2M has six. The models with the largest deviation from the reanalyses ranges tend to be those models with the three lowest resolutions; the exception is the high top model MRI-ESM1 (mixed centroid projected onto cluster one) a model outside the figure 6 rectangle (because it has so few cluster two events). The lowest resolution model has the four (out of nine) largest deviations from the reanalyses; no other model has more than one.

As stated above, different paths arrive at a similar LSMP at event onset, especially so over the key area near the NW coast of California. The discussion above examines the models' treatment of the two types of LSMPs that tend to occur prior to onset. Since the LSMPs at onset are so similar, they can be lumped together when examining that key region. *Grotjahn* [2011] develops, describes, and tests a 'circulation index' (hereafter CI). To obtain CI one calculates then combines un-normalized projections of the 850 hPa temperature and 500 hPa meridional velocity anomalies of each individual event upon corresponding ensemble averages of events in NNRA1 data over key regions. The key regions are where the NNRA1 ensemble members consistently have large anomalies in these two variables. Hence CI measures how similar a given event is to the ensemble average of events in NNRA1 data over these key areas. The stronger the event, the larger CI tends to be. Negative values of CI mean the event has anomaly patterns that have opposite sign to the ensemble average of NNRA1 data over more of the key areas than not. *Grotjahn* [2011] also discusses the physical basis for the CI, in that it samples quantities and areas related to amplifying CCV heat: by suppressing the sea breeze and lowering the subsidence inversion. *Grotjahn* [2011] found CI values to be a good proxy for how strong the larger environment is to develop a CCV heat wave.

Figure 7 shows histograms of CI values at onset for all the events in a given dataset. For reference, CI values for each of the three reanalyses during this period are calculated. Histograms provide a useful visual impression of the distribution; quantitative information relating to this figure is provided in Table 3. As with onset data, the ranges in the reanalyses of CI properties will be used as a conservative estimate of the range of acceptable model values.

With the chosen bin ranges in CI, all three reanalyses have a prominent central peak range (CI=1.0 to 1.25) with the next higher interval having slightly more members than the next lower interval. Of the better performing half of the models (based on the rectangle of figure 6) prior to onset, three of these models have distributions that match these properties of the reanalyses. (The three being CNRM-CM5, HadGEM2-CC, and bcc-csm1-1.) Two models, GFDL-ESM2G and GFDL-ESM2M have the most number of their events in the next higher range. GFDL-ESMG shows a positive skew, which is opposite to NNRA1. For GFDL-ESM2M, only the interquartile range (IQR) is unusually high; i.e. the CI ranges chosen may be exaggerating the visual difference seen in figure 7. Two models have notably different distributions: CCSM4 and bcc-csm1-1-m; each having a hint of a bimodal distribution. CCSM4 has two ranges with the most number of events; though the model has the highest average value and third quartile (Table 3) the model has two other scores (#<0.9 and skew) that are within the acceptable range. The bimodal structure of CCSM4 results from the model handling cluster two differently from cluster one. Of the two peaks in figure 7, the lower valued peak matches the peak in cluster two CI values and the upper-valued one matches the peak in cluster one CI values. However, the largest CI value is a cluster two event. So, especially with these small numbers of events, one cannot say that one cluster is generally larger than the other at onset in CCSM4. No other systematic difference was found between cluster one and cluster two in CI values. The two highest resolution models have the two highest mean values of CI among the models. Higher mean values of CI are generally consistent with larger separation between the cluster means (the ordinate of figure 6). No model has more than four distinct parameters in Table 3 that are within the acceptable range. Two of the models selected by the figure 6 rectangle (bcc-csm1-1-m and GFDL-ESM2M) have none of their seven distinct parameters within the reanalyses ranges. All but the two highest resolution models have too many events with CI <0.9 at onset; low values are consistent with the model temperature anomaly being centered too far onshore instead of in the key region. IQR and standard deviation are two ways of measuring the spread of CI values, these quantities are quite consistent in NCEP reanalyses, but less so for ERA-Interim. Models (except one: MIROC-ESM) find the IQR to be larger than one standard deviation, but how much larger varies a lot; most models are within the IQR range of the reanalyses and only one model (CNRM-CM5) is in the range of the standard deviation. The two NCEP reanalyses differ rather strongly in the skew, with ERA-Interim having an intermediate value. Four models are in the reanalyses range of skew, and three out of the four are in the rectangle of figure 6.

4. Summary

A prior study (LG2015) found two distinct ways that the large scale meteorological patterns (LSMPs) develop prior to the onset of heat waves affecting the California Central Valley (CCV). This study examines how well these two clusters of development are simulated in 14 climate models as well as how the patterns vary between three reanalyses prior to and at event onset. Another study [*Grotjahn*, 2013] examined how a single climate model captured the LSMPs at the onset of CCV heat waves; this study expands that analysis to thirteen additional models.

Individual events are identified for each model and with a few exceptions, most models develop a similar number of events during a 34 year period of historical simulation as for comparable periods in reanalyses. Pattern projection coefficients are calculated by projecting each event onto each cluster mean found for the NCEP-NCAR reanalysis (NNRA1) data. The pattern projection was calculated for times prior to onset over regions that strongly distinguish the two cluster means. The relative sizes of the projections for each cluster mean defined whether an event matched cluster one, cluster two, or was some hybrid structure labeled ‘mixed’. The two clusters are distinct, having small, negative projection upon each other.

No model outperforms other models in all (or even most) quantities tested here. *All* models produce some events belonging to each cluster. However, the proportions of events in the three categories (cluster one, cluster two, mixed) varied. All three reanalyses have slightly more members of cluster one, in which onset develops rapidly as parcel trajectories cross the north Pacific and sink off (and adjacent to) the northern California coast, a region critical for development of the CCV heat wave [*Grotjahn*, 2011; LG2015]. All three reanalyses have somewhat fewer cluster two events (than cluster one), cluster two events are characterized by southward expansion of a pre-existing, strong, hot anomaly centered in SW Canada and NW US. The reanalyses contain a small number of ‘mixed’ events that cannot be clearly classified into either cluster. The ratio of cluster type varies from model to model. For several models, the ratio between clusters one and two is comparable to those in the reanalyses. For a few models, one cluster occurs notably more often than the other. There is a systematic tendency for the cluster means in the models to have smaller projections than any of the reanalyses, indicative of a tendency for the models to have LSMPs that are either too weak or shifted horizontally. In the latter case, examination of the lower tropospheric temperature patterns finds the models tend to center the warm anomaly onshore instead of just offshore.

While no model clearly outperforms all others, there is a tendency for models with higher resolution to perform better. Generally, higher resolution models: have better fractions of their events in the two cluster categories, have larger separation between the mean projections of the members of each cluster, and have larger amplitude of their LSMPs. The larger amplitudes were seen for both individual cluster projections prior to onset and a circulation index applicable at event onset. Quantities intended to measure how well models capture the two types of clusters were used to divide the models into better performing and less well performing halves. Most of the models in the better half tended to do well at onset as well. However, a few exceptions to these conclusions are found. There is some indication that HT models do not simulate the cluster means as well as LT models but again exceptions occur. For example, the highest resolution HT model was outperformed (by the measures used here) by LT models with much lower resolution, however one HT model did make the ‘better half’ group.

Acknowledgements

This research was funded in part by NSF (U.S. National Science Foundation) Grant 1236681 and also supported by the USDA National Institute of Food and Agriculture, Hatch project CA-D-LAW-4264-H. NCEP Reanalysis and Reanalysis 2 data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/. ECMWF ERA-Interim data used in this project have been obtained from the ECMWF data server. 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 data are available from the above sources. Data used for projections are available from http://grotjahn.ucdavis.edu/EWEs and from the authors.

References

Ahlström, A., G. Schurgers, A. Arneth, and B. Smith (2012), Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections, *Environmental Research Letters*, *7*(4), 044008. doi:10.1088/1748-9326/7/4/044008

Bumbaco, K. A., K. D. Dello, and N. A. Bond (2013), History of Pacific Northwest Heat Waves: Synoptic Pattern and Trends*, *Journal of Applied Meteorology and Climatology*, *52*(7), 1618-1631, doi:10.1175/JAMC-D-12-094.1.

Charlton‐Perez, A. J., M. P. Baldwin, T. Birner, R. X. Black, A. H. Butler, N. Calvo, N. A. Davis, E. P. Gerber, N. Gillett, and S. Hardiman (2013), On the lack of stratospheric dynamical variability in low‐top versions of the CMIP5 models, *Journal of Geophysical Research: Atmospheres*, *118*(6), 2494-2505. doi: 10.1002/jgrd.50125

Chen, F., C. E. Konrad II (2006) A synoptic climatology of summertime heat and humidity in the piedmont region of North Carolina. *Journal of Applied Meteorology and Climatology*, 45(5), 674-685. doi:10.1175/JAM2345.1

Dai, A. (2013), Increasing drought under global warming in observations and models, *Nature Climate Change*, *3*(1), 52-58. doi:10.1038/nclimate1633

Dee, D., S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, and P. Bauer (2011), The ERA‐Interim reanalysis: Configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553-597. doi**:** 10.1002/qj.828

Fischer, E. M., S. Seneviratne, P. Vidale, D. Lüthi, and C. Schär, 2007: Soil moisture-atmosphere interactions during the 2003 European summer heat wave. *Journal of Climate*, **20,** 5081-5099. doi: http://dx.doi.org/10.1175/JCLI4288.1

Gershunov, A., D. Cayan, and S. Iacobellis (2009), The great 2006 heat wave over California and Nevada: Signal of an increasing trend, *Journal of Climate*, 22, 6181–6203, doi:10.1175/2009JCLI2465.1.

Gershunov, A., K. Guirguis (2012) California heat waves in the present and future. *Geophysical Research Letters*, 39, L18710, doi: 10.1029/2012GL052979

Gent, P. R., et al. (2011), The Community Climate System Model Version 4, *Journal of Climate*, *24*(19), 4973-4991, doi:10.1175/2011JCLI4083.1.

Gillett, N., and J. Fyfe (2013), Annular mode changes in the CMIP5 simulations, *Geophysical Research Letters*, *40*(6), 1189-1193. doi: 10.1002/grl.50249

Grossman-Clarke, S., J. A. Zehnder, T. Loridan, and C. S. B. Grimmond, 2010: Contribution of Land Use Changes to Near-Surface Air Temperatures during Recent Summer Extreme Heat Events in the Phoenix Metropolitan Area.

Journal of Applied Meteorology and Climatology, **49,** 1649-1664. doi: 10.1175/2010JAMC2362.1

Grotjahn, R. (2011), Identifying extreme hottest days from large scale upper air data: a pilot scheme to find California Central Valley summertime maximum surface temperatures, *Climate dynamics*, *37*(3-4), 587-604. doi: 10.1007_s00382-011-0999-z

Grotjahn, R. (2013), Ability of CCSM4 to simulate California extreme heat conditions from evaluating simulations of the associated large scale upper air pattern, *Climate dynamics*, *41*(5-6), 1187-1197. doi: 10.1007/s00382-013-1668-1

Grotjahn, R. (2015) Western North American extreme heat, associated large scale synoptic-dynamics, and performance by a climate model. *Dynamics and predictability of large-scale, high-impact weather and climate events*, edited by J. Li et al., Cambridge University Press, Cambridge, 198-209. ISBN 978-1-107-07142-1

Grotjahn, R., et al. (2015), North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends, *Climate Dynamics*, 1-34, doi:10.1007/s00382-015-2638-6.

Grotjahn, R., and G. Faure (2008), Composite predictor maps of extraordinary weather events in the Sacramento, California, Region*, *Weather and Forecasting*, *23*(3), 313-335. doi: 10.1175/2007WAF2006055.1

Hirschi, M., and Coauthors, 2011: Observational evidence for soil-moisture impact on hot extremes in southeastern Europe. *Nature Geoscience*, **4,** 17-21. doi:10.1038/ngeo1032

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, and J. Woollen (1996), The NCEP/NCAR 40-year reanalysis project, *Bulletin of the American Meteorological Society*, *77*(3), 437-471.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter (2002), NCEP–DOE AMIP-II Reanalysis (R-2), *Bulletin of the American Meteorological Society*, *83*(11), 1631-1643, doi:10.1175/BAMS-83-11-1631.

Kawatani, Y., and K. Hamilton (2013), Weakened stratospheric quasibiennial oscillation driven by increased tropical mean upwelling, *Nature*, *497*(7450), 478-481, doi:10.1038/nature12140

http://www.nature.com/nature/journal/v497/n7450/abs/nature12140.html - supplementary-information.

Kawazoe, S., and W. J. Gutowski Jr (2013), Regional, very heavy daily precipitation in CMIP5 simulations, *Journal of Hydrometeorology*, *14*(4), 1228-1242. doi: 10.1175/JHM-D-12-0112.1

Kim, S. T., and J. Y. Yu (2012), The two types of ENSO in CMIP5 models, *Geophysical Research Letters*, *39*(11). doi: 10.1029/2012GL052006

Lau, N.-C., and M. J. Nath (2012), A model study of heat waves over North America: Meteorological aspects and projections for the twenty-first century, *Journal of Climate*, *25*(14), 4761-4784. doi:10.1175/JCLI-D-11-00575.1

Lee, J.-Y., and B. Wang (2014), Future change of global monsoon in the CMIP5, *Climate Dynamics*, *42*(1-2), 101-119. doi: 10.1007/s00382-012-1564-0

Lee, Y.-Y., and R. X. Black (2013), Boreal winter low‐frequency variability in CMIP5 models, *Journal of Geophysical Research: Atmospheres*, *118*(13), 6891-6904. doi: 10.1002/jgrd.50493

Lee, Y.-Y., and R. X. Black (2015), The Structure and Dynamics of the Stratospheric Northern Annular Mode in CMIP5 Simulations, *Journal of Climate*, *28*(1), 86-107. doi: 10.1175/JCLI-D-13-00570.1

Lee, Y-Y. and R. Grotjahn (2015) California Central Valley summer heat waves form two ways. *J. Clim.*, DOI: 10.1175/JCLI-D-15-0270.1.

Li, L. J., and Coauthors (2013) The flexible global ocean atmosphere land system model: Version g2: FGOALS-g2. *Adv. Atmos. Sci.*, 30(3), 543–560, doi: 10.1007/s00376-012-2140-6.

Loikith, P. C., and A. J. Broccoli (2012), Characteristics of observed atmospheric circulation patterns associated with temperature extremes over North America, *Journal of Climate*, *25*(20), 7266-7281. doi: 10.1175/JCLI-D-11-00709.1

Loikith, P.C., D. E. Waliser, H. Lee, J. D. Neelin, B. R. Lintner, S. McGinnis, L. O. Mearns, and J. Kim (2015) Evaluation of large-scale meteorological patterns associated with temperature extremes in the NARCCAP regional climate model simulations. *Climate Dynamics*, doi: 10.1007/s00382-015-2537-xManzini, E., A. Y. Karpechko, J. Anstey, M. Baldwin, R. Black, C. Cagnazzo, N. Calvo, A. Charlton‐Perez, B. Christiansen, and P. Davini (2014), Northern winter climate change: Assessment of uncertainty in CMIP5 projections related to stratosphere‐troposphere coupling, *Journal of Geophysical Research: Atmospheres*, *119*(13), 7979-7998. doi: 10.1002/2013JD021403

Martin, G. M., N. Bellouin, W. J. Collins, I. D. Culverwell, P. R. Halloran, S. C. Hardiman, T. J. Hinton, C. D. Jones, R. E. McDonald, A. J. McLaren, F. M. O'Connor, M. J. Roberts, J. M. Rodriguez, S. Woodward, M. J. Best, M. E. Brooks, A. R. Brown, N. Butchart, C. Dearden, S. H. Derbyshire, I. Dharssi, M. Doutriaux-Boucher, J. M. Edwards, P. D. Falloon, N. Gedney, L. J. Gray, H. T. Hewitt, M. Hobson, M. R. Huddleston, J. Hughes, S. Ineson, W. J. Ingram, P. M. James, T. C. Johns, C. E. Johnson, A. Jones, C. P. Jones, M. M. Joshi, A. B. Keen, S. Liddicoat, A. P. Lock, A. V. Maidens, J. C. Manners, S. F. Milton, J. G. L. Rae, J. K. Ridley, A. Sellar, C. A. Senior, I. J. Totterdell, A. Verhoef, P. L. Vidale, and A. Wiltshire (2011) The HadGEM2 family of Met Office Unified Model climate configurations, *Geosci. Model Dev., 4*, 723-757, doi:10.5194/gmd-4-723-2011, 2011.Menne, M.J., C.N. Williams, and R.S. Vose, 2009: The United States Historical Climatology Network Monthly Temperature Data - Version 2. *Bulletin of the American Meteorological Society*, **90**, 993-1107. doi: 10.1175/2008BAMS2613.1

Mishra, V., D. Kumar, A. R. Ganguly, J. Sanjay, M. Mujumdar, R. Krishnan, and R. D. Shah (2014), Reliability of regional and global climate models to simulate precipitation extremes over India, *Journal of Geophysical Research: Atmospheres*, *119*(15), 9301-9323. doi: 10.1002/2014JD021636

Purich, A., T. Cowan, W. Cai, P. van Rensch, P. Uotila, A. Pezza, G. Boschat, and S. Perkins (2014), Atmospheric and Oceanic Conditions Associated with Southern Australian Heat Waves: A CMIP5 Analysis, *Journal of Climate*, *27*(20), 7807-7829, doi:10.1175/JCLI-D-14-00098.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bulletin of the American Meteorological Society*, *93*(4), 485-498. doi: 10.1175/BAMS-D-11-00094.1

Wang, M., X. Yan, J. Liu, and X. Zhang (2013), The contribution of urbanization to recent extreme heat events and a potential mitigation strategy in the Beijing–Tianjin–Hebei metropolitan area. *Theoretical and Applied Climatology*, **114,** 407-416. doi: 10.1007/s00704-013-0852-x

Woollings, T., B. Harvey, and G. Masato (2014), Arctic warming, atmospheric blocking and cold European winters in CMIP5 models, *Environmental Research Letters*, *9*(1), 014002. doi:10.1088/1748-9326/9/1/014002

Figure Captions

Figure 1 A map the CCV NCDC stations considered in this study. Shading indicates topographic elevation, with the scale indicated on the right edge of the plot.

Figure 2 Cluster composite means of three anomaly fields in NNRA1 data. These three combinations of variable, level, and time before onset are used to sort heat wave events of three Reanalyses and 14 CMIP5 simulations: U500 (-2day), T850 (-2day), T850 (-1day). The domain enclosed by purple box (150W-100W, 20N-60N) is used in a projection analysis. Contours are only drawn using grid points having sign count fractions with magnitude over 1/3 of cluster member numbers are plotted. Contour intervals are 2 m/s and 1K.

Figure 3 Composite means of T850 for the events of cluster one. NCEP-NCAR Reanalysis is considered as a reference. Bottom row shows composites at event onset; middle row are composites one day before onset; top row is two days before onset. Only grid points having sign count fractions with magnitude over 1/3 of the number of cluster members are colored (sign count fraction of -1/3 means two thirds of the members have the same negative sign; sign count fraction of 0.9 means that 95% of the cluster members have the same positive sign). In this cluster the strong warm temperature anomaly develops in the ‘key area’ off the N California coast only in the last day before event onset.

Figure 4 Same as Fig. 3 except for the events of cluster two. This cluster has a pre-existing strong warm temperature anomaly in SW Canada and NW US.

Figure 5 Scatter plots of the two cluster projection coefficients for each of the events in three reanalyses and 14 individual models. Projections are calculated with respect to the cluster means in the NNRA1 (‘NCEP-NCAR’) data. A dot marks each event in cluster one, a circle marks each event in cluster two, and mixed events are marked with a '+' symbol. Mixed events lie within the parallel dashed lines, whereby the projection coefficients are too similar to distinguish the event as being one cluster type or the other.

Each triangle represents the mean or ‘centroid’ of the projection coefficients in each cluster. Grey solid lines connect two triangles, which represents the distance between the two cluster mean locations. Dotted lines mark zero projection onto each cluster mean. Text labels indicate the fraction of all events in that dataset that are present in each of the three categories, ‘clust1’ refers to cluster one (e.g. figure 3); ‘clust2’ refers to cluster two (e.g. figure 4); ‘mixed’ refers to mixed events.

Figure 6 Scatter plot of frequency ratio of cluster one and two versus and distance between cluster one and two. The center of each cluster is an average of projection coefficients for all events within a cluster (as shown in Figure 5). A multiplication symbol marks NNRA1 reanalysis (‘NCEP-NCAR’) data. Black/grey dots are for other reanalyses and models. Grey dots are drawn for models when their mixture cluster does not have the fewest members. Grey dashed rectangle is drawn centered on NCEP-NCAR. Models in that rectangle are the half of the models that separate the two clusters relatively clearly. Smaller (larger) number before model name represents higher (coarser) horizontal resolution while “H” or “L” after model names represents high-top/low-top models. (Recall Table 1.)

Figure 7: Histograms of the number of events having CI (circulation index, see Grotjahn [2011]) values on the event onset dates within 0.25 ranges for the three reanalyses and each model. The interval(s) with the largest number of events in each dataset are shaded. A small triangle is placed on the abscissa to mark the average CI value at onset for each dataset. Information presented here is interpreted in concert with Table 3. Generally, the models have similar average CI values as the reanalyses. The distributions of four of the seven models in the rectangle of figure 6 (emphasizing projections prior to onset) have similar CI distributions at onset similar to the reanalyses.

Figure S1: Grid points used for identifying heat wave events in each of the 14 models. Surface elevations: 300m, 700m, 1100m, and 1500m are contoured for selected models and high resolution California topography.

Table 1. Models and NCEP-NCAR reanalysis properties and heat wave events

Model	Horizontal resolution (lon. by lat.)[a]	HT vs LT[b]	CV Grid #[c]	Min grid #[d]	Avg. Duration[e]	Events[f]
NCEP-NCAR	-	-	15	6	**4.07**	**28**
CCSM4	1:288x192	LT	4	2	3.75	32
MRI-ESM1	2:320x160	HT	5	3	3.64	*33
bcc-csm1-1-m	2:320x160	LT	5	3	4.16	31
CNRM-CM5	3:256x128	LT	3	2	3.87	30
HadGEM2-CC	4:192x144	HT	4	2	4.38	*34
inmcm4	5:180x120	LT	2	1	4.64	*45
NorESM1-M	6:144x96	LT	2	1	4.04	*50
GFDL-CM3	7:144x90	HT	3	2	3.48	29
GFDL-ESM2G	7:144x90	LT	3	2	4.14	29
GFDL-ESM2M	7:144x90	LT	3	2	4.38	29
bcc-csm1-1	8:128x64	LT	1	1	4.24	*33
MIROC-ESM	8:128x64	HT	1	1	3.97	30
MIROC-ESM-CHEM	8:128x64	HT	1	1	4.32	25
FGOALS-g2	9:128x60	LT	1	1	4.16	*37
Model Average					**4.08**	**33**

[a] grid size of data provided, [b] stratospheric representation with a high top (HT) or low top (LT) model, [c] number of grid points (or NCDC stations) associated with the CCV, [d] minimum number of grid points (stations) needed to exceed threshold for an event day, [e] average length of heat waves (each event must be ≥3 days), [f] number of events during 1977-2010 in NCEP-NCAR, 1972-2005 in models (34 summers), * Number of events that differ from observed by more than ±15%.

Table 2. Average projection coefficients between clusters; spread around and between cluster means

Table 2: Average projection coefficients between clusters; spread around and between cluster means									
Cluster Avg.	C1 centroid[a]		C2 centroid		Mixture centroid		Mean event distance from cluster centroid		Distance between two centroids
Projection Coefficient onto	C1	C2	C1	C2	C1	C2	C1	C2	
NCEP-NCAR	1.00	-0.19	-0.31	1.00	0.37	0.39	0.36	0.49	1.77
ERA-Interim	1.15	-0.20	-0.09	1.27	0.58	0.54	0.48	0.59	1.92
NCEP-DOE	1.20	-0.18	-0.11	1.17	0.70	0.50	0.44	0.56	1.88
CCSM4	1.03	-0.18	-0.04	0.84	0.49	0.56	0.40	0.51	1.48
MRI-ESM1	0.93	-0.09	-0.27	0.87	0.28	0.42	0.44	0.33	1.55
bcc-csm1-1-m	1.16	0.04	-0.09	1.08	0.48	0.49	0.43	0.52	1.62
CNRM-CM5	0.84	-0.29	-0.26	0.72	0.53	0.44	0.47	0.53	1.50
HadGEM2-CC	1.05	-0.28	0.00	0.84	0.55	0.45	0.58	0.35	1.53
inmcm4	0.90	-0.15	0.03	0.73	0.36	0.32	0.46	0.32	1.24
NorESM1-M	0.72	-0.12	-0.17	0.71	0.33	0.27	0.33	0.41	1.21
GFDL-CM3	0.80	0.20	0.07	0.89	0.37	0.41	0.49	0.46	1.00
GFDL-ESM2G	0.98	-0.08	-0.15	0.93	0.51	0.44	0.37	0.52	1.51
GFDL-ESM2M	1.10	-0.10	0.16	1.05	0.43	0.49	0.41	0.52	1.49
bcc-csm1-1	0.88	-0.13	-0.38	0.80	0.49	0.39	0.49	0.42	1.57
MIROC-ESM	0.77	-0.15	-0.07	0.78	0.32	0.25	0.42	0.41	1.25
MIROC-ESM-CHEM	0.93	-0.04	-0.06	0.78	0.58	0.51	0.54	0.39	1.28
FGOALS-g2	0.61	-0.08	-0.18	0.55	0.36	0.36	0.23	0.23	1.01

[a] ‘centroids’ = grey triangles in Figure 5

Table 3: Statistical properties of CI values at event onset (companion to Figure 7).									
Reanalysis	Average value	Value minus NNRA1	# <0.9	% <0.9	First quartile	Third quartile	IQR	Standard Deviation	Skew
NCEP-NCAR	1.14	0	2	7	1.04	1.28	0.23	0.24	-0.25
ERA-Interim	1.20	0.06	2	8	1.03	1.45	0.42	0.27	-0.19
NCEP-DOE	1.19	0.05	1	4	1.07	1.34	0.27	0.22	0.01
Model									
CCSM4	1.36	0.22	2	6	1.14	1.63	0.48	0.31	0.00
MRI-ESM1	1.32	0.18	2	6	1.15	1.55	0.40	0.32	-0.27
bcc-csm1-1-m	1.24	0.10	4	13	0.95	1.53	0.58	0.33	-0.41
CNRM-CM5	1.09	-0.05	9	30	0.89	1.26	0.38	0.26	-0.02
HadGEM2-CC	1.14	0.0	8	24	0.99	1.32	0.33	0.30	-0.06
inmcm4	0.99	-0.15	15	33	0.86	1.20	0.34	0.29	-0.39
NorESM1-M	1.02	-0.13	16	32	0.85	1.24	0.39	0.30	-0.38
GFDL-CM3	1.20	0.06	6	21	1	1.45	0.45	0.33	-0.11
GFDL-ESM2G	1.27	0.13	4	14	1.03	1.44	0.41	0.31	0.06
GFDL-ESM2M	1.22	0.08	6	21	0.91	1.50	0.59	0.35	-0.38
bcc-csm1-1	1.16	0.02	7	21	0.96	1.38	0.41	0.33	-0.54
MIROC-ESM	1.16	0.02	5	17	1.04	1.41	0.38	0.45	-1.28
MIROC-ESM-CHEM	1.20	0.06	8	32	0.88	1.47	0.59	0.43	0.43
FGOALS-g2	1.05	-0.09	11	30	0.85	1.26	0.41	0.29	-0.47

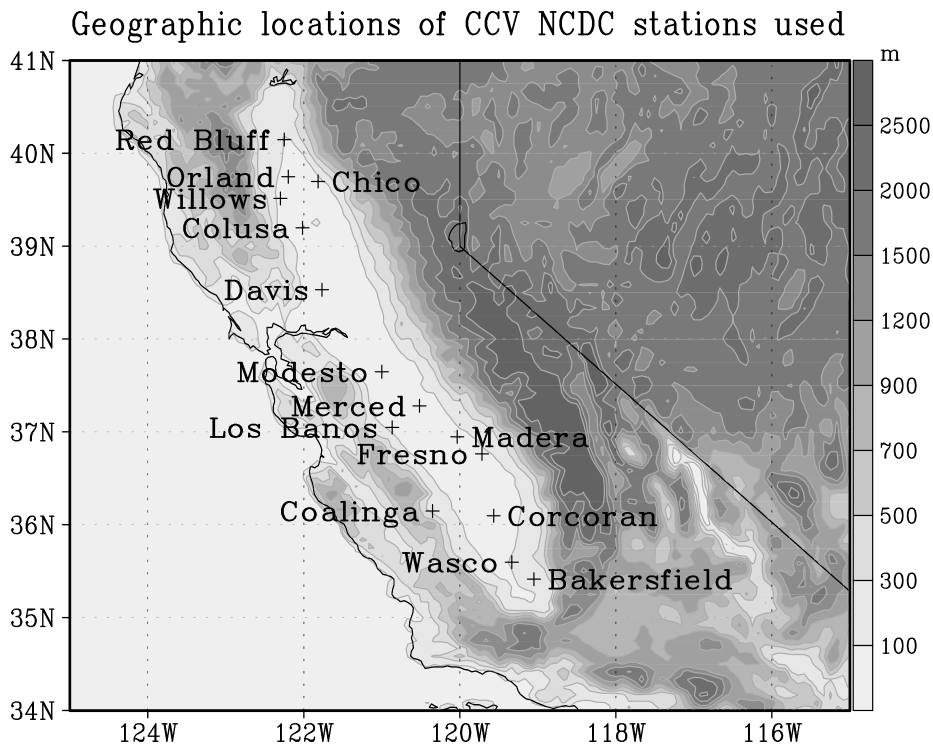

Figure 1: A map the CCV NCDC stations considered in this study. Shading indicates topographic elevation, with the scale indicated on the right edge of the plot.

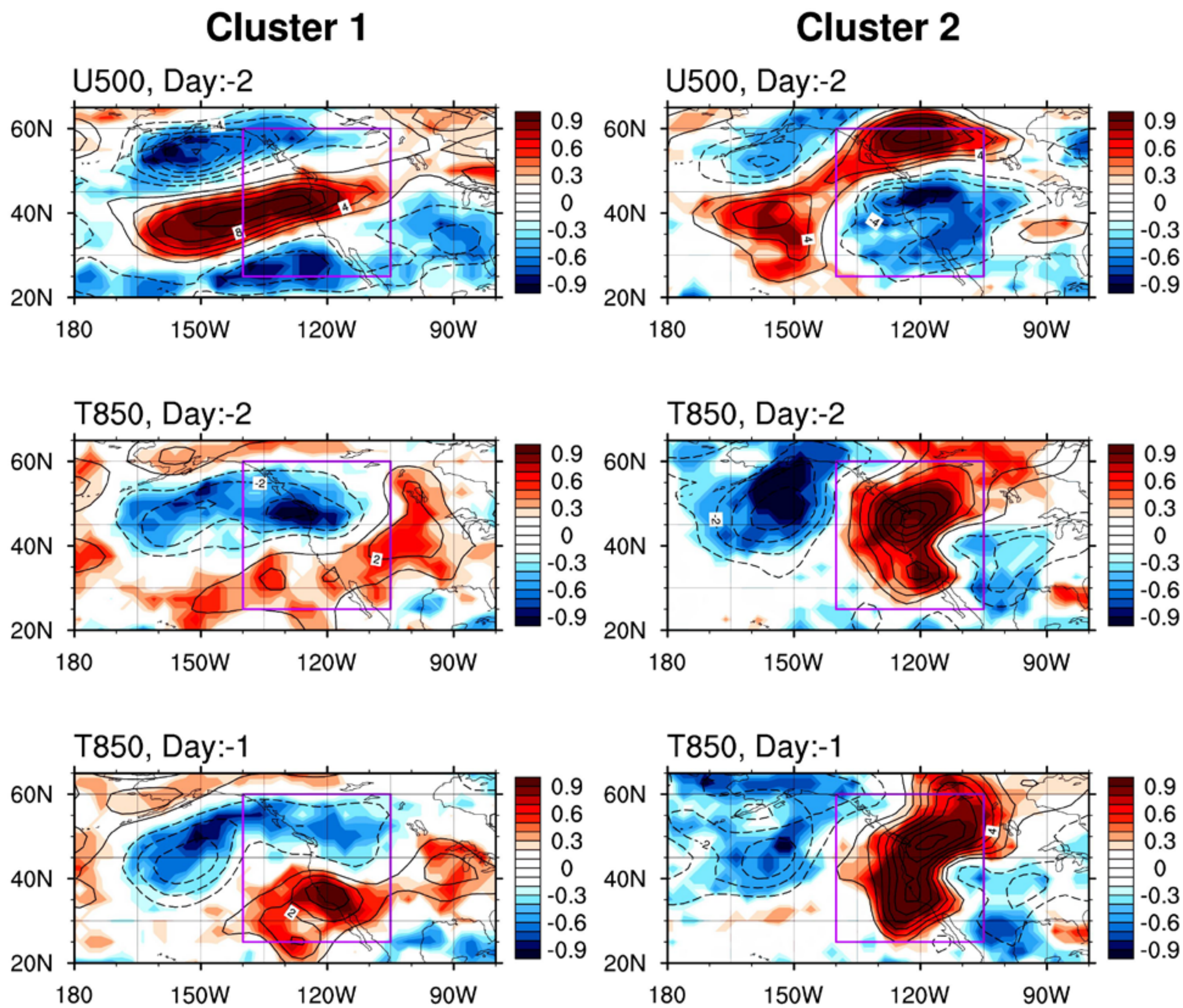

Figure 2 Cluster composite means of three anomaly fields in NNRA1 data. These three combinations of variable, level, and time before onset are used to sort heat wave events of three reanalyses and 14 CMIP5 simulations: U500 (-2day), T850 (-2day), T850 (-1day). The domain enclosed by purple box (150W-100W, 20N-60N) is used in a projection analysis. Contours are only using grid points having sign count fractions with magnitude over 1/3 of cluster member numbers. Contour intervals are 2 m/s and 1K.

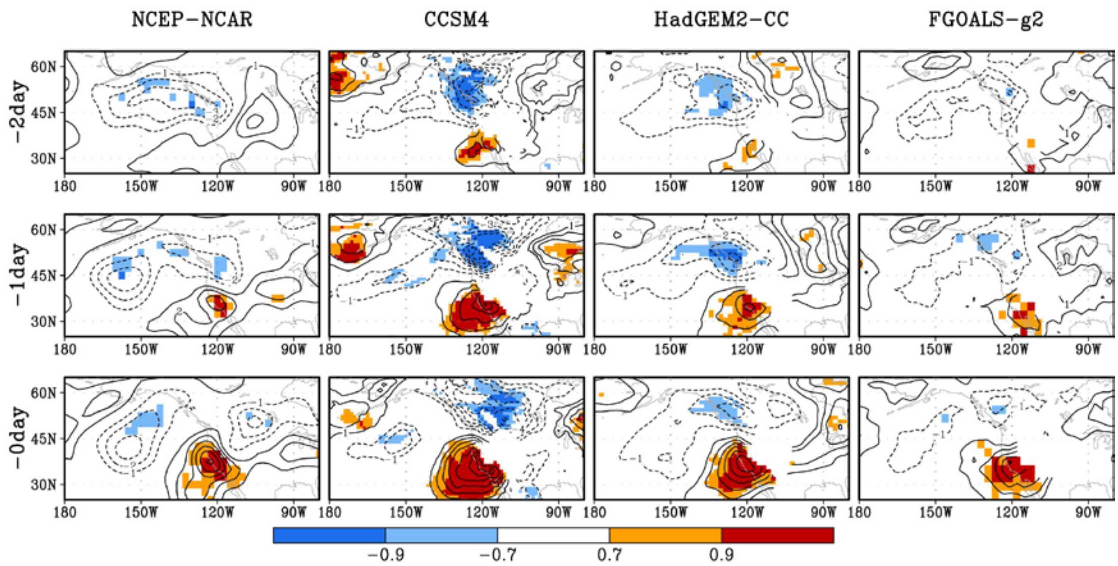

Figure 3 Composite means of T850 for the events of cluster one. NCEP-NCAR Reanalysis is considered as a reference. Bottom row shows composites at event onset; middle row are composites one day before onset; top row is two days before onset. Only grid points having sign count fractions with magnitude over 1/3 of the number of cluster members are colored (sign count fraction of -1/3 means two thirds of the members have the same negative sign; sign count fraction of 0.9 means that 95% of the cluster members have the same positive sign). In this cluster the strong warm temperature anomaly develops in the 'key area' off the N California coast only in the last day before event onset.

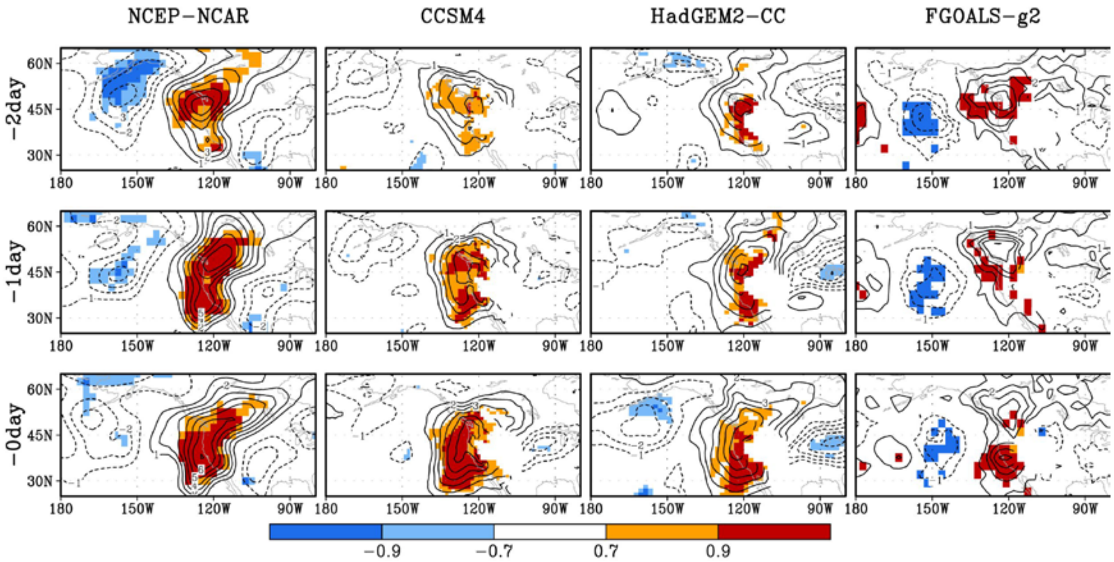

Figure 4 Same as Fig. 3 except for the events of cluster two. This cluster has a pre-existing strong warm temperature anomaly in SW Canada and NW US.

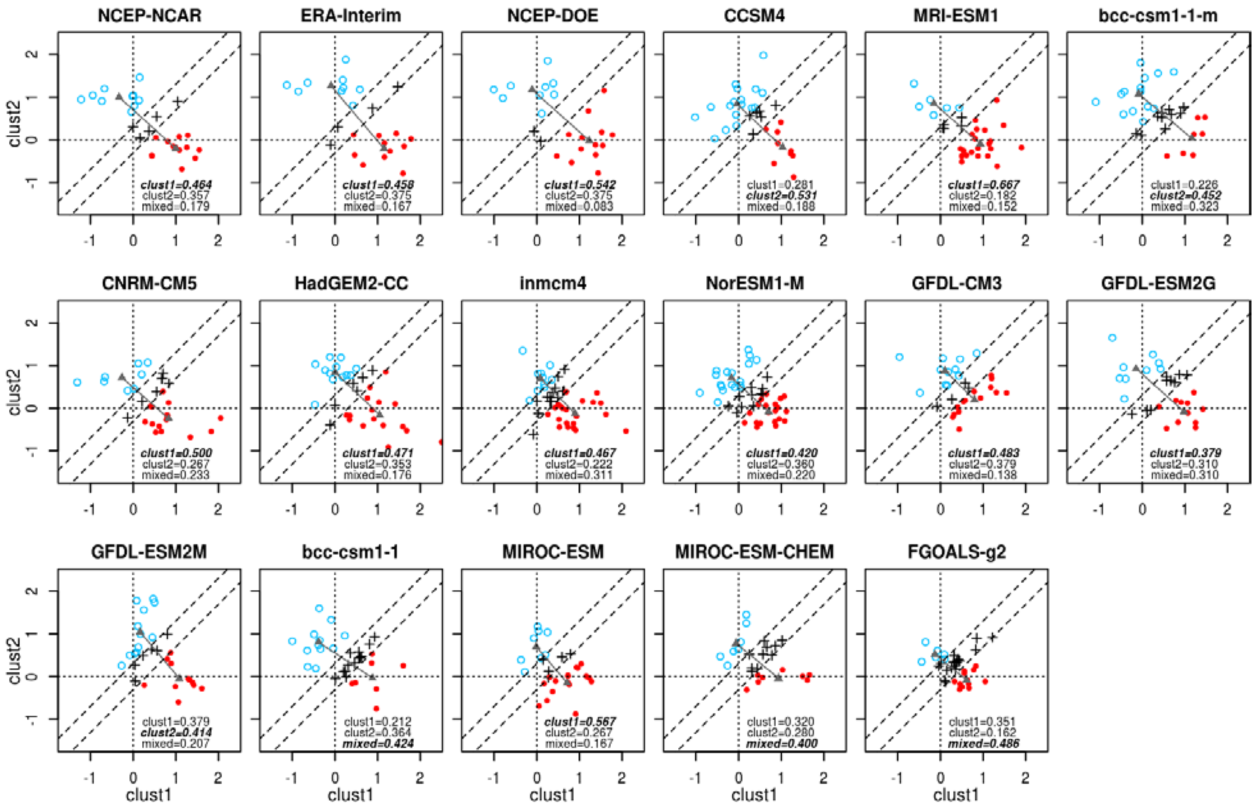

Figure 5 Scatterplots of the two cluster projection coefficients for each of the events in three reanalyses and 14 individual models. Projections are calculated with respect to the cluster means in the NNRA1 ('NCEP-NCAR') data. A dot marks each event in cluster one, a circle marks each event in cluster two, and mixed events are marked with a '+' symbol. Mixed events lie within the parallel dashed lines, whereby the projection coefficients are too similar to distinguish the event as being one cluster type or the other. Each triangle represents the mean or 'centroid' of the projection coefficients in each cluster. Grey solid lines connect two triangles, which represents the distance between the two cluster mean locations. Dotted lines mark zero projection onto each cluster mean. Text labels indicate the fraction of all events in that dataset that are present in each of the three categories, 'clust1' refers to cluster one (e.g. figure 3); 'clust2' refers to cluster two (e.g. figure 4); 'mixed' refers to mixed events.

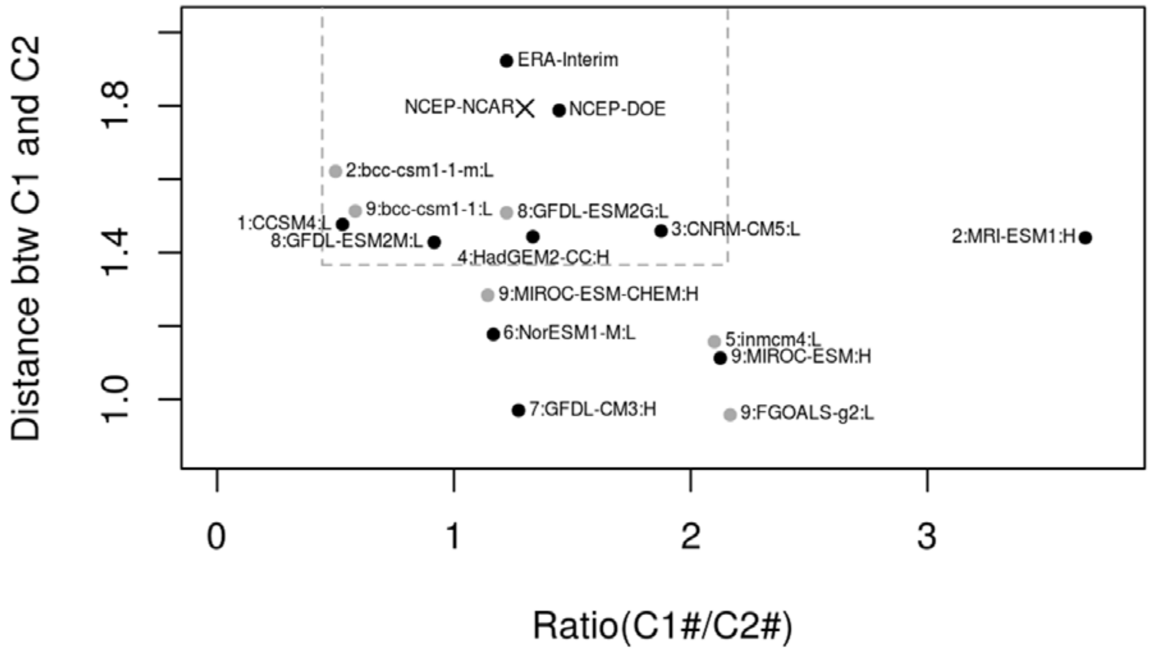

Figure 6 Scatter plot of frequency ratio of cluster one and two versus and distance between cluster one and two. The center of each cluster is an average of projection coefficients for all events within a cluster (as shown in Figure 5). A multiplication symbol marks NNRA1 reanalysis ('NCEP-NCAR') data. Black/grey dots are for other reanalyses and models. Grey dots are drawn for models when their mixture cluster does not have the fewest members. Grey dashed rectangle is drawn centered on NCEP-NCAR. Models in that rectangle are the half of the models that separate the two clusters relatively clearly. Smaller (larger) number before model name represents higher (coarser) horizontal resolution while "H" or "L" after model names represents high-top/low-top models. (Recall Table 1.)

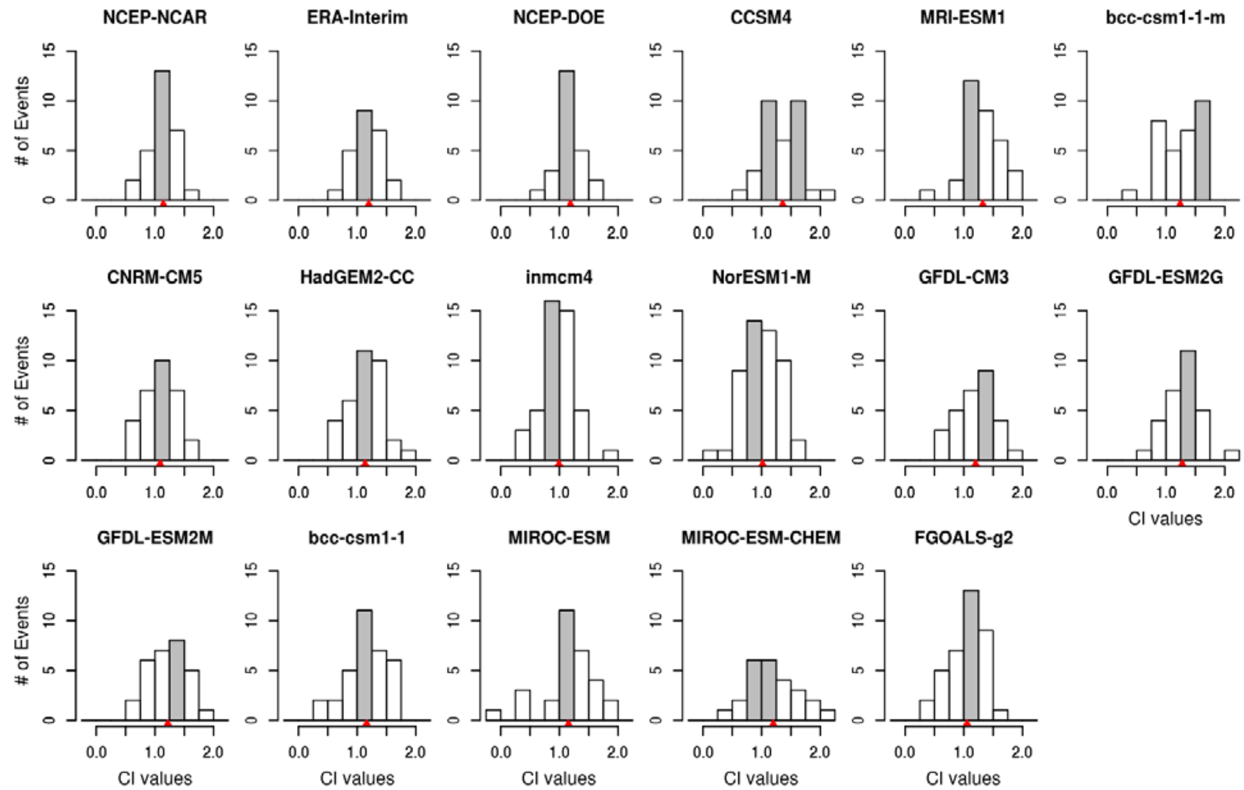

Figure 7: Histograms of the number of events having CI (circulation index, see Grotjahn [2011]) values on the event onset dates within 0.25 ranges for the three reanalyses and each model. The interval(s) with the largest number of events in each dataset are shaded. A small triangle is placed on the abscissa to mark the average CI value at onset for each dataset. Information presented here is interpreted in concert with Table 3. Generally, the models have similar average CI values as the reanalyses. The distributions of four of the seven models in the rectangle of figure 6 (emphasizing projections prior to onset) have similar CI distributions at onset similar to the reanalyses.

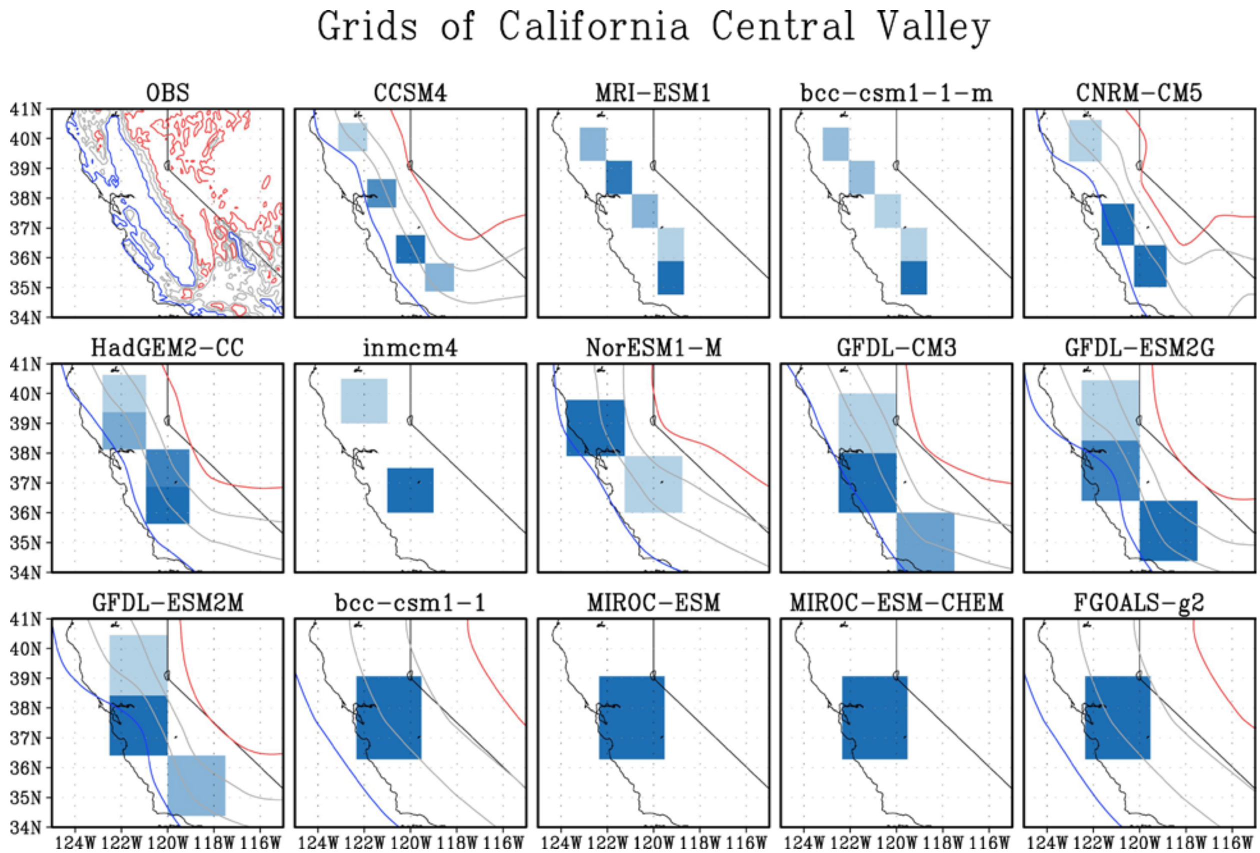

Figure S1. Grid points used for identifying heat wave events in each of the 14 models. Surface elevations: 300m, 700m, 1100m, and 1500m are contoured for selected models and high resolution California topography.