Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA

Citation	Rupp, D. E., J. T. Abatzoglou, K. C. Hegewisch, and P. W. Mote (2013), Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA, Journal of Geophysical Research: Atmospheres, 118, 10,884–10,906. doi:10.1002/jgrd.50843
DOI	10.1002/jgrd.50843
Publisher	American Geophysical Union
Version	Version of Record
Citable Link	http://hdl.handle.net/1957/47925
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA

David E. Rupp,[1] John T. Abatzoglou,[2] Katherine C. Hegewisch,[2] and Philip W. Mote[1]

Received 24 April 2013; revised 8 August 2013; accepted 13 September 2013; published 10 October 2013.

[1] Monthly temperature and precipitation data from 41 global climate models (GCMs) of the Coupled Model Intercomparison Project Phase 5 (CMIP5) were compared to observations for the 20th century, with a focus on the United States Pacific Northwest (PNW) and surrounding region. A suite of statistics, or metrics, was calculated, that included correlation and variance of mean seasonal spatial patterns, amplitude of seasonal cycle, diurnal temperature range, annual- to decadal-scale variance, long-term persistence, and regional teleconnections to El Niño Southern Oscillation (ENSO). Performance, or credibility, was assessed based on the GCMs' abilities to reproduce the observed metrics. GCMs were ranked in their credibility using two methods. The first simply treated all metrics equally. The second method considered two properties of the metrics: (1) redundancy of information (dependence) among metrics, and (2) confidence in the reliability of an individual metric for accurately ranking models. Confidence was related to how robust the estimate of the metric was to ensemble size, given that for most of the models only a small number of ensemble members (i.e., realizations of the 20th century) were available. A cursory comparison with 24 CMIP3 models revealed few differences between the two generations of models with respect to the statistics analyzed.

Citation: Rupp, D. E., J. T. Abatzoglou, K. C. Hegewisch, and P. W. Mote (2013), Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA, *J. Geophys. Res. Atmos.*, *118*, 10,884–10,906, doi:10.1002/jgrd.50843.

1. Introduction

[2] Over the last several years, climate change impacts assessments at regional and local scales have used 21st century climate projections from global climate models (GCMs) participating in the World Climate Research Programme's Coupled Model Intercomparison Project Phase 3 (CMIP3). With Phase 5 (CMIP5) now well underway, and most simulations from the new generation of GCMs already available, many impact assessments and other applications are beginning to use projections from CMIP5. The question, then, of how well the CMIP5 GCMs simulate climate at regional scales is of great interest to both researchers and resource managers.

[3] Two primary goals motivate the evaluation of GCMs. The first is the principal goal of model developers' evaluation efforts: to identify model deficiencies and potential processes responsible for the deficiencies. The second, and the one that motivates this paper, is more application driven: to provide information about model uncertainty beyond that associated with climate projections. The latter evaluation is critical as these models provide descriptions of climate change and are used in impacts modeling. There are a variety of schools of thought about the use of model evaluations for applications, ranging from "model democracy" (e.g., *Knutti* [2010]) which posits that each model simulation presents an equally valid and equally likely depiction of the future, to evaluation that is provided for informational purposes but not used to modify projections of the future (e.g., *Mote and Salathé* [2010]), to model weighting or culling, in which a model's performance at simulating 20th century climate (its "reliability" or "credibility") is taken into account numerically in future projections (e.g., *Giorgi and Mearns* [2002]). Model culling is effectively weighting with binary weights. The justification for weighting or culling models—necessarily an untestable hypothesis—is that a model that fails to reproduce aspects of the past climate will be less likely to produce a correct projection of future climate. *Mote and Salathé* [2010] found that weighting models made little difference in projected seasonal means of temperature and precipitation in the Northwest, though for other regions the same, or similar, approach made a bigger difference [*Giorgi and Mearns*, 2002; *Brekke et al.*, 2008; *Pitman and Perkins*, 2008]. Others have demonstrated that the appropriate determination of model weights is not trivial and that weighting may simply serve to increase uncertainty [*Christensen et al.*, 2010; *Weigel et al.*, 2010]. In addition to quantifying the mean projected changes, though, it is often of interest to quantify the uncertainty, and for these purposes the model evaluation may have a bigger impact simply by reducing the number of potential outlier models.

Additional supporting information may be found in the online version of this article.

[1] Oregon Climate Change Research Institute, College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon, USA.

[2] Department of Geography, University of Idaho, Moscow, Idaho, USA.

Corresponding author: D. E. Rupp, Oregon Climate Change Research Institute, College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, USA. (david.rupp@oregonstate.edu)

2169-897X/13/10.1002/jgrd.50843

Table 1. CMIP5 Models Used in This Study and Some of Their Attributes

Model	Center	Number of Ensemble Members:T/ P/ Tmin/ Tmax/	Atmospheric Resolution (Lon. × Lat.)	Vertical Levels in Atmosphere
BCC-CSM1-1	Beijing Climate Center, China Meteorological Administration	3/ 3/ 3/ 3	2.8 × 2.8	26
BCC-CSM1-1-M	Beijing Climate Center, China Meteorological Administration	3/ 3/ 3/ 3	1.12 × 1.12	26
BNU-ESM	College of Global Change and Earth System Science, Beijing Normal University, China	1/ 1/ 1/ 1	2.8 × 2.8	26
CanESM2	Canadian Centre for Climate Modeling and Analysis	5/ 5/ 5/ 5	2.8 × 2.8	35
CCSM4	National Center of Atmospheric Research, USA	6/ 6/ 6/ 6	1.25 × 0.94	26
CESM1-BGC	Community Earth System Model Contributors	1/ 1/ 1/ 1	1.25 × 0.94	26
CESM1-CAM5	Community Earth System Model Contributors	3/ 3/ 3/ 3	1.25 × 0.94	26
CESM1-FASTCHEM	Community Earth System Model Contributors	3/ 3/ 3/ 3	1.25 × 0.94	26
CESM1-WACCM	Community Earth System Model Contributors	1/ 1/ 1/ 1	2.5 × 1.89	66
CMCC-CESM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1/ 1/ 1/ 1	3.75 × 3.71	39
CMCC-CM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1/ 1/ 1/ 1	0.75 × 0.75	31
CMCC-CMS	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1/ 1/ 1/ 1	1.88 × 1.87	95
CNRM-CM5	National Centre of Meteorological Research, France	10/ 10/ 10/ 10	1.4 × 1.4	31
CNRM-CM5-2	National Centre of Meteorological Research, France	1/ 1/ 1/ 1	1.4 × 1.4	31
CSIRO-Mk3-6-0	Commonwealth Scientific and Industrial Research Organization/ Queensland Climate Change Centre of Excellence, Australia	10/ 10/ 10/ 10	1.8 × 1.8	18
EC-EARTH	EC-EARTH consortium	5/ 7/ 4/ 4	1.13 × 1.12	62
FGOALS-g2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	5/ 5/ 5/ 5	2.8 × 2.8	26
FGOALS-s2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	3/ 3/ 3/ 3	2.8 × 1.7	26
FIO-ESM	The First Institute of Oceanography, SOA, China	3/ 3/ 3/ 3	2.81 × 2.79	26
GFDL-CM3	NOAA Geophysical Fluid Dynamics Laboratory, USA	5/ 5/ 5/ 5	2.5 × 2.0	48
GFDL-ESM2G	NOAA Geophysical Fluid Dynamics Laboratory, USA	3/ 3/ 1/ 1	2.5 × 2.0	48
GFDL-ESM2M	NOAA Geophysical Fluid Dynamics Laboratory, USA	1/ 1/ 1/ 1	2.5 × 2.0	48
GISS-E2-H	NASA Goddard Institute for Space Studies, USA	4/ 4/ 4/ 4	2.5 × 2.0	40
GISS-E2-H-CC	NASA Goddard Institute for Space Studies, USA	1/ 1/ 1/ 1	2.5 × 2.0	40
GISS-E2-R	NASA Goddard Institute for Space Studies, USA	2/ 2/ 2/ 2	2.5 × 2.0	40
GISS-E2-H-CC	NASA Goddard Institute for Space Studies, USA	1/ 1/ 1/ 1	2.5 × 2.0	40
HadCM3	Met Office Hadley Center, UK	10/ 10/ 10/ 10	3.75 × 2.5	19
HadGEM2-AO	Met Office Hadley Center, UK	1/ 1/ 1/ 1	1.88 × 1.25	38
HadGEM2-CC	Met Office Hadley Center, UK	1/ 1/ 1/ 1	1.88 × 1.25	60
HadGEM2-ES	Met Office Hadley Center, UK	5/ 5/ 5/ 5	1.88 × 1.25	38
INMCM4	Institute for Numerical Mathematics, Russia	1/ 1/ 1/ 1	2.0 × 1.5	21
IPSL-CM5A-LR	Institut Pierre Simon Laplace, France	6/ 6/ 1/ 1	3.75 × 1.8	39
IPSL-CM5A-MR	Institut Pierre Simon Laplace, France	3/ 3/ 1/ 1	2.5 × 1.25	39
IPSL-CM5B-LR	Institut Pierre Simon Laplace, France	1/ 1/ 1/ 1	3.75 × 1.8	39
MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	5/ 5/ 5/ 5	1.4 × 1.4	40
MIROC-ESM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	3/ 3/ 3/ 3	2.8 × 2.8	80
MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	1/ 1/ 1/ 1	2.8 × 2.8	80
MPI-ESM-LR	Max Planck Institute for Meteorology, Germany	3/ 3/ 3/ 3	1.88 × 1.87	47
MPI-ESM-MR	Max Planck Institute for Meteorology, Germany	3/ 3/ 3/ 3	1.88 × 1.87	95
MRI-CGCM3	Meteorological Research Institute, Japan	5/ 5/ 5/ 5	1.1 × 1.1	48
NorESM1-M	Norwegian Climate Center, Norway	3/ 3/ 3/ 3	2.5 × 1.9	26

[4] Our goal here is not to cull or weight models to narrow or refine future projections, but rather to evaluate model performance in order to make informed recommendations to those who may use these model outputs. Downscaled climate data from these models will be used as inputs to impacts models, including models of forest and range dynamics, crop growth, and hydrology, and these "downstream" modelers may want to know how well these GCMs simulate particular properties of the regional climate. For those who have the capacity to run only a few scenarios, this paper may guide the selection of which GCMs to use as inputs.

[5] *Hawkins and Sutton* [2009, 2011] have nicely illustrated the contributions of three sources of uncertainty to regional- and global-scale projections, and these are addressed well in the formulation of CMIP5: uncertainties in global forcing (chiefly greenhouse gases), physical response as represented by model formulation, and internal or unforced variability. CMIP5 handles the first by the use of several different "Representative Concentration Pathways" (RCPs). The second is the primary motivation for using a large number of global models available through CMIP5 (*Mote et al.* [2011] recommend at least 10) in describing future climate or running impacts models. The third is the primary reason that many modeling centers have contributed multiple "ensemble members" to CMIP5—simulations whose boundary conditions and model formulation are the same, but which differ typically by having different initial conditions. In our model evaluation, the first source of uncertainty is irrelevant (since global forcing for the recent past is certain and well quantified) but the second and third sources of uncertainty are important

Table 2. CMIP3 Models Used in This Study and Some of Their Attributes

Model	Center	Number of Ensemble Members (T and P):	Atmospheric Resolution (Lon. × Lat.)	Vertical Levels in Atmosphere
bccr_bcm2_0	Bjerknes Centre for Climate Research	1	1.9 × 1.9	31
ccma_cgcm3_1_t47	Canadian Centre for Climate Modelling and Analysis	5	2.8 × 2.8	31
ccma_cgcm3_1_t63	Canadian Centre for Climate Modelling and Analysis	1	1.9 × 1.9	31
cnrm_cm3	Centre National de Recherches Météorologiques	1	1.9 × 1.9	45
csiro_mk3_0	Commonwealth Scientific and Industrial Research Organisation	3	1.9 × 1.9	18
csiro_mk3_5	Commonwealth Scientific and Industrial Research Organisation	4	1.9 × 1.9	18
gfdl_cm2_0	Geophysical Fluid Dynamics Laboratory	3	2.5 × 2	24
gfdl_cm2_1	Geophysical Fluid Dynamics Laboratory	3	2.5 × 2	24
giss_aom	Goddard Institute for Space Studies	2	4 × 3	12
giss_e_h	Goddard Institute for Space Studies	5	5 × 4	20
giss_e_r	Goddard Institute for Space Studies	9	5 × 4	20
iap_fgoals1_0_g	Institute of Atmospheric Physics	3	2.8 × 2.8	26
ingv_echam4	Instituto Nazionale de Geofisica e Vulcanologia	1	2.8 × 2.8	19
inmcm3_0	Institute of Numerical Mathematics	1	5 × 4	21
ipsl_cm4	Institut Pierrre Simon Laplace	1	3.75 × 2.5	19
miroc3_2_hires	Center for climate System Research	1	1.1 × 1.1	56
miroc3_2_medres	Center for climate System Research	3	2.8 × 2.8	20
miub_echo_g	Meteorological Institute University of Bonn	5	3.9 × 3.9	19
mpi_echam5	Max Planck Institute for Meteorology	4	1.9 × 1.9	31
mri_cgcm2_3_2a	Meteorological Research Institute	5	2.8 × 2.8	30
ncar_ccsm3_0	National Center for Atmospheric Research	7	1.4 × 1.4	26
ncar_pcm1	National Center for Atmospheric Research	4	2.8 × 2.8	26
ukmo_hadcm3	Hadley Centre for Climate Prediction and Research	2	3.75 × 2.75	19
ukmo_hadgem1	Hadley Centre for Climate Prediction and Research	2	1.9 × 1.3	38

and we address aspects of them in formulating our approach to model evaluation.

[6] Our objective is to evaluate the CMIP5 models in terms of their ability to recreate statistics of the 20th observed climate for a particular region. Our region of interest is the northwestern contiguous United States, also known regionally as the Pacific Northwest (PNW); therefore, this paper in one sense serves as an extension of the CMIP3 evaluation of *Mote and Salathé* [2010]. However, this study differs from theirs in not only the generation of GCMs, but we also expand the number and variety of statistics uses as evaluation metrics, give explicit consideration to the third source of uncertainty described above (i.e., internal variability), and rank the models by two methods. We also briefly compare the performance of CMIP3 and CMIP5 models.

2. Data and Methods

2.1. Data

[7] We chose to focus our evaluation on temperature and precipitation, which are both the most commonly observed variables and the most widely used in impacts modeling. While other candidate variables have also been evaluated in other papers, some (e.g., 500 mb height) are mainly used as diagnostics for understanding errors in temperature and precipitation, and for others (e.g., solar radiation) it is difficult to obtain high quality gridded data. Simulated values of near-surface temperature (T), daily minimum (Tmin) and maximum (Tmax) temperature, and precipitation rate (P) were acquired from 41 GCMs (see Table 1) of the CMIP5 "historical" experiment [*Taylor et al.*, 2012]. The historical experiment includes both natural and anthropogenic forcings for the years 1850–2005, though precise start and end dates vary by modeling group. For a given GCM, the number of members per ensemble varied from 1 to 10, differing only by initial conditions. The data were obtained at the monthly frequency, with the exception of Tmin and Tmax for three GCMs. IPSL-CM5A-LR, IPSL-CM5A-MR, and IPSL-CM5B-LR had known problems with monthly mean Tmin and Tmax, so monthly means were calculated from the acquired daily data.

[8] Monthly mean temperature and precipitation rates from 24 GCMs were also acquired from the corresponding CMIP3 historical experiment known as "20cm3" that uses natural and anthropogenic forcings for the years 1860–2000 (see Table 2). Our study did not examine daily maximum and minimum temperatures from CMIP3.

[9] Models were validated against observations; however, there is not a definitive source of observed data and differences exist among available data sets. These differences arise for many reasons: the different stations included, the different spatial resolution, the different methods of interpolation that account for topographic effects, and the different approaches to coping with nonclimatic driven changes in the climate record (e.g., station relocation, instrumentation change, urban heat island effect, and time of observation bias). Rather than pick one "best" data set, we used five gridded data sets of monthly means of the following variables: near-surface daily minimum, maximum, and average temperature, and surface precipitation rate. The data sets were (1) University of East Anglia Climatic Research Unit (CRU) TS3.10.01, 0.5° × 0.5°, 1901–2009 [*Harris et al.*, 2013], (2) Parameter-elevation Regressions on Independent Slopes Model (PRISM), 2.5′ × 2.5′, 1895–2012 [*Daly et al.*, 2008], (3) University of Delaware Air Temperature and Precipitation (UDelaware) v.3.01, 0.5° × 0.5°, 1901–2010 (Willmott and Matsuura, 2012, Terrestrial air temperature: 1900–2010 gridded monthly time series, version 3.01, http://climate.geog.udel.edu/~climate/html_pages/Global2011/README.GlobalTsT2011.html; Willmott and Matsuura, 2012, Terrestrial precipitation: 1900–2010 gridded monthly time series version, 3.02, http://climate.geog.udel.edu/~climate/html_pages/Global2011/Precip_revised_3.02/README.GlobalTsP2011.html), (4) National Center for Environmental Prediction/National Center for Atmospheric Research Reanalysis (NCEP), ~1.9° × 1.9°, 1948–2012

Table 3. Definitions of Performance Metrics and the Confidence in the Metrics for Model Ranking

Metric[a]	Confidence Category	Description
Mean-T	Highest	Mean annual temperature (T) and precipitation (P), 1960–1999
Mean-P	Highest	
DTR-*MMM*[c]	Highest	Mean diurnal temperature range, 1950–1999
SeasonAmp-T	Highest	Mean amplitude of seasonal cycle as the difference between warmest and coldest month (T), or wettest and driest month (P). Monthly precipitation calculated as percentage of mean annual total, 1960–1999.
SeasonAmp-P	Higher	
SpaceCor-*MMM*-T	Highest	Correlation of simulated with observed the mean spatial pattern, 1960–1999.
SpaceCor-*MMM*-P[c,b]	Higher	
SpaceSD-*MMM*-T	Highest	Standard deviation of the mean spatial pattern, 1960–1999. All standard deviations are normalized by the standard deviation of the observed pattern.
SpaceSD-*MMM*-P[c,b]	Higher	
TimeVar.1-T	Lower	Variance of temperature calculated at frequencies (time periods of aggregation) ranging for $N = 1$ and 8 years, 1901–1999.
TimeVar.8-T	Lowest	
TimeCV.1-P	Lower	Coefficient of variation (CV) of precipitation calculated at frequencies (time periods of aggregation) ranging for $N = 1$ and 8 water years, 1902–1999.
TimeCV.8-P	Lowest	
Trend-T	Lower	Linear trend of annual temperature and precipitation, 1901–1999.
Trend-P	Lowest	
ENSO-T	Lower	Correlation of winter temperature and precipitation with Niño3.4 index, 1901–1999.
ENSO-P	Lowest	
Hurst-T	Lowest	Hurst exponent using monthly difference anomalies (T) or fractional anomalies (P), 1901–1999.
Hurst-P	Lowest	

[a] Unless otherwise noted, metrics are average over PNW domain: 124.5°W – 110.5°W, 41.5°– 49.5°N.
[b] Expanded domain: 165°W – 100°W, 20°N – 60°N.
[c] *MMM* is the season designation: DJF, MAM, JJA, and SON.

[*Kalnay et al.*, 1996], and (5) European Centre for Medium-Range Weather Forecasts 40 Year Re-analysis (ERA40), ~2.5° × 2.5° mid-1957 to mid-2002 [*Uppala et al.*, 2005]. While CRU, PRISM, and UDelaware are data sets based on surface station observations, NCEP and ERA40 are reanalysis data sets based on a numerical model of the atmosphere that assimilates observations to update model states. Some consider reanalysis data sets to be a fairer benchmark with which to evaluate GCMs [*Mote and Salathé*, 2010].

[10] It is known that NCEP contains a spurious pattern in the winter precipitation fields at high latitudes, most notable poleward of about 50° [*Sheffield et al.*, 2004]. Though adjustments have been made to NCEP data sets to improve the precipitation fields [*Sheffield et al.*, 2004, 2006], these adjustments were made only over land. Moreover, these adjusted data sets rely heavily on CRU in their correction procedure, and we already included CRU in our analyses. Therefore, we chose to use NCEP as is, while keeping in mind the problems with its representation of precipitation.

[11] CRU, UDelaware, NCEP, ERA40, and CMIP data sets were regridded to a common resolution of 1° × 1° using an inverse-distance-weighting interpolation algorithm. PRISM data sets were regridded by averaging all native cells within the coarser 1° × 1° cell. Grid cell centers were located on the whole degree.

2.2. Performance Metrics

[12] A variety of metrics have been proposed for evaluating climate models, some of which summarize features of the climatological mean state, while some others summarize temporal variability. Common metrics include the correlation between the modeled and observed climatological mean global or regional field, and the second spatial moment of these same fields (e.g., *Gleckler et al.* [2008]). Other metrics include the statistical moments of the time series averaged over a given area; typically, the first or second moments are utilized, though a small number of studies include the third moment (e.g., *Brekke et al.* [2008]) or effectively all moments through the use of the entire probability density function (pdf) (e.g., *Perkins et al.* [2007]). Another means of evaluating models is to quantify how well they simulate internal modes of climate variability, such as the El Niño Southern Oscillation (ENSO), and how well they reproduce regional teleconnections of these modes [*Joseph and Nigam*, 2006; *Brekke et al.*, 2008; *Pierce et al.*, 2009; *Mo et al.*, 2012]. Some other metrics are long-term linear trends [*Brekke et al.*, 2008; *Mote and Salathé*, 2010] and the amplitude and phase of the intraannual cycle [*Brekke et al.*, 2008; *Pierce et al.*, 2009].

[13] Lacking any established standard methodology for evaluating climate models at a regional scale [*Gleckler et al.*, 2008], we proceeded in the spirit of *Brekke et al.* [2008], and selected a number of metrics that consider both properties of the regionally averaged time series and larger-scale patterns having regional influence. Metrics were selected on the basis of having theoretical merits as well as being relevant for impacts modeling. For temperature and precipitation, we examined the following metrics to assess the performance of the GCMs for the PNW (see also Table 3 for details):

[14] 1. Climatological mean of annual value (Mean).

[15] 2. Mean seasonal amplitude (SeasonAmp).

[16] 3. Spatial standard deviation (SpaceSD) of the climatological mean field, by season.

[17] 4. Spatial correlation (SpaceCor) of the observed to modeled climatological mean fields, by season.

[18] 5. Linear time trend of annual values (Trend).

[19] 6. Time series variance (TimeVar) for temperature and coefficient of variation (TimeCV) for precipitation: calculated at frequencies ranging from 1 to 10 years.

[20] 7. Persistence (Hurst) measured using the Hurst exponent.

[21] 8. Strength of ENSO teleconnection (ENSO) in winter.

[22] Also, but for temperature only, we calculated one additional metric:

[23] 9. Mean diurnal temperature range (DTR), by season.

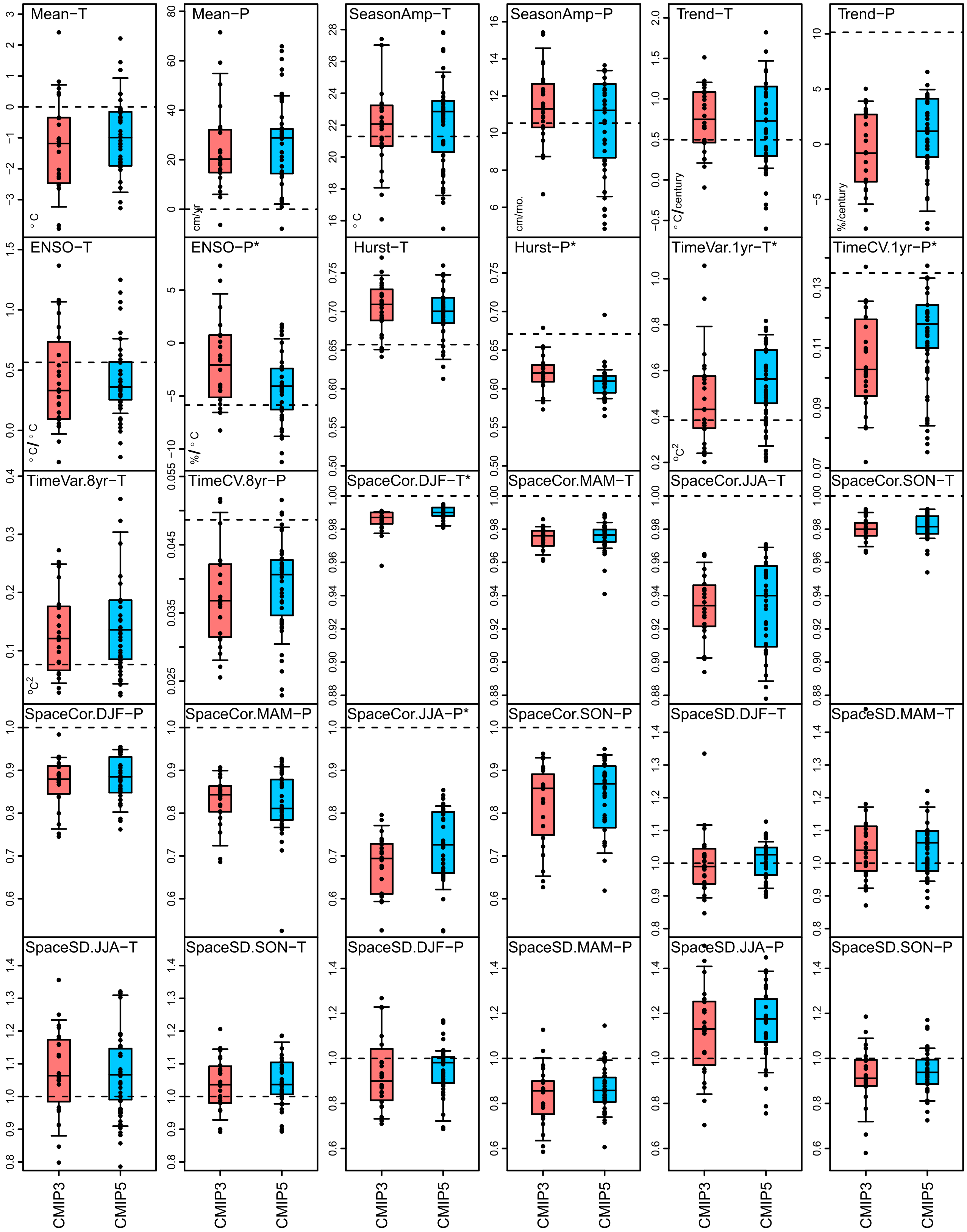

Figure 1. Metrics for 25 CMIP3 and 41 CMIP5 GCMs. Points show the mean of the ensemble of values from each GCM, the box-whisker plots give the 5th, 25th, 50th, 75th, and 95th percentiles, and the horizontal dashed lines are the observed value (or mean where more than one observation data set was used). Metric names followed by "*" indicate a statistically significant difference in the mean between CMIP3 and CMIP5 (p-value < 0.1).

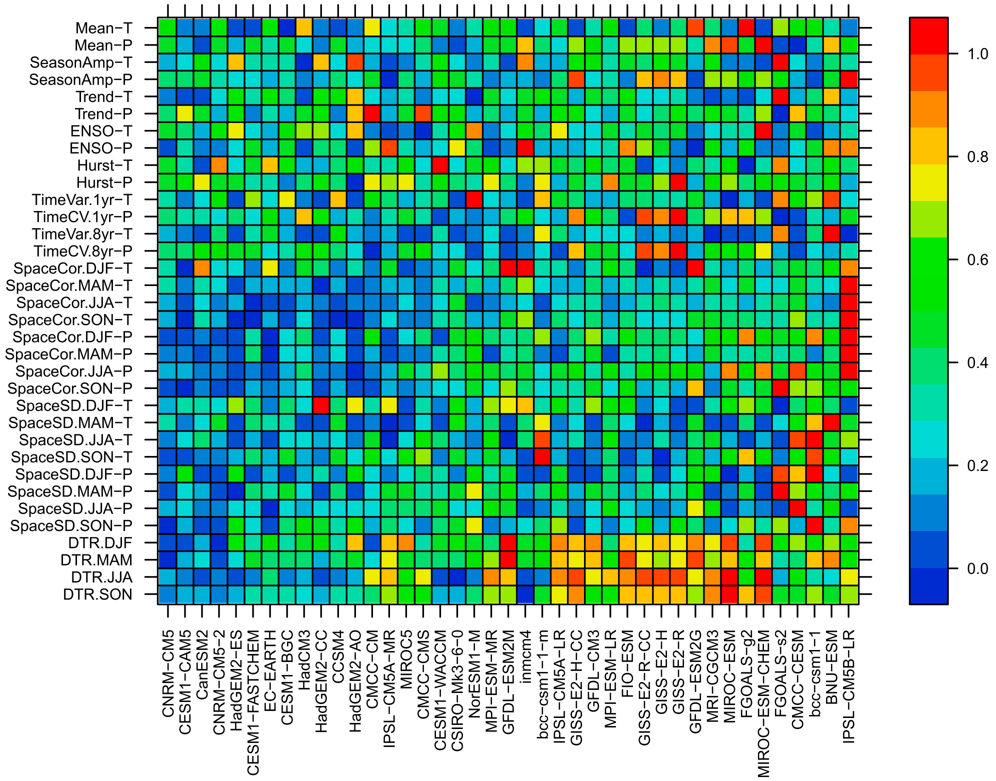

Figure 2. Relative error of the ensemble mean of each metric for each CMIP5 GCM. Models are ordered from least (left) to most (right) total relative error, where total relative error is the sum of relative errors from all metrics.

[24] The previous evaluation of CMIP3 GCMs for the PNW by *Mote and Salathé* [2010] considered only features of climatologic means and long-term trends of annual values (metrics 1, 3–5). In this study we expand our evaluation to include other properties of the climate. For one, interannual to decadal variability in mean annual temperature and water year (October–September) precipitation were examined to assess models' ability to represent the magnitude of natural variability at 1–10 year time scales irrespective of factors and processes that drive that variability. We quantified variability with the standard deviation for temperature and the coefficient of variation (CV) for precipitation. Both metrics are justifiable for time series with a standard Gaussian distribution, and may not accurately represent the true distribution.

[25] Furthermore, we characterized persistency in the monthly anomaly time series, with the mean seasonal cycle removed. The chosen persistence metric is a dimensionless scaling exponent *H* first described by *Hurst* [1951] in his

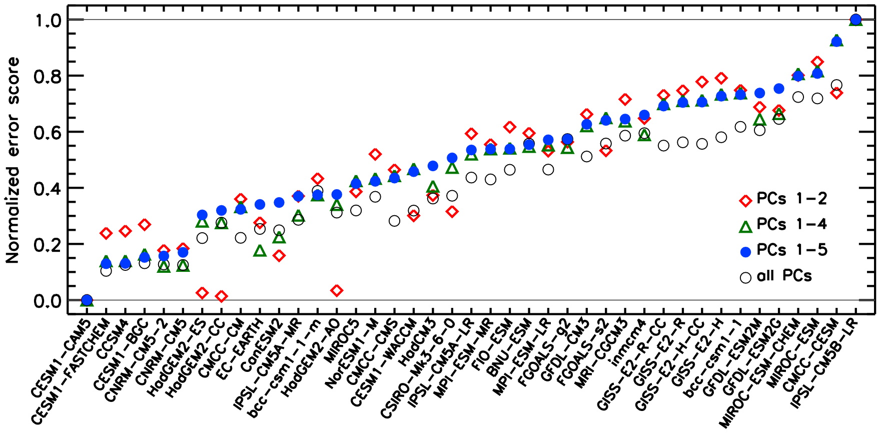

Figure 3. Forty-one CMIP5 GCMs ranked according to normalized error score from empirical orthogonal function (EOF) analysis of 18 performance metrics. Ranking is based on the first five principal components (filled blue circles). The open symbols show the models' error scores using the first two, four, and all principal components (PCs). The best scoring model has a normalized error score of 0.

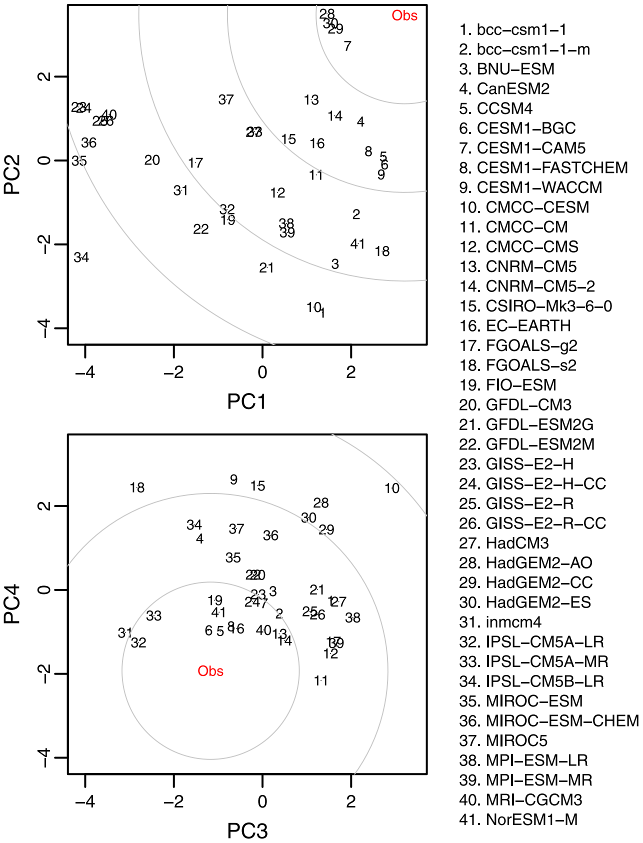

Figure 4. Loadings of the first four principal components (PC1, PC2, PC3, PC4) from EOF analysis of 18 evaluation metrics (includes diurnal temperature range) and 41 CMIP5 GCMs. "Obs" indicates the observation data set.

study of river flows. The Hurst exponent takes on value between 0 and 1, where $H > 0.5$ signifies long-term positive autocorrelation, $H = 0.5$ signifies no autocorrelation, and $H < 0.5$ signifies anticorrelation. We estimated the Hurst exponent with the rescaled range method [*Hurst*, 1951]. See Appendix B for details.

[26] Finally, ENSO teleconnections are important for characterizing the relationships between sea surface temperatures (SSTs) and seasonal climate over the PNW. Note that we did not evaluate a model's ability to simulate ENSO itself, which is beyond the scope of this study and is already the subject of recent papers [*Kim and Yu*, 2012; *Bellenger et al.*, 2013]. Previous studies have shown that CMIP3 models generally do well at reproducing its mean spatial pattern [*Pierce et al.*, 2009] and temporal variance [*Joseph and Nigam*, 2006; *Mo et al.*, 2012]; however, accurate modeling of the ENSO evolution has been mixed [*Joseph and Nigam*, 2006]. As we expect much ongoing and future examination of ENSO within the CMIP5 models by the research community, we have excluded any such analysis here. Early results do indicate that CMIP5 models perform better at simulating some characteristics of ENSO [*Kim and Yu*, 2012; *Bellenger et al.*, 2013].

[27] ENSO teleconnections to the PNW were quantified by linearly regressing winter (January–February–March; JFM) average temperature and total precipitation against the Niño3.4 index averaged over the months of November through March (NDJFM). The Niño3.4 index is an average of sea surface temperature (SST) in the region bounded by 120°W–170°W and 5°S–5°N. For the observations, we used the Niño3.4 index derived from the HadISST1 global sea surface temperature data set [*Rayner et al.*, 2003]. For each CMIP5 simulation, we calculated the Niño3.4 index from the simulated near-surface air temperature (*tas*) instead of SST. This was to avoid downloading and processing the corresponding SST data sets for each simulation, not all of which were available at the time of writing. The use of near-surface air temperature (T) instead of SST should have only a very minor effect on the slope of the linear regression because monthly T follows SST closely in the Niño3.4 domain. Using, for example, the four-member ensemble of HadGEM2-ES to calculate the Niño3.4 index using both T and SST, we find the two Niño3.4 series to be very strongly correlated ($r = 0.99$) with nearly identical standard deviation (the ratio of the standard deviation using T to the standard deviation using SST is 0.97).

[28] We evaluated most metrics over the PNW, defined here as the area bounded in longitude by 124.5° and 110.5° W, and in latitude by 41.5° and 49.5°N. This domain covers the states of Oregon, Washington, Idaho, western Montana, and small slices of adjacent states and British Columbia. However, because (1) the spatial resolution of the models is such that they represent the PNW with as few as 24 grid points (4 × 6) (which gives little spatial detail over the region), and (2) the climate of the PNW is driven by larger-scale oceanic and atmospheric patterns that we want to be faithfully simulated, the spatial variance and correlation metrics were examined over a large domain (165°W – 100°W, 20°N – 60°N). This expanded domain covers a large portion of western North America and the eastern Pacific Ocean and therefore includes a part of the Pacific Ocean with regionally relevant ocean circulation features such as the Alaskan Gyre and California Current.

[29] We calculated several metrics (Mean, SeasonAmp, SpaceSD, SpaceCor) over the latter four decades of the 20th century (1960–1999), and DTR over 1950–1999, in order

Table 4. Loadings by Metric of the Leading Five Empirical Orthogonal Functions (EOFs)[a]

Metric	EOF 1	EOF 2	EOF 3	EOF 4	EOF 5
Mean-T	0.04	0.15	***−0.42***	***0.42***	***−0.42***
Mean-P	***−0.31***	0.10	−0.04	−0.23	0.07
SeasonAmp-T	***0.31***	0.10	0.04	***0.39***	0.07
SeasonAmp-P	***0.37***	−0.06	−0.04	0.19	0.14
Trend-T	0.13	***−0.33***	***−0.30***	−0.23	0.05
ENSO-T	0.07	−0.23	***−0.40***	***−0.31***	−0.10
TimeVar.1-T	***0.28***	***−0.32***	−0.16	−0.04	−0.16
TimeCV.1-P	***0.34***	−0.17	−0.14	0.06	−0.15
DTR-DJF	***0.26***	0.10	0.05	***−0.31***	−0.07
DTR-JJA	***0.34***	0.13	−0.10	0.15	0.09
SpaceCor-DJF-T	0.03	***0.27***	0.07	−0.17	***−0.78***
SpaceCor-JJA-T	0.26	0.16	0.07	−0.41	−0.03
SpaceCor-DJF-P	0.19	***0.36***	−0.11	−0.05	0.17
SpaceCor-JJA-P	0.24	***0.35***	−0.07	***−0.28***	0.18
SpaceSD-DJF-T	0.21	0.04	***0.45***	−0.06	−0.18
SpaceSD-JJA-T	0.24	***−0.28***	***0.33***	−0.06	−0.05
SpaceSD-DJF-P	−0.04	***0.45***	−0.06	0.01	0.05
SpaceSD-JJA-P	0.05	−0.18	***0.42***	0.14	−0.12

[a] Strongest loadings (absolute values ≥ 0.260) are in bold italics.

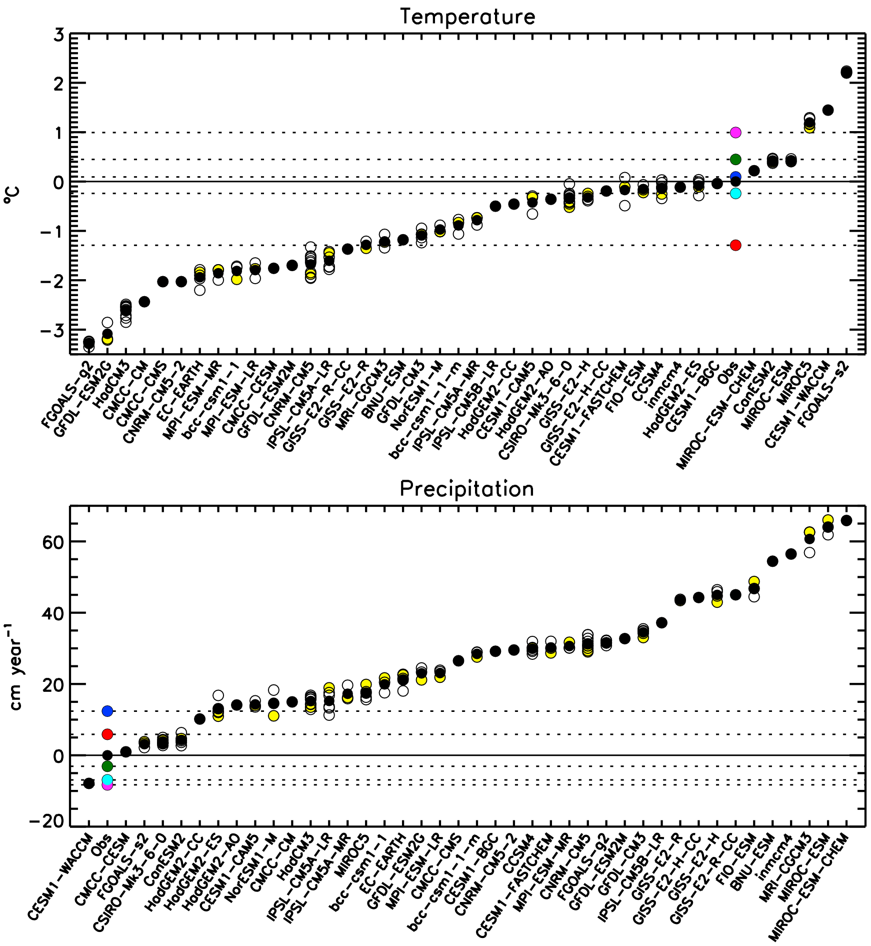

Figure A1. PNW mean annual temperature and precipitation bias for 41 CMIP5 GCMs. For each GCM, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from NCEP (red), ERA40 (magenta), CRU (dark green), PRISM (blue), UDelaware (Cyan), and average of observations (black).

to include the shorter NCEP and ERA40 data sets in the analysis. However, those metrics that are more sensitive to record length (i.e., those that do not simply describe the mean state of the time series) were calculated over the 20th century (1901–1999) and consequently only for CRU, PRISM, and UDelaware.

[30] In addition to calculating each metric for each ensemble member of each model, we also calculated a "multimodel mean" value. Given that models have different numbers of ensemble members, those with larger ensembles will generally give better estimates of a particular statistic than those models with smaller ensembles. However, for simplicity, we gave each model equal weight when calculating a mean. We also tested weighting each model by the square root of the number of members in a model's ensemble (inspired by the definition of standard error), but found it made little difference with respect to our conclusions.

2.3. Model Ranking by Overall Performance

[31] While a large number of metrics helps to elucidate the different strengths and weaknesses of models, it also presents a challenge for selecting a subset of more credible models, for two reasons. For one, some metrics may be more important than others for a particular application, and the rankings may depend on which set of metrics are selected (e.g., *Santer et al.* [2009]). For another, there may be redundancy in the metrics, given that not all are independent, either statistically or physically. In either situation, treating all metrics equally might be inadvisable.

[32] We applied two methods for ranking the models. The first simply treated all metrics equally, while the second did not. We describe both methods below.

[33] The first method included all performance metrics and assigned equal weight to each metric. For a given model *i* and metric *j*, we first defined an error $E_{i,j}$ as

$$E_{i,j} = \left|x_{obs,j} - x_{i,j}\right| \quad (1)$$

where x_{obs} and x_i are the observed metric and simulated ensemble mean metric, respectively. Application of equation

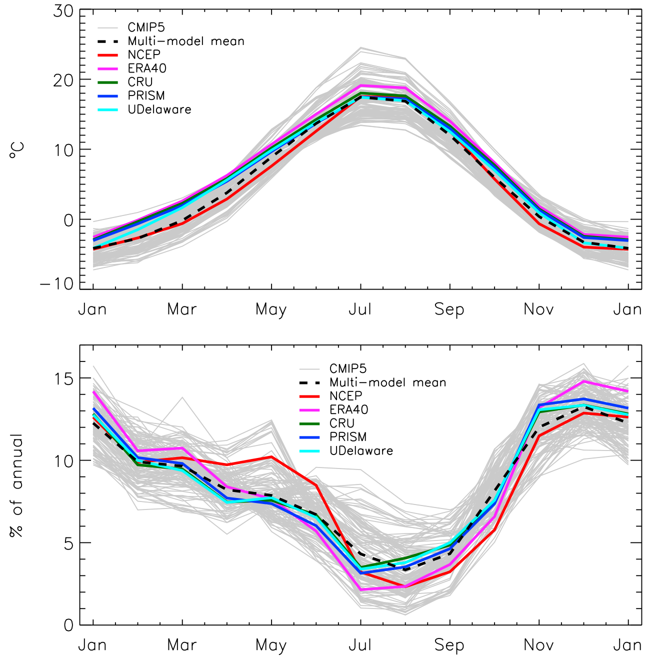

Figure A2. Mean seasonal cycle of temperature (upper panel) and relative precipitation (lower panel) averaged over the PNW. Monthly means calculated from gridded observation data sets (NCEP, ERA40, CRU, PRISM, UDelaware) and from all ensemble members from 41 CMIP5 GCMs.

(1) included correlations (where x_{obs} necessarily equaled 1). Furthermore, we defined a relative error $E^*_{i,j}$ as

$$E^*_{i,j} = \frac{E_{i,j} - \min\left(E_{i,j}\right)}{\max\left(E_{i,j}\right) - \min\left(E_{i,j}\right)} \qquad (2)$$

and then summed the relative error across all m metrics:

$$E^*_{i,\,tot} = \sum_{j=1}^{m} E^*_{i,j} \qquad (3)$$

to get the total relative error $E^*_{i,tot}$ per model. Ordering the models by their respective total relative error determined the ranking.

[34] The second, more considered approach to ranking took into account redundancy in information among metrics and in the confidence in the rankings of the individual metrics. To address the latter, we first excluded those metrics that were identified as not being robust. This exclusion of metrics is described in detail in section 3.2. Also, so as not to so heavily weight those metrics calculated for each of four seasons, we used only the winter (DJF) and summer (JJA) values.

[35] To address the matter of information redundancy, we conducted an empirical orthogonal function (EOF) analysis on the remaining metrics following *Pierce et al.* [2009]. This allowed us to reduce the large number of metrics, some of which covary and others of which add little information, down to a reduced number of orthogonal and more consequential metrics. The EOF analysis was done on the values x, which include both observed (x_{obs}) and simulated (x_i) values for each metric. In other words, the observations were treated like another model. Note that for the EOF analysis, we normalized the values by the mean and standard deviation for a particular metric. North's "rule of thumb" [*North et al.*, 1982] was used to objectively determine which EOFs were statistically distinct; this provided the basis for selecting a relatively small number of leading EOFs to use as the final performance measures.

[36] While the leading EOFs will provide a greatly reduced number of metrics, to arrive at a single ranking of overall model performance, we simply calculated the Euclidean distance from the observations to each model in EOF space across all dimensions of the leading EOFs. We used this distance as the overall error score per GCM, and normalized it to range from 0 (least error) to 1 (most error).

3. Results and Discussion

3.1. CMIP3 and CMIP5

[37] A detailed discussion of the skill of the CMIP5 models with regards to each performance metric is given in Appendix A. Here we provide merely a cursory comparison of the performance of the models in CMIP3 and CMIP5.

[38] Overall, we found few outstanding differences between the CMIP3 and CMIP5 multimodel ensembles across 30 performance metrics (Figure 1). We used a two-tailed Wilcoxon rank sum test to compare CMIP3 and CMIP5 multimodel means. We found that SpaceCor.DJF-T, Hurst-P, TimeCV.1yr-P, SpaceCor.JJA-P, TimeVar.1yr-T, and ENSO-P, in order, showed differences at a significance level of 0.9.

[39] Based on the above comparison of means and visual inspection of the distributions, the most notable change from CMIP3 to CMIP5 was the increases in temporal variability of both precipitation and temperature at the annual scale. Also of note was the apparently stronger response of PNW precipitation to ENSO in CMIP5, indicating an improvement in PNW teleconnections to ENSO. This result is consistent with the improvement in CMIP5, reported by *Polade et al.* [2013], in reproducing the observed covariance of the main mode of SST variability in the North Pacific (which has contributions from both ENSO and the Pacific Decadal Oscillation) with the main mode of precipitation variability over North America. In contrast, the change in precipitation persistence as represented by the Hurst exponent was actually in the direction away from the observed value, but only slightly: the CMIP3 and CMIP5 multimodel mean Hurst exponents were still effectively similar (0.62 and 0.61, respectively).

[40] A more extensive comparison of CMIP3 and CMIP5 would quantify differences in the spread and shape of the distributions, not simply changes in the mean. Additionally, one could isolate those models that have been improved, changed, and/or enhanced between CMIP3 and CMIP5 from those that are new arrivals in CMIP5 [e.g., *Knutti et al.*, 2013; *Polade et al.*, 2013]. We leave such analyses to future studies.

3.2. Intramodel and Intermodel Variability

[41] Internal variability may hinder the evaluation of model performance for some metrics. The importance of internal variability is exemplified by intramodel variability for a given performance metric. Given an adequate ensemble size for each model, a performance metric may be estimated with a high degree of confidence (i.e., low standard error) and

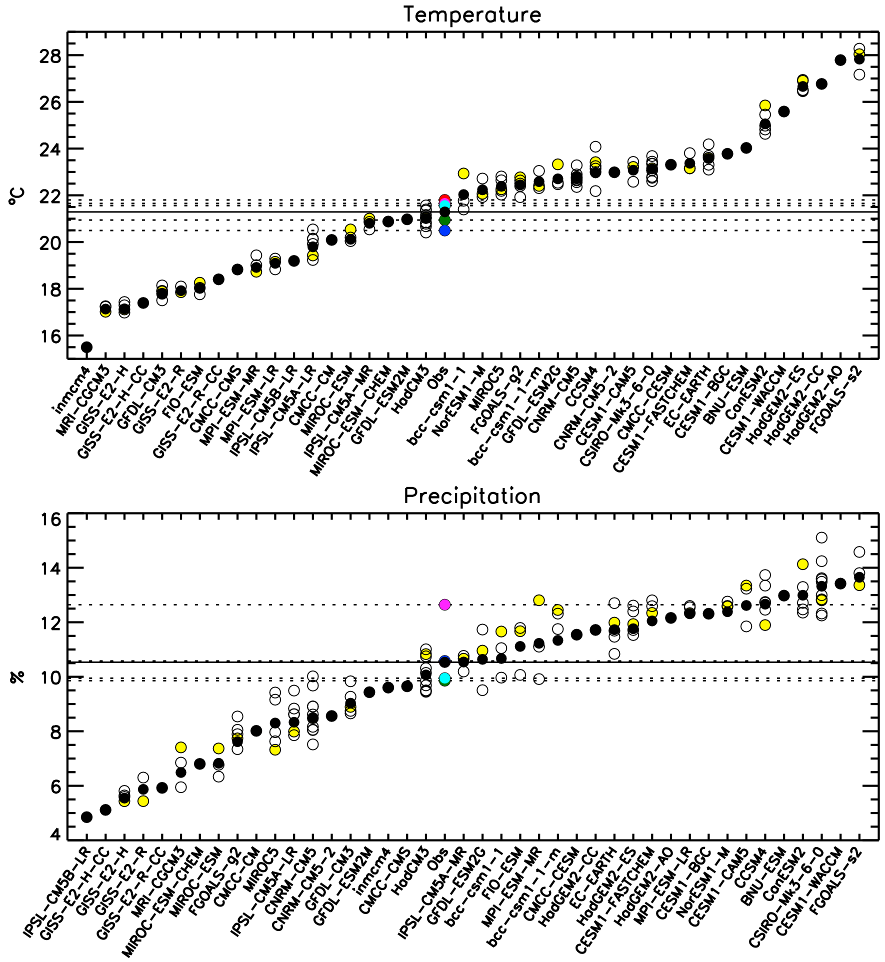

Figure A3. PNW mean seasonal cycle amplitude in temperature and relative precipitation. For each CMIP5 GCM, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from NCEP (red), ERA40 (magenta), CRU (dark green), PRISM (blue), UDelaware (Cyan), and average of observations (black). Monthly precipitation is calculated as a percentage of the mean annual total, so the amplitude is the difference of percentages.

consequently models can be ranked with a degree of confidence. For a given metric, an adequate ensemble size can be related to the spread of the ensembles of the individual models relative the spread across all models. Compare, for example, the intramodel spread with the intermodel range for mean annual temperature (Figure A1, upper panel) and for 20th century temperature trend (Figure A10). Visual inspection of Figures A1 and A10 illustrates that adequate ensemble size varies greatly with the particular metric. For example, the 40 year climatological mean annual temperature over the PNW shows little intramodel spread relative to total model spread. The consequence is that even though we only have one ensemble member for about one third of the models, and three or fewer members for roughly two third of the models, the ranking for this statistic would change little with additional ensemble members. In other words, a model's position in the ranking would likely only move at most one or two positions in either direction. This is not the case with long-term temperature trend, however. Below we describe how we apply this attribute of our metrics to rank their robustness.

[42] The most robust set of rankings are for the climatological mean annual temperature and precipitation (Figure A1), the amplitude of the annual temperature cycle (Figure A3, upper panel), the mean diurnal temperature range (Figure A6), and the spatial standard deviation and spatial correlation of seasonal temperature (Figure A9, upper panel). A second set of rankings is somewhat less robust: the amplitude of the annual precipitation cycle (Figure A3, lower panel) and the spatial standard deviation and spatial correlation of seasonal precipitation (Figure A9, lower panel).

[43] Note that all of the metrics above represent climatological averages computed over a 40 or 50 year period and thus should not, in principle, be very sensitive to temporal (i.e., internal) variability. However, our confidence in the model rankings decreases when we use metrics that summarize properties of the time series other than the long-term means. For example, the 20th century temperature trend (Figure A10) shows the intramodel range (for those models having five or more ensemble members) to be roughly 40% of the intermodel range. Choosing, for the sake of argument,

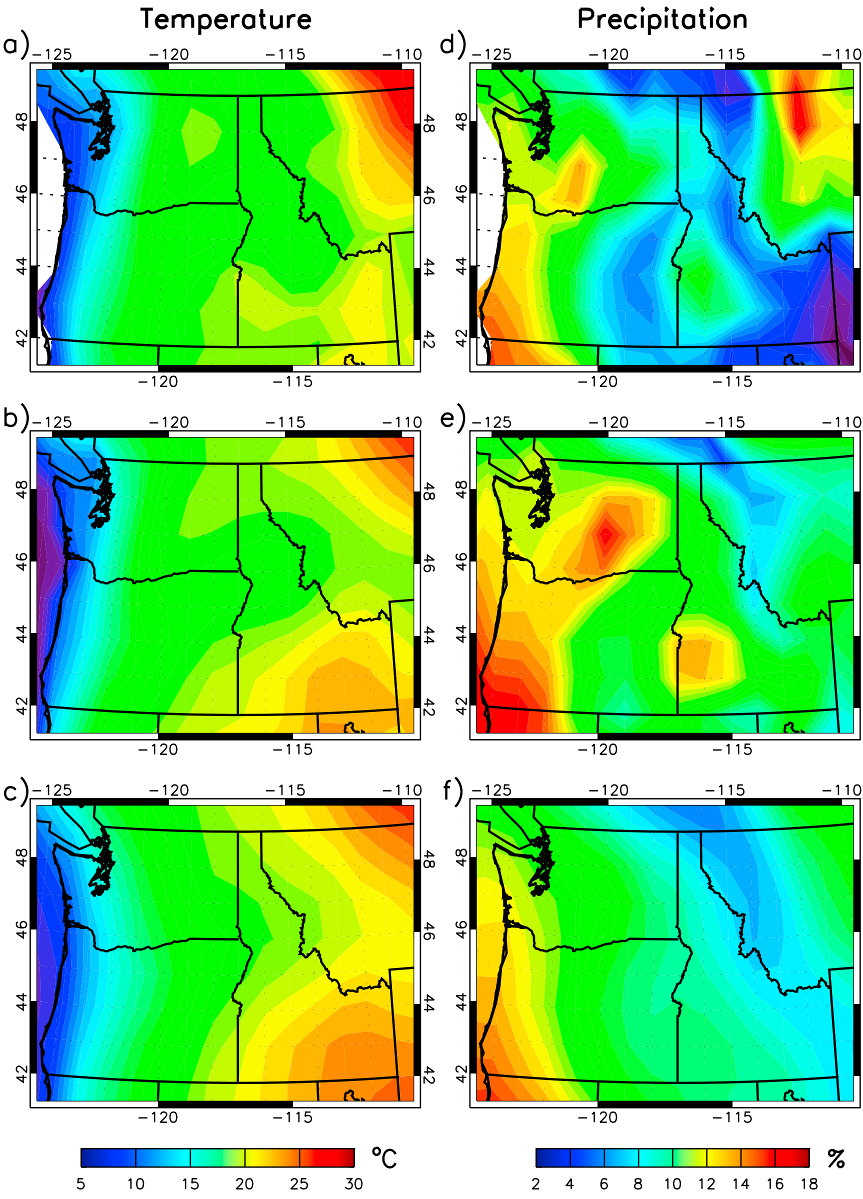

Figure A4. Mean seasonal cycle amplitude of temperature from (a) CRU, (b) ERA40, and the (c) CMIP5 multimodel mean, and mean season cycle amplitude of relative precipitation from (d) CRU, (e) ERA40, and the (f) CMIP5 multimodel mean.

those time series metrics for which the intramodel range is influential but still roughly less the 50% of the intermodel range leaves us with a third subset that contains annual temperature trend (Figure A10), standard deviation of annual temperature (Figure A12, upper panel), coefficient of variation of annual precipitation (Figure A13, upper panel), and, debatably, winter temperature response to ENSO (Figure A15, upper panel). The fourth, and final, subset includes those metrics where the intramodel variability is too high to comfortably rank models: annual precipitation trend (ranking not shown), standard deviation of octadal temperature (Figure A12, lower panel), coefficient of variation of octadal precipitation (Figure A13, lower panel), Hurst exponent for temperature and precipitation (ranking not shown), and winter precipitation response to ENSO (Figure A15, lower panel). This final subset of metrics might be best used to examine the behavior of the CMIP5 GCMs as whole, and not to rank one model against another, except to possibly identify the few models that lie farthest from the observations, or to compare only the small number of models with the most ensemble members.

[44] Based on the above four groupings of metrics, we defined four categories of robustness, or confidence, in rankings: “highest,” “higher,” “lower,” “lowest.” Table 3 lists within which category each metric falls.

3.3. Model Ranking by Overall Performance

[45] The models, ranked using the simple method, are listed in Figure 2. Also shown are the relative errors for the individual metrics. Each model scored well in at least one metric, but there were about 11 models that scored low (i.e., relative error > 0.6) in no more than two metrics.

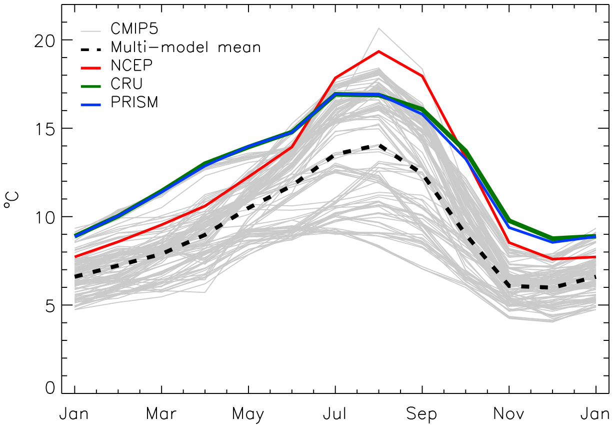

Figure A5. Mean seasonal cycle of diurnal temperature range averaged over the PNW. Monthly means calculated from gridded observation data sets (NCEP, CRU, and PRISM) and from all ensemble members from 41 CMIP5 GCMs.

Overall, the highest-ranked model was CNRM-CM5. The CESM1/CCSM4 family of models also stood out as "best" performers (with the exception of CESM1-WACCM). Other high scoring models were CanESM2, CNRM-CM5-2, the four models from the Hadley Center, and EC-EARTH.

[46] Prior to ranking the models based on the EOF-based method, we first excluded those metrics identified in the previous section as not being robust ("Lowest" category in Table 3). Furthermore, retaining only summer and winter for those metrics calculated seasonally left 18 of the 34 metrics. Using this subset of metrics, the leading five EOFs were significantly distinct. They cumulatively explained 30%, 48%, 59%, 68%, and 74% of the variance, respectively.

[47] The models, ranked in order using the first five EOFs, are shown in Figure 3. Using fewer EOFs (2 or 4, for example) would not dramatically change the overall picture, either (e.g., no models would move from the bottom one third to the top one third), though there would be some reordering of models. For example, the three models from the HadGEM2 family would place in the top 4 using only the first two EOFs.

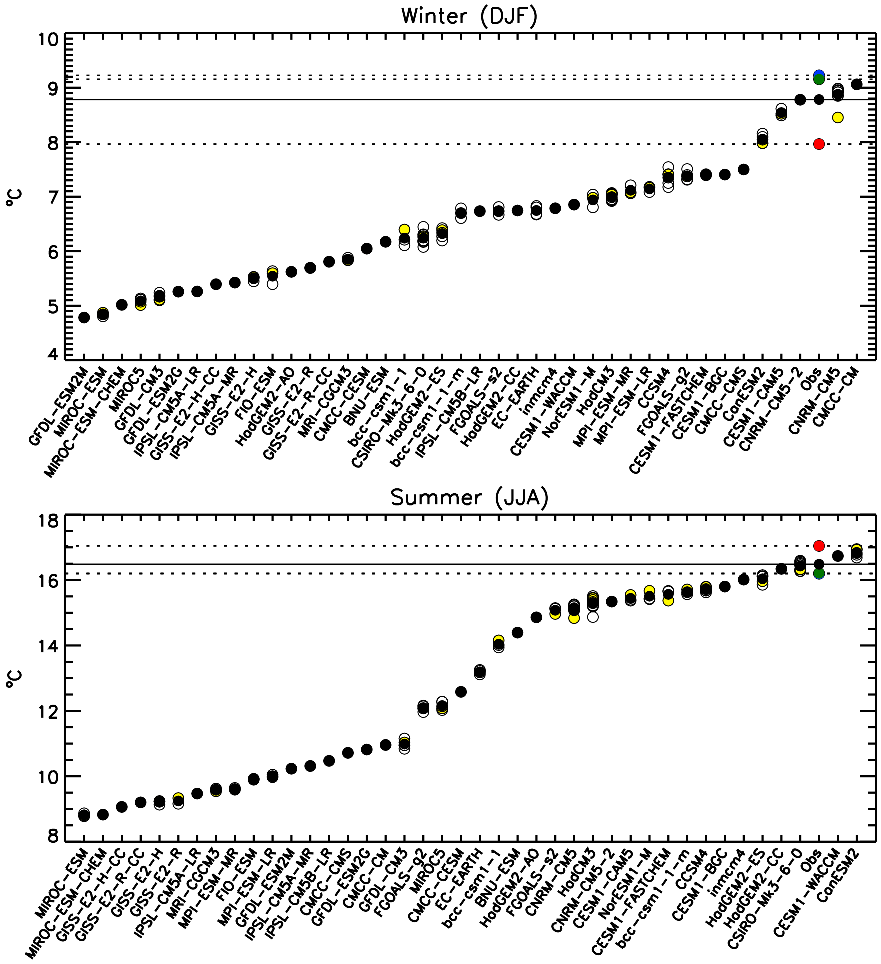

Figure A6. PNW mean diurnal temperature range in winter (DJF) and summer (JJA). For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from NCEP (red), CRU (dark green), PRISM (blue), and average of observations (black).

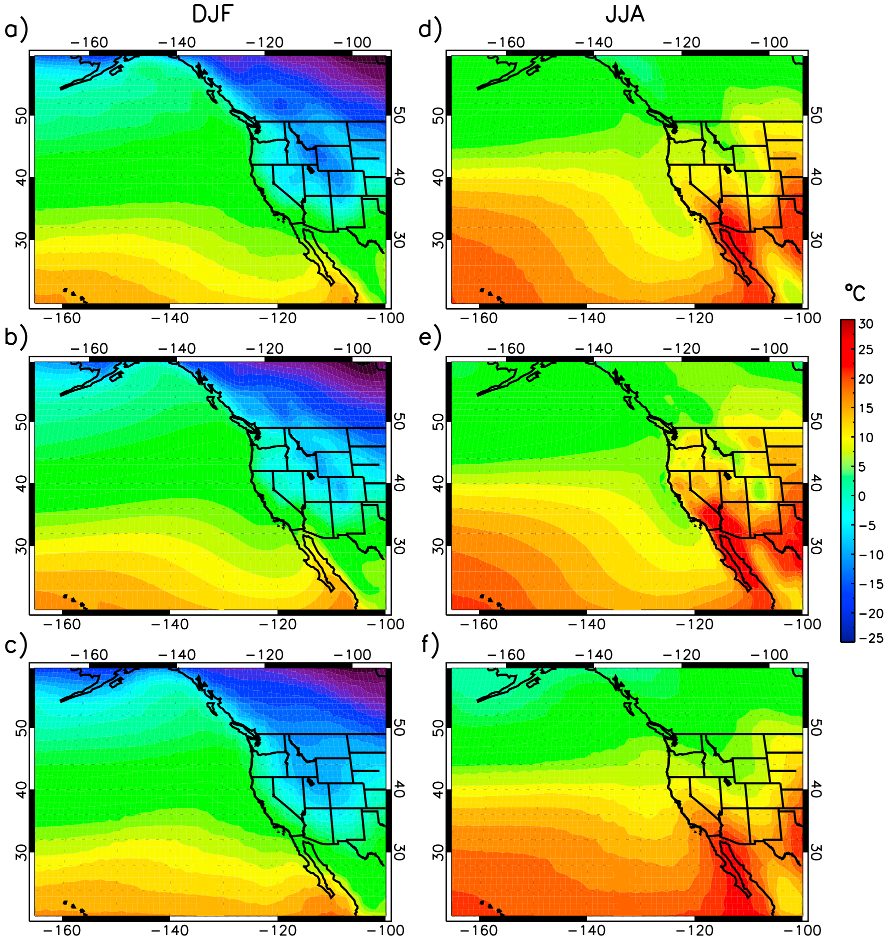

Figure A7. Mean winter (DJF) temperature from (a) NCEP, (b) ERA40 and the (c) CMIP5 multimodel mean, and mean summer (JJA) temperature from (d) NCEP, (e) ERA40, and the (f) CMIP5 multimodel mean.

[48] Comparing the results from our initial simpler ranking to the more complex EOF-based analysis reveals minor differences. Nearly no models occupied precisely the same position in each method, but the general order was similar. For example, nearly all the same models placed within the top 15 using either model. For a few models, the differences were substantial (e.g., BNU-ESM moved up 16 positions using the EOF method).

[49] As a visual aid toward differentiating models from observations, we examined their positions in EOF "space," by plotting the first four principal components (PCs) against each other (PC1 versus PC2 and PC3 versus PC4) (Figure 4). For example, PC1 distanced the MIROC-ESM and GISS-E2 families of the models from the observations, while PC2 revealed the relatively large errors in CMCC-CESM and bcc-csm1-1. Another outcome of the EOF analysis is that it highlighted the similarities of the models that come from the same modeling center. Note, for example, the intramodel groupings of the model roots HadGEM2, MIROC-ESM, and MPI-ESM in the upper panel of Figure 4. While it may not be surprising that variants of a base model from an individual center scored similarly given they model many processes in the same way (e.g., *Masson and Knutti* [2011]), individual metrics often separated similar models (e.g., PNW mean annual temperature differed by a 1°C between GISS-E2-R and GISS-E2-H; see Figure A1, upper panel); this analysis demonstrates that the variability introduced by model variations is much less than intermodel variability.

[50] An examination of the leading EOFs suggests possible groupings of dominant metrics (i.e., those with the largest loadings) (Table 4) that would aid in interpreting the axes in Figure 4. In the first EOF (which determines the PC1-axis), three pairs of metrics stood out as dominant. These were (1) seasonal amplitude of precipitation and temperature, (2) annual variability of precipitation and temperature, and (3) mean winter and summer diurnal temperature ranges. Another strong loading came from the mean annual precipitation. Excluding the last metric, the first EOF can be viewed largely as a measure of temporal variations across temporal scales (daily, seasonal, and interannual).

[51] In the second EOF, five of the strongest loadings pertain to the spatial patterns of climatological mean temperature and precipitation over the northwestern Pacific/western US. Though long-term temperature trend and variance of annual temperature also appeared as strong metrics, the second EOF can be viewed as largely characterizing spatial variability. We did not see suggestions of ways to assign a characteristic or property to the remaining leading EOFs.

[52] It is important to note that we have not quantified the sensitivity of the rankings to the choice of observational data

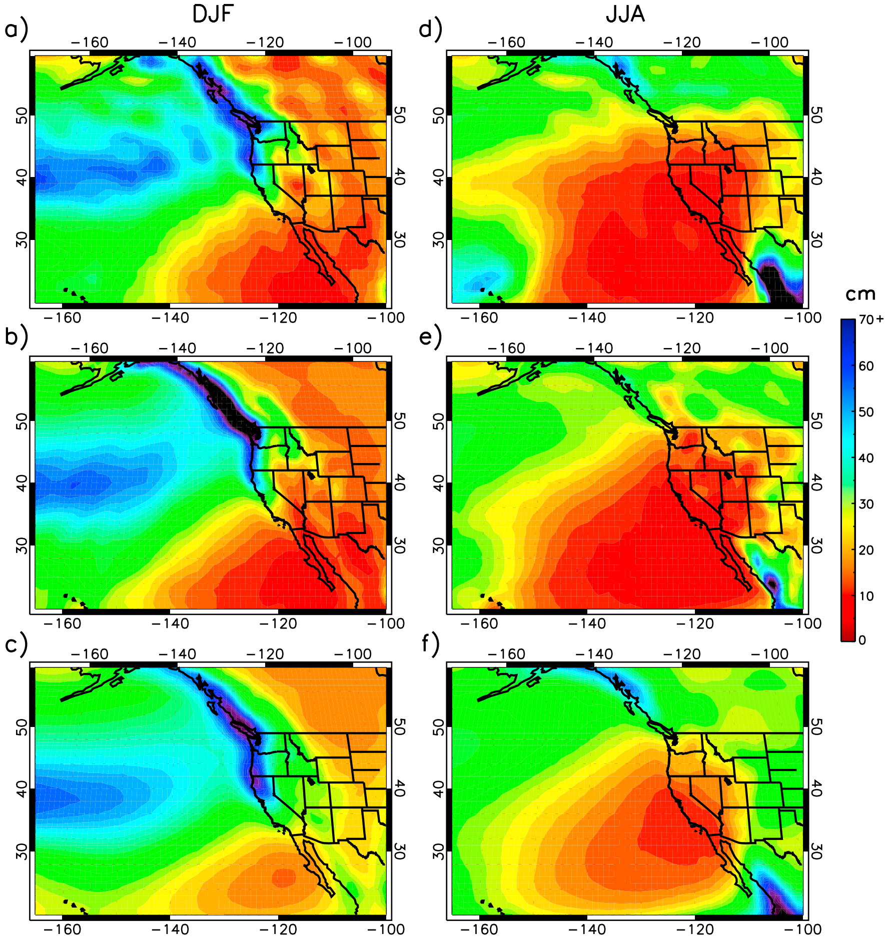

Figure A8. Mean winter (DJF) precipitation from (a) NCEP, (b) ERA40, and the (c) CMIP5 multimodel mean, and mean summer (JJA) precipitation from (d) NCEP, (e) ERA40, and the (f) CMIP5 multimodel mean.

set or to ensemble size. We expect that the effect of observational data set selection is minor given that intermodel spread for most metrics was much greater than the spread among observational data sets (though admittedly for some metrics, particularly mean temperature, the spread among observed quantities was sizable). Ensemble size, however, is an important issue, and it may be wise to consider giving some degree of preference to models with larger ensembles. One could also conduct a sensitivity analysis, for example, by generating multiple rankings based on a random sampling of the ensemble members.

[53] Last, the spread in model performance, as measured in this study, will largely be dependent on how each model parameterizes subgrid physics in the atmosphere, ocean, and land domains. To some extent, however, it may also be simply a function of the degree of discretization of the modeling domain [e.g., *Polade et al.*, 2013]. In fact, we found the EOF-based model ranking to be significantly correlated to the horizontal resolution of atmosphere (Pearson's correlation coefficient $r = 0.52$), that is the higher ranked models tended to have finer grid spacing. Among the individual metrics, six of those most correlated to spatial resolution were spatial correlations (SpaceCor-) JJA-P, -SON-T, SON-P, JJA-T, MAM-T, and DJF-P with $r = 0.59$, 0.54, 0.46, 0.45, 0.35, and 0.35, respectively. The only other two metrics that showed significant correlation (significance level = 0.95) were SpaceSD.MAM-P ($r = 0.36$) and DTR.DJF ($r = 0.33$). It is probably not coincidental that six metrics that were most correlated to the spatial resolution were themselves correlations of the simulated to observed mean spatial pattern.

4. Summary and Conclusions

[54] We evaluated 41 GCMs from CMIP5 with respect to their performance at simulating 20th century climate for the PNW. As a group, the models closely reproduced observations for a wide variety of temperature-based metrics. Individually, however, the models generated a wide range of values for many metrics, suggesting good reason for evaluating and ranking the models. The models, as a group, performed less well as judged by the precipitation-based metrics, though still reproduced observations for many dominant features of the system.

[55] Inadequate ensemble sizes for most models prevented us from ranking models with high confidence for many metrics. This was generally true more for precipitation than for temperature, and more for metrics that summarized properties of time series than properties of spatial fields. We therefore, reemphasize the need already stated by others [e.g., *Pierce et al.*, 2009] of generating enough realizations to reduce the effects of internal climate variability on certain metrics.

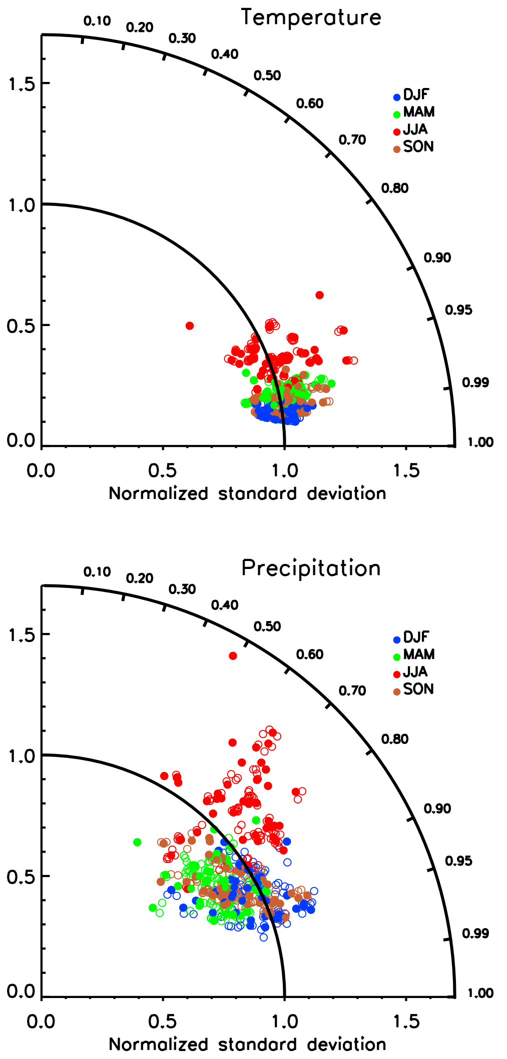

Figure A9. Normalized standard deviations (radius) and correlation coefficients (angle) by season for the climatological mean fields of temperature and precipitation from CMIP5. The spatial domain is approximately that shown in Figures A7 and A8. For each variable, the reference field for the normalization and the correlation is ERA40 reanalysis. Filled circles show the first ensemble members from each model and open circles show remaining ensemble members. Note that a perfect simulation would have both a normalized standard deviation and a correlation coefficient equal to unity.

[56] A wide range of values within an ensemble also implies a large standard error on the estimate of the observed climate statistic taken over the same length of record (given the simulated and observed internal variabilities are similar). In such a case, the uncertainty of the observed statistic is high, and therefore its value as a performance target is low. Increasing the ensemble size does not alleviate this problem, as it is a longer record length of the observations that is required. As record lengths are limited to the currently available observations, this supports our approach of assigning at least a qualitative measure of confidence to individual performance metrics.

[57] After excluding those metrics for which we had the least confidence in the estimates, we used EOF analysis to reduce the remaining metrics down to five performance measures (the leading EOFs). The loadings in the first two EOFs suggested they could be roughly classified as pertaining to "temporal variability" and "spatial variance/correlation," respectively. Given that the first EOF is dominated by temporal variability highlights the importance of temporal variability in providing the power to discriminate among models.

[58] Though we have provided an overall ranking of the models, separate rankings could be calculated using a subset of those metrics that are more relevant to different types of impact assessments. For example, *Brekke et al.* [2008] used different metric sets for water supply, hydropower, and flood control. For some applications, models could be evaluated using only those metrics that are not readily addressed using bias-correction methods, though one should be aware that these metrics tend to be those for which we have less confidence. It would also be desirable to assess the CMIP5 GCMs on their ability to adequately reproduce phenomena that affect daily weather in the PNW, such as the position and intensity of storm tracks, and formation of atmospheric rivers. For this initial evaluation, however, we chose to use only monthly data sets of temperature and precipitation, which directly excluded phenomena that occur on finer time scales.

[59] Last, when the objective is to assess the impact of climate change over the next several decades or more, model performance is not the sole consideration in developing estimates of future change and its uncertainty. Model performance should be balanced by the need to adequately sample the yet only very roughly known distribution of climate trajectories.

Appendix A: CMIP5 Models' Skill by Performance Metric

A1. Climatologic Mean

[60] The simulated mean annual temperature of the PNW differed by 5.5°C from the coolest to the warmest model (Figure A1, upper panel). The observation data sets also differed notably amongst themselves; NCEP was 2.4°C, 1.9°C, 1.5°C, and 1.2°C cooler than ERA40, CRU, PRISM, and UDelaware, respectively. Taken as an average, the five observation data sets were 0.8°C warmer than the median of the simulated mean annual temperatures. Moreover, 23 of 41 models fell within range of the five observational values. (Note: henceforth the observed values will be reported as the average of the observation data sets used, unless specifically stated otherwise).

[61] Mean annual precipitation was less well reproduced by the models as all but one model generated more precipitation than observed (Figure A1, lower panel). The range across models was large: 75 cm yr^{-1} difference between the wettest and driest model. Compared with the estimated 76 cm of annual precipitation received in the PNW, the wettest model produced nearly 80% too much precipitation in the region. Only six models fell within the range of the observed values.

A2. Seasonal Cycle

[62] All the models reproduced the phase and general shape of the seasonal cycle of temperature (Figure A2, upper panel), though the amplitude of the seasonal cycle varied widely among models (Figure A3), ranging from 15.5°C to 27.8°C. The median of the modeled amplitude was 22.2°C, which was within 1°C of the observed amplitude.

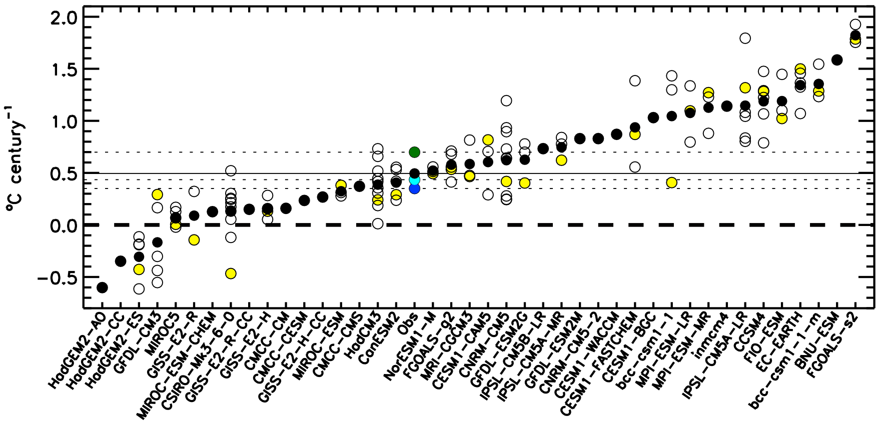

Figure A10. PNW-averaged trends in annual temperature over the 20th century for all simulations and observations. For each of 41 CMIP5 GCMs, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

[63] All models generated more precipitation in winter than summer, which is characteristic of the PNW (Figure A2). However, the models generated a wide range of amplitudes of the seasonal precipitation cycle, with a handful of model severely under-simulating the strength of the seasonal variation. Calculating mean monthly precipitation as a percentage of the mean annual total, seasonal precipitation amplitude ranged from as small as 4.8% to as large as 13.7%. In comparison, the observed mean amplitude was 10.5%, which is essentially identical to median of all the models. (Note that with mean monthly precipitation calculated as a percentage of annual, the percentages reported above are the differences of percent precipitation between the wettest and driest months).

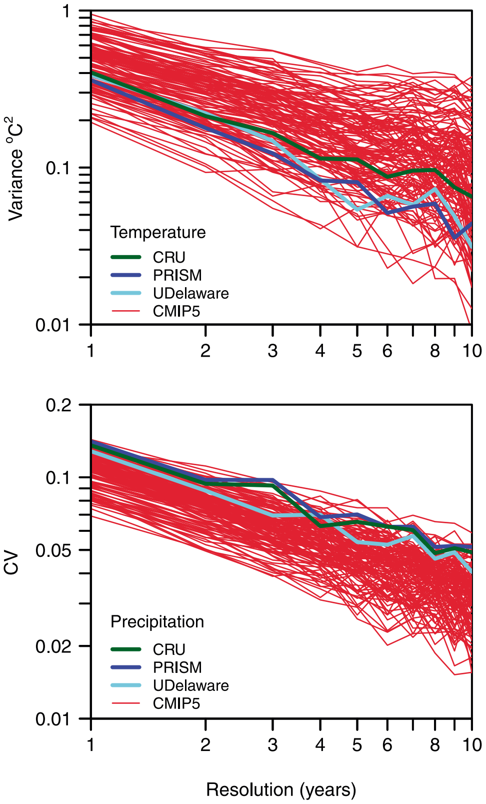

Figure A11. (Upper panel) Variance of temperature anomalies and (lower panel) coefficient of variation of precipitation against temporal resolution for the PNW-averaged time series. Red lines show results from all ensemble members from 41 CMIP5 GCMs.

[64] Though the statistics above were averaged over the entire PNW, there are strong regional gradients in the amplitudes of the seasonal cycle of both the temperature and precipitation. Visual inspection showed good agreement between the observed (CRU and ERA40) and the multimodel mean spatial pattern of the cycle temperature amplitude overall (Figure A4, left panels). The models as a whole accurately reproduced the strong west-to-east gradient, though the multimodel mean generated a larger seasonal amplitude (by ~ 5°C) in southeastern Idaho than was evident in CRU. This discrepancy was not as stark when compared to PRISM or UDelaware (not shown).

[65] The dominant gradient in the observed (CRU and ERA40) precipitation amplitude in the region is from the southern coast of Oregon toward the Rocky Mountains (Figure A4, right panels). The multimodel mean reproduces this feature, but not surprisingly, lacks some finer-scale features. This may be due in part to the smoothing from averaging across all the models of varying native resolution, but also reflects an incomplete representation of key topographical features in the GCMs. For one, CRU (and PRISM and UDelaware, though not shown) all gave a weaker seasonal cycle in the Rocky Mountains, particularly in southeastern Idaho, than did the multimodel mean.

A3. Diurnal Temperature Range

[66] Mainly due to much greater cloud cover in winter, the winter diurnal temperature range (DTR) is much smaller than summer DTR: 9°C and 17°C in January and July, respectively, for both CRU and PRISM (Figure A5). The seasonal cycle of multimodel mean DTR largely resembles the seasonal cycle of DTR from CRU and PRISM, while NCEP and a large group of models showed a bigger difference between summer and winter DTR. However, simulated

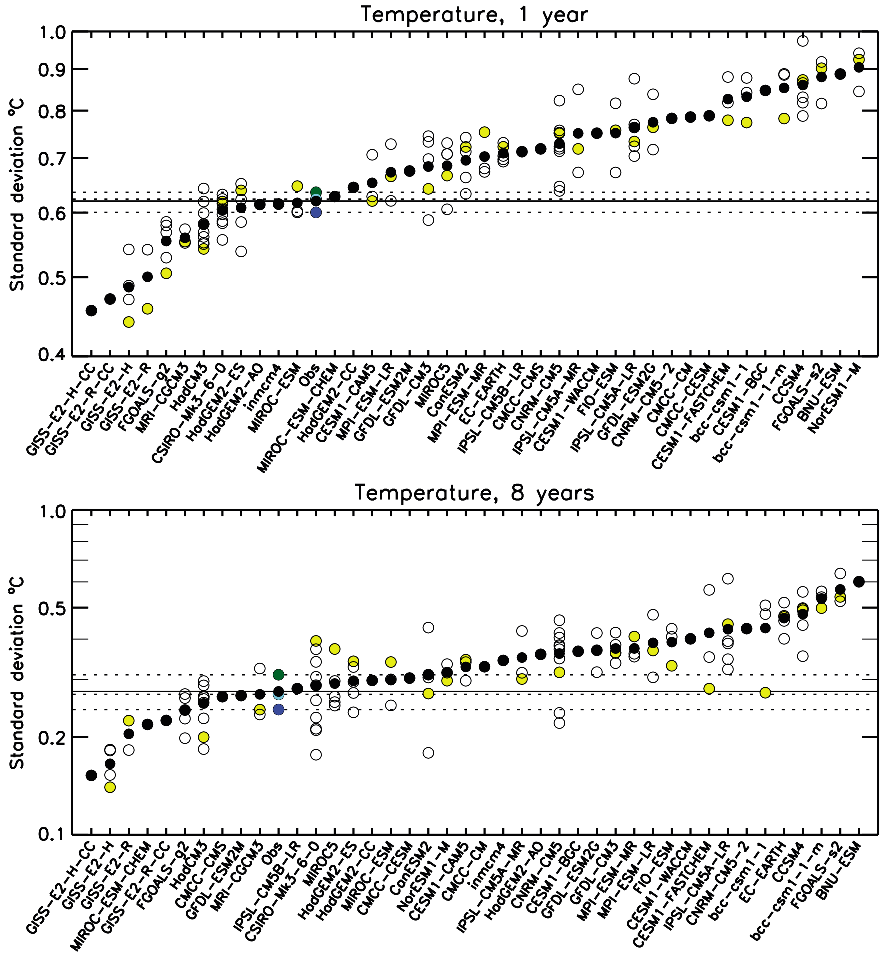

Figure A12. Standard deviation of temperature anomalies at resolutions of 1 year and 8 years. Values were averaged over the PNW domain. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

DTR tended to be about 2.5–3.5°C too low throughout the year compared to CRU and PRISM, or 1–5°C compared to NCEP. With very few exceptions, the individual GCMs generated a diurnal temperature range that was too small, irrespective of season (Figure A6). Intermodel variability also changed throughout the year, with greater differences among models occurring in summer. While 18 GCMs had a summer DTR that was within about 2°C of the observed summer DTR, model skill for this attribute dropped precipitously for the remaining models (Figure A6, lower panel). Though we have not done so here, it may be worth investigating if this apparent bimodal distribution of the models can be explained by some shared properties of the models with less, or more, error.

[67] One should be aware that observations of Tmin and Tmax are relatively instantaneous, whereas simulated Tmin and Tmax have been averaged over some time step that varies by GCM and therefore are effectively biased toward lower DTR. We have not attempted to account for this bias here.

[68] Another point worth noting is western most grid cells of the PNW domain contained some influence of ocean cells in the models and in NCEP, unlike CRU and PRISM. We might expect this to suppress DTR somewhat in both the models and NCEP, as compared to CRU and PRISM. As a test, we reduced the size of the PNW domain by 1° on the westward side and recalculated the PNW-average DTR. While this slightly increased DTR across all data sets, it only negligibly affected the relative values of DTR; thus, the ranking of models based on DTR alone (Figure A6) did not change. Also, the discrepancy between NCEP and CRU/PRISM remained unchanged, implying that the differences among observational data sets seen in Figure A5 are not related to the influence of ocean cells.

A4. Large-Scale Spatial Patterns

[69] The multimodel mean temperature field over the North Pacific/western North America accurately reproduced the ERA40 climatological fields (Figure A7), with correlation coefficients (r) of 0.995 and 0.951 in winter and summer, respectively (for NCEP, $r = 0.996$ and 0.972, respectively). Much of this high correlation resulted from simply matching the general latitudinal temperature gradient, but more interesting features of the climatological

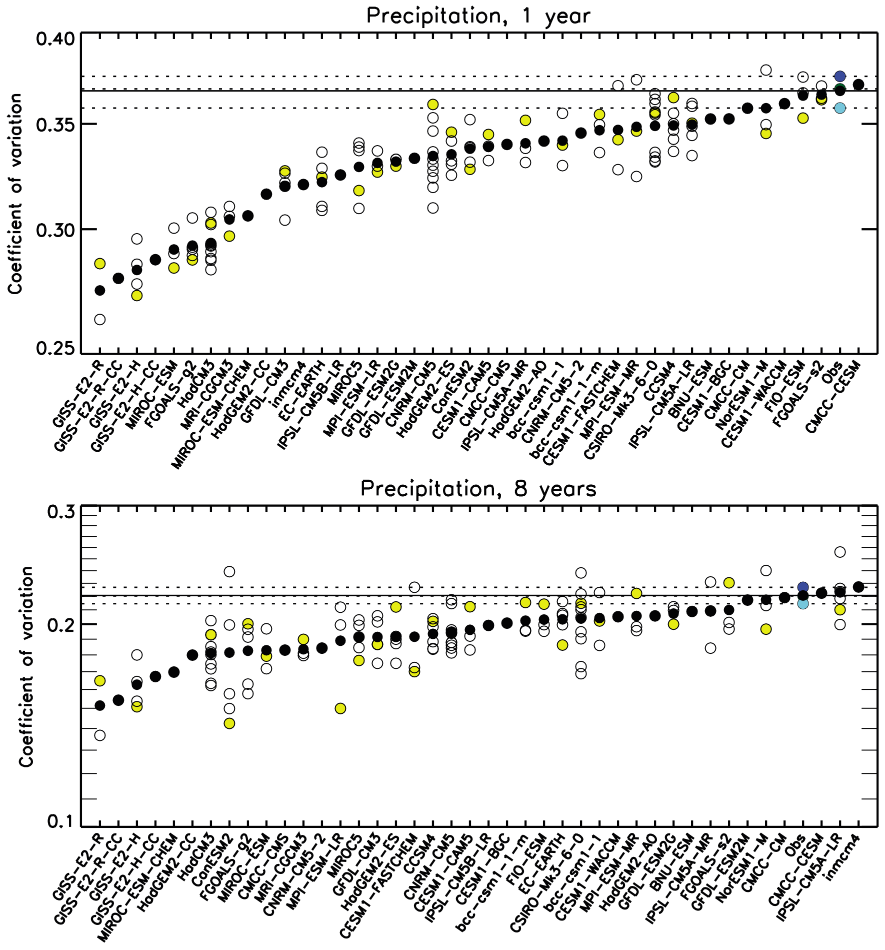

Figure A13. Coefficient of variation of precipitation at resolutions of 1 year and 8 years. Values were averaged over the PNW domain. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

fields were also reproduced. Of particular interest to the PNW is the temperature pattern over the eastern North Pacific. Most notably in the summer, the southward-moving California Current brings cooler sea surface temperatures (SSTs) near the west coast of the United States, while the Alaskan Current moves warmer water up the coast of Canada and southeast Alaska. The effect of these currents on the near ocean surface air temperature patterns is clearly evident in the observations and simulations, though the location of the divergence of these two currents, which occurs roughly at the latitude of the US-Canada boundary, as shown most clearly in summer temperature pattern, is located further south in the multimodel mean. Moreover, the "tongue" of cooler air along the California coast does not extend as far south in the multimodel mean as in ERA40 and NCEP. Other notable features that are reproduced by the GCMs as whole are the cooler air temperatures over the Rocky Mountain Range and the tongue of warm air extending northward up the Gulf of California. Individually, all models were very highly correlated to observations in winter ($0.981 \leq r \leq 0.995$) and highly correlated in summer ($0.89 \leq r \leq 0.99$) but for one outlier ($r = 0.78$). The variances of the modeled fields were also similar to the observed variance, though more so winter, when all standard deviations were within ±13% of the observed standard deviations. In summer, all GCMs were within +32% of observations (Figure A9, upper panel).

[70] The multimodel mean precipitation field over the North Pacific/western North America reproduced the main large-scale climatological features of the ERA40 field in winter ($r = 0.94$), though was less faithful in summer ($r = 0.83$) (Figure A8) (for NCEP, $r = 0.93$ and 0.82, respectively). In winter, the most prominent features of the observed precipitation field were (1) the band of heavy precipitation across the central North Pacific midlatitudes that weakened as it progressed eastward, and (2) high precipitation along the coast of North America from northern California to southeast Alaska (Figure A8, left panels). The multimodel mean reproduced these two features both in their location and extent. However, over much of western North America and over the ocean west of Mexico, the multimodel mean gave too much precipitation. In summer, the dominant feature is the dry zone that covers western United States and extends into the eastern Pacific

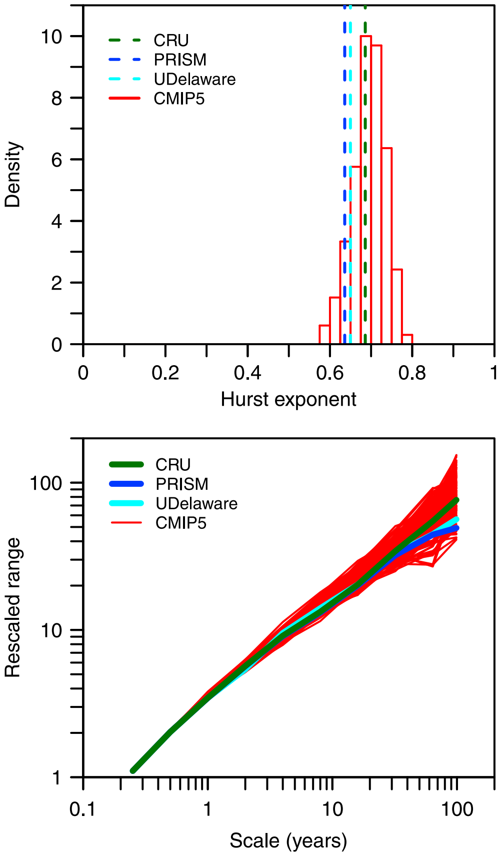

Figure A14. Upper panel: histogram of the Hurst exponent for PNW average temperature in all CMIP5 simulations. The vertical dashed lines indicate the Hurst exponent estimated from observations. Lower panel: the rescaled range against the time scale calculated from the observations (heavy lines) and simulations (thin red lines).

(Figure A8, right panels). This dry zone was present in the multimodel mean field, but was wetter than the both ERA40 and NCEP. Individually, the spatial patterns of all models correlated well with the spatial pattern of ERA40 precipitation in winter ($0.76 \le r \le 0.96$) while the correlations weakened in spring and fall, and were weakest in summer ($0.48 \le r \le 0.85$). Normalized standard deviations ranged from 0.58 to 1.24 across all simulations and all seasons, with a large majority of models simulating too little spatial variability in spring and too much variability in summer (Figure A9, lower panel).

A5. 20th Century Trend

[71] The average annual temperature in the PNW increased during the 20th century by an estimated 0.70°C based on CRU, while the trend calculated from PRISM and UDelaware was 0.35°C and 0.44°C per century. Of 41 CMIP5 models, 37 also produced a positive trend in temperature over the 20th century (Figure A10), with a multimodel mean warming of 0.61°C. Simulations did not produce large differences among seasons in warming, ranging from 0.53°C in fall to 0.74°C in summer. Moreover, there was no consistency in seasonal differences between simulations and observations, i.e., the seasons with greater observed warming were not those with greater simulated warming.

[72] The linear trend in observed regional mean annual precipitation, while calculated to be about +10% over the 20th century (calculated as percent change from the mean, 1901–1999), was not statistically significant. In fact, the sign of the trend is sensitive to the period of record, and is negative if one begins in, for example, 1940 [e.g., (J. T. Abatzoglou et al., Understanding seasonal climate variability and change in the Pacific Northwest of the United States, submitted to *Journal Climate*, 2013)]. Models produced ensemble-average trends ranging from −8% to +7% per century, while only three individual ensemble members from three GCMs exceeded the observed 10% per century (results not shown). The multimodel mean trend in annual precipitation was only +0.5% per century, and seasonal differences were minor, ranging from −1.4% to +1.9% per century in JJA and SON, respectively. In summary, the lack of a strong, modeled trend in precipitation is consistent with the lack of a statistically significant trend in observed precipitation.

A6. Temporal Variability

[73] Overall, the CMIP5 models tended to produce too much interannual-to-decadal variability in PNW-averaged times series of temperature relative to the observations (Figure A11, upper panel). At the annual scale, simulated standard deviations ranged by a factor of 2, from 0.46°C to 0.90°C (Figure A12). At the octadal (i.e., 8 year) scale, simulated values ranged from 0.15°C to 0.6°C, or a factor of 4. Still, many models did not fall very far from the observations. For example, 21 of 41 models (~50%) had standard deviations that were within ±18% of the observed standard deviation at the annual scale, while ~50% of the models were within ±25% at the octadal scale. A general similarity in the scaling of the variance among models and observations meant that a model that was, for example, too variable at the annual scale was also likely too variable at the decadal scale.

[74] In the case of precipitation, nearly all models generated less temporal variability than seen in observations, and this was consistent across scales (Figure A11, lower panel). Though both the simulations and observations showed apparent power law scaling of the CV, the simulated variances in general decreased too rapidly with increasing scale. Despite this tendency to under-represent the variability, ~50% of models had CVs that were within ±8% and ±10% of the observed CV at the annual scale and octadal (8 year) scale, respectively (Figure A13). The cause(s) of the differences among models in simulating both the variability in precipitation and temperature may be related, as suggested by the significant correlation between the CV of precipitation and the standard deviation of temperature and at both the annual ($r = 0.79$) and octadal scale ($r = 0.49$).

A7. Long-Term Persistence

[75] The Hurst exponent of the observed temperature anomalies ranged from 0.64 to 0.69, depending on the data set (CRU, PRISM, or UDelaware). Though the causes of observed Hurst exponent are not explored here, these values could, for example, indicate long-term memory or nonstationarity in the mean [*Klemes*, 1974]. In either case, the Hurst exponent > 0.5 implies that the processes that determine temperature over the PNW occur over a wide

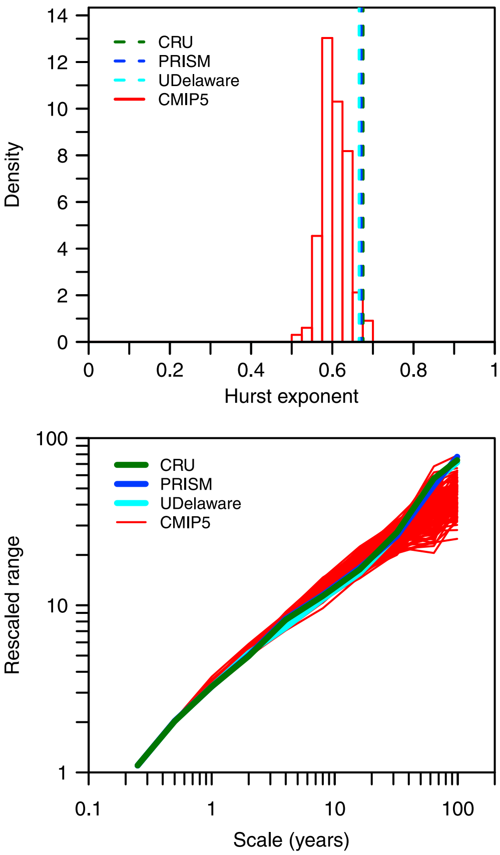

Figure A15. Upper panel: histogram of the Hurst exponent for all PNW-average precipitation CMIP5 simulations. The vertical dashed lines indicate the Hurst exponent estimated from observations. Lower panel: the rescaled range against the time scale calculated from the observations (heavy lines) and simulations (thin red lines).

range of scales [*Tessier et al.*, 1996]. The mean Hurst exponent averaged over either all ensemble members, or all models, was 0.69. Individual simulations showed Hurst exponents all greater than 0.5 ($0.59 \le H \le 0.78$) with 90% of values falling between 0.62 and 0.75 (Figure A14, upper panel).

[76] The estimated Hurst exponent of the observed precipitation anomalies ranged from 0.67 and 0.68, similar to that for temperature. The mean simulated Hurst exponent was 0.61, with 90% of values falling between 0.55 and 0.67 (Figure A15, upper panel). All but three ensemble members gave Hurst exponents that were less than the observed value. Even so, the Hurst exponents estimated from the simulated precipitation were all greater than 0.5.

[77] We repeated the above analyses after removing the long-term linear trend from both the temperature and precipitation records. This was in an attempt to remove the possible anthropogenic influence on the Hurst phenomenon, if one assumes that the linear trend, if present, was predominantly driven by GHG concentrations. The results were minor differences of −0.01 to −0.03 (<5%) in H for both observed temperature and precipitation, respectively. The simulations showed similar decreases in H (i.e., <5%).

[78] Our results for precipitation appear to contradict those of *Kumar et al.* [2013] who estimated Hurst exponents, averaged over 19 CMIP5 models, that were not significantly different from 0.5 over any land mass, including the PNW (see their Figure A10). This is important because their results, unlike ours, suggest that the CMIP5 multimodel ensemble fails to capture the observed persistence in precipitation observations anywhere in the globe. However, the two studies used different techniques for estimating the Hurst exponent, which may explain some of the discrepancy. For this reason, we elaborate on our calculation of the Hurst exponent in Appendix B.

A8. ENSO Teleconnections

[79] The PNW has long been known to exhibit a correlation with ENSO with respect to both temperature and precipitation, particularly in winter to early spring (e.g., *Ropelewski and Halpert* [1986]; *Mote et al.* [2003]). Consistent with CRU observations, a positive PNW temperature response to ENSO was apparent in the GCMs: all but three models had a positive response of winter (JFM) temperature to ENSO (Figure A16, upper panel). Moreover, the multimodel mean response was a 0.41°C increase in PNW winter temperature for every 1°C increase in the Niño3.4 index, which is close to the observed response of 0.57°C °C^{-1}. The agreement in the spatial pattern of the ENSO response is remarkable, particularly west of the continental divide (Figure A17, upper panels), where the transition between a positive to negative response occurs through southern California and Nevada and central Utah in both the observed (see also *Yu et al.* [2012]) and multimodel mean fields.

[80] A precipitation response to ENSO was also apparent in most models, with 35 of 41 showing reduced JFM precipitation with warmer tropical Pacific temperatures (Figure A16, lower panel). The multimodel mean response was −4.4% °C^{-1}, compared with the observed response of −5.9% °C^{-1}. The spatial patterns of the observed and mean simulated ENSO precipitation response were also similar in the PNW, with the transition of the sign of the response occurring near the southern Oregon and Idaho borders in both cases (Figure A17, lower panels). Absent from the simulated response, however, was the tongue of observed positive (wetter) response that extends northward through eastern Oregon and Washington. This discrepancy may be due to the GCMs not resolving prominent topographic features (i.e., the Cascades) and therefore not simulating the changing rain-shadow effect that occurs with a changing ENSO phase [*Siler et al.*, 2013].

Appendix B: Calculation of the Hurst Exponent

[81] The Hurst exponent was estimated from the rescaled range. Shortcomings of the rescaled range method have been documented and various alternatives for estimating the Hurst exponent have been developed [*Caccia et al.*, 1997; *Simonsen et al.*, 1998; *Kantelhardt et al.*, 2003; *Chamoli et al.*, 2007]. However, the rescaled range has a long history of use and produces errors less than 10% for $H > 0.5$ for data sets on the order of 10^3 points [*Bassingthwaighte and Raymond*, 1994; *Chamoli et al.*, 2007; *Hamed*, 2007], therefore suited our needs.

[82] Following *Hamed* [2007], the Hurst exponent H is related to the rescaled range by

$$E(R/S) = (m/2)^H \tag{B1}$$

where $E(R/S)$ is expected value of the rescaled range (R/S) and m is the number of consecutive values from a time

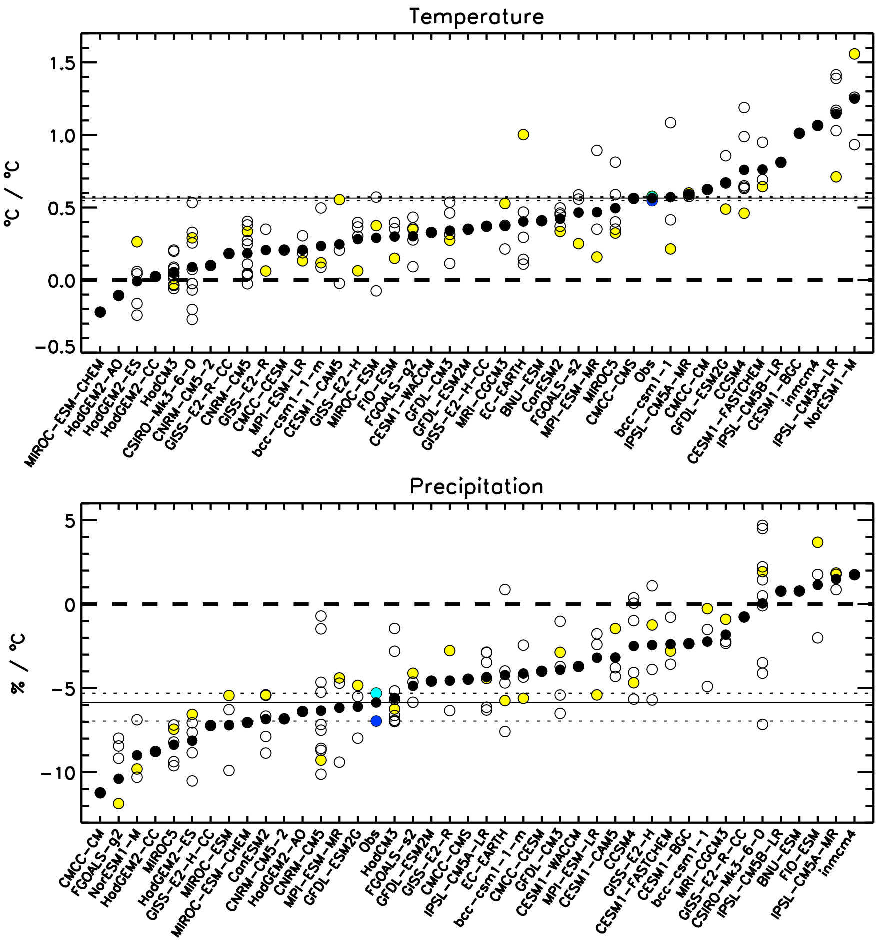

Figure A16. Response of PNW winter (JFM) temperature and precipitation to the Niño3.4 index averaged over NDJFM. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

series (i.e., the "scale" in Figures A14 and A15). The rescaled range is given by

$$\frac{R}{S} = \frac{\max\limits_{1 \le k \le m}(D_k) - \min\limits_{1 \le k \le m}(D_k)}{S} \tag{B2}$$

where D_k is the cumulative sum of the deviations from the mean of the sample up to the k^{th} element of m:

$$D_k = \sum_{i=1}^{k}(y_i - \bar{y}) \tag{B3}$$

[83] The sample mean $\bar{y}$ and biased standard deviation S are given, respectively, by

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{B4}$$

and

$$S = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{B5}$$

[84] The rescaled range is calculated for M number of samples of size m that make up the total time series, and the expected value $E(R/S)$ is taken. The procedure is repeated over a range of m.

[85] It is worth noting that the estimate of Hurst exponent may vary with the time scales (i.e., length m of the interval) across which the rescaled range statistic is calculated. We calculated the rescaled range for intervals ranging from 3 months to 99 years (see Figure A14, lower panel) and estimated H as the slope of a linear fit of the log-rescaled range against log(m). In the case of temperature, it is evident from Figure A14 (lower panel) that the overall results would vary somewhat, though not by much, if we used instead some subset of the scales we used (e.g., 1 to 32 years). However, in the case of precipitation, an arguably anomalous increase in the observed rescaled range occurs between scales of 32 and 64 years (Figure A15, lower panel). If we limit the range to scales of 32 years and under, the observed H becomes 0.64 and falls well within the 5% and 95% percentiles of the simulated values ($H = 0.60$ and 0.68, respectively).

[86] Last, it is evident for both temperature and precipitation that the relationship from both the observed and simulated time series is not log-log linear, but that the curves

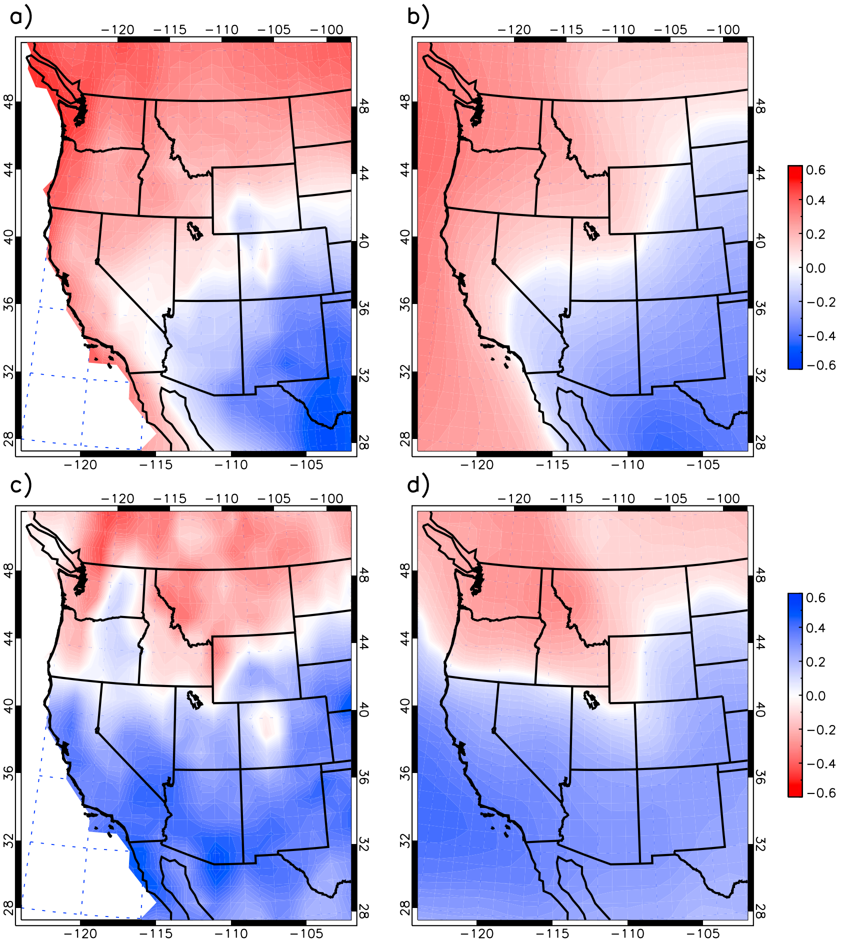

Figure A17. Correlation of CRU winter (JFM) (a) temperature and (c) precipitation with the Niño3.4 index averaged over NDJFM, and mean correlation of simulated winter (b) temperature and (d) precipitation to the same index. Simulations were from 41 CMIP5 models. Note that the legends have been reversed between the upper and lower plots so that red implies both warmer and dryer conditions in the PNW with a higher Niño3.4 index (i.e., El Niño conditions).

show convexity (Figures A14 and A15, lower panels), whereas the Hurst phenomenon implies a strictly log-log linear behavior. Consequently, the absolute values of *H* reported here are dependent on the particular range of scales we have chosen.

[87] **Acknowledgments.** This research was funded by the U.S. Department of Agriculture via the Regional Approaches to Climate Change - Pacific Northwest Agriculture (REACCH PNA, USDA-NIFA grant 2011-68002-30191) project, the NOAA Climate Impacts Research Consortium (RISA for the Northwest), the U.S. Department of the Interior via the Northwest Climate Science Center, and the Willamette Water 2100 project (NSF grant EAR-1039192). We thank Rachel Calmer and Darrin Sharp for assistance with data acquisition and processing, Nick van de Giesen for his thoughts on the Hurst exponent, and three anonymous reviewers.

References

Bassingthwaighte, J. B., and G. M. Raymond (1994), Evaluating rescaled range analysis for time series, *Ann. Biomed. Eng.*, *22*, 432–444, doi:10.1007/BF02368250.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard (2013), ENSO representation in climate models: From CMIP3 to CMIP5, *Clim. Dyn.*, doi:10.1007/s00382-013-1783-z.

Brekke, L. D., M. D. Dettinger, E. P. Maurer, and M. Anderson (2008), Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments, *Clim. Change*, *89*, 371–394.

Caccia, D., D. Percival, M. Cannon, G. Raymond, and J. Bassingthwaighte (1997), Analyzing exact fractal time series: Evaluating dispersional analysis and rescaled range methods, *Physica A*, *246*, 609–632.

Chamoli, A., A. R. Bansal, and V. P. Dimri (2007), Wavelet and rescaled range approach for the Hurst coefficient for short and long time series, *Comput. Geosci.*, *33*, 83–93, doi:10.1016/j.cageo.2006.05.008.

Christensen, J. H., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen (2010), Weight assignment in regional climate models, *Clim. Res.*, *44*, 179–194, doi:10.3354/cr00916.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. A. Pasteris (2008), Physiographically-sensitive mapping of temperature and precipitation across the conterminous United States, *Int. J. Climatology*, *28*, 2031–2064, doi:10.1002/joc.1688.

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

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models. *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Hamed, K. H. (2007), Improved finite-sample Hurst exponent estimates using rescaled range analysis, *Water Resour. Res.*, *43*, W04413, doi:10.1029/2006WR005111.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister (2013), Updated high-resolution grids of monthly climatic observations – The CRU TS3.10 Dataset, *Int. J. Climatol.*, doi:10.1002/joc.3711.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*, 1095–1107, doi:10.1175/2009BAMS2607.1.

Hawkins, E., and R. Sutton (2011), The potential to narrow uncertainty in projections of regional precipitation change, *Clim. Dyn.*, *37*, 407–418, doi:10.1007/s00382-010-0810-6.

Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 770–799.

Joseph, R., and S. Nigam (2006), ENSO evolution and teleconnections in IPCC's twentieth-century climate simulations: Realistic representation?, *J. Clim.*, *19*, 4360–4377.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–470, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kantelhardt, J. W., D. Rybski, S. A. Zschiegner, P. Braun, E. Koscielny-Bunde, V. Livina, S. Havlin, and A. Bunde (2003), Multifractality of river runoff and precipitation: Comparison of fluctuation analysis and wavelet methods, *Physica A*, *330*, 240–245.

Kim, S.T., and J.-Y. Yu (2012), The two types of ENSO in CMIP5 models, *Geophys. Res. Lett.*, *39*, L11704, doi:10.1029/2012GL052006.

Klemes, V. (1974), The Hurst phenomenon a puzzle?, *Water Resour. Res.*, *10*, 675–688.

Knutti, R. (2010), The end of model democracy?, *Clim. Change*, *102*, 395–404, doi:10.1007/s10584-010-9800-2.

Knutti, R., D. Masson, D., and A. Gettelman (2013), Climate model genealogy: Generation CMIP5 and how we got there, *Geophys. Res. Lett.*, *40*, 1–6, doi:10.1002/grl.50256.

Kumar, S., V. Merwade, J. Kinter III, and D. Niyogi (2013), Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 20th century climate simulations, *J. Clim.*, doi:10.1175/JCLI-D-12-00259.1, in press.

Masson, D., and R. Knutti (2011), Climate model genealogy, *Geophys. Res. Lett.*, *38*, L08703, doi:10.1029/2011GL046864.

Mo, K. C., L. N. Long, and J.-K. E. Schemm (2012), Characteristics of drought and persistent wet spells over the United States in the Atmospheric-Land-Ocean Coupled Model Experiments, *Earth Interact.*, *16*, doi:10.1175/2012EI000437.1.

Mote, P. W., and E. P. Salathé Jr. (2010), Future climate in the Pacific Northwest, *Clim. Change*, *102*, 29–50.

Mote, P. W., et al. (2003), Preparing for climatic change: The water, salmon, and forests of the Pacific Northwest, *Clim. Change*, *61*, 45–2003.

Mote, P. W., L. Brekke, P. Duffy, and E. Maurer (2011), *Guidelines for Constructing Climate Scenarios*, EOS, Transactions, Amer. Geophys. Union, *92*, doi:10.1029/2011EO310001.

North, G. R., T. L. Bell, and R. F. Cahalan (1982), Sampling errors in the estimation of empirical orthogonal functions, *Mon. Weather Rev.*, *110*, 699–706.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney (2007), Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions, *J. Clim.*, *20*, 4356–4376, doi:10.1175/JCLI4253.1.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler (2009), Selecting global climate models for regional climate change studies, *Proc. Natl. Acad. Sci. U. S. A.*, *106*, 8444–8446.

Pitman, A. J., and S. E. Perkins (2008), Regional projections of future seasonal and annual changes in rainfall and temperature over Australia based in skill-selected AR4 models, *Earth Interact.*, *12*, 1–50, doi:10.1175/2008EI260.1.

Polade, S. J., A. Gershunov, D. R. Cayan, M. D. Dettinger, and D. W. Pierce (2013), Natural climate variability and teleconnections to precipitation over the Pacific-North American region in CMIP3 and CMIP5 models, *Geophys. Res. Lett.*, *40*, 2296–2301, doi:10.1002/grl.50491.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *J. Geophys. Res.*, *108*(D14), 4407, doi:10.1029/2002JD002670.

Ropelewski, C. F., and M. S. Halpert (1986), North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO), *Mon. Weather Rev.*, *114*, 2352–2362.

Santer, B. D., et al. (2009), Incorporating model quality information in climate change detection and attribution studies, *Proc. Natl. Acad. Sci. U. S. A.*, *106*, 14,778–14,783.

Sheffield, J., A. D. Ziegler, E. F. Wood, and Y. Chen (2004), Correction of the high-latitude rain day anomaly in the NCEP–NCAR reanalysis for land surface hydrological modeling, *J. Clim.*, *17*, 3814–3828.

Sheffield, J., G. Goteti, and E. F. Wood (2006), Development of a 50-year resolution global dataset of meteorological forcings for land surface modeling, *J. Clim.*, *19*, 3088–3111.

Siler, N., G. Roe, and D. Durran (2013), On the dynamical causes of variability in the rain-shadow effect: A case study of the Washington Cascades, *J. Hydrometeor.*, *14*, 122–139, doi:10.1175/JHM-D-12-045.1.

Simonsen, I., A. Hansen, and O. M. Nes (1998), Determination of the Hurst exponent by use of wavelet transforms, *Phys. Rev. E*, *58*, 2779–2787.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tessier, Y., S. Lovejoy, P. Hubert, D. Schertzer, and S. Pecknold (1996), Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions, *J. Geophys. Res.*, *101*(D21), 26,427–26,440, doi:10.1029/96JD01799.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Quart. J. R. Meteorol. Soc.*, *131*, 2961–3012, doi:10.1256/qj.04.176.

Weigel, A., R. Knutti, M. A. Liniger, and C. Appenzeller (2010), Risks of model weighting in multimodel climate projections, *J. Clim.*, *23*, 4175–4191, doi:10.1175/2010JCLI3594.1.

Yu, J.-Y., Y. Zou, S. T. Kim, and T. Lee (2012), The changing impact of El Niño on US winter temperatures, *Geophys. Res. Lett.*, *39*, L15702, doi:10.1029/2012GL052483.