

# Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments

Levi D. Brekke · Michael D. Dettinger ·  
Edwin P. Maurer · Michael Anderson

Received: 1 December 2006 / Accepted: 6 November 2007 / Published online: 28 February 2008  
© Springer Science + Business Media B.V. 2007

**Abstract** Ensembles of historical climate simulations and climate projections from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset were investigated to determine how model credibility affects apparent relative scenario likelihoods in regional risk assessments. Methods were developed and applied in a Northern California case study. An ensemble of 59 twentieth century climate simulations from 17 WCRP CMIP3 models was analyzed to evaluate relative model credibility associated with a 75-member projection ensemble from the same 17 models. Credibility was assessed based on how models realistically reproduced selected statistics of historical climate relevant to California climatology. Metrics of this credibility were used to derive relative model weights leading to weight-threshold culling of models contributing to the projection ensemble. Density functions were then estimated for two projected quantities (temperature and precipitation), with and without considering credibility-based ensemble reductions. An analysis for Northern California showed that, while some models seem more capable at recreating limited aspects twentieth century climate, the overall tendency is for comparable model performance when several credibility measures are combined. Use of these metrics to decide which models to include in density function development led to local adjustments to function shapes, but led to limited affect

---

L. D. Brekke (✉)

Technical Service Center 86-68520, U.S. Bureau of Reclamation, Denver, CO 80225-0007, USA  
e-mail: lbrekke@do.usbr.gov

M. D. Dettinger

U.S. Geological Survey and Scripps Institution of Oceanography, La Jolla, CA 92093-0224, USA  
e-mail: mddettin@usgs.gov

E. P. Maurer

Civil Engineering Department, Santa Clara University, Santa Clara, CA 95053-0563, USA  
e-mail: emaurer@enr.scu.edu

M. Anderson

Division of Flood Management, California Department of Water Resources, Sacramento,  
CA 95821-9000, USA  
e-mail: manderso@water.ca.gov

on breadth and central tendency, which were found to be more influenced by “completeness” of the original ensemble in terms of models and emissions pathways.

## 1 Introduction

Resource managers currently face many questions related to potential climate changes, in particular how climate may change and what regional impacts would ensue. One of the most pressing questions is whether contemporary resource-management decisions might increase or decrease future impacts. To address these questions, analysts typically compile global climate projections and then spatially downscale them for impacts assessment at regional scales relevant to a given decision. Given that there are more than 20 global climate models currently in operation, producing simulations of future climate under several different greenhouse gas (GHG) emission scenarios (Meehl et al. 2005), it is not surprising that results drawn from such impacts assessments depend on the particular GHG forcing scenarios and climate models present in these compilations.

Focusing on a few projected scenarios (e.g., bookend analyses) can lead to significant divergence among the projected future impacts. While such analyses are useful for illustrating what may be at stake under different future scenarios, they provide limited guidance for management responses in the present (e.g., Brekke et al. 2004; Cayan et al. 2006; Hayhoe et al. 2004; Vicuna et al. 2007). Rather, it is important for managers to understand the distributed and consensus nature of projected climate change impacts (Dettinger 2005; Maurer 2007) so that managers can begin to consider the relative likelihood of future impacts rather than just isolated examples of potential impacts. In line with this philosophy, there has been a trend in recent impacts assessments to base the studies on larger multi-model projection ensembles, as for instance in recent studies of potential hydrologic impacts in California’s Central Valley (Maurer and Duffy 2005; Maurer 2007), the Colorado River Basin (Milly et al. 2005; Christensen and Lettenmaier 2006), and in other locations (Wilby and Harris 2006; Zierl and Bugmann 2005).

Expanding regional impacts analyses to consider larger climate projection ensembles creates the opportunity to address and communicate impacts in the terms of risks rather than isolated examples of possible impacts. The difference between risk and impact assessments is that risk goes beyond scenario definition and analysis of associated impacts to also address relative scenario likelihoods. Framing regional assessments in terms of risk is attractive from a management perspective because risk information is better suited for strategic planning, where responses can be formulated based upon weighted prospects of different impacts. This approach helps to guide the timely use of limited available funds that might support responses to inherently uncertain future conditions.

The product of a risk analysis is a presentation of distributed impacts from a collection of scenarios weighted by their estimated likelihoods. In the climate change context, *absolute* scenario likelihoods cannot be identified. However, *relative-* or *consensus-based* likelihoods of various scenarios can be estimated from an ensemble of climate projections by fitting a climate projection density function. Granted, such density functions only represent a limited portion of the climate change uncertainties, because elements such as social and physical factors affecting global GHG sources and sinks in the future, climate response and interactions with these GHG sources and sinks, and alternative climate model structures are not included. However, the ensemble density functions provide a more complete basis for using and interpreting the elements of the available ensembles than is afforded by scenario analyses without such context. Using projection density functions to

infer relative scenario likelihoods promotes strategic response planning, framed by perception of which outcomes are – at present – projected to be more likely among projected possibilities, which is a step forward from not considering scenario likelihoods at all.

Several methods for generating climate projections density functions have been proposed (Tebaldi et al. 2005; Dettinger 2006). The ensembles employed can include several greenhouse gas emission pathways, multiple climate models, and multiple “runs” of a given pathway-model combination differing by initial conditions. In applying these methodologies, it is natural to ask whether all members from a projection ensemble should be valued equally. Put another way, there are more than 20 coupled atmosphere–ocean climate models informing the last assessment from IPCC (2007): in the context of regional response analysis, are the projections from all of these models equally credible?

This latter thought motivates the two questions considered in this paper: (1) How does apparent model credibility at a regional scale, translated into relative model weighting and subsequent model culling, affect estimates of climate projection density functions?, and (2) How are relative scenario likelihoods, derived from the density function, affected when model credibility is considered when estimating the function?

To explore the first question, a philosophy is adopted that relative model credibility in projecting twenty-first century climate can be estimated from relative model accuracy in simulating twentieth century climate. Some studies have found little effect of weighting future climate projections by perceived differences between model completeness (e.g., Dettinger 2005, weighted models based on whether they required “flux corrections” or not to avoid climate drift and found little difference in estimated density functions). However, several interpretations of projection ensembles have been designed around the concept of weighting results by perceived historical accuracies (AchutaRao and Sperber 2002, 2006; Bader et al. 2004; Phillips et al. 2006). Following this approach, a procedure is developed here, beginning with a model credibility analysis to produce relative model credibility indices, as a basis for model culling, followed by a nonparametric procedure for estimating climate projection density functions, with and without consideration of model credibility results.

In the latter step, an “Uncertainty Ensemble” of climate projections is used to fit the density functions. In relation to the second question, a *subset* of the Uncertainty Ensemble is identified (i.e. Impacts Ensemble) for which relative scenario weights are derived from the density functions fit with and without consideration of model credibility. The nested nature of this Impacts Ensemble within the Uncertainty Ensemble represents a typical situation in regional risk assessment where the feasible size of an Impacts Ensemble (i.e. scenarios studied in detail for impacts in multiple resource areas) is less than what can be considered when fitting the climate projection density function because of the computational intensity of the impact calculations. The remainder of this paper presents the methodologies for model credibility and climate projection density analyses (section 2), results from applying these methods to a particular region (i.e. California’s Central Valley) where the concern is water resources impacts (section 3), a discussion of method limitations and areas of potential improvement (section 4), and a summary of major conclusions (section 5).

## 2 Methodology

The analytical sequence features two primary parts. The first part involves a model credibility analysis based on how well selected parts of the historical climate are simulated by the various models, and includes the following steps: (1) choosing relevant climate

variables and reference data, (2) choosing performance metrics and computing measures of model-to-observation similarities, and (3) deriving weights and culled model groups based on these measures. The second part is an analysis of climate projections where projection density functions are fit with and without consideration of results from the model credibility analysis.

The analyses are based on simulated climate variables from coupled atmosphere–ocean general circulation models (i.e. WCRP CMIP3 models) used to produce (a) twenty-first century climate projections under both SRES A2 and B1 emissions pathways (IPCC 2001) and (b) simulations for the “climate of the twentieth century experiment (20C3M),” conducted by CMIP3 participants (Covey et al. 2003). The Lawrence Livermore National Laboratory’s Program for Climate Model Diagnosis and Intercomparison (PCMDI) hosts a multi-model archive for 20C3M historical simulations, twenty-first century projections, and other scenario and control run datasets. Among the models producing 20C3M simulations, the number of available “runs” varied per model, with runs differing by initialization decisions. An attempt was made to focus on models that had simulated both A2 and B1 on the grounds that these pathways represent a broad and balanced range of SRES possibilities (IPCC 2001). However, this criterion was relaxed for two of the selected models from which only A2 or B1 simulations were available.

In total, 17 climate models were represented in our survey (Table 1). Collectively they were used to produce 59 20C3M simulations (used for the model credibility analysis) and 75 climate projections comprised of 37 SRES A2 simulations and 38 SRES B1 simulations. The set of 75 climate projections served as the Uncertainty Ensemble, mentioned in section 1, and was used in climate projection density analysis. From the Uncertainty Ensemble, *subsets* of 11 SRES A2 and 11 SRES B1 projections were identified as a 22-member Impacts Ensemble (Table 1). The selected 22 projections are the same projections considered in a previous study on potential hydrologic impacts uncertainty within the Sierra Nevada (Maurer 2007).

## 2.1 Credibility analysis: choosing simulated and references climate variables

The first step in the model credibility analysis involved choosing simulated climate variables relevant to the geographic region of interest (i.e. Northern California in this case study), and identifying climate reference data to which simulated historical climates could be compared (i.e. 20C3M results listed in Table 1). Three types of variables were used: local variables that define Northern California climatology, distant variables that characterize global-scale climatic processes, and variables that describe how global processes relate to the local climatology (i.e. teleconnections). This mix was chosen because the first interest for this regional scale assessment was how the local climate variables might change in the future. However, in the projections considered here, those changes are driven and established by global scale forcings (GHGs) and resulting processes. Thus, both local and global performances are important. Furthermore, the connection between global and local processes must be accurately recreated in the models (indicated by either the presence or absence of significant inter-variable correlation, i.e. teleconnections) in order for their simulated local responses to global forcings to be considered reliable.

Two local variables were used to describe the regional climatology: surface air temperature and precipitation conditions (i.e. NorCalT and NorCalP near {122W, 40N}). Global-scale phenomena driving the regional climatology via teleconnections include pressure conditions over the North Pacific (related to mid-latitude storm track activity upwind of North America) and the phase of the El Niño Southern Oscillation (ENSO;

**Table 1** Climate projections and models included in this case study

| WCRP CMIP3<br>model I.D. <sup>a</sup> | Model abbreviation in<br>this study | Model<br>number | Projection run numbers <sup>b</sup> |       |         |    | 20C3m<br>Ensemble |
|---------------------------------------|-------------------------------------|-----------------|-------------------------------------|-------|---------|----|-------------------|
|                                       |                                     |                 | Uncertainty                         |       | Impacts |    |                   |
|                                       |                                     |                 | A2                                  | B1    | A2      | B1 |                   |
| CGCM3.1(T47)                          | cccma_cgcm31                        | 1               | 1...5                               | 1...5 |         |    | 1...5             |
| CNRM-CM3                              | cnrm_cm3                            | 2               | 1                                   | 1     | 1       | 1  | 1                 |
| CSIRO-Mk3.0                           | csiro_mk30                          | 3               | 1                                   | 1     | 1       | 1  | 1...3             |
| GFDL-CM2.0                            | gfdl_cm20                           | 4               | 1                                   | 1     | 1       | 1  | 1...3             |
| GFDL-CM2.1                            | gfdl_cm21                           | 5               | 1                                   | 1     |         |    | 1...3             |
| GISS-ER                               | giss_model_er                       | 6               | 1                                   | 1     | 1       | 1  | 1...9             |
| INM-CM3.0                             | inmcm3_0                            | 7               | 1                                   | 1     | 1       | 1  | 1                 |
| IPSL-CM4                              | ipsl_cm4                            | 8               | 1                                   | 1     | 1       | 1  | 1                 |
| MIROC3.2(hires)                       | miroc32_hires                       | 9               |                                     | 1     |         |    | 1                 |
| MIROC3.2(medres)                      | miroc32_medres                      | 10              | 1...3                               | 1...3 | 1       | 1  | 1...3             |
| ECHAM5/MPI-OM                         | mpi_echam5                          | 11              | 1...3                               | 1...3 | 1       | 1  | 1...3             |
| MRI-CGCM2.3.2                         | mri_cgcm232a                        | 12              | 1...5                               | 1...5 | 1       | 1  | 1...5             |
| CCSM3                                 | ncar_ccsm30                         | 13              | 1...5                               | 1...8 |         |    | 1...8             |
| PCM                                   | ncar_pcm1                           | 14              | 1...4                               | 2...3 | 1       | 2  | 1...4             |
| UKMO-HadCM3                           | ukmo_hadcm3                         | 15              | 1                                   | 1     | 1       | 1  | 1...2             |
| UKMO-HadGEM1                          | ukmo_hadgem1                        | 16              | 1                                   |       |         |    | 1...2             |
| ECHO-G                                | miub_echo-g                         | 17              | 1...3                               | 1...3 |         |    | 1...5             |
| Total Runs                            |                                     |                 | 37                                  | 38    | 11      | 11 | 59                |

<sup>a</sup>From information at Lawrence Livermore National Laboratory's Program for Climate Model Diagnosis and Intercomparison (PCMDI), September 2006: <http://www-pcmdi.llnl.gov>

<sup>b</sup>Run numbers assigned to model- and pathway-specific SRES projections and model-specific 20c3m simulations in the WCRP CMIP3 multi-model data archive at PCMDI.

affecting Pacific-region interannual climate variability and beyond). Two measures of these phenomena were used in this study, respectively: the North Pacific Index (NPI), describing mean sea level pressure within {30N–65N, 160E–140W} and the Nino3 Index describing ENSO-related mean sea surface temperature within {5S–5N, 150W–90W}.

Monthly time series of each evaluation variable were extracted from each 20C3M simulation for the latter half of the twentieth century (1950–1999). Likewise, monthly 1950–1999 “reference” data were also obtained. For NPI, NorCalP, and NorCalT, the reference data were extracted from the NCEP Reanalysis (Kalnay et al. 1996, updated and provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at <http://www.cdc.noaa.gov/>), which is a data set of historical observations modified and interpolated with the use of an atmospheric climate model and which describes climate conditions at roughly the same scale as the coupled climate models used in the historical simulations. The reference data for Nino3 were obtained from the Monthly Atmospheric and SST Indices archive provided by the NWS Climate Prediction Center, Camp Springs, Maryland, USA, from their Web site at <http://www.cpc.noaa.gov/data/indices/>.

Simulated and reference time series were compared during the 1950–1999 period. It is arguable whether this 50-year historical period should be shorter and more recent. The decision to consider 1950–1999 as opposed to a shorter period was driven by recognition that 20C3M simulations express interdecadal variability that might be out of phase with that

observed during the twentieth century, leading to exaggerated simulation-reference differences or similarities based solely on choice of sampling overlap period.

## 2.2 Credibility analysis: choosing performance metrics and teleconnections

The next step was to choose performance metrics that describe statistical aspects of the local and global variables and their teleconnections. For each variable, a set of six metrics was evaluated (Table 2): the first three statistical moments of annual conditions (mean, variance, skewness), an interdecadal variance describing lower frequency variability (Table 2, note 4), amplitude of seasonality defined by the range of mean monthlies, and phase of seasonality defined by correlation between simulated and reference mean monthlies. For the local variables (i.e. NorCalT and NorCalP), the characteristics of extreme positive anomalies were considered (i.e. the annual maximum monthly value exceeded in 10% of years). Seasonal correlations (teleconnections) between the two local and two global variables were considered, as well as seasonal and annual correlations between the two global variables. For NorCalT, the 50-year trend was used as an additional variable-specific metric. For NorCalP, a metric of drought recurrence and severity was also included and framed by knowledge of relevant 1950–1999 droughts in the case study region. Because the most significant sustained drought in the Northern California instrumental record was about 6 years in duration (1987–1992), the drought metric was defined as the running 6-year precipitation total exceeded by 90% of the 6-year spells within each 50-year time series. Finally, as a measure of how temporally realistic the simulated ENSO processes were, and because local interannual variability is influenced by ENSO, a metric describing El Niño reoccurrence was defined as the spectral power of the 50-year Nino3 time series concentrated in the 2-to-7 year range.

This is a fairly extensive array of metrics, and it is difficult to know which of the metrics are most pertinent to the projection of future impacts of increasing GHGs. Application of this methodology thus will involve consideration of which metrics are the most relevant measures of model credibility. In applications of this approach to other regions, decisions to include other, or additional, metrics beyond those discussed herein should depend on the region and climate in question. Although no definitive statements can be made as to which metrics are the more relevant, it is reasonable to expect that different impacts assessment perspectives might gravitate toward different metrics. For illustration purposes, three perspectives are defined and used here to explore sensitivity of impressions to this question. The perspectives are loosely termed Water Supply, Hydropower, and Flood Control. For each perspective, the 49 metrics of Table 2 were reduced to six arbitrarily chosen metrics as being “more relevant.”

- For the Water Supply perspective, managers are assumed to have the ability to seasonally store precipitation-runoff and thus might be more concerned about how climate models reproduce the past precipitation in terms of long-term mean and seasonality phase, past temperature in terms of long-term trend and seasonality phase, multi-year drought severity, and global teleconnections relevant to the precipitation season (e.g., Nino3-NorCalP during Winter).
- For the Hydropower perspective, concerns were assumed to be similar to those of Water Supply (e.g., precipitation long-term mean, temperature long-term trend), but with more concern shifted to other types of global teleconnections with local climate (e.g., NPI correlation with NorCalP during autumn and winter) and on global interannual variability in general as it might affect hydro-energy resources from a larger

**Table 2** Variables and performance metrics used in this model credibility study

| Performance metrics                            | Metrics by climate variable, 1950–1999 monthly data, describing:<br>[A] global, [B] local, [C] teleconnections |                                |                   |       |
|--|--|--------------------------------|-------------------|-------|
|  | NPI <sup>a</sup>   | NorCalP                        | NorCalT           | Nino3 |
| Mean <sup>b</sup>                              | [A]  | <b><u>[B]</u></b> <sup>m</sup> | [B]               | [A]   |
| Variance <sup>c</sup>                          | [A]  | <u>[B]</u>                     | [B]               | [A]   |
| Interdecadal variance <sup>d</sup>             | [A]  | [B]                            | [B]               | [A]   |
| Skewness <sup>c</sup>                          | [A]  | <u>[B]</u>                     | [B]               | [A]   |
| Seasonality amplitude <sup>e</sup>             | [A]  | <u>[B]</u>                     | [B]               | [A]   |
| Seasonality phase <sup>f</sup>                 | [A]  | <b>[B]</b>                     | <b>[B]</b>        | [A]   |
| 6-year mean, 90%exc <sup>g</sup>               |  | <b>[B]</b>                     |                   |       |
| Annual maximum month, 10% exc <sup>h</sup>     |  | <u>[B]</u>                     | [B]               |       |
| Trend in annual mean (50-year)                 |  |                                | <b><u>[B]</u></b> |       |
| El Niño reoccurrence <sup>i</sup>              |  |                                |                   | [A]   |
| Correlation with Nino3 during OND <sup>j</sup> | [C]  | [C]                            | [C]               |       |
| Correlation with Nino3 during JFM              | [C]  | <b>[C]</b>                     | [C]               |       |
| Correlation with Nino3 during AMJ              | [C]  | [C]                            | [C]               |       |
| Correlation with Nino3 during JAS              | [C]  | [C]                            | [C]               |       |
| Correlation with NPI during OND                |  | <u>[C]</u>                     | [C]               |       |
| Correlation with NPI during JFM                |  | <u>[C]</u>                     | [C]               |       |
| Correlation with NPI during AMJ                |  | [C]                            | [C]               |       |
| Correlation with NPI during JAS                |  | [C]                            | [C]               |       |
| Correlation with Nino3, annually <sup>k</sup>  | [C]  |                                |                   |       |

OND October through December, JFM January through March, AMJ April through June, JAS July through September

<sup>a</sup> NPI (North Pacific Index) is defined as monthly mean sea level pressure within (30N–65N, 160E–140W), Nino3 is defined as monthly mean sea surface temperature within (5S–5N, 150W–90W), and NorCalP and NorCalT are monthly precipitation and surface air temperatures near 122W and 40N.

<sup>b</sup> Mean annual total for NorCalP; mean annual average for other variables.

<sup>c</sup> Computed on annual total or mean values (see note b).

<sup>d</sup> Computed on annual values smoothed by a 9-year moving average.

<sup>e</sup> Computed on monthly means, identifying the difference between maximum and minimum values.

<sup>f</sup> Computed as correlation between simulation and climate reference monthly means (see note l).

<sup>g</sup> 90% exceedence value in sorted series of running 6-year mean-annual values.

<sup>h</sup> 10% exceedence value in sorted series of annual maximum month values.

<sup>i</sup> Based on spectral analysis of the time series phase variability (see note m), identifying the average power in 2- to 7-year period band.

<sup>j</sup> Computed as correlation between seasonal mean conditions between the indicated variable pair (row and column; see abbreviations).

<sup>k</sup> Same as note j, but computed as correlation between annual mean conditions.

<sup>l</sup> Climate reference for NPI, NorCalP, and NorCalT from NCEP/NCAR Reanalysis monthly data products at NOAA Climate Diagnostic Center; Climate reference conditions for Nino3 from monthly index values at NOAA Climate Prediction Center.

<sup>m</sup> Processing involves removing monthly means, and then scaling total power by normalizing the mean-removed time series to unit variance.

<sup>n</sup> Metric Subsets: Water Supply (bold), Hydropower (italics), Flood Control (underline).

- regional hydropower market encompassing the runoff region of interest (e.g., Nino3 El Niño reoccurrence).
- For the Flood Control perspective, more focus was assumed to be placed on how the climate models recreate more extreme aspects of precipitation climatology (e.g., precipitation skewness, seasonality amplitude, annual maximum month that is exceeded in 10% of the 50-year evaluation period).

Notably, historical simulations were not rated in terms of whether they reproduced precipitation trends of the past 50 years. As noted previously, the region in question is subject to large, significant and persistent multidecadal climate fluctuations historically, under the influence of multidecadal climate processes over the Pacific Ocean basin and beyond (e.g., Mantua et al. 1997; McCabe et al. 2004). Historically the multidecadal Pacific Ocean influence has resulted in significant and persistent climatological differences between the 1948–1976 period and the 1977–1999 period. Although these long-term differences may very well be mostly natural, random and reversible, they have imposed a trend-like character on the Northern California climate during 1950–1999. Even a skillful coupled ocean–atmosphere model, initiated much earlier in the nineteenth or twentieth centuries, would not be expected to reproduce the timing of such natural multidecadal fluctuations in a way that would reproduce the trend-like halving of the 1950–1999 window considered here. Thus the presence or absence of a regionalized precipitation trend in observations and apparent difference in historical simulated trend did not seem to be a good measure of the simulation skill in the particular study region considered here.

### 2.3 Credibility analysis: deriving model weights and model culling

After computing simulation metrics, run-specific calculations of simulated-minus-reference metric differences were pooled by model and averaged to produce 17 model-representative differences. A distance-based methodology was then used to measure overall model-to-reference similarities for each set of metrics. Under the distance-based philosophy, a distance is computed within a “similarity space” defined along “metric dimensions.” For example, the similarity space could be a seven-dimensional space spanned by the seven NPI performance metrics, or a 49-dimensional space spanned by all performance metrics in Table 2. Given a space definition, distance can be computed using one of several distance formulas. Euclidean or Manhattan distance formulas were explored in this study (Black 2006), with focus ultimately placed on Euclidean distance. Results were found to be insensitive to choice of distance formula, primarily because metric differences were scaled to have unit variance across models for each metric, prior to distance calculation, so that metric differences generally all had magnitudes near or less than one. Such magnitudes aggregate into similar distances using the Euclidean and Manhattan formulas.

The purpose of scaling metric differences was to prevent metrics measured in large units from dominating the computed distance (e.g., the El Niño reoccurrence metric differences have values on the order of  $10^3$  where as the seasonality and teleconnection correlation-metrics have differences on the order of  $10^{-1}$  to  $10^{-2}$ ). A disadvantage of scaling the metric differences is that it can exaggerate a metric’s influence on model discrimination even though pre-scaled metric differences were quite similar (e.g., simulated NorCaT seasonality phase and difference from reference).

For a given set of metrics, the procedure results in computation of 17 model-representative distances from references. Relative model weights were then computed as the inverse of this distance. Finally, a threshold weight criterion was used to cull models from

consideration in the subsequent climate projection density analysis. The model-culling depends on the metric set used and the threshold model weight selected to differentiate between models that will be retained and those that will not. For illustration, in this study, the weight threshold was defined to be median among the 17 weights, so that the nine highest weighted models were retained from among the 17 considered.

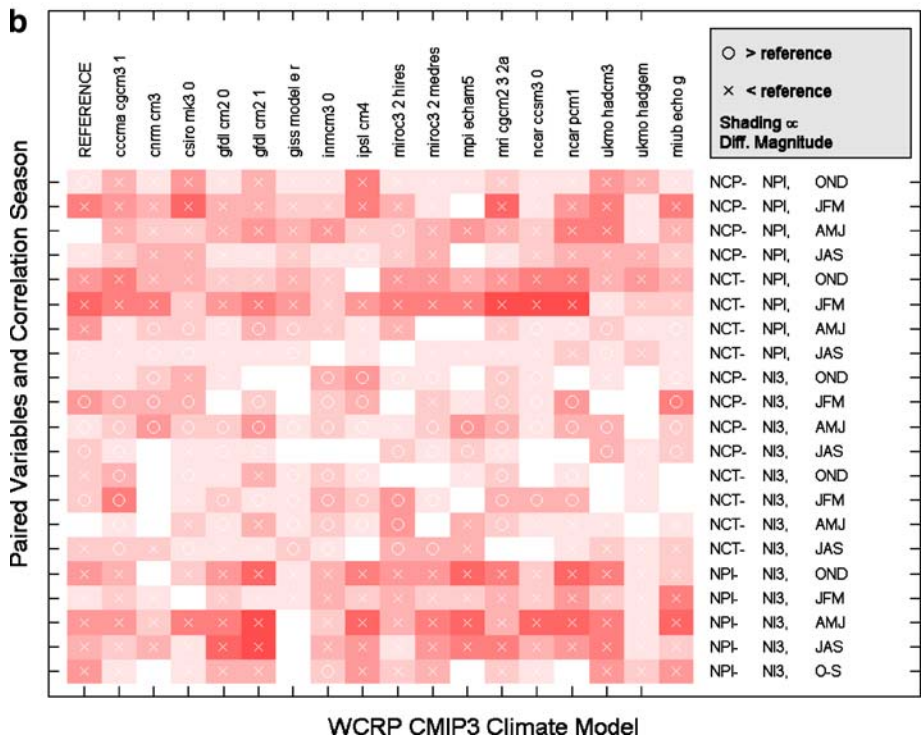
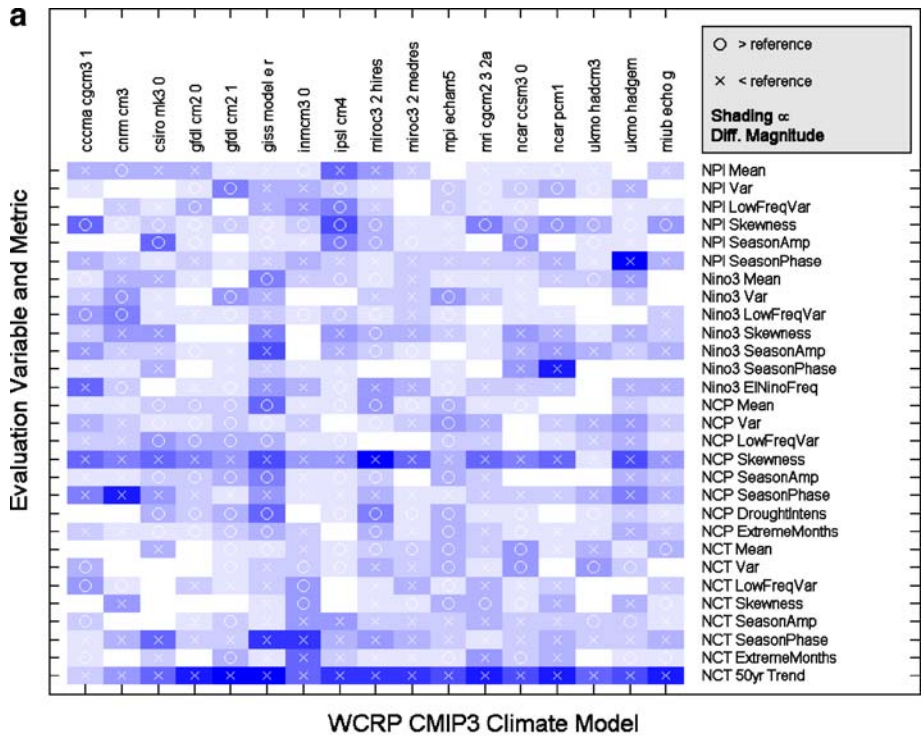
Looking ahead to the climate projection density analysis, the use of these credibility analysis results could have involved proportional weighting of the models rather than culling of models. All models could have been retained and various methods to proportionally represent model contribution in the projection density functions could have been used. This alternate approach was explored, with density functions fit using nonparametric techniques (section 2.4). However, it led to excessively multi-modal, “peaky” density functions, set up by the interspersed positions of fitting data (i.e. specific projections) from “less credible” models and “more credible” models, and was abandoned for the culling-based approach used here.

#### 2.4 Climate projection density analysis, with and without model credibility

Density functions were constructed for projected anomalies of 30-year average “annual total precipitation” and “annual mean surface air-temperature” [i.e.  $d(P)$  and  $d(T)$ , respectively], evaluated for the 2010–2039 and 2040–2069 periods relative to a 1950–1999 base period. Several methodologies have been proposed for developing density functions that describe likelihoods of univariate or multivariate climate projections (Tebaldi et al. 2005; Dettinger 2006). An empirical procedure is used here, involving nonparametric density estimation using Gaussian kernels (Scott 1992; Wilks 1995) with optimized bandwidths (Silverman 1986). For the multivariate case of jointly projected anomalies of temperature and precipitation, a product-kernel extension of the univariate approach is used (Scott 1992). Nonparametric density estimation has been applied in numerous statistical-hydrology studies (e.g., Lall et al. 1996; Piechota et al. 1998). It will be shown that very similar joint density functions are obtained by using another estimation approach (Dettinger 2006). The emphasis here is on the common motivation underlying these methods: to consolidate projection information into distributions that help focus planning attention on ensemble consensus rather than extremes (Dettinger 2006), and whether relative model credibility should be factored into this consolidation.

A key decision in estimating the density functions was how to deal with the various numbers of simulations available from a given pathway-model combination (e.g., model CCSM3 contributes eight SRES B1 simulations whereas model PCM contributes two). In the present study, all simulations from all contributing models have been treated as equals. Just as the culling approach used here could have been replaced with a weighting of all the models, this assignment of equal weights to all of the simulations could have been replaced by weightings of the contributions from various simulations that avoided overemphasis of simulations from the more prolific modeling groups. Such weightings were explored in this study and tended to yield results similar to the distributions shown herein, especially with respect to the central tendencies and spans of the density functions.

In applications of the product kernel for bivariate density estimation, relative variable scales and choices for variable-specific domain resolution and range can influence results. To account for this influence, each univariate function contributing to the product kernel was fit to anomalies scaled by their respective standard deviations. After constructing the bivariate density function from these scaled data, the function values relative to each scaled anomaly position were mapped back into their unscaled values.



**Fig. 1** **a** Scaled model-specific average difference between multiple 20c3m run results and Reference for metric types [A] and [B] in Table 2. Scaling involves pooling run- and metric-specific differences across models and scaling collectively to unit variance. *Shading* shows magnitude of scaled difference, with *darker shading* showing greater magnitude of difference. “X” and “O” symbols are used to indicate the sign of difference from reference. **b** Similar to **a**, but for metric type [C] in Table 2

### 3 Case study results – Northern California

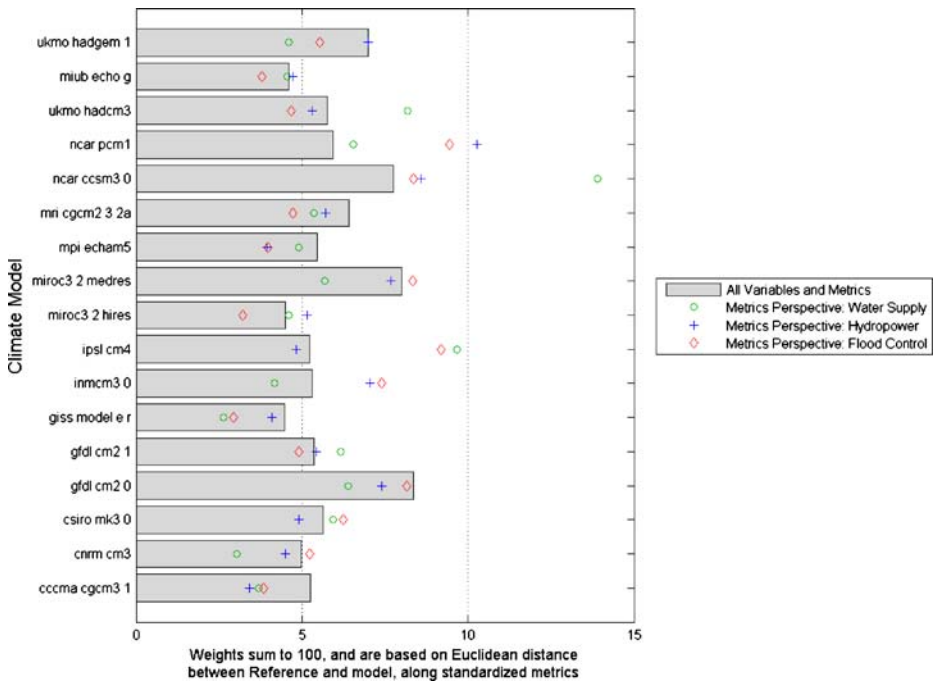
#### 3.1 Credibility analysis

As mentioned in section 2, the case study region for this study was Northern California, leading to the focus on two relevant local climate variables for credibility analysis (NorCalP and NorCalT), two global variables influential on local variables (NPI and Nino3), and respective global–local teleconnections. Summaries of scaled, model-representative, metric differences between simulated 20C3M results and observational references are shown on Fig. 1a and b. The figures qualitatively indicate relative differences among models for each metric. They do not indicate *specific* differences for a given model and metric. For example, consider differences between each models’s average-20C3M NPI Mean and Reference NPI Mean (Fig. 1a, top row). The figure shows shading that scales from light to dark as the *magnitude of a difference* increases; the *sign* of the difference is indicated by “x” for negative and “o” for positive. Results suggest that “mpi echam5” and “ukmo hadgem” generally did a better job reproducing Reference NPI Mean. As another example, consider the bottom row of Fig. 1a, which shows that the models consistently underpredicted the Reference trend in NorCalT during 1950–1999. However, what isn’t shown on Fig. 1a (because specific differences are not shown) is that all models correctly simulated a *warming* trend, just not enough warming compared to Reference.

Figure 2 indicates relative model weights derived for each model based on the metric values indicated in Table 2 (i.e. All Variables and Metrics, and metric sets related to the Water Supply, Hydropower, and Flood Control perspectives). The latter three sets were defined and discussed in section 2.2. For the “All Variables and Metrics” case, which incorporated all 49 metrics in Table 2, the relative model weight varies among models by roughly a factor two. Projections from the “gfdl cm2 0,” “mirc3 2 medres,” and “ncar ccs3 0” models would be granted more credibility in this context (Fig. 2). For the “gfdl cm2 0” and “mirc3 2 medres” models, their greater weights stem from scoring well in multiple variable-specific subsets. The greater weight for the “ncar ccs3 0” model was obtained more by scoring well in the NorCalP subset. For the three perspectives, which focused on considerably fewer metrics, the range of relative model weights grows to a factor of 3 to 4.

Retaining models having weight greater than or equal to the median weight among the 17 model-specific values, Table 3 shows groups of retained models based on each set of model weights from Fig. 2. The mix of retained models differs depending on which variables were used to decide credibility. This is particularly the case if credibility is determined by fewer simulation metrics. The ensemble of coupled climate models providing projections includes fairly wide ranges of credibility when individual metrics are considered, but have more similar credibility when the intercomparison is made across a suite of simulated variables. That is, generally, a model may do very well on one metric but not another, and overall these differences average out for most model-to-model comparisons when several dozen metrics are brought to bear.

The effect on deciding model retention of various choices of metric sets was explored, with a subset of results illustrated on Fig. 3, which shows how model retention varies for



**Fig. 2** Model Weights computed based on different metric sets (see Table 3). For a given metric set, model weights are scaled collectively to sum to 100

each of the combinatorial possibilities of one- to eight-metric sets from the type [B] metrics associated with NorCalP (Table 2). Results show that, while model retention varies with the metric set used, some models would be more frequently retained and thus are could be considered to be the relatively more credible models for simulating NorCalP (e.g., “ukmo hadgem1,” “ncar pcm,” “ncar ccs3 0,” “miroc3 2 medres,” “ipsl cm4,” “inmcm3 0,” and “gfdl 2 0”). Next generation models (e.g., “gfdl cm 2 1” compared to “gfdl cm 2 0”) and higher resolution models (e.g., “miroc3 2 hires” compared to “miroc3 2 medres”) do not necessarily fare better than their predecessors in such credibility evaluations.

### 3.2 Climate projection density functions

The results from Table 3 were carried forward to the construction of climate projection density functions. Prior to fitting density functions, ensembles of projected time series (Table 1) for surface air temperature and precipitation anomalies, as simulated near {122W, 40N}, were extracted from the A2 and B1 simulations retained in the previous step. Anomalies were computed as deviations of the projected monthly values from the model’s 1950–1999 20C3M monthly means. Projected anomalies were then bias-corrected to account for model tendencies relative to observations on the projected quantities (i.e. NorCalP and NorCalT from Reanalysis). Bias-correction was performed on a month-specific basis by multiplying projected anomalies by the ratio of Reanalysis monthly means to the model’s 20C3M monthly means. After bias-correction, monthly anomalies were consolidated into annual mean surface air temperature anomalies and annual total precipitation anomalies for each the 75 projections (Fig. 4).

**Table 3** Model membership in projection ensemble after culling by model credibility using different metric sets<sup>a,b</sup>

| WCRP CMIP3 Model I.D. <sup>c</sup> | Metric Set <sup>b</sup>   |                      |                    |                       |
|------------------------------------|---------------------------|----------------------|--------------------|-----------------------|
|                                    | All variables and metrics | Water supply metrics | Hydropower metrics | Flood control metrics |
| CGCM3.1(T47)                       |                           |                      |                    |                       |
| CNRM-CM3                           |                           |                      |                    | x                     |
| CSIRO-Mk3.0                        | x                         | x                    |                    | x                     |
| GFDL-CM2.0                         | x                         | x                    | x                  | x                     |
| GFDL-CM2.1                         |                           | x                    | x                  |                       |
| GISS-ER                            |                           |                      |                    |                       |
| INM-CM3.0                          |                           |                      | x                  | x                     |
| IPSL-CM4                           |                           | x                    |                    | x                     |
| MIROC3.2(hires)                    |                           |                      |                    |                       |
| MIROC3.2(medres)                   | x                         | x                    | x                  | x                     |
| ECHAM5/MPI-OM                      | x                         |                      |                    |                       |
| MRI-CGCM2.3.2                      | x                         | x                    | x                  |                       |
| CCSM3                              | x                         | x                    | x                  | x                     |
| PCM                                | x                         | x                    | x                  | x                     |
| UKMO-HadCM3                        | x                         | x                    | x                  |                       |
| UKMO-HadGEM1                       |                           |                      |                    |                       |
| ECHO-G                             | x                         |                      | x                  | x                     |

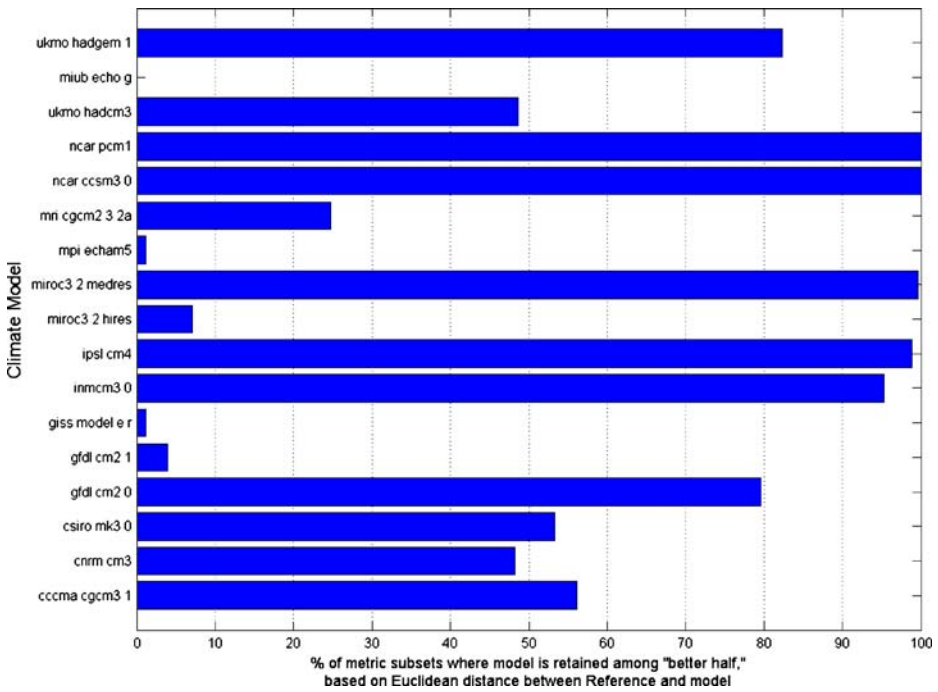
<sup>a</sup>Based on evaluation of models' 20c3m Euclidean similarity to Reference (Table 2, note 1).

<sup>b</sup>First column considers all variables and metrics from Table 2. Remaining three columns consider six metrics chosen as relevant to three impacts perspectives (Table 2, note n).

<sup>c</sup>WCRP CMIP3 Model I.D. explained in Table 1, note a.

Each time series of projected annual anomalies was averaged over the periods 2010–2039 and 2040–2069, leading to two 75-member pools of projected “30-year mean” anomalies (i.e. projected “climatological” anomalies) for density function fitting. Density functions for projected climatological temperature anomalies [ $d(T)$ ] and precipitation anomalies [ $d(P)$ ] are shown on Fig. 5a and b, respectively. Density functions were constructed for each projected quantity and period for five cases: “No Model Culling,” meaning that functions were fit to all 75 projections listed in Table 1, and the four basis metric sets used for model culling (Table 3, columns 2–5). Anomaly positions of the 22-member Impacts Ensemble (sections 1 and 2) are also shown on the horizontal axes of each figure. Density functions for jointly projected climatological anomalies for temperature and precipitation [ $d(T,P)$ ] are shown on Fig. 6a and b, for the 2040–2069 period only and respectively for the “No Model Culling” and “Cull Basis: Water Supply” (Table 3, column 3). Also shown on Fig. 6a and b are two density surfaces, one estimated using product kernel technique described in section 2, and another using a second estimation method described by Dettinger (2006). The similarity of the estimated surfaces suggests that choice of estimation methods is not crucial here.

Focusing on how  $d(T)$  varies with the choice of retained models, it is clear that the choice of models led to some changes in density magnitudes within the functions. However, comparison of the functions for the non-culled and culled cases shows that the general spread and central tendencies of the density functions are not drastically affected by the how model credibility assessment was used to cull models. Moreover, the positions of the dominant modes are generally consistent. It seems that the 75-member ensemble of



**Fig. 3** Sensitivity of model culling results to choice of NorCalP metrics subset (Table 2). All combinations of one to eight NorCalP metrics are considered. For each metrics set, relative model weights were computed, and a “greater than or equal to median weight” criterion was used to determine model-membership in the projection ensemble. Model-membership frequency was then assessed across metric sets, shown here as a percent-frequency

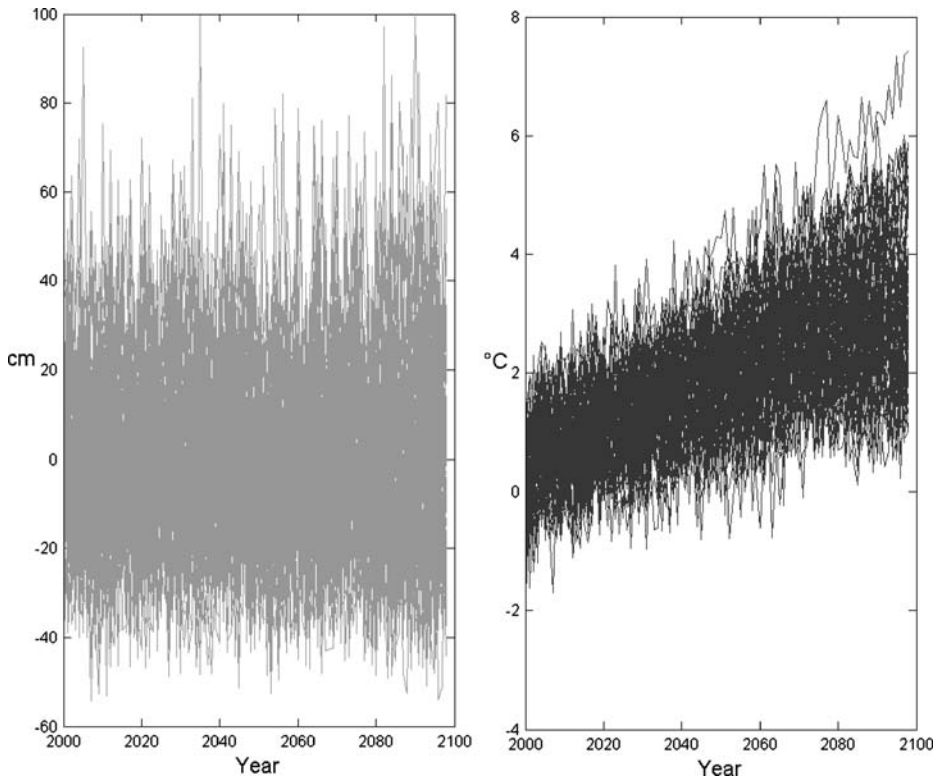
projections included sufficient scatter and structure so that the spread and central tendency of  $d(T)$  could be captured with any of a large number of possible subsets and weightings.

For  $d(P)$ , the decision of which models to retain or emphasize was more influential. The central tendency of  $d(P)$  shifted to a more negative anomaly values compared to the “no change” central value obtained from the full 75-member ensemble. That said, once the less credible models were dropped, the choice on which metric basis to use for culling seemed to be less significant, and the central tendencies and spread of  $d(P)$  functions were relatively similar.

Comparison of  $d(T,P)$  based on “No Model Culling” and “Cull Basis: Water Supply” reflects a combination of the impressions drawn from the various  $d(T)$  and  $d(P)$  functions. Like  $d(T)$ , the breadth and central tendency of the  $d(T,P)$  relative to the  $T$ -axis is relative unaffected by decision to cull models. And like  $d(P)$ , the decision to cull models using the Water Supply perspective causes the peak of the density surface to shift toward a more negative anomaly position.

### 3.3 Using climate projection density functions to derive scenario weights

Having fit climate projection density functions, the focus now shifts to the nested set of projection members that might be studied for detailed impacts (i.e. the Impacts Ensemble, described in sections 1 and 2), and their respective plotting positions within each of the density functions. The purpose is to assign relative scenario weights based on scenario

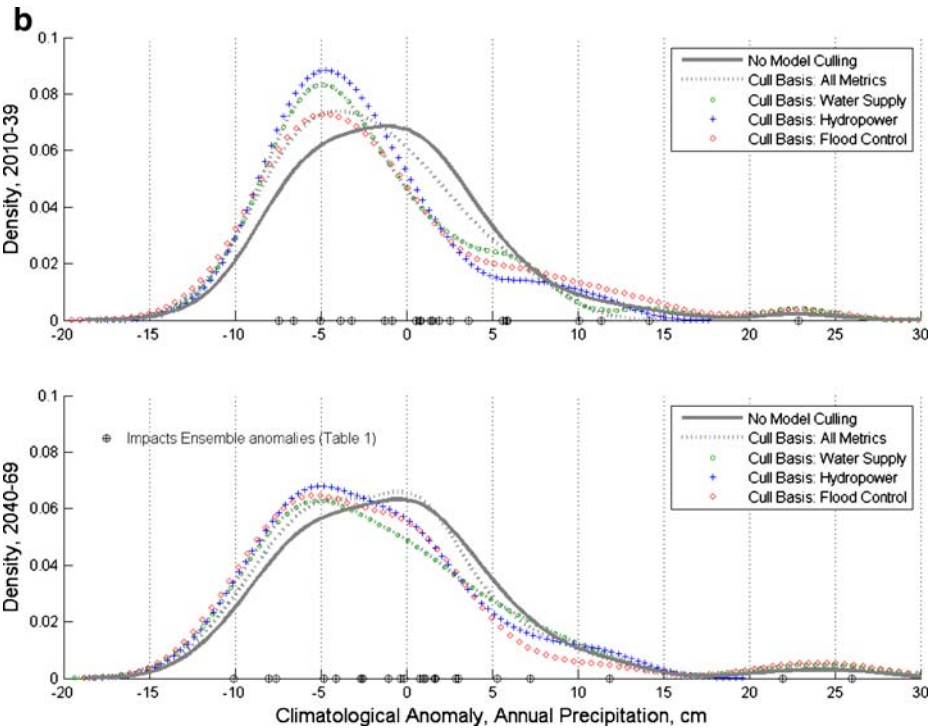
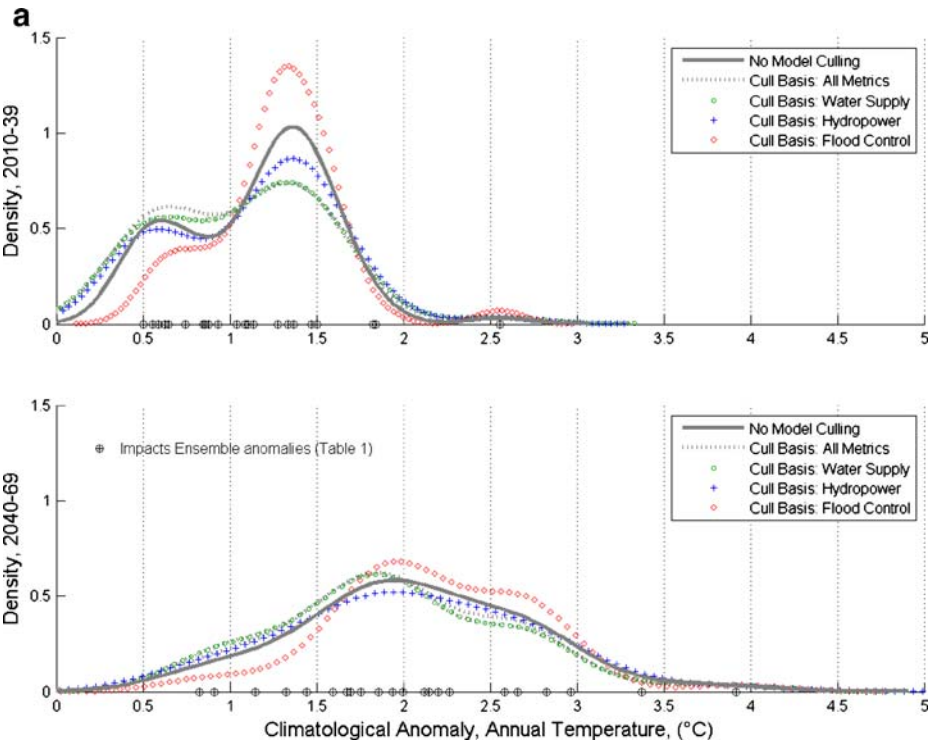


**Fig. 4** Projected annual anomaly time series, computed relative to 1950–1999 NCEP Reanalysis (Kalnay et al. 1996) annual total precipitation (cm) and mean annual surface air temperature ( $^{\circ}\text{C}$ ) in Northern California near  $\{122\text{W}, 40\text{N}\}$ . Time series are from the 75 projection ensemble listed in Table 1

densities within either  $d(T)$ ,  $d(P)$ , or  $d(T,P)$ . As mentioned, the Impacts Ensemble includes 22 of the 75 projection members used to estimate the density functions. The positions of those 22 members are shown on the horizontal axes of Fig. 5a and b, and as circle-cross symbols overlaying the larger “x” symbols on Fig. 6a and b.

Scenario-specific point-densities were identified from six projection distributions:  $d(T)$  (from Fig. 5a),  $d(P)$  (from Fig. 5b), or  $d(T,P)$  (Fig. 6a and b), from both the “No Model Culling” and “Cull Basis: Water Supply” functions. These point-densities were then considered in aggregate to imply *relative* scenario likelihoods in the context of a detailed and computationally intensive risk assessment based on these 22 scenarios. Each of the six sets of scenario densities were translated into corresponding sets of scenario weights (Fig. 7) by rescaling each set of 22 densities so that they sum to 22 (i.e. default scenario weight would be one, and a set of 22 density-based weights would have a mean of one).

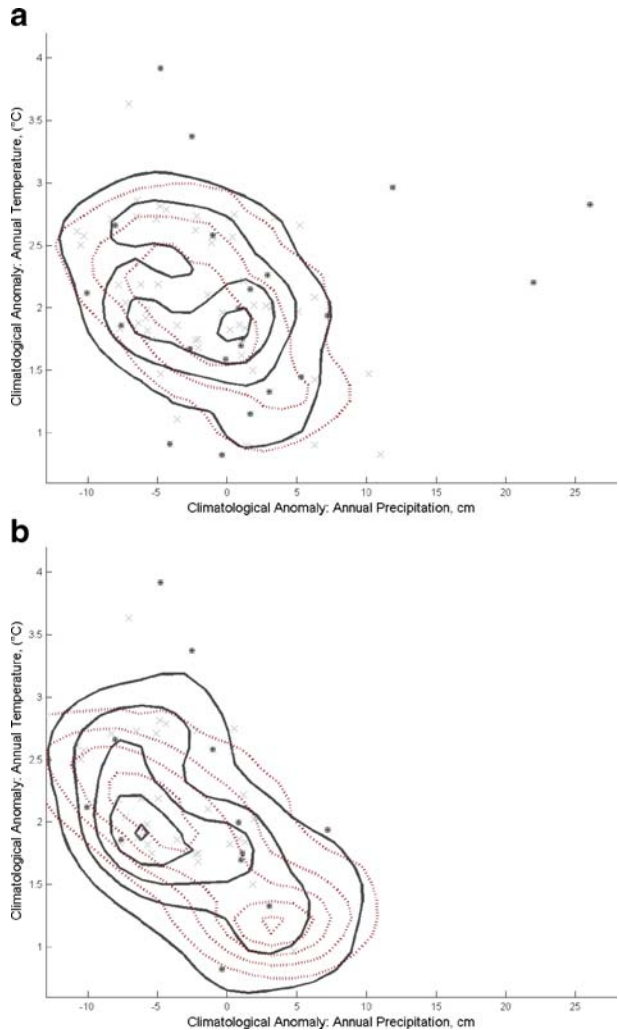
When focus is placed on a projected quantity or joint-quantities, particular choices of models included in the density estimation process had minimal effect on the relative magnitudes of scenario weights [e.g., compare weights from  $d(T)$  fit with models from “No Model Culling” versus models from “Cull Basis: Water Supply”]. More significantly, however, the choice of projected quantity was very significant in determining relative scenario weights [e.g., compare weights from  $d(T)$  relative to weights from  $d(P)$  or  $d(T,P)$ ]. Questions remain as to which projected quantities should steer integration of impacts for the assessment of risk. Addressing this question may be a more important decision for the risk

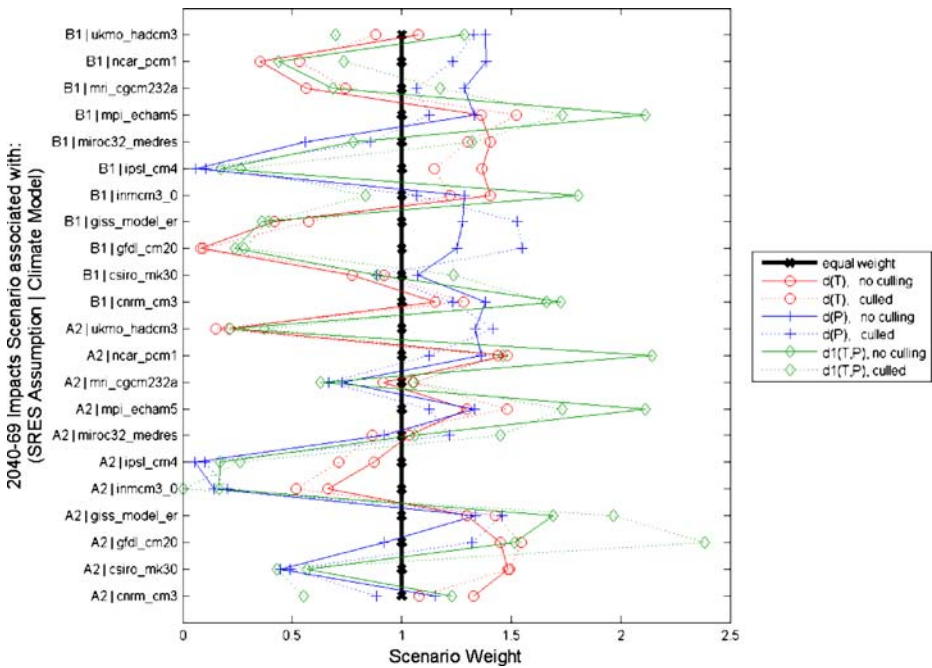


**Fig. 5 a** Density functions for projected climatological surface air temperature anomaly (i.e. change in projected 30-year mean from 1950 to 1999 mean in Northern California near (122W, 40N)) evaluated for the 2010–2039 and 2040–2069 periods. “No Model Culling” implies density function fit to all 75 projections listed in Table 1. Other legend labels correspond to subsets of these 75 projections, where model-contribution to the subsets is indicated by the credibility-based model subsets listed in Table 3 (columns 2 through 5). *Circle-cross symbols on horizontal axis* show anomaly positions of a 22-member subset (i.e. “Impacts Ensemble”) of the 75-member set of fitting projections. **b** Same as **a**, but for projected climatological precipitation anomaly

assessment than the decision on whether to consider model filtering when constructing the climate projection density function. Conceptually, if both projected temperature and precipitation changes are considered in the risk assessment, then perhaps  $d(T,P)$  might offer the preferred information. Moreover, if the projected temperature and precipitation trends are correlated, then  $d(T,P)$  would also be preferred.

**Fig. 6 a** Density function for jointly projected climatological surface air temperature and precipitation anomalies (i.e. change in projected 30-year mean from 1950 to 1999 mean in Northern California near (122W, 40N)) evaluated for the 2040–2069 period. *Dashed line* shows 0.05 interval (ascending in value from ~0 at plot perimeter). *Solid contours* show the density surface estimated using the nonparametric technique. *Dashed contours* show the density surface estimated using the second technique (Dettinger 2006). *Light-colored cross symbols* show positions of the joint-anomalies from the 75 fitting projections in Table 1. *Circle-cross symbols on horizontal axis* show anomaly positions of a 22-member subset (i.e. “Impacts Ensemble”) of the 75-member set of fitting projections, and overlie the “x” symbols marking these same members as they’re part of the 75-member fitting ensemble. **b** Same as **a**, but with the density function fit to a retained-model subset of “Uncertainty Ensemble” projections (explaining why there are fewer “x” and circle-cross fitting data relative to **a**). Model culling reflected the Water Supply perspective (Table 2) and associated model membership (Table 3)

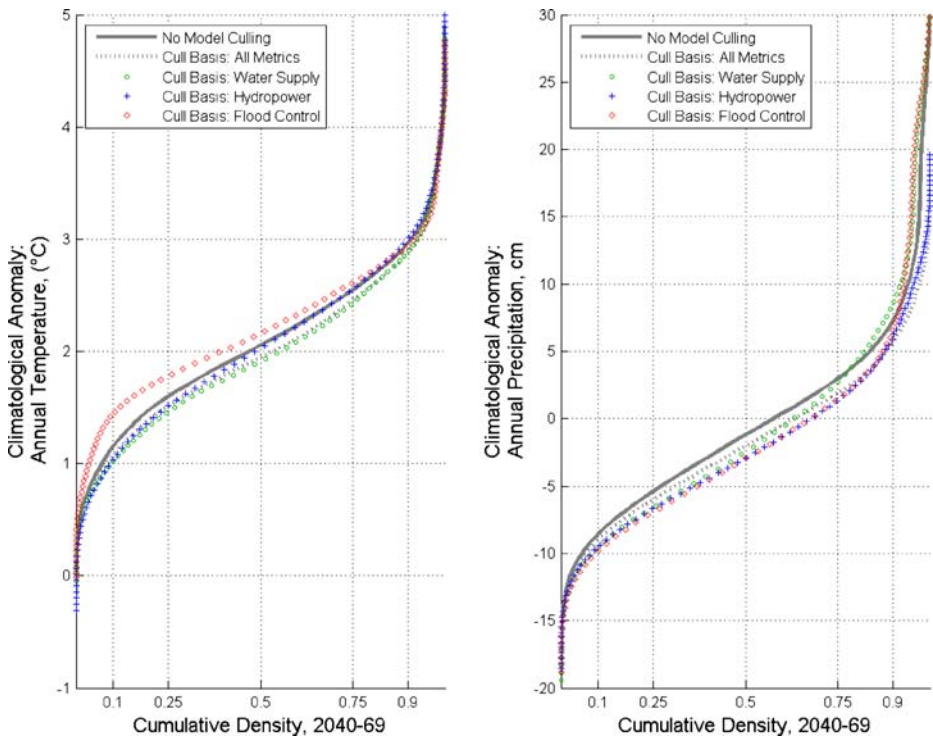




**Fig. 7** Sampled densities from six density functions function coordinates corresponding to projected “Impacts Ensemble” anomalies for the 2040–2069 period. Six functions correspond to three projected conditions, fit to a projection ensemble assembled with or without model filtering. Conditions are projected climatological surface air temperature anomaly (Fig. 5a), precipitation anomaly (Fig. 5b), and joint anomalies for both (Fig. 6a, b). Model culling is based on the Water Supply perspective (Table 2) and associated model membership (Table 3)

### 3.4 Discussion

Revisiting the density functions, it is notable that the functions are not smooth and indeed tend to be multi-modal, contrasting from parametric density functions that might have been constructed from the same data. The multi-modal aspects of  $d(T)$ ,  $d(P)$ , and  $d(T,P)$  are introduced by the nonparametric density estimation technique used in this case study (which might be more easily interpreted as constructing a smoothed “histogram-like” functions from the fitting data). These effects are somewhat muted when the information from the density functions are presented in terms of cumulative densities, or cumulative distribution functions [e.g.,  $D(T)$  and  $d(P)$  derived for the 2040–2069 period from  $d(T)$  and  $d(P)$ , respectively, shown on Fig. 8]. For decision-makers, it may be preferable to show scenario possibilities in terms of cumulative distributions or quantiles rather than density functions. For example, decision-makers might hold the risk “value” that planning strategies should accommodate a range of projected climate conditions up to a threshold change exceeded by a minor fraction of projections (e.g., 10%). Applying this hypothetical decision criterion using results from this study, planning would be done to accommodate changes up to [–] deg C or [–] cm of annual precipitation, considering the various  $D(T)$  and  $d(P)$  functions on Fig. 8. If the decision criterion were modified to consider jointly projected occurrence of temperature and precipitation anomalies, then information from  $D(T)$  and  $d(P)$  would have to be replaced by an estimate of  $D(T,P)$  using density



**Fig. 8** Cumulative distribution functions developed from the density functions of projected climatological surface air temperature and precipitation anomalies (Fig. 5a and b, respectively) evaluated for 2040–2069 period

information in either Fig. 6a or b. However, switching focus back to the task of conducting climate change risk assessment, it is necessary to assign relative scenario likelihoods to individual impacts scenarios. For this objective, density functions, rather than cumulative distributions, are needed given that the former reveal how a specific projection member is positioned within the context of projection consensus and breadth.

Finally, on the matter of how density function form may be sensitive to the fitting technique, the sensitivity of derived scenario weights to fitting technique was explored by reconstruction of  $d(T,P)$  for the 2040–2069 period and “No Model Culling” and “Cull Basis: Water Supply,” using the principal component (PC) resampling technique described in Dettinger (2006). Figure 6a and b show density contours from this technique, which can be compared to those developed using the nonparametric technique. As mentioned, comparison of these two surfaces shows that choice of technique had minimal effect on the function’s central tendency and breadth. The distributions obtained from the two methods also share the strong correlation that tends to pair wetter scenarios with (relatively) cooler scenarios and drier scenarios with the warmest scenarios. This correlation, however, was much muted when the resampling approach was adjusted to avoid weighting the more prolific model groups more than those that provided only single realizations of each model/emissions scenario combination (not shown). Further comparisons of the two sets of contours will indicate that only the general features of these distributions can be estimated in a confident, methods-independent way.

## 4 Limitations

The present analysis is subject to several limitations. First, note that these methods provide information on climate projection consensus and not the true probability of climate change. Understanding this limitation will be important in a decision-making context where decision-makers may not anticipate the complex appearance of the density functions, which are as stated are essentially smoothed, multi-modal, “histogram-like” functions. The appearance of these functions is set up by use of nonparametric techniques to fit the functions rather than imposing parametric forms (e.g., Gaussian), and that the function was fit to a limited and not necessarily homogeneous projection sample. Once the decision-makers get used to these odd looking distributions, it will be equally important that they not be over-interpreted; that is, some of the multimodality of these distributions is surely artifact rather than signal.

The correct interpretation of such density functions is that they indicate projection consensus within the ensemble of projections considered. Although there may be an inclination to use the density functions to guide statements on “climate change probability,” such application should be avoided. The reason is that key climate change uncertainties are not represented within the spectrum of currently available climate projections. To illustrate, consider that for this case study a 75-member projection ensemble served as the basis for fitting density functions, representing information from a heterogeneous mix of 17 coupled ocean–atmosphere climate models under two emissions pathways, reflecting various states of modeling capability and a crude cross section of the uncertainties concerning future emissions. Not represented among these projections are the uncertainties associated with the many factors not included in current climate models or in the pathways considered here (e.g., assumed global technological development, distributed energy-technology portfolios, resultant spatial distribution of GHG sources and sinks through times, and biogeochemical interaction with GHG sources and sinks, and many others). For these reasons, it is important to interpret the “climate projection density” functions featured in this analysis as being a characteristic of the ensemble considered and not the full range of uncertainties. In the end, “climate projection densities” are expected to be distinctly different from climate-change probabilities.

It also bears mentioning that the historical 20c3m climate simulations included in the WCRP CMIP3 archive and used here are not strictly comparable, which introduces uncertainty surrounding credibility analysis results and climate projection initial conditions. Although the 20c3m simulations all shared the same primary anthropogenic GHG forcings, the exact combinations of natural radiative forcings and some secondary anthropogenic influences varied from modeling group to modeling group. This, along with the issue of simulating low-frequency natural climate variations discussed earlier, limits our ability to interpret relative model differences meant to be revealed by the credibility analysis.

Other limitations stem from the absence of basic features that are generally required of statistical frameworks, including: (1) requirement to account for the uncertainties of the Reference climate definitions framing the model credibility analysis, (2) a preference for the credibility analysis to be focused *only* on past simulation of the projected quantity, and (3) requirement to account for the interdependence among credibility analysis variables and metrics (i.e. “dimensions” in the distance-similarity framework). Attribute (1) limits the results produced from the present analysis so that it does not fully represent yet another aspect of the uncertainties associated with the projections, in this case, the uncertainty as to

how well the models really do represent the real-world climate. It will be beneficial if future work can be recast to factor in such uncertainties.

Attribute (2) points to a matter of philosophy in the analytical design herein: whether to frame credibility analysis on a model's ability to recreate only past simulation of projected quantities or a *mix* of regionally relevant local and global climate variables influencing the projected quantities, along with their teleconnections (including the projected quantity). When weighing these options, a real-world limitation emerges in that the projected quantities in question include many historical influences besides the GHG trends that motivate development of GHG-based projections. The complex nature of the climate system is also a factor, as projected quantities depend on the fate and evolution of many other variables within the models. The analytical choice to focus only on past simulation of the projected quantities is reasonable if it can be assumed that credibility in projecting a given quantity is informed *completely* by understanding the model's capability in simulating past values of that quantity. However, in the case of regional climate projection, there is recognition that models can produce "correct answers" for different climate variables and specific regional locations for the "wrong reasons." This fact, although contradictive to the preceding philosophy, promotes consideration for a broader mix of variables and metrics in the credibility analysis, on the idea that ability to recreate a mix of regionally relevant variables and metrics during past simulation should be a good indicator of a model's ability to project an embedded quantity (or quantities) within that mix.

Considering the mix of regionally relevant climate variables and metrics used to define model credibility, it is reasonable to assume that inter-variable and inter-metric correlations exist, in defiance of consideration (3), because they are sampled from a common modeled or observed climate system. Nevertheless, such variables and metrics are treated herein as being independent dimensions when computing distance-based model-to-reference similarity. Perhaps future work could focus on modifying the credibility analysis to be framed around a more limited set or transformed set of regionally relevant variables and metrics that are essentially uncorrelated, thereby avoiding the issue of inter-variable and inter-metric correlations affecting interpretation of computed similarity distance.

Finally, focusing on greater numbers of variables and metrics tended to work against the reasonable objective of using credibility analysis to reduce perceived projection uncertainty by focusing on scenarios produced by a set of "best" models. Our results showed that the cumulative differences between models became more muted as more variables and metrics were considered. This particular case study was framed with a goal to identify a "more credible half" of the available models upon which to focus attention (much like the approach used by Milly et al. (2005) and to explore how such model selections affect density function development and density-based scenario weights.

## 5 Summary and conclusions

A methodology has been developed for use in regional assessments, to evaluate the relative credibility of models providing twenty-first century climate projections based on their relative accuracies in simulating past climate conditions. The method rests on the philosophy that the relative credibility of a given model's climate projections among those of other models can be inferred from the model's performance in recreating twentieth century climatology compared to other models. A distance-similarity approach was used to compare among models, where modeled twentieth century climate differences were measured from reference observations of

several regionally relevant climate variables and statistical metrics. Computed distances were then translated into relative model weights, which were then used to select (and possibly weight) among models when estimating climate projection densities.

Case study application of these methods for the Northern California region indicates that:

- Credibility analysis based on multiple climate variables and metrics allows models to be distinguished according to more comprehensive simulation performance. However, use of a greater number of variables and metrics led to less apparent distance-based differences among models.
- Credibility analysis based on a more limited set of variables and metrics led to greater apparent distance-based differences among models. However, the resultant model weights and subsequently use of weights to filter models produced model-culling decisions that depend greatly on the (somewhat arbitrary) choice of metrics.
- Using credibility analysis results to cull models and affect construction of climate projection density functions led to some change in the local aspects of the density functions. For functions describing projected temperature change, results showed that the overall function spread and central tendency tended to be more influenced by how inclusive and extensive the original ensemble was (i.e. Uncertainty Ensemble from Table 1) compared to the influence of deciding whether to filter down to a “better half” of models before fitting the functions. That is, the various culling of the projections used to estimate the distributions did relatively little to change either the central tendencies or ranges of the distributions obtained. For functions describing projected precipitation change, results lead to similar impressions, except that the central tendency of the projected precipitation anomalies’ were more sensitive to choice of whether to consider model-culling, but not so much to choice of which cull basis to use among the three “perspectives” considered (Table 3).

Revisiting the motivating question of whether relative scenario weights derived from credibility-based density functions (as framed by these methods) were significantly different than those derived from density functions that do not consider model culling, our results suggest that:

- Accounting for model credibility through model-culling prior to fitting the density function has some influence on the relative scenario weights, which could translate into effects on the subsequent risk assessment.
- Perhaps more significantly, the relative scenario weights are relatively more sensitive to the choice of projected quantity (e.g.,  $d(T)$ ,  $d(P)$ , or  $d(T,P)$ ) than to the chosen cull basis (Table 3) prior to estimating the density function describing that quantity.

**Acknowledgments** This project was funded by multiple sources, including the U.S. Bureau of Reclamation Science and Technology Program sponsored by the Reclamation Research Office, through direct contributions from the Reclamation Mid-Pacific Region Office, and in-kind contributions from California Department of Water Resources, U.S. Geological Survey and Santa Clara University. We thank staff at Scripps Institute of Oceanography for compiling and processing projection datasets used in these analyses (with funding provided by the California Energy Commission’s California Climate Change Center at Scripps). We acknowledge the modeling groups for making their climate simulations available for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the CMIP3 model output, and the WCRP’s Working Group on Coupled Modelling (WGCM) for organizing the model data analysis activity. The WCRP CMIP3 multi-model dataset is supported by the Office of Science, U.S. Department of Energy.

## References

- AchutaRao K, Sperber KR (2002) Simulation of El Niño Southern Oscillation: results from the coupled model intercomparison project. *Clim Dyn* 19:191–209
- AchutaRao K, Sperber KR (2006) ENSO simulation in coupled ocean–atmosphere models: are the current models better. *Clim Dyn* 27:1–15
- AchutaRao KM, Covey C, Doutriaux C, Fiorino M, Gleckler P, Phillips T, Sperber K, Taylor K (2004) An appraisal of coupled climate model simulations, D Bader (ed), Rep. UGRL-TR-202550, 183pp, Program for climate model diagnosis and intercomparison. Lawrence Livermore National Lab., Livermore, California
- Black PE (2006) (eds) Dictionary of algorithms and data structures. U.S. National Institute of Standards and Technology, Gaithersburg, MD
- Brekke LD, Miller NL, Bashford KE, Quinn NWT, Dracup JA (2004) Climate change impacts uncertainty for water resources in the San Joaquin River basin. *Calif, J Am Water Resour Assoc* 40:149–164
- Cayan DR, Maurer EP, Dettinger MD, Tyree M, Hayhoe K (2006) Climate change scenarios for the California region. *Climatic Change* (in press)
- Christensen N, Lettenmaier DP (2006) Climate change and Colorado River Basin: Implications of the FAR Scenarios for Hydrology and Water Resources. poster presentation at the Third Annual Climate Change Research Conference co-sponsored by the California Energy Commission and California Environmental Protection Agency, September 13–15 2006
- Covey C, AchutaRao KM, Cubasch U, Jones P, Lambert SJ, Mann ME, Phillips TJ, Taylor KE (2003) An overview of results from the Coupled Model Intercomparison Project (CMIP). *Glob Planet Change* 37:103–133
- Dettinger MD (2005) From climate change spaghetti to climate change distributions for 21st century. *San Franc Estuary Watershed Sci* 3(1):1–14
- Dettinger MD (2006) A component-resampling approach for estimating probability distributions from small forecast ensembles. *Clim Change* 76:149–168
- Hayhoe K, Cayan D, Field C, Frumhoff P, Maurer E, Miller N, Moser S, Schneider S, Cahill K, Cleland E, Dale L, Drapek R, Hanemann RM, Kalkstein L, Lenihan J, Lunch C, Neilson R, Sheridan S, Verville J (2004) Emissions pathways, climate change, and impacts on California. *Proc Natl Acad Sci (PNAS)* 101(34):12422–12427
- IPCC (Intergovernmental Panel on Climate Change) (2001) Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the third assessment report of the IPCC, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds) Cambridge University Press, 881 pp
- IPCC (2007) Climate Change 2007 – The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC, Solomon S, Qin D, Manning M, Marquis M, Averyt K, Tignor MMB, Miller HL, Chen Z (eds) Cambridge University Press, 996 pp
- Kalnay E, Kanamitsu M, Kistler R et al (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc* 77:437–471
- Lall U, Rajagopalan BR, Tarboton DG (1996) A nonparametric wet/dry spell model for resampling daily precipitation. *Water Resour Res* 32:2803–2823
- Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC (1997) A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bull Am Meteorol Soc* 78:1069–1079
- Maurer EP (2007) Uncertainty in hydrologic impacts of climate change in the Sierra Nevada, California under two emissions scenarios. *Clim Change* 82:309–325
- Maurer EP, Duffy PB (2005) Uncertainty in projections of streamflow changes due to climate change in California. *Geophys Res Lett* 32(3):L03704
- McCabe GJ, Palecki MA, Betancourt JL (2004) Pacific and Atlantic Ocean influences on multidecadal drought frequency in the United States. *Proc Natl Acad Sci* 101:4136–4141
- Meehl GA, Covey C, McAvaney B, Latif M, Stoufer RJ (2005) Overview of the coupled model intercomparison project. *Bull Am Meteorol Soc* 86:89–93
- Milly PCD, Dunne KA, Vecchia AV (2005) Global pattern of trends in streamflow and water availability in a changing climate. *Nature* 438:347–350
- Phillips TJ, AchutaRao K, Bader D, Covey C, Doutriaux CM, Fiorino M, Gleckler PJ, Sperber KR, Taylor KE (2006) Coupled climate model appraisal: a benchmark for future studies. *EOS Trans* 87:185
- Piechota TC, Chiew FHS, Dracup JA, McMahon TA (1998) Seasonal streamflow forecasting in eastern Australia and the El Niño-Southern Oscillation. *Water Resour Res* 34:3035–3044

- Scott DW (1992) Multivariate density estimation: theory, practice, and visualization. Probability and mathematical statistics. Wiley, New York
- Silverman BW (1986) Density estimation for statistics and data analysis. Monographs on statistics and applied probability. Chapman and Hall, New York
- Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a bayesian approach to the analysis of multi-model ensembles. *J Climate* 18:1524–1540
- Vicuna S, Maurer EP, Joyce B, Dracup JA, Purkey D (2007) The sensitivity of California water resources to climate change scenarios. *J Am Water Resour Assoc* 43(2):482
- Wilby RL, Harris I (2006) A framework for assessing uncertainties in climate change impacts: low-flow scenarios for the River Thames, UK. *Water Resour Res* 42:W02419
- Wilks DS (1995) Statistical methods in the atmospheric sciences. Academic Press, New York
- Zierl B, Bugmann H (2005) Global change impacts on hydrological processes in Alpine Catchments. *Water Resour Res* 41:W02028