Title

Implications of climate model selection for projections of decision-relevant metrics: A case study of chill hours in California

Authors

Jagannathan, Kripa
Jones, Andrew D
Kerr, Amber C

Publication Date

2020-04-01

DOI

10.1016/j.cliser.2020.100154

Peer reviewed

Climate Services 18 (2020) 100154

Contents lists available at ScienceDirect

Climate Services

journal homepage: www.elsevier.com/locate/cliser

Original research article

Implications of climate model selection for projections of decision-relevant metrics: A case study of chill hours in California

Kripa Jagannathan[a,b,*], Andrew D. Jones[b], Amber C. Kerr[c]

[a] University of California, Berkeley, Energy and Resources Group, Berkeley, CA 94720, USA
[b] Lawrence Berkeley National Laboratory, Earth & Environmental Sciences Area – Berkeley Lab, 1 Cyclotron Road, Berkeley, CA 94720, USA
[c] University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Keywords:
Climate change
Adaptation
Climate models
California
Model Skill
Chill hours

ABSTRACT

Decision-makers today have relatively easy web-based access to climate projections from several different models and downscaled datasets. Yet, there is minimal guidance on the credibility and appropriate use of such models and projections for specific adaptation contexts. The few studies that provide recommendations on model choice are often based on evaluations of broad physical climate metrics (such as temperature averages or extremes) at regional scales, without additional examination of local-scale decision-relevant climatic metrics (such as growing degree days or chill hours) that underpin the adaptation action. While such broad regional skill may be considered necessary for the overall credibility of models, it is not clear whether it is sufficient to ensure good skill for decision applications. This paper evaluates the skill of different Global Circulation Models (GCMs) in predicting the decision-relevant metric of chill hours in Fresno, California, and examines how model selection impacts future projections. We find that good skill in predicting broader physical climate metrics in California does not guarantee skill in prediction of chill hours in Fresno. In fact, the models with good regional climatic skill were mutually exclusive of the ones with good skill for chill hours, which leads to some counterintuitive results for this unique metric. Since many decision-relevant metrics are non-linear derivations of primary physical quantities (like the chill hour metric), more such decision-relevant model evaluations are needed to provide better insights on model credibility and choice for adaptation decisions.

Practical implications

There are currently about 60 different Global Circulation Models (GCMs) that can provide projections of future climate. For a decision-maker looking to utilize climate data, synthesizing these vast range of possibilities can be a formidable task as there is limited guidance on which set of models and projections are appropriate for their specific adaptation context (Barsugli et al., 2013; Jones et al., 2016; Maurer et al., 2014; Moss et al., 2019; Snover et al., 2013). Literature suggests that for a realistic representation of the future, a sample of at least several models should be used. However, there is no consensus on how this sample should be chosen for decision-relevant applications (Overland et al., 2011). The few studies that provide guidance on model choice are based on evaluations of models' historical performance for broad physical climate metrics (such as temperature averages or extremes) at regional scales, without an additional evaluation of decision-relevant climatic metrics (such as growing degree days or chill hours) at local scales that a user may find more relevant. This raises a question as to whether these models that perform well for broad regional climate, i.e. have 'broad regional-skill' will also perform well for specific decision-relevant metrics, i.e. have 'specific local-skill'. And relatedly, how future projections may differ based on whether models are picked for broad regional or specific local-skill. This study assesses this question by evaluating the skill of different climate models in predicting the decision-relevant metric of chill hours in Fresno - California, and examining the extent to which the choice of GCMs alters chill hour projections for the future. We also highlight the similarities and differences in projections based on whether models are chosen for skill in broad-scale physical climatic metrics for California or for skill in predicting chill hours in Fresno.

Chill hours (defined as cumulative hours below 45°F or 7.2 °C from November 1 to February 28 or 29) is one of the most important decision-relevant climate metrics for temperate fruit and nut tree crops. Observed data shows that from 1971 to 2012, chill hours in Fresno have been decreasing at the rate of −8.4 chill hours per year (ch/yr). This negative chill hours slope/trend, can

* Corresponding author.
E-mail addresses: kripajagan@berkeley.edu (K. Jagannathan), adjones@lbl.gov (A.D. Jones), ackerr@ucdavis.edu (A.C. Kerr).

https://doi.org/10.1016/j.cliser.2020.100154
Received 30 September 2018; Received in revised form 31 July 2019; Accepted 27 January 2020
Available online 10 February 2020

be a cause for concern to farmers and adaptation practitioners, as reduced winter chill strongly impacts crop yield and quality (Kerr et al., 2018; Lobell and Field, 2011). Estimates of future chill could help growers better anticipate for additional management costs, as well as assist in choosing the right crop species, varieties, or rootstocks that are more adapted to future climate change (Luedeling et al., 2009a; Pathak et al., 2018).

In our skill evaluation, we analyzed the performance of both raw GCMs as well as the popular downscaled dataset used by the State of California (i.e. the LOCA dataset which is based on the Localized Climate Analogues downscaling method). We found that models that perform well for broad physical climatic metrics of temperature, precipitation and El-Nino patterns, do not necessarily perform well for chill hours. Hence, computing future projections of chill hours using a broad regional-skill based sampling approach, provides some counterintuitive results. This may be due to the fact that the relationship between the temperature and chill hour metric is non-linear, as chill hours only accumulate below a certain threshold temperature. Although it is difficult to know what types of sampling are more realistic or appropriate, the results strongly indicate that different skill-based sampling approaches can have important repercussions for the analysis of future chill hours, and perhaps also other such decision-relevant metrics.

Overall, we find that the peculiarities of specific decision-relevant metrics – such as this non-linear threshold-based chill hour metric – can lead to counterintuitive findings that question the validity of some generally accepted recommendations on climate model selection for impact and adaptation studies. We find that broad regional climate skill of models is not always sufficient to ensure skill for some decision-relevant metrics, and an additional layer of decision-relevant model evaluation may be needed to better understand how models perform on the eventual metric of relevance to the user. Since many crucial adaptation decisions in agriculture, energy, water management, and other fields are made based on similar threshold-based metrics (such as growing degree days, heating or cooling degree days, and days over 100°F), more such model evaluations can help to better understand model credibility in specific decision-contexts. Further, there is a critical need for more nuanced research on model selection strategies for decision applications, to ensure that adaptation action is based on the best available climate projections of the future.

1. Introduction

The past couple of decades have seen a large proliferation of different climate projections. The Coupled Model Intercomparison Project (CMIP5) involves over 25 modeling centers across the world, and currently there are about 60 General Circulation Models (GCMs) that can provide projections of future climate (WCRP, 2017). For decision-makers looking to use climate information for future planning (such as adaptation practitioners, farm managers, crop advisors, and water managers), this gives rise to what Barsugli et al. (2013) term the 'practitioner's dilemma': how to synthesize the large number of projections, assess their credibility, characterize their uncertainties, and use them wisely for a particular decision context (Maurer et al., 2014; Moss et al., 2019; Mote et al., 2011; Snover et al., 2013). Since GCMs were not originally developed with decision-makers' needs in mind (Jones et al., 2016), the reliability of these numerous model projections for local and regional scales, and the limits to their utility for specific impact and adaptation questions, is an ongoing and challenging field of inquiry (Barsugli et al., 2013; Mendlik and Gobiet, 2016; Mote et al., 2011; Overland et al., 2011).

In the past, a simple arithmetic average of all available model projections was often used to represent the "best" estimate of future projections (Flato et al., 2013; Herger et al., 2018). However, recent research has shown this approach to have several limitations, as it implicitly assumes that each individual model is independent from others and has equal abilities. Neither of these are completely valid assumptions, especially for regional and local scale assessments (Hayhoe et al., 2017; Knutti et al., 2010; Mendlik and Gobiet, 2016; Overland et al., 2011; Sanderson et al., 2017). Researchers have since developed more systematic approaches for model selection, weighting, and averaging that account for model interdependence as well as the relative abilities of climate models in simulating the regional climatology of relevance (Hayhoe et al., 2017; Herger et al., 2018; Mendlik and Gobiet, 2016; Sanderson et al., 2017, 2015). Newer studies such as the Fourth National Climate Assessment (NCA4), and the California's Fourth Climate Change Assessment (CCCA4), have preferred to use these skill-based approaches as an attempt to provide more refined projections for their specific geographic areas of interest.

While there have been several advancements in developing skill-based model selection and weighting approaches for regional studies, there are still unanswered questions when it comes to model selection for sector-specific climate adaptation problems that often need projections of a particular decision-relevant climatic metric at a specific location (Barsugli et al., 2013; Mote et al., 2011; Overland et al., 2011). In such cases, it remains unclear whether model selection and weighting needs to account for the skill of the models for this specific local-scale climatic metric, in addition to their skill for average climatological variables over larger regions (Overland et al., 2011). Further, most model skill evaluation studies tend to focus solely on metrics of broad physical climate phenomena (e.g. large-scale atmospheric patterns and averages/extremes/anomalies in temperature and precipitation), and there are very few studies that evaluate model skill for specific decision-relevant metrics associated with particular adaptation decisions (e.g. growing degree days[1] and chill hours[2] in the agriculture sector, or heating or cooling degree days in the energy sector) (Moss et al., 2019). Since climate models predict different metrics with varying skill (Girvetz et al., 2013; Snover et al., 2013), GCM skills for several of these decision-relevant metrics[3] remain largely unknown. The question remains whether model selections based on broad regional climatological evaluations can sufficiently ensure robust and reliable projections of decision-relevant climatic metrics that also account for the full range of scientific uncertainties associated with the specific metric.

The goal of this paper is to examine how model selection impacts projections for a particular decision-relevant metric, and to highlight the similarities and differences in results based on whether models are chosen for skill in broad-scale physical climatic metrics or for skill in the decision-relevant metric. We focus on chill hours at a specific location (Fresno, CA) as a case study. Chill hours are defined as the cumulative hours below 7.2 °C from November 1 to February 28 or 29, and is one is one of the most important decision-relevant climate metrics for several high-value temperature fruit and nut tree crops. The non-linear relationship between temperature and chill hours makes it an interesting threshold-based metric for analyzing model skill. We first evaluate the skill of different GCM-derived datasets (both downscaled and raw data) in predicting historical chill hours. We then explore whether models with good skill for broad physical climatic metrics in California are also skilled in predicting chill hours in Fresno, and we investigate whether and to what extent the spread of chill hour

[1] Refers to the number of degrees by which daily average temperature falls above a threshold temperature.

[2] Calculated as the cumulative sum of hours below a stated base temperature

[3] Decision-relevant metrics are often computed using specialized algorithms for which physical climate metrics serve as inputs. While we make a distinction between 'physical climate metrics' and 'decision-relevant metrics' in this paper, we also acknowledge that there may be some overlap between the two. For example, some basic physical climate metrics such as monthly average temperature or seasonal precipitation patterns can also be decision-relevant metrics.

projections in the future differ based on model selection. Our objective is not to make a case for model sub-selection that is based on specific decision-relevant metrics nor to suggest a new selection or weighting method. Rather, this paper highlights the repercussions of using model selection or weighting approaches that are based on broader regional skill without additionally assessing model skill for the local decision-relevant climatic metric of interest. In doing so, we provide an improved understanding of broad versus specific skill of the current generation of climate models.

2. Model weighting, selection and averaging approaches

In order for projections of climate change to be robust and reliable, they need to provide a balanced and unbiased estimate of the entire distribution of potential future changes (Mendlik and Gobiet, 2016). The goal is to maximize model diversity and capture the true uncertainty in regional change, while also assuring good model performance (Hayhoe et al., 2017; Mendlik and Gobiet, 2016). The equally weighted multi-model mean (MMM) has often been regarded as the gold standard for synthesizing projected changes from large model ensembles, as the averaging can lead to cancellation of offsetting errors in individual global models (Hayhoe et al., 2017; Herger et al., 2018; Knutti et al., 2010; Mendlik and Gobiet, 2016). This method is predominantly used in global climate change studies, such as the Intergovernmental Panel on Climate Change (IPCC)'s Fifth Assessment Report (AR5). However, this "one model, one vote" approach has increasingly been called into question, including in AR5, due to co-dependencies between individual models (Flato et al., 2013; Sanderson et al., 2017). Several GCMs are known to share structural components, sections of code, and representations of certain features; hence, they may have similar errors (Flato et al., 2013; Hayhoe et al., 2017; Knutti et al., 2010; Mendlik and Gobiet, 2016; Overland et al., 2011; Sanderson et al., 2015). Therefore, each model run cannot be deemed to represent an independent projection estimate, and a simple multi-model mean can lead to biases and double counting (Herger et al., 2018; Mendlik and Gobiet, 2016). Further, the MMM may not always present offsetting errors for specific variables or at regional scales (Knutti et al., 2010). In order to avoid these potential biases, newer studies are using more systematic approaches to arrive at regional projections of climate change (Hayhoe et al., 2017; Pierce et al., 2016).

In lieu of the MMM, other approaches for synthesizing projected changes from multi-model ensembles include sophisticated model weighting or model sub-selections, that are based on three main criteria: ensuring good model performance in the past, maintaining the spread of the climate change signal to accurately represent uncertainties, and accounting for model interdependence (Hayhoe et al., 2017; Sanderson et al., 2017, 2015). These approaches select or assign weights to models based on model dependence studies and performance evaluations that compare 20th-century hindcast simulations to observations of certain key variables and regions of interest. Model sub-selection is essential for many user applications to limit computational demand and make data handling more manageable (Barsugli et al., 2013; Herger et al., 2018; Mendlik and Gobiet, 2016; Pierce et al., 2016). A critical consideration in model selection is that the eventual sub-set of models maintains the key properties of the full ensemble, such as the extent of the spread in future projections (which provides a measure of uncertainty) and the statistical properties of the climate change signal (Herger et al., 2018; Mendlik and Gobiet, 2016; Pierce et al., 2016). Therefore, it is recommended that a sample size of at least several models is maintained, and that performance evaluations be used to detect and account for severely unrealistic models that cannot be trusted for clearly argued reasons, rather than to select a handful of 'best' performing models (McSweeney et al., 2014; Mendlik and Gobiet, 2016; Overland et al., 2011).

Despite the growing focus on weighting and model selection, there are still some criticisms of these approaches. Depending on how they are used, some model selection approaches could lead to underestimations or inflations of uncertainties in regional projections (Hayhoe et al., 2017; Madsen et al., 2017). If performance measures for model sampling are narrowly defined, it can lead to subjective rankings and over-determination of GCM accuracy, where models may be getting the right answers but not for the right reasons (Mendlik and Gobiet, 2016). Some studies have also suggested that there may not be a strong correlation between a model's past performance and its ability to predict the future, thereby questioning the very premise of skill-based model selection (Hayhoe et al., 2017; Knutti et al., 2010). Others maintain that depending on the scientific question, a MMM of all available model simulations (unselected and unweighted) might still be appropriate (Madsen et al., 2017; Pierce et al., 2009).

Nevertheless, most of the recent literature suggests that evaluating models' past skill for the region and the climatic metric of interest is an essential prerequisite, even if it does not guarantee accurate model projections under new climate states (Flato et al., 2013; Hayhoe et al., 2017; Knutti et al., 2010; Overland et al., 2011). While a good representation of the overall climatology of a region (i.e. mean, variability and trends in broad climatic metrics), is often regarded as a 'necessary' condition, there are still questions on whether such broad skill is 'sufficient' to ensure skill in specific local-scale decision-relevant climatic variables, or if an additional layer of decision-relevant model evaluation is required to ensure that projections are adequately representing the problem at hand.

Notwithstanding these challenges and complexities, many new assessment studies including the NCA4 and the CCCA4 have preferred to move away from the MMM, and use model weighting and selection strategies for providing regional projections (Hayhoe et al., 2017; Pierce et al., 2016). For example, the state of California used a three-tiered model performance evaluation (which included assessment of model skill for global climatology, western U.S. climate and hydrology, and the California state hydrology and climate extremes) to identify 10 models that provide reasonably realistic simulations of the state's historical and current climate (Cal-Adapt, 2017; CATRWG, 2017; CCTAG, 2015). These models were evaluated for several broad physical climatic metrics that are of importance to the region, such as seasonal minimum and monthly mean temperatures, monthly and seasonal maximum precipitation, correlation with El Niño teleconnections, etc. (Brekke et al., 2008; CCTAG, 2015; Pierce et al., 2009; Rupp et al., 2013). In cases where using output from 10 models is still unwieldy for the user, the state has identified a further-reduced set of 4 models which have been shown to substantially cover the range of projections represented by the larger 10-model ensemble for broad physical climatic metrics (CCTAG, 2015; Pierce et al., 2016). Although these climate models have not specifically been evaluated for various decision-relevant climatic metrics at local scales, the state's Climate Action Team has recommended that impact and adaptation studies in California use this subset of 10 or 4 models (CATRWG, 2017). The question remains whether this set of 10 (or 4) models that have been evaluated to meet the criteria of good past skill and representation of spread in projections for multiple physical climatic metrics across broader regions, will also meet the same criteria for decision-relevant metrics such as winter chill hours.

With this background and context, this study aims to assess whether models that are skilled in predicting California's overall climate are also skilled in predicting the decision-relevant metric of chill hours in Fresno. Further, we also examine whether and to what extent different samplings of GCMs (MMM versus other skill-based samplings) alters the results of chill hour projections. Our aim is not to propose a new model selection approach based on such evaluations; rather, we hope to address the question of how broad skill compares to specific skill of climate models and provide insights on the implications of different model selections for decision-relevant metrics. Our focus on the metric of chill hours can also provide useful insights for agricultural adaptation in California. Many fruit and nut trees must meet a certain winter chilling

requirement for the flowers and fruits to develop properly and for the trees to attain optimum yields (Luedeling et al., 2011). Reduced winter chill can have a significant negative impact on crop yield and quality, and sustained decline of chill hours may impact the viability of some perennial crops in areas where they were once widespread (Baldocchi and Wong, 2007; Kerr et al., 2018; Lobell et al., 2006; Luedeling et al., 2009a; Medellín-Azuara et al., 2011; Pathak et al., 2018). Estimates of future chill could help growers better anticipate for additional management costs, as well as assist in choosing the right crop species, varieties, or rootstocks that are more adapted to future climate change (Luedeling et al., 2009a; Pathak et al., 2018).

3. Methods

3.1. Calculation of observed chill hours

In order to evaluate the implications of model choice for chill hours, we first analyzed long-term historical trends in observed annual chill hours for the Fresno station of the National Weather Service Cooperative Network (lat/lon: 36.78, −119.72). Fresno was chosen as the location of interest because the county is one of California's major production centers for fruit and nut crops, such as almonds, grapes, pistachios, cherries and peaches (CDFA, 2016; County of Fresno, 2015). In addition, the Fresno weather station had a long record (1971–2012) of daily time-series temperature data, which is required for chill hour computation. Since results of the evaluation can be sensitive to the time period chosen, we chose the entire time period for which daily time-series temperature record available.

For this paper, we computed chill hours as the cumulative number of hours between 0 and 7.22 °C (32 and 45°F) in the winter months of November, December, January and February (Baldocchi and Wong, 2007). This is known as the Chilling Hours model. Though there are several different methods of computing chilling requirements for fruit and nut trees, some of which may be regarded as more precisely representing tree physiology (Luedeling and Brown, 2011), the Chilling Hours model is the simplest and most transparent model which is most widely used by practitioners. Different chill calculation methods (such as the Dynamic model, the Utah model, and the Positive Utah model) may yield different predictions for how chill hours will change in future climate (Darbyshire et al., 2011; Luedeling et al., 2009b; Luedeling and Brown, 2011). However, since the results of our paper rely mainly on a comparison between the observed and modeled chill hours, we limited the scope of this project to using one consistent chill hour calculation method while evaluating multiple GCMs. Although some specific results may differ, we anticipate that our broad findings and discussion points will be relevant to other chill hour models, which we hope will be tested by other researchers.

Since hourly data is not available for most GCMs, we used daily minimum and average temperature to estimate daily chill hours, following the trigonometric approximation method using an idealized mean diurnal temperature course (developed by Baldocchi & Wong, 2007). These authors found that their daily-to-hourly interpolation produced fairly accurate and unbiased chill hour estimates for this region (e.g., for Zamora, CA, another Central Valley location, the correlation between chill hours computed from actual hourly data versus interpolated daily data was $R^2 = 0.887$). Further, a comparison of chill hours from actual hourly data from the Fresno weather station versus the trigonometrically interpolated daily data show that the two methods provide very similar results ($R^2 = 0.98$) for both mean and slope of historical chill hours. Details of the trigonometric approach and the equations used are provided in Fig. A.1 and Eq. A.1 of Appendix A.

To understand our chosen location within the larger regional context, we also computed chill hours for all other weather stations that are located within 75 km of the Fresno station and that have a comparable long-term temperature record. Five other NWS COOP weather stations matched these criteria: Madera, Friant, Hanford, Lemoore and Visalia.

3.2. Model skill evaluation for chill hours

In order to assess the difference between models' relative skill for broad regional climatology versus that of chill hours at a local scale, we first assess the skill of various GCMs for chill hours, and then compare our results with other regional climatology evaluation studies conducted for California and the Western US. For the model skill evaluation, we compared both downscaled and raw GCM data with observed historical chill hours for Fresno. Here we note that while the analysis of downscaled data can tell us how well this dataset is able to capture chill hours in Fresno, the results would not provide a true measure of quality of the GCM. This is because downscaled data from all GCMs are bias-corrected to the same observed dataset, thereby making several aspects of the GCMs statistically indistinguishable (Maurer et al., 2014). Hence, in addition to downscaled data, we also assessed raw GCM data to evaluate the skill of different GCMs in predicting chill hours.

We used raw climate model data from the Coupled Model Intercomparison Project Phase 5 (CMIP5) archives. For the downscaled dataset, we used the high-resolution data (1/16° or about 6 km spatial resolution and daily time-scale) derived using the Localized Climate Analogues or LOCA method, since the LOCA data is regarded as one of the better datasets for representing California's climate (Pierce et al., 2016). Temperature data for the grid cell containing the Fresno weather station was collected from raw and downscaled CMIP5 models (and corresponding ensemble members) for which daily average and minimum near-surface air temperature (T_{av} and T_{min}) outputs were available for both the historical run and the Representative Concentration Pathway (RCP) 8.5 future projections[4]. The historical runs begin in 1850 and end in 2005 and are driven by standardized greenhouse gas concentrations, aerosols, and land-use change forcing datasets. We chose RCP 8.5 in order to use projections with the largest potential signal compared to internal variability (McSweeney et al., 2014). We also note that choice of RCP is not expected to greatly influence model results for relatively near-term projections such as 2050 (CATRWG, 2017; Hawkins and Sutton, 2009).

We assessed downscaled data from a total of 32 GCMs and raw data from 29 GCMs. Wherever available, data from different initial-condition ensemble runs (model runs beginning with different initial conditions in 1850) of the GCMs were also included. For the downscaled dataset, only one ensemble run was available for each GCM. For the raw data, eight GCMs had daily time-series data for more than one ensemble run, while the remaining 21 only had daily data for a single run of the model. We identified a total of 57 model runs from the raw GCM dataset for further analysis (henceforth termed 'raw GCMs' dataset), and 32 model runs from the downscaled dataset (or 'LOCA' dataset) (Table A.1 and A.2, Appendix A). Out of these model runs, 26 runs were common to both datasets. That is, for 26 models, raw data and LOCA downscaled data, were available for the same ensemble run, allowing for one-to-one comparison.

Two evaluation criteria were used to assess model skill for computing chill hours: (a) comparison between modeled and observed multi-year mean annual chill hours for the historical period of 1971–2012; and (b) comparison between modeled and observed slope/trend in chill hours (1971–2012). We chose these two criteria because, if models are to be used to project the future effect of climate change, it is crucial to test not only their ability to capture the mean state but also the historical trend (i.e. slope of chill hours), the latter of which indicates the models' historical response to climate forcing. Also, depending on the choice of method, both mean and slope of chill hours can be relevant for computing future chill hours from climate model data. For example, future chill hours can be obtained directly by taking

[4] RCP 8.5 represents a high pathway for future greenhouse gas emissions for which radiative forcing reaches greater than 8.5 Wm^{-2} by 2100 and continues to rise (IPCC, 2013).

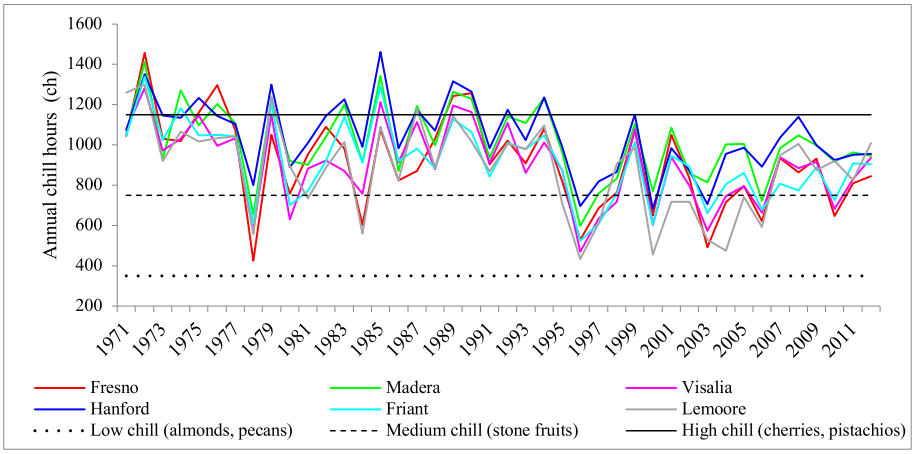

Fig. 1. Observed annual chill hours for the historical period 1971–2012, computed from 6 weather station records in the vicinity of Fresno, CA. For reference, three annual chill hour thresholds are shown as straight lines. These are broad indications of thresholds, lower than which many varieties of the listed crops will be negatively impacted.

the mean annual value from the climate models for the period under consideration. Alternately, future chill hours for the period can also be computed by taking a future slope estimate from the climate models and using it to extrapolate the current observed baseline mean (Daniels et al., 2012; Hawkins et al., 2013). We note that mean annual chill hours and slope of chill hours over time are related metrics, and they are not fully independent of each other.

The comparison between modeled and observed data was conducted by using p-values from a Student's *t*-test comparing the observed results to the modeled outcomes obtained for the different model runs. The p-value provides a means of identifying which model runs differ from observations in a statistically significant manner. In this case, a model with a low p-value can be said to be definitely unskillful; i.e., the modeled and observed results differ significantly. However, a high p-value does not guarantee skillful performance of the model. A high p-value can also reflect a noisy observational signal or a high degree of variability in the model, either of which may make it difficult to reject the null hypothesis that the modeled and observational data in fact reflect the same distribution. We also acknowledge that since a large number of pairwise *t*-test comparisons are being performed here simultaneously, there is a potential of Type I error, or the "false positive" error of rejecting a null hypothesis when it is actually true (Weisstein, n.d.). However, because our criterion for a model's 'success' is failing to reject the null hypothesis, we believe that using an uncorrected p-value of 0.05 makes our approach more conservative and stringent. In any case, we also performed a sensitivity by applying the Bonferroni correction that rectifies the Type I error, and found that the correction did not change our overall results significantly.

Table 1
Summary of results from the comparison of model performance against historically observed chill hours from 1971 to 2012. Rows are the different types of datasets that were evaluated, and columns present multi-model average predictions for annual mean and slope of historical chill hours for two different model samplings. "All models" refers to the average prediction from all the model runs in the dataset (also referred to sometimes as "multi-model mean"). "Sample based on chill hours skill" refers to the mean prediction from only those model runs that have a p-value of greater than 0.05 when modeled results were compares to historical observations. The observed mean and slope are also provided for comparison.

	Annual Mean Chill Hours (ch)		Chill Hours Slope (ch/yr)	
Dataset type (below)	All models	Sample based on chill hours skill	All models	Sample based chill hours skill
Observed	910	910	−8.4	−8.4
LOCA	850	860	−1.9	−3.5
Raw GCMs	1290	950	−3.6	−5.3

4. Results

4.1. Observed historical chill hours show a negative trend over time

Fig. 1 shows annual chill hours from 1971 to 2012 from the 6 weather stations. The multi-year mean annual chill for the historical period ranged from 900 to 1040 chill hours (ch) for the different stations. For this 41-year time period, a statistically significant negative trend in chill hours (at 95% confidence) was observed in all locations. The trend ranged from −6 to −8 chill hours per year (ch/yr). There was significant variability at annual to decadal scales, and hence shorter time intervals did not always demonstrate a statistically significant decline in chill.

To put these results in a decision-making context, the fruit and nut crops of California require anywhere between 200 and 2000 ch annually for optimal yields. Most almond and pecan cultivars have a relatively low annual chilling requirement of 200 to 500 ch, whereas walnuts, plums, peaches, and nectarines require medium chill of at least 650 ch. Many cultivars of cherries, pistachios, apples, and pears have a higher chill requirement of over 1000 ch (Baldocchi and Wong, 2007; UCANR, 2018). The observed data shows that in the last two decades, the Fresno region has seen an increase in the number of years with chill lower than 700 ch.

To rule out the possibility that data from the Fresno weather station (near Fresno Airport) were influenced by the urban heat island (UHI) effect, we cross-checked data at this weather station with gridded observed data from a grid area of 2000 km^2 around the Fresno region to examine potential biases (if any). We found that the chill hour estimates from the grid cell were very similar to our point location estimates (Fig. A.2 and Table A. 3, Appendix A). In addition, the mean and slope of historical chill hours from the 5 other nearby weather stations (Fig. 1 and Table A. 4, Appendix A) were also comparable to that of Fresno. This suggests that the UHI effect did not significantly bias our chosen dataset.

4.2. Model skill for chill hours

Table 1 provides a summary of results from the historical skill evaluation conducted for the two climate model datasets: LOCA and Raw GCMs.

4.2.1. LOCA data accurately predicts historical mean but underestimates trend

The historical mean chill hour predictions from the LOCA downscaled dataset were largely accurate, with a majority of GCMs having p-values greater than 0.05, meaning that they were statistically indistinguishable from the observations. Since the inter-model variability in chill hours mean was not very high, higher p-values can be inferred to

represent models that predict historical mean chill hours close to the observed data from the Fresno station. The overall annual mean prediction from all the 32 LOCA model runs (i.e. the multi-model mean) was 850 ch, which aligns fairly closely with the observed annual mean of 910 ch between 1971 and 2012. The spread in the multi-model mean predictions was low, with the highest annual estimate being 900 ch and the lowest 800 ch (Refer Fig. A.3, Appendix A). If models were to be sampled based on skill in predicting mean chill hours, then the sample size becomes 25 model runs (representing models with p-value greater than 0.05). However, the average results from these runs was very similar to the overall multi-model mean, suggesting that model choice does not strongly influence prediction of historical mean chill hours from LOCA data. This result is unsurprising, as the downscaled data is designed to closely correspond to the observed data that provided the basis for LOCA bias correction.

On the other hand, the prediction of chill hour slope is more sensitive to the type of model selection. The multi-model mean prediction of chill slope from all 32 LOCA models was −1.9 ch/yr, which is significantly less than the observed slope of −8 ch/yr. If models were sampled for skill in chill hour slope predictions, then the sample size becomes 15 model runs that had a p-value greater than 0.05. This skill based sampling resulted in a slope prediction of −3.5 ch/yr, which is slightly closer to the observed slope as compared to the multi-model slope prediction, but still on the lower side. Surprisingly, none of the LOCA model runs predicted a mean slope that was equal to or higher in magnitude than the observed mean slope of −8 ch/yr, showing that the multi-model spread in slope was also not very high. In fact most LOCA model runs (23 of 32) predicted the mean historical slope to be less steep than even the lower confidence interval of the observed slope (lower than −3.2 ch/yr). Overall, the LOCA dataset systematically underestimated the significant negative slope of chill hours that was observed in the historical period across a variety of locations in and around Fresno (Refer Fig. A.4, Appendix A).

The average results from chill hours skill-based model sampling are presented in this and subsequent sub-sections for comparison purposes only, and not to suggest that this sampling be used for decision applications. The implications of these results are not straightforward, and hence are further detailed in the discussion section.

4.2.2. Raw GCMs tend to overestimate mean and show highly variable historical trend predictions

Because the LOCA data predicted chill hours trends that are systematically different from observations, this raises a question as to whether this bias originated in the GCMs themselves or in the downscaling process. Our skill evaluation of the raw GCMs provided insights into how the LOCA process transformed the raw model data.

In contrast with the accurate historical mean predictions of the LOCA dataset, the multi-model mean annual chill hour predictions from all the 57 model runs of the raw GCMs was 1290 ch, which is significantly higher than observed values. These raw GCMs seemed to have biases in the same direction (colder than observed) as compared to the historical observed mean chill hours, and the multi-model mean was even colder than the coolest decades of our historical dataset. If the models were sampled for skill in raw GCMs for prediction of mean chill hours, then the sample size became quite small with only 4 of the 29 raw GCMs (or 12 of 57 model runs) having model runs with a p-value greater than 0.05 (Fig. 2). The historical mean annual chill as predicted by this skill based sampling of model runs was 950 ch. While this prediction is more in line with the observed value of 910 ch, it is limited by the small sample size. Only one model underestimated the mean, while 27 of 57 model runs predicted the mean as greater than the upper confidence interval of the observed annual chill hours (greater than 1350 ch), again showcasing that the raw GCM biases seemed to be in the same direction rather than offsetting each other (as would usually be expected in multi-model ensembles). The intra-model variability in prediction of mean chill was not very high; different runs of a GCM gave similar results.

Here, we note that the statistical correspondence between models and observations is not expected to be perfect even in the case of a hypothetical perfect model, because we are comparing larger GCM grid cells (of varying sizes and locations) to point observations. In particular, inclusion of the nearby Sierra Nevada mountain range in the grid cell could explain the overestimation of chill hours in some cases. However, when we examined the grid cell locations of the models closely, we realized that even the models that did not include the Sierra Nevada range exhibited this overestimation. The grid cell locations of the 4 high-ranked models were also notably different, indicating that other model characteristics are predominantly driving the chill hour predictions, rather than the size and location of the grid cells.

Fig. 3 compares the observed historical slope in chill hours with simulated results. The results show that the raw GCMs predicted the slope of historically observed chill hours more accurately than the LOCA data, although both approaches tended to underestimate the magnitude of the slope. The multi model mean prediction of slope from all 57 model runs was only −3.6 ch/yr, which was more accurate than the LOCA prediction of −1.9 ch/yr, but which was still less than half the magnitude of the observed chill slope (-8.4 ch/yr). However, the raw GCM slopes seemed to have a wider spread in results than the LOCA slopes, with individual model runs having slopes ranging from −12.4 to + 4.5 ch/yr. Eleven out of 57 model runs actually predicted an increasing trend in chill slope, while 6 runs predicted a decreasing slope steeper than −6 ch/yr.

The raw GCM results also exhibited high variability even within ensemble members of the same model, indicating that inter-annual to decadal scale internal variability plays a large role in slope predictions. For example, the slopes for the 10 runs of CSIRO-Mk3-6-0 ranged from −10.4 to +4.3 ch/yr, with one run ranked as the second best predictor of historical slope and another ranked as the second worst. Other models such as EC-EARTH and MIROC5 also showed high intra-model variability in prediction of slope. These results suggest that the prevalent inability of the p-value score to reject the null hypothesis could be due to the fact that the declining trend in chill hours may be strongly influenced by internal variability in the climate system in addition to forcing, making it difficult to discern true model skill for chill slope. Hence, even though 42 out of 57 model runs had p-values greater than 0.05 (i.e., they could not be statistically distinguished from observations), it is difficult to say whether or not they are truly skilled in capturing chill slope.

Because raw GCM slopes were overall more accurate than LOCA slopes and showed a better distribution around the observed slope of −8.4 ch/yr, GCM performance did not explain the systematic underestimation of historical chill hours slope in the LOCA data. We further clarified this by comparing the GCMs to the LOCA data in a pairwise fashion, examining pre- and post- downscaling slopes for the same model runs (Fig. 4). When we aggregated LOCA data for the same spatial area as the GCM grid cells, we found that these flat slopes were observed over the entire grid cell. This indicates that the systematic bias in the LOCA slopes most likely developed during the downscaling process and was not due to the scale difference between the GCMs and LOCA datasets. On the other hand, the LOCA downscaling was able to eliminate the cold bias in the model predictions of historical mean chill hours.

4.2.3. Comparing broad regional skill versus specific local skill of GCMs

In order to explore our main objective of understanding the difference between broad regional and specific local skill of climate models, Fig. 5 compares historical mean and slope predictions of chill hours from different model samplings: all models, models with good skill for California's regional climate, and models with good skill for chill hours in Fresno. As detailed in Section 2, for models with good regional skill in predicting California's climate, the state has conducted a series of detailed model evaluation studies (CCTAG, 2015; Pierce et al., 2016)

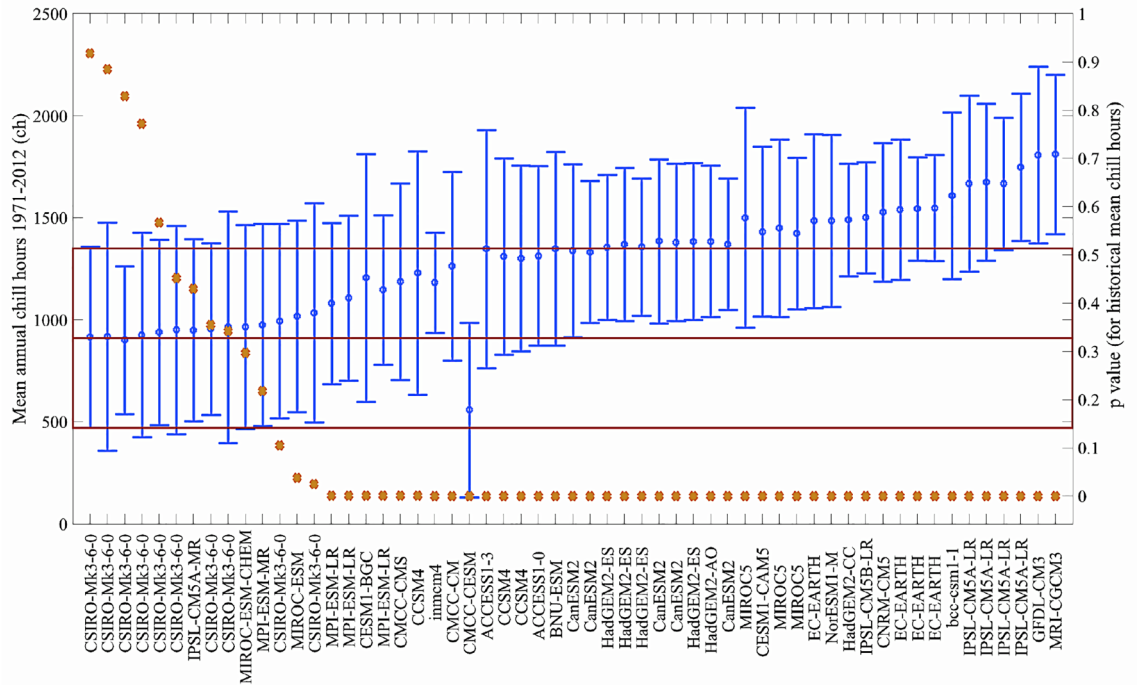

Fig. 2. Statistical analysis comparing model performance to the historically observed mean annual chill hours from 1971 to 2012: On the X-axis, models are ranked by accuracy in predicting historical mean annual chill hours, with the best models on the left and worst on the right. The red box represents the range of the observed mean chill for the Fresno weather station (95% confidence interval). The blue dots represent the predicted historical mean annual chill hours from each GCM simulation, and the blue error bars are the standard error of the mean (95% confidence interval). The orange markers are the p-values (for mean chill hours) derived from Student's *t*-test comparing each GCM simulation to the observed data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

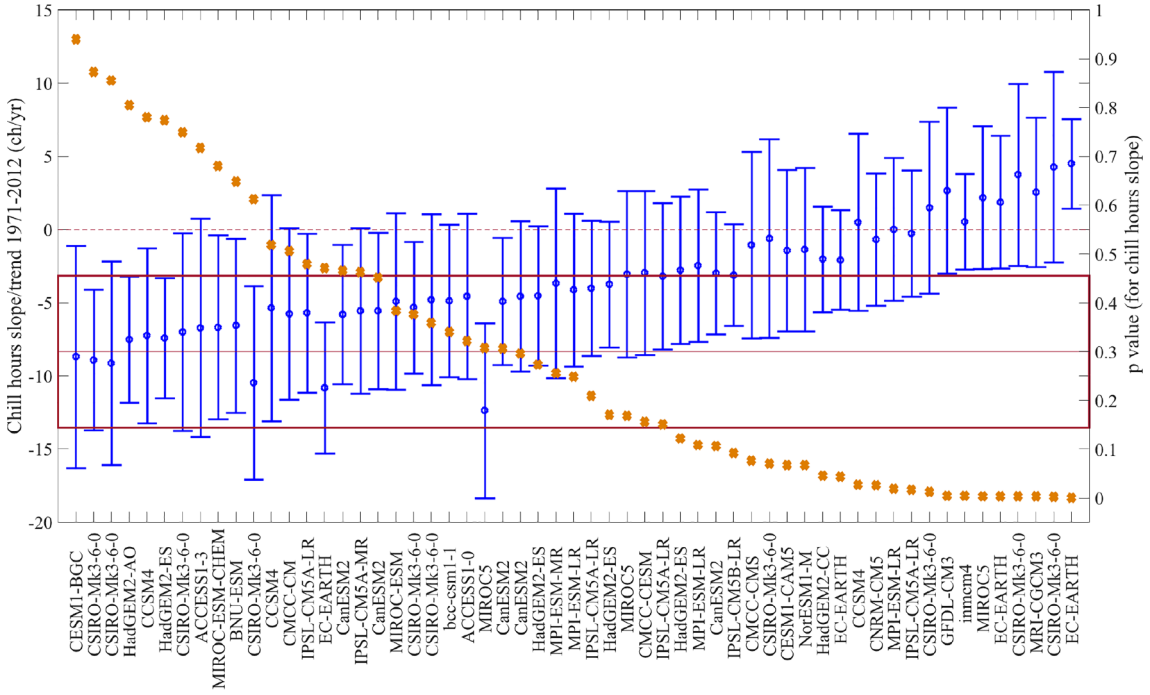

Fig. 3. Statistical analysis comparing model performance to the historically observed chill hours slope from 1971 to 2012: On the X axis, models are ranked based on accuracy in predicting historical slope in chill hours, with the best models on the left and worst on the right. The red box represents the 95% confidence interval of the observed chill slope for the Fresno weather station. The red dashed line indicates 0 slope, i.e., no change in annual mean historical chill hours over time. The blue dots represent the mean chill hours slope from each GCM simulation, and the error bars represent the 95% confidence interval of the slope. The orange markers are the p-values for the slope derived from Student's *t*-test comparing each GCM simulation to the observed data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

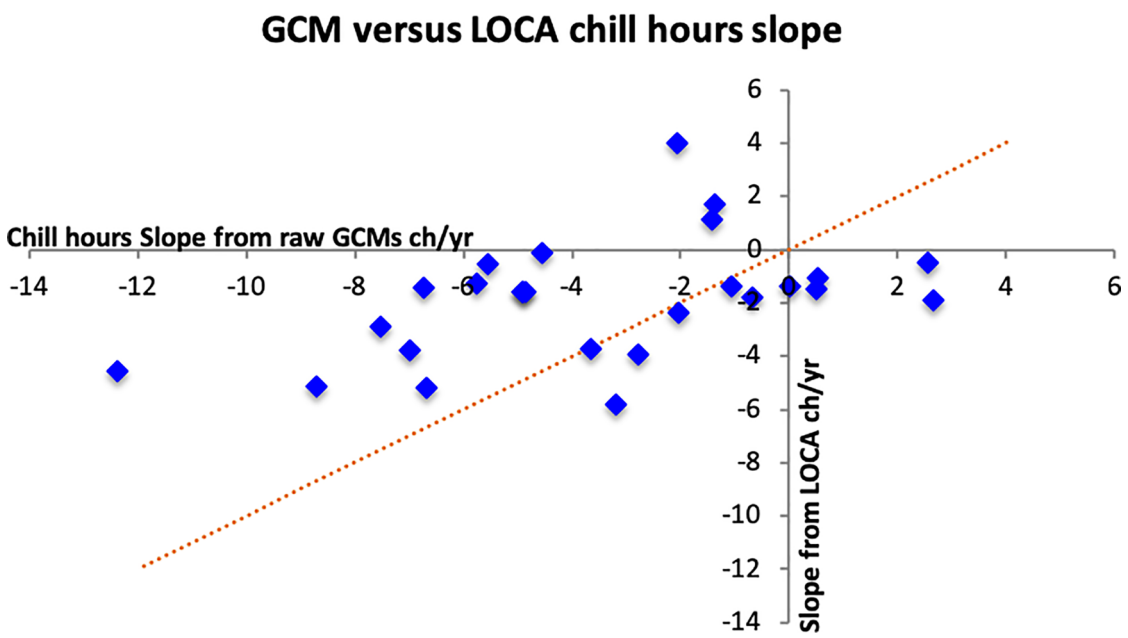

Fig. 4. Pair-wise comparison between chill hours slope predictions from raw GCMs (X-axis), and from LOCA (Y-axis) illustrating pre- and post- downscaling slopes for the same model runs. Each of the blue diamond markers represents slope predictions from a single GCM run pre- and post- LOCA downscaling. The orange dotted line is presented as a reference linear line. The figure shows that the magnitude of the negative slope of the raw GCMs is systematically higher than the LOCA (i.e. more points are further to the top of the linear line than the bottom). The spread of the slopes is also much larger for the raw GCMs than the LOCA slopes, indicating that the bias may have originated during the downscaling process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and identified a subset of 10 and 4 models (henceforth termed 'CAL10′ or 'sample based on regional skill' and 'CAL4′ or 'reduced sample based on regional skill') as good predictors of multiple broad physical climatic metrics for California (such as seasonal temperature, precipitation, and El Niño teleconnections). We use the CAL 10 and CAL 4 models to represent models with good regional skill. For the models with good skill for chill hours in Fresno or 'sample based on chill hours skill' we use those models that have a $p > 0.05$ in our historical skill evaluation.

For historical mean chill hours, we see that none of the raw CAL10 GCMs had a p-value > 0.05, and Fig. 5 (a) shows that this sampling based on regional skill of raw GCMs has in fact eliminated those models that have a good local skill for mean chill hours. Therefore, while the tail end of the lower range of the multi-model sampling for raw GCMs (black error bar) still represents those raw GCMs with good local skill in predicting mean chill hours (red error bar), the regional samplings (blue error bars) completely eliminates this range that is more in line with the observed mean chill hours. On the other hand, mean chill hours from LOCA seem to be less sensitive to model choice, with most model samplings presenting similar historical predictions.

For chill hours slope Fig. 5 (b), the regional sampling of raw GCMs (dark blue error bar) seems to have slightly underestimated the average negative trend as compared to both the multi-model sampling (black error bar) and the local skill-based sampling (red error bar). The regional skill-based samplings exhibit a more pronounced underestimation of slope for the LOCA data (comparing blue error bars of LOCA data with the red). However, since the LOCA data was shown to systematically underestimate historical slope, almost all samplings have a smaller average slope, with only the local skill-based sampling giving results which are reasonably closer to the historical slope.

In general, we did not find the models with good regional skill to be better skilled at predicting Fresno chill hours than the other GCM samplings. Fig. 5 shows that the CAL10 and CAL4 models (the blue bar plots) do not systematically perform better at predicting historical mean or slope, as compared to the overall multi-model mean or the high-ranked models.

4.3. Future projections of chill hours: Implications of skill-based model samplings

The range of skills that we observed across models and the two datasets raised questions about the extent to which skill in predicting historical chill affects predictions of future chill. For a climate data user, there are several ways of getting to a projection of the future. These include a variety of choices in selecting a dataset (raw or downscaled), choosing different types of samples of GCMs, and identifying a calculation method for obtaining future projections. Using the specific example of mid-century (2050) projections of chill hours, we highlight the similarities and differences in results based on these different choices. We illustrate our results using the raw and downscaled climate model datasets, four model samplings (also used in Section 4.2.3), and two calculation methods for future projections. Table 2 describes each of these choices, and Fig. 6 shows results for 2050 chill hours from the different sets of choices. In addition, we also computed chill hour projections for three time periods: Short-term (2010–30), Medium-term (2010–50), and Long-term (2010–80) which are presented as Fig. A.3 and A.4 in Appendix A.

Although it is unlikely that climate data users would choose to use raw GCM datasets directly, we explore projections from this dataset as it provides a means to assess the impact of model skill on future projections (see Section 3.2). Our results suggest that using the raw direct mean from all models and the regional skill-based samples (CAL10 and CAL4) provide the most unrealistic prediction method (Cluster a, Fig. 6), as it implausibly claims that chill in 2050 will be higher than the historically observed chill of 910 ch. We see that the cold-bias in the raw GCMs (described in 4.2.2) continues into the future, and most models continue to have biases in a single direction. And since the regional skill-based sample does not contain any models that had a good local skill for chill hours (also described in 4.2.2), the range of projections from the regional skill-based sample completely eliminates the range of projections from the models with good mean chill hours skill (that is, in Cluster a, Fig. 6, the blue error bars do not have any ranges that are in common with the red error bars). The projections from the

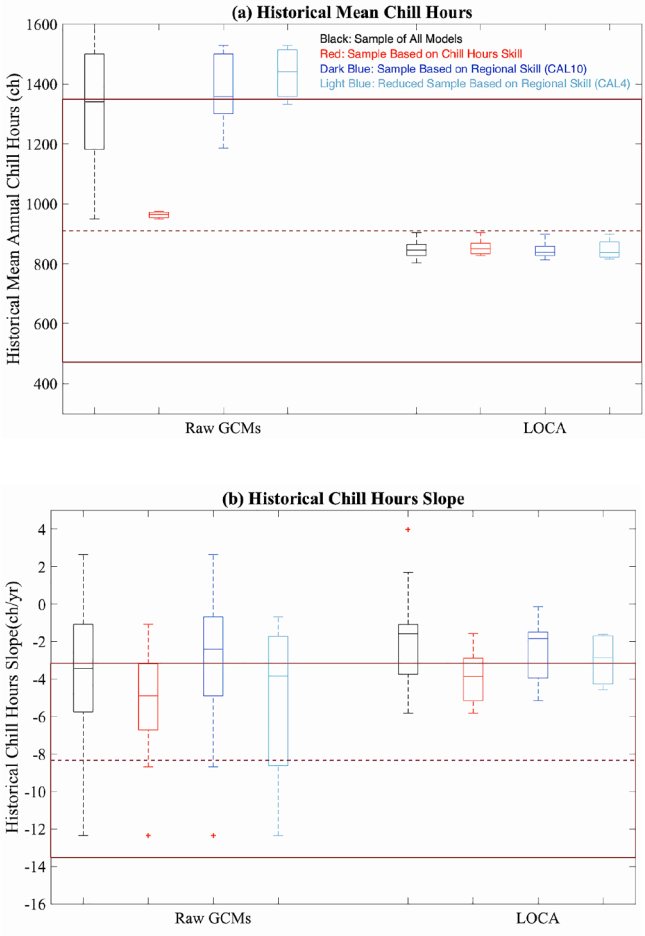

Fig. 5. Historical mean chill hours and chill hours slope predictions from different model samplings. The error bars show the historical predictions from different sampling and the magenta box represents the observed range (95% confidence interval). The figure shows that model samples based on good broad regional skill for predicting California's climate (dark and light blue plots) do not perform better than the overall mean or the sample based on skill for chill hours, i.e., the predictions from the light and dark blue plots are no closer to the observed mean or slope (magenta dashed line) than the other model samplings (red and black plots). To avoid an uneven comparison, these box plots represent only those 26 model runs for which both raw GCM and LOCA data was available. Therefore, these results may be slightly different than the results in Sections 4.2.1 and 4.2.2, which included all of the model runs evaluated in each dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

raw GCM sample based on chill hours skill (red error bar Cluster a, Fig. 6), fall within a more plausible range; they are lower than historically observed mean chill hours and are also more in line with estimates derived from the LOCA dataset. On the other hand, future projections using chill slopes from raw GCMs (Cluster b, Fig. 6) are not very sensitive to model samplings, as there is not a lot of difference in results represented by the different colored error bars.

The LOCA dataset warranted close examination, as it a preferred dataset for use in impact and adaptation studies in California (CATRWG, 2017). We note from clusters c and d of Fig. 6 that the LOCA dataset (irrespective of model choice) leads to higher projections of chill hours in 2050, which could perhaps indicate that the historical underestimation of the negative slope by the LOCA dataset is continuing into the future. Our preliminary evaluation indicates that these lower slope predictions could be related to a fundamental issue of the LOCA data underestimating the warming trend of winter season (November-December-January-February) $T_{min}$and T_{av}, as compared to the trend observed in the raw GCMs (Refer Fig. A.6, Appendix A).

With regards to the implications of model selection, we observe that the regional skill based samples of CAL10 and CAL 4 models for LOCA (dark blue box plots in clusters c and d of Fig. 6) fail to capture the range of projections represented by either the multi-model ensemble or the sample based on chill hours skill (black and red box plots in clusters c and d). Both the black and red box plots extend further in the direction of lower future chill hours, a direction of greater potential concern to growers. The reduced regional skill based sample of CAL 4 models shows an even narrower range of future chill projections compared to the CAL 10 models when direct LOCA means are used (difference between light and dark blue error bars in cluster c of Fig. 6). Further, the CATRWG classifies the CAL4 models as HadGEM2-ES being the "warm & dry" model, CNRM-CM5 is "cool & wet", CanESM2 has projections in the "middle", and MIROC5 provides "complementary" projections or "covers a range of outputs" for certain broad physical climate metrics (CATRWG, 2017). However, we found that these classifications are not valid for the chill hours metric; for example, HadGEM2-ES was not the warmest model, and CanESM2 did not provide average projections of chill hours (Appendix B: Sheet titled 'Data for Figs. 4, 6 & 7'). Overall, clusters c and d of Fig. 6 indicate that the chill hour skill-based sampling for the LOCA data, in fact, covers a larger spread in future projections while the regional skill-based samples show a much smaller range. These results strongly indicate that different skill based sampling approaches can have important repercussions for the analysis of future chill hours.

5. Discussion

This paper explores the implications of model choice for chill hours, focusing on two key criteria for model selection: (1) past performance of models for the metric of interest, and (2) whether the models are able to capture a full range or legitimate spread of potential future projections. In terms of past performance, we were surprised to find that none of the raw GCMs that performed well for key physical climatic metrics in California, showed good skill for chill hours in Fresno. In terms of spread, since the regional and local skill-based model samplings were mutually exclusive, there was no overlap in the range of projections predicted by the two samples of raw GCMs. For the downscaled LOCA dataset, the regional skill-based sample of 10 and 4 models, showed a much smaller range/spread in future projections than the multi-model ensemble or the sample based on chill hours skill. This result is interesting because these 10 GCMs have been proven to "represent the magnitude and spread of temperature and precipitation change over the 21st century similar to those of the full set of 31 CMIP5 GCMs" (CCTAG, 2015) (pp 52), and the reduced set of 4 models are shown to "substantially cover the results from the set of 10" (Pierce et al., 2016) (pp 1). However we find that, they do not behave the same way to capture the range of projections for the chill hour metric. While it is difficult to say whether or not this narrower range of projections is appropriate, nevertheless, it illustrates the complex relationship between broad regional versus specific local skill, and questions the appropriateness of such regional skill-based samplings for specific decision applications at a local scale.

To provide an additional and slightly more focused comparison of broad regional versus specific local skill, we used Rupp et al.'s (2013) model evaluation for the US Pacific Northwest – to identify a sample of models that ranked well (based on the authors' ranking criteria) for several broad temperature metrics that were most related to chill hours (e.g. diurnal temperature range (DTR) in December-January-February (DJF) and spatial correlation of the observed to modeled climatological mean DJF temperature). Again, we found no relationship between this sample of 7 GCMs with good regional temperature skill and the models that ranked well for the chill hour metrics in our analysis, further illustrating that regional skill does not always ensure specific local skill.

If regional skill-based sampling is complicated and does not provide

Table 2
Description of different climate datasets, model samplings and projection calculation methods that were used to illustrate the differences in future projections in chill hours due to model choice.

	Choices	Description
Climate Model Dataset	"Raw GCMs"	Dataset with the uncorrected GCMs.
	"LOCA"	Dataset with GCMs that were downscaled based on the Localized Climate Analogues method.
Model Sampling	"Sample of All Models"	Refers to all of the GCMs (and model runs) within a particular dataset.
	"Sample Based on Chill Hours Skill"	Refers to only those GCM runs that had a p-value greater than 0.05 from the historical skill evaluation (where mean is used to compute chill, the p-value score for mean was used, and where slope is used the p-value for slope was used).
	"Sample Based on Regional Skill"	Refers to the CAL 10 models identified by the State of California as having good predictions of California's climate.
	"Reduced Sample Based on Regional Skill"	Refers to the smaller subset of CAL 4 models that encompass the range of projections that the CAL10 models produce.
Projection Calculation Method	"Direct mean" method	Refers to directly using the mean chill from the sample GCMs for the required time period. To avoid random errors due to use of a single year, we used the average annual chill from 2040 to 2050 as projections of 2050.
	"Slope" method	Refers to indirectly calculating future chill using a historical observed baseline of mean chill, and applying the chill slope (from sample GCMs) for the future period (i.e.) 2050 mean annual chill hours = Fresno observed mean annual chill hours 2001–2010 – (40 * chill slope 2010–50)

a clear advantage, is it appropriate to simply use the multi-model mean? Many studies have shown that the MMM is often the best predictor for several temperature metrics, including seasonal temperature metrics such as JFM (January-February-March) T_{min} ,T_{av} and variability in California (Pierce et al., 2009), and monthly T_{av} in the Western US (Rupp et al., 2013). This is because the multi-model mean often leads to cancellation of offsetting errors in the individual global models. By contrast, our research finds that most raw GCMs have errors/biases in the same direction for our metric of interest, leading to a systematic overestimation of mean chill hours (Section 4.2.2). This is in line with other studies that suggest that the MMM might not provide a better estimate when assessing single variables (Knutti et al., 2010). Therefore, the multi-model raw GCM mean may not increase the accuracy of predictions of mean chill hours, and hence cannot be used uncritically.

Overall, this paper suggests that the peculiarities of specific decision-relevant metrics – such as this non-linear threshold-based chill

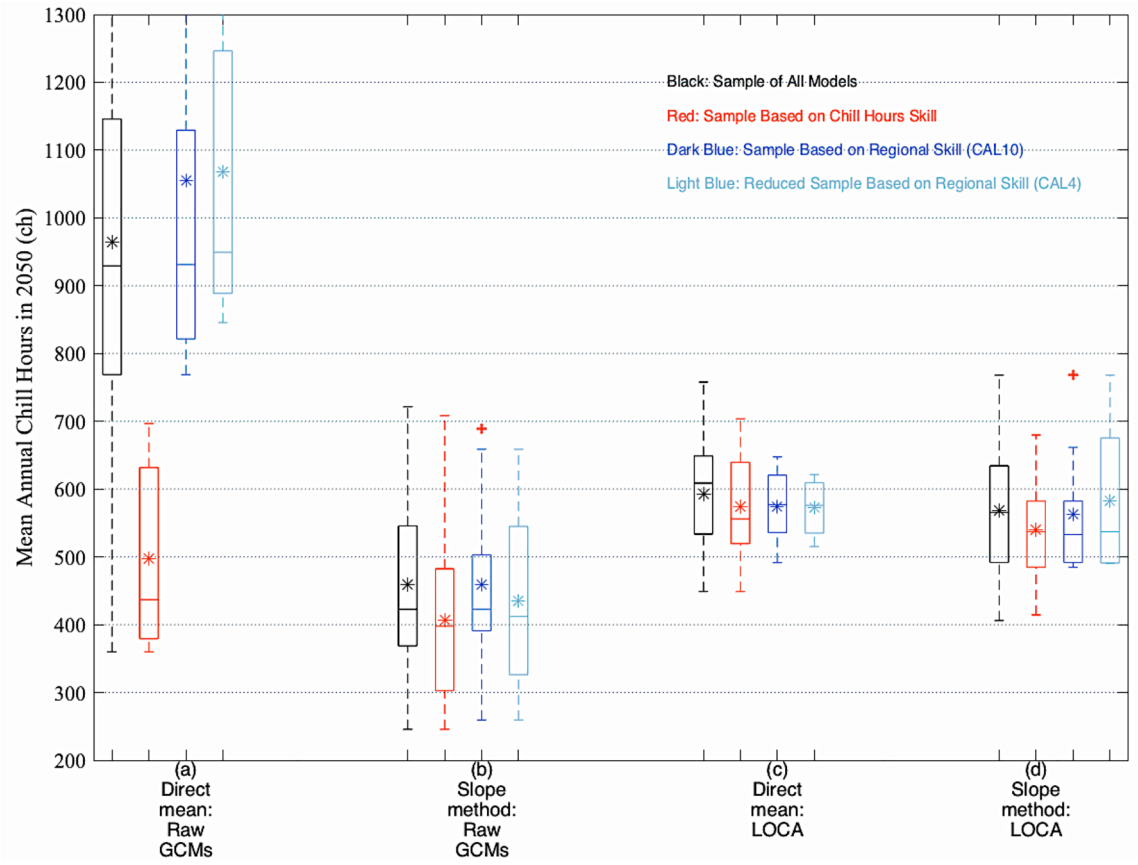

Fig. 6. 2050 chill hours computed using different climate datasets, model samplings, and computation methods. Each cluster of box plots represents 2050 chill hours computed through one method (direct mean or slope method), and for one dataset (raw GCM or LOCA). The asterisks show the mean values, and the red plus signs represent the outliers. To make the details legible, we chose a Y-axis limit of 1300 ch that excludes some of the more extreme results; the un-cropped version of the graph is in Appendix A, Figure A.5. To avoid an uneven comparison, these box plots represent only those 26 model runs for which both raw GCM and LOCA data was available. Therefore these results may be slightly different than the results in Sections 4.2.1 and 4.2.2, which included all of the model runs evaluated in each dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hour metric – can lead to counterintuitive findings that question the validity of some generally accepted recommendations on climate model selection for impact and adaptation studies. Chill hours only accumulate when temperature is below 7.2 °C, which means there is a non-linear threshold relationship between temperature and chill hours (Fig. A.7, Appendix A). This means that the relationship between model skill for temperature versus that of chill hours is not straightforward: some physical errors may be amplified and others may be dampened often in non-intuitive ways.

Our intent is not to suggest that model sampling be solely based on a single decision-relevant metric as such a narrow focus may lead to over-determination the accuracy of the current generation of GCMs. Rather we highlight that broad regional skill alone may also not be sufficient to guarantee specific localized skill, which implies that there are additional physical processes that are important for decision-relevant metrics, and that are not well represented in regional skill evaluations. Moreover, we also highlight that there may be cases where broad regional skill and specific local skill are at odds with one another. Further research is needed to critically understand why models are not performing well for crucial decision-relevant processes, and additionally on how models must be selected when broad and specific skill do not align.

6. Conclusion

Adaptation studies often tend to use model sampling that is either based on convenience, or based on model skill for predicting broad regional climatic metrics, without fully understanding the implications of such model choices for decision-relevant metrics at local scales. This paper highlights the implications of model choice for decision-relevant metrics, which can be useful to both climate researchers and climate information users who are interested in, or using different model selection approaches. Using the case of the decision-relevant metric of chill hours, we find that broad regional climate skill of models alone is insufficient, and does not always ensure that models are skilled for decision-relevant metrics. Hence, an additional layer of decision-relevant model evaluation may be necessary, particularly if the metric has a non-linear relationship with primary physical quantities. Since many crucial adaptation decisions in agriculture, energy, water management, and other fields are made using threshold-based metrics (such as growing degree days, heating or cooling degree days, and days over 100°F), further such evaluations are needed to better understand how models perform on the eventual metric of relevance to decisions. Such assessments can also help identify the relevant physical climatic processes that most robustly represent decision-relevant metrics, and ensure that future research focuses on refining model representation of the processes that are most relevant to stakeholder applications.

Further, even after such evaluations are conducted, there remain many unanswered questions on how to choose models based on broad regional and decision-relevant local evaluations. Do models need to perform well on both evaluations? What if there are no models with both broad regional and specific local skill? Decision-makers around the world face an urgent need to implement science-based adaptation measures. Hence, there is a critical need for more nuanced research on model selection strategies for decision applications, to ensure that adaptation actions are based on the best available climate projections for the specific context.

Acknowledgments and Data

The authors would like to acknowledge Margaret Torn and Peter Nico for their critical guidance and support throughout the project. The authors would also like to thank Tapan Pathak and David Doll of the University of California (UC) Cooperative Extension for their valuable perspectives on the potential use of climate projections for farm-level adaptation decisions. This work was supported by the UC Office of the President (Global Food Initiative, contract number AWD00000169) as part of a strategic partnership funding to Berkeley Lab, and by the Office of Science, Office of Biological and Environmental Research, Climate and Environmental Science Division, of the U.S. Department of Energy under contract DE-AC02-05CH11231 as part of the Hyperion Project (An Integrated Evaluation of the Simulated Hydroclimate System of the Continental US, award number DE-SC0016605). The project also received support from the U.S. Department of Agriculture's California Climate Hub in the form of a student fellowship. Background data is in Appendix B of the supplementary material. Weather station records of daily temperature can be accessed from https://wrcc.dri.edu/. The CMIP5 model output data and the LOCA downscaled data used in this work can be downloaded from https://esgf-node.llnl.gov/search/cmip5/ and https://gdo-dcp.ucllnl.org/, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cliser.2020.100154.

References

Baldocchi, D., Wong, S., 2007. Accumulated winter chill is decreasing in the fruit growing regions of California. Clim. Change 87, 153–166. https://doi.org/10.1007/s10584-007-9367-8.

Barsugli, J.J., Guentchev, G., Horton, R.M., Wood, A., Mearns, L.O., Liang, X.-Z., Winkler, J.A., Dixon, K., Hayhoe, K., Rood, R.B., Goddard, L., Ray, A., Buja, L., Ammann, C., 2013. The practitioner's dilemma: How to assess the credibility of downscaled climate projections. Eos. Trans. Am. Geophys. Union 94, 424–425. https://doi.org/10.1002/2013EO460005.

Brekke, L.D., Dettinger, M.D., Maurer, E.P., Anderson, M., 2008. Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. Clim. Change 89, 371–394. https://doi.org/10.1007/s10584-007-9388-3.

Cal-Adapt, 2017. Underlying Data and Model Selection in Cal- Adapt 2.0 [WWW Document]. URL https://cal-adapt.org/blog/2017/underlying-data-and-model-selection-in-cal-adapt-2-0/ (accessed 2.7.18).

CATRWG, 2017. Projected Climate Scenarios Selected to Represent a Range of Possible Futures in California Description. Sacramento, CA.

CCTAG, 2015. Perspectives and Guidance for Climate Change Analysis. Sacramento, CA.

CDFA, 2016. California Agricultural Statistics Review 2015–2016. Sacramento, CA.

County of Fresno, 2015. 2015 Fresno County Annual Crop & LIvestock Report. Fresno, CA.

Daniels, A.E., Morrison, J.F., Joyce, L.A., Croosktion, N.L., Chen, S.C., McNulty, S.G., 2012. Climate projections FAQ, Gen. Tech. Rep. RMRS-GTR-277WWW. Fort Collins, CO.

Darbyshire, R., Webb, L., Goodwin, I., Barlow, S., 2011. Winter chilling trends for deciduous fruit trees in Australia. Agric. For. Meteorol. 151, 1074–1085. https://doi.org/10.1016/j.agrformet.2011.03.010.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., Rummukainen, M., 2013. Evaluation of Climate Models. Clim. Chang. 2013 Phys. Sci. Basis. Contrib. Work. Gr. I to Fifth Assess. Rep. Intergov. Panel Clim. Chang. 741–866. https://doi.org/10.1017/CBO9781107415324.

Girvetz, E.H., Maurer, E., Duffy, P., Ruesch, A., Thrasher, B., Zganjar, C., 2013. Making Climate Data Relevant to Decision Making: The important details of Spatial and Temporal Downscaling 43.

Hawkins, E., Osborne, T.M., Ho, C.K., Challinor, A.J., 2013. Calibration and bias correction of climate projections for crop modelling: An idealised case study over Europe. Agric. For. Meteorol. 170, 19–31. https://doi.org/10.1016/j.agrformet.2012.04.007.

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

Hayhoe, K., Edmonds, J., Kopp, R.E., LeGrande, A.N., Sanderson, B.M., Wehner, M.F., Wuebbles, D.J., 2017. Climate Models, Scenarios, and Projections, in: Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Dokken, D.J., Stewart, B.C., Maycock, T.K. (Eds.), Climate Science Special Report: Fourth National Climate Assessment, Volume I. U.S. Global Change Research Program, Washington, DC, USA, pp. 133–160. https://doi.org/10.7930/J0WH2N54.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., Sanderson, B.M., 2018. Selecting a climate model subset to optimise key ensemble properties. Earth Syst. Dyn. 9, 135–151. https://doi.org/10.5194/esd-9-135-2018.

IPCC, 2013. Annex III: Glossary, in: Stocker, T.., Qin, D., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1447–1466.

Jones, A., Calvin, K., Lamarque, J.-F., 2016. Climate Modeling with Decision Makers in Mind. Eos (Washington. DC) 97, 2–5. https://doi.org/10.1029/2016EO051111.
Kerr, A., Dialesandro, J., Steenwerth, K., Lopez-Brody, N., Elias, E., 2018. Vulnerability of California specialty crops to projected mid-century temperature changes. Clim. Change 148, 419–436. https://doi.org/10.1007/s10584-017-2011-3.
Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G.A., 2010. Challenges in combining projections from multiple climate models. J. Clim. 23, 2739–2758. https://doi.org/10.1175/2009JCLI3361.1.
Lobell, D.B., Field, C.B., 2011. California perennial crops in a changing climate. Clim. Change 109, 317–333. https://doi.org/10.1007/s10584-011-0303-6.
Lobell, D.B., Field, C.B., Cahill, K.N., Bonfils, C., 2006. Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties. Agric. For. Meteorol. 141, 208–218. https://doi.org/10.1016/j.agrformet.2006.10.006.
Luedeling, E., Brown, P.H., 2011. A global analysis of the comparability of winter chill models for fruit and nut trees. Int. J. Biometeorol. 55, 411–421. https://doi.org/10.1007/s00484-010-0352-y.
Luedeling, E., Girvetz, E.H., Semenov, M.A., Brown, P.H., 2011. Climate change affects winter chill for temperate fruit and nut trees. PLoS One 6, e20155. https://doi.org/10.1371/journal.pone.0020155.
Luedeling, E., Zhang, M., Girvetz, E.H., 2009a. Climatic changes lead to declining winter chill for fruit and nut trees in California during 1950–2099. PLoS One 4, 1–9. https://doi.org/10.1371/journal.pone.0006166.
Luedeling, E., Zhang, M., Luedeling, V., Girvetz, E.H., 2009b. Sensitivity of winter chill models for fruit and nut trees to climatic changes expected in California's Central Valley. Agric. Ecosyst. Environ. 133, 23–31. https://doi.org/10.1016/j.agee.2009.04.016.
Madsen, M.S., Langen, P.L., Boberg, F., Christensen, J.H., 2017. Inflated Uncertainty in Multimodel-Based Regional Climate Projections. Geophys. Res. Lett. 44, 11606–11613. https://doi.org/10.1002/2017GL075627.
Maurer, E.P., Brekke, L., Pruitt, T., Thrasher, B., Long, J., Duffy, P., Dettinger, M., Cayan, D., Arnold, J., 2014. An enhanced archive facilitating climate impacts and adaptation analysis. Bull. Am. Meteorol. Soc. 95, 1011–1019. https://doi.org/10.1175/BAMS-D-13-00126.1.
McSweeney, C.F., Jones, R.G., Lee, R.W., Rowell, D.P., 2014. Selecting CMIP5 GCMs for downscaling over multiple regions. Clim. Dyn. 5, 3237–3260. https://doi.org/10.1007/s00382-014-2418-8.
Medellín-Azuara, J., Howitt, R.E., MacEwan, D.J., Lund, J.R., 2011. Economic impacts of climate-related changes to California agriculture. Clim. Change 109, 387–405. https://doi.org/10.1007/s10584-011-0314-3.
Mendlik, T., Gobiet, A., 2016. Selecting climate simulations for impact studies based on multivariate patterns of climate change. Clim. Change 135, 381–393. https://doi.org/10.1007/s10584-015-1582-0.
Moss, R.H., Avery, S., Baja, K., Burkett, M., Chischilly, A.M., Dell, J., Fleming, P.A., Geil, K., Jacobs, K., Jones, A., Knowlton, K., Koh, J., Lemos, M.C., Melillo, J., Pandya, R., Richmond, T.C., Scarlett, L., Snyder, J., Stults, M., Waple, A., Whitehead, J., Zarrilli, D., Fox, J., Ganguly, A., Joppa, L., Julius, S., Kirshen, P., Kreutter, R., McGovern, A., Meyer, R., Neumann, J., Solecki, W., Smith, J., Tissot, P., Yohe, G., Zimmerman, R., 2019. A framework for sustained climate assessment in the United States. Bull. Am. Meteorol. Soc. 100, 897–907. https://doi.org/10.1175/BAMS-D-19-0130.
Mote, P., Brekke, L., Duffy, P.B., Maurer, E., 2011. Guidelines for constructing climate scenarios. Eos (Washington. DC) 92, 257–264. https://doi.org/10.1029/2011EO310001.
Overland, J.E., Wang, M., Bond, N.A., Walsh, J.E., Kattsov, V.M., Chapman, W.L., 2011. Considerations in the selection of global climate models for regional climate projections: The Arctic as a case study. J. Clim. 24, 1583–1597. https://doi.org/10.1175/2010JCLI3462.1.
Pathak, T., Maskey, M., Dahlberg, J., Kearns, F., Bali, K., Zaccaria, D., 2018. Climate Change Trends and Impacts on California Agriculture: A Detailed Review. Agronomy 8, 25. https://doi.org/10.3390/agronomy8030025.
Pierce, D.W., Barnett, T.P., Santer, B.D., Gleckler, P.J., 2009. Selecting global climate models for regional climate change studies. Proc. Natl. Acad. Sci. 106, 8441–8446. https://doi.org/10.1073/pnas.0900094106.
Pierce, D.W., Cayan, D.R., Dehann, L., 2016. Creating Climate Projections to support the 4th California Climate Assessment. Sacramento, CA. https://doi.org/Docket #: 16-IEPR-04.
Rupp, D.E., Abatzoglou, J.T., Hegewisch, K.C., Mote, P.W., 2013. Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest US. J. Geophys. Res. 118, 1–23. https://doi.org/10.1002/jgrd.50843.
Sanderson, B.M., Knutti, R., Caldwell, P., 2015. A representative democracy to reduce interdependency in a multimodel ensemble. J. Clim. 28, 5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1.
Sanderson, B.M., Wehner, M., Knutti, R., 2017. Skill and independence weighting for multi-model assessments. Geosci. Model Dev. 10, 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017.
Snover, A.K., Mantua, N.J., Littell, J.S., Alexander, M.A., Mcclure, M.M., Nye, J., 2013. Choosing and Using Climate-Change Scenarios for Ecological-Impact Assessments and Conservation Decisions. Conserv. Biol. 27, 1147–1157. https://doi.org/10.1111/cobi.12163.
UCANR, 2018. Tree Selection, The California Backyard. Orchard [WWW Document].
WCRP, 2017. CMIP5 - Data Access - Availability [WWW Document]. Res. Program. - Coupled Model Intercomp. Proj. C, World Clim https://cmip.llnl.gov/cmip5/availability.html (accessed 12.17.17).
Weisstein, E.W., n.d. Bonferroni Correction [WWW Document]. MathWorld–A Wolfram Web Resour. URL http://mathworld.wolfram.com/BonferroniCorrection.html.