Bayesian Model Averaging of Climate Model Projections Constrained by Precipitation Observations over the Contiguous United States

E. C. MASSOUD AND H. LEE

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California

P. B. GIBSON

Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

P. LOIKITH

Department of Geography, Portland State University, Portland, Oregon

D. E. WALISER

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, and Joint Institute for Regional Earth System Science and Engineering, University of California, Los Angeles, Los Angeles, California

(Manuscript received 31 October 2019, in final form 17 July 2020)

ABSTRACT

This study utilizes Bayesian model averaging (BMA) as a framework to constrain the spread of uncertainty in climate projections of precipitation over the contiguous United States (CONUS). We use a subset of historical model simulations and future model projections (RCP8.5) from the Coupled Model Intercomparison Project phase 5 (CMIP5). We evaluate the representation of five precipitation summary metrics in the historical simulations using observations from the NASA Tropical Rainfall Measuring Mission (TRMM) satellites. The summary metrics include mean, annual and interannual variability, and maximum and minimum extremes of precipitation. The estimated model average produced with BMA is shown to have higher accuracy in simulating mean rainfall than the ensemble mean (RMSE of 0.49 for BMA versus 0.65 for ensemble mean), and a more constrained spread of uncertainty with roughly a third of the total uncertainty than is produced with the multimodel ensemble. The results show that, by the end of the century, the mean daily rainfall is projected to increase for most of the East Coast and the Northwest, may decrease in the southern United States, and with little change expected for the Southwest. For extremes, the wettest year on record is projected to become wetter for the majority of CONUS and the driest year to become drier. We show that BMA offers a framework to more accurately estimate and to constrain the spread of uncertainties of future climate, such as precipitation changes over CONUS.

KEYWORDS: Climate change; Satellite observations; Bayesian methods; Optimization; Climate models; Ensembles

1. Introduction

Annual precipitation averaged across the contiguous United States (CONUS) has increased approximately 4% from 1901 to 2015 (USGCRP 2017). Changes in precipitation are one of the most important potential outcomes of climate change because precipitation is a critical factor for the functioning of societies and ecosystems. Regional differences are apparent, as the Northeast, Midwest, and Great Plains have had

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-19-0258.s1>.

Corresponding author: Elias Massoud, elias.massoud@jpl.nasa.gov

DOI: 10.1175/JHM-D-19-0258.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

increases while parts of the Southwest and Southeast have had decreases in precipitation (Walsh et al. 2014; Easterling et al. 2017). Seasonal differences are also apparent, as the northern United States received more precipitation in the spring relative to historical climate, parts of the southwestern United States received less in the winter and spring, and the northeast United States received more in the summer and fall (Peterson et al. 2013). Furthermore, heavy precipitation events in most parts of CONUS have increased in both intensity and frequency (Dettinger 2011; Janssen et al. 2014, 2016; Easterling et al. 2017). In particular, mesoscale convective systems, which are the main mechanism for warm season precipitation in the central part of the United States, have increased in occurrence and precipitation amounts since 1979 (Feng et al. 2016). Additionally, atmospheric rivers, which are narrow jets of integrated water vapor transport that have large impacts on local weather and regional hydrology, have also increased in frequency and intensity in recent decades (Dettinger 2011; Dettinger and Cayan 2014). While uncertainty remains in the magnitude, and in some places, the sign of projected change in mean precipitation, future increase in the frequency and intensity of extreme events is broadly projected across most parts of CONUS (Dettinger 2011; Pierce et al. 2013; Janssen et al. 2014, 2016; Payne and Magnusdottir 2015; Warner et al. 2015; Gao et al. 2015; Radić et al. 2015; Hagos et al. 2016; Shields and Kiehl, 2016a,b; Espinoza et al. 2018; Massoud et al. 2019a).

Changes in precipitation in a warmer climate are governed by many factors. A primary physical mechanism for increases in precipitation is the enhanced water vapor content in a warmer atmosphere, which enhances moisture convergence into storms (Wang et al. 2015). Although energy constraints can be used to understand global changes in precipitation (Shepherd 2014), projecting regional changes is much more difficult because of uncertainty in projecting changes in the large-scale circulation. For CONUS, future changes in seasonal average precipitation will include a mix of increases, decreases, or little change, depending on location and season (Easterling et al. 2017; USGCRP 2017). Over the globe, high-latitude regions are generally projected to become wetter, whereas the subtropical regions are projected to become drier. Since CONUS lies between these two regions, there is significant uncertainty about the sign and magnitude of future changes to precipitation in much of the region, particularly in the middle latitudes of the nation. However, since atmospheric water vapor will increase with increasing temperatures, confidence is high that precipitation extremes in the

form of heavy rainfall will increase in frequency and intensity in the future throughout CONUS.

Through the Coupled Model Intercomparison Project (CMIP), the climate modeling community provides a suite of model simulations and projections (e.g., CMIP5; Taylor et al. 2012). These models are used to characterize climate projection uncertainty arising from model differences and large ensemble simulations (cf. Massoud et al. 2019b, 2020), as well as to characterize uncertainty inherent in the climate system due to internal variability (e.g., Kay et al. 2015). These ensembles provide an important resource for examining and evaluating the models that cause uncertainties in future climate projections (Lee et al. 2018). Often, when creating multi-model averages, projections of the future from each model are considered to be equally likely, without accounting for model skill or for the fact that some models are very similar to other models in the archive, which could lead to a biased weighting (Collins et al. 2013; Espinoza et al. 2018; Massoud et al. 2018). Owing to different model performances with respect to observations (Knutti and Sedláček 2013; Hidalgo and Alfaro, 2015; Gibson et al. 2019) and the lack of independence among models (Annan and Hargreaves 2011; Sanderson and Knutti 2012; Sanderson et al. 2015), there is evidence that giving equal weight to each available model projection is suboptimal (Knutti et al. 2010; Wenzel et al. 2014; Abramowitz and Bishop 2015; Alexander and Easterbrook 2015; Sanderson et al. 2015, 2017; Knutti et al. 2017; Eyring et al. 2019; Massoud et al. 2019a). The concept of model dependence becomes an issue when there is “double counting” of models that are simulated with common parameterizations or tuning practices, and therefore the assumption is that models that demonstrate dependence with other models should be penalized. These underlying assumptions have been challenged by a number of studies over recent years (Masson and Knutti 2011; Pennell and Reichler 2011; Knutti and Sedláček 2013; Sanderson et al. 2015; Knutti et al. 2017; Herger et al. 2018; Massoud et al. 2019a). Therefore, current advances in model averaging have focused on the concept of model skill as well as model independence (Knutti et al. 2010; Melillo et al. 2014; Herger et al. 2018; Massoud et al. 2019a).

There have been numerous studies that investigate the concept of model averaging based on model skill or independence. A number of studies have attempted to weigh models (only) accounting for model skill. For example, Tebaldi and Knutti (2007) proposed an ensemble averaging scheme that increased the weight of models with low observational biases, but the method potentially discounts outlier projections and does not consider model dependence. Bishop and Abramowitz (2013) proposed a

method that produced a set of statistically independent model averages from the original archive, and applied this method to CMIP5 projections in [Abramowitz and Bishop \(2015\)](#). While their method minimizes the error of the model average compared to an observed target and is by definition optimal, the coefficients of each model can be positive or negative, and negative weights can no longer be directly interpreted as physical entities that conserve mass or energy. Another study is [Langenbrunner and Neelin \(2017\)](#), which used sampled equally weighted models created by a mixture distribution strategy and defined pareto optimal solutions to project changes in precipitation over California. However, this method also did not consider model independence. [Sanderson et al. \(2015\)](#) present a weighting strategy for use with climate model ensembles, which considers both skill in the climatological performance of models as well as the interdependency of models. This strategy introduced by Sanderson et al. has been implemented in several studies so far (e.g., [Sanderson et al. 2015, 2017](#); [Knutti et al. 2017](#); [Massoud et al. 2019a](#)), as well as in the Fourth National Climate Assessment report ([Easterling et al. 2017](#); [USGCRP 2017](#)).

[Eyring et al. \(2019\)](#) thoroughly discuss the state of climate model evaluation and constraining the spread of uncertainty in future projections, and they show there is a need for better understanding advanced weighting methods to produce more tightly constrained climate projections. Therefore, there is a need for model averaging tools to more judiciously combine multimodel climate projections while considering the interdependency of models, such as the Bayesian model averaging (BMA) method ([Hoeting et al. 1999](#)). BMA is an approach that produces a multimodel average created from optimized model weights, which correspond to a distribution of weights for each model, such that the BMA-weighted model ensemble average for the historical simulation closely matches the observational reference constraint (see [Fig. 1](#)). In essence, the close fit to observations is a consequence of applying higher weights on more skillful models. Furthermore, since the BMA method estimates a distribution of model weights, various model combinations become possible, which explicitly takes care of the model dependence issue. To our best knowledge, BMA has not been widely used to constrain uncertainties in future projections from global climate models, with the exception of [Massoud et al. \(2019a\)](#) which used BMA to constrain the spread of uncertainty in future projections of global atmospheric rivers. There are other studies that used similar methods, such as [Olson et al. \(2016\)](#) and [Fan et al. \(2017\)](#) which used BMA to make probabilistic projections of temperature in southeast Australia from a suite of

regional climate models. [Olson et al. \(2019\)](#) used a quasi-Bayesian method to weight climate model projections of summer mean maximum temperature change over Korea. Thus, in our study, we use BMA as a framework for model weighting that assesses model skill and independence, and as a result constrains the spread of uncertainty in future climate projections from a set of global climate models.

We implement BMA to estimate future precipitation changes over CONUS using a subset of CMIP5 models (historical and representative concentration pathway 8.5, or RCP8.5). The models are evaluated for fidelity with the Tropical Rainfall Measuring Mission (TRMM; [Kummerow et al. 1998](#)), which provides observed information on precipitation. In this study, we utilize five precipitation metrics, including mean daily rainfall, annual and interannual variability of the mean daily rainfall, and wet and dry extremes represented as the mean daily rainfall during the wettest and driest years on record, respectively (depicted in [Fig. 2](#)). These metrics are chosen to assure that we are not only rewarding models that have high skill in representing mean precipitation, but also other aspects of the precipitation distribution such as variability or extremes. This is one of the first studies to focus on and estimate future precipitation changes over CONUS using a weighted model strategy that is based on skill and independence (cf. [Sanderson et al. 2017](#)), and the first to use BMA as a framework to produce a multimodel average that constrains the spread of uncertainty in end-of-century precipitation projections over CONUS.

2. Data and methods

a. Observations: TRMM and CPC precipitation

TRMM was a joint U.S.–Japan satellite mission for monitoring tropical and subtropical precipitation ([Kummerow et al. 1998](#)). The satellite includes a number of precipitation-related instruments, such as a precipitation radar, a visible and infrared sensor, and Special Sensor Microwave Imager microwave imager ([Kummerow et al. 2000](#); [Huffman et al. 2007](#); [Almazroui 2011](#)). We used the “daily accumulated precipitation” variable from the “TRMM 3B42v7 daily” product for our study, for the years 1998–2016. All TRMM estimates are regridded on a common grid ($0.5^\circ \times 0.5^\circ$) using bilinear interpolation. Then, all the precipitation summary metrics applied in this study are computed, i.e., long-term mean daily precipitation, annual cycle and interannual variability in daily precipitation, and the wettest and driest years on record representing the extremes at each grid (explained in [Fig. 2](#)). The estimates for

FIG. 1. Schematic illustration of model averaging using a six-member ensemble and a precipitation metric to be projected. The simulation of each model is displayed with the solid colored shapes, and the model average and its uncertainty are depicted with the open circle and the box-and-whisker plot, respectively. The verifying observation is indicated separately with the brown “x” symbol, but this is only shown in (b) since observed data are not used to train the ensemble mean. (a) The ensemble mean (EnsM) strategy showing each arrow is the same size for all models, indicating the same weight is applied for each model. All of the models have the same shape (circle), since model independence is not considered for this strategy. (b) The BMA method shows the weighted model average and its uncertainty range closely match the observation. For the BMA method, there are higher weights on skillful models, and even higher weights for independent models (i.e., red and black circles, blue and purple diamonds are not independent models in the schematic, represented with similar shapes). This method considers the total order effect of each model’s skill and independence when estimating the model weights.

CONUS computed for each of the metrics obtained from TRMM are shown in Fig. 3 (note that some parts of Canada and Mexico are also shown in the maps).

Gibson et al. (2019) showed that climate model evaluation for precipitation indices over CONUS can be highly sensitive to the reference observational products used, such as in situ, satellite, or reanalysis data. As a result of this, it was important to test the sensitivity of the BMA weights trained using TRMM as a reference product against the use of other reference observational products. For this, we use the Climate Prediction Center (CPC) station data from the U.S. unified rain gauge dataset, which is composed of multiple sources (Higgins et al. 2000). The data record of the CPC observations is

from 1950 to 2005. We apply the BMA model weight estimation to various time periods, including 1998–2005 (7 years), 1991–2005 (14 years), 1977–2005 (28 years), and 1950–2005 (55 years). We use the various time periods to minimize the influence of the results to climate variability.

b. Model data: CMIP5 suite of models

Available climate models from the CMIP5 suite of models are used (Hibbard et al. 2007; Meehl and Hibbard 2007; Meehl et al. 2009) from the historical experiment and the future projected changes (RCP8.5). The models used in our study are listed in Table 1. All model simulations are regridded on a common grid

FIG. 2. The summary metrics (or objective functions) that are used to train the various BMA-weighted model ensemble averages are displayed in this figure. The mean precipitation (OF1) shows average daily rain rates, the annual cycle variation (OF2) shows the strength of the seasonal cycle, the interannual variability (OF3) shows the strength of year-to-year variance, the maximum annual precipitation accumulation (OF4) is the strength of the most extreme wet year on record, and the minimum annual precipitation accumulation (OF5) is the strength of the most extreme dry year on record.

($0.5^\circ \times 0.5^\circ$) using bilinear interpolation to match the resolution of the regridded TRMM estimates, which will allow direct comparison and model weighting. Then, all the precipitation summary metrics investigated in this study are computed. Once the models are prepared for evaluation against the TRMM observations, the BMA model weights can be estimated. All CMIP5 models used in this study are the “r1i1p1” version of the model simulations for the years 1998–2005, which were determined as the years of the simulations that overlap the TRMM data (i.e., TRMM data starts in 1998 and CMIP5 simulations end in 2005, so the period 1998–2005 is the maximum overlapping period that can be used for this analysis). The precipitation variable that was extracted from each model is the “precipitation_flux” which represents precipitation in units of kilograms per square meter per second ($\text{kg m}^{-2} \text{s}^{-1}$) but were converted to millimeters per day (mm day^{-1}) for this study.

c. Precipitation metrics

In this study, we utilize five precipitation metrics to apply the BMA weighting. All metrics are depicted in Fig. 2. The first is mean daily rainfall, which is the long-term average of daily precipitation. Then the second metric is the annual variability of the mean daily rainfall and represents the annual cycle, which is a way to describe how much variance in precipitation occurs throughout the course of the year. The third metric is interannual variability, which describes the year-to-year variability. Then to describe precipitation extremes, the fourth metric is the mean daily rainfall during the

wettest year on record, and the fifth metric is the mean daily rainfall during the driest year on record.

d. Bayesian model averaging

Bayesian model averaging has been widely used in previous literature, and Hoeting et al. (1999) provide a comprehensive overview of its different variants. BMA differs from other model averaging methods as it explicitly estimates the weights and associated uncertainties of the weights for each model by maximizing a specified likelihood function, i.e., BMA obtains model weights which produce model combinations that have the maximum likelihood of matching historical observations compared to other model combinations. Using these optimized weights, BMA constructs the mean and uncertainty distribution of the performance metrics (or objective function) of interest. Applications of BMA have been described in works such as Raftery et al. (2005), Gneiting and Raftery (2005), Duan et al. (2007), Vrugt and Robinson (2007), Vrugt et al. (2008), Bishop and Shanley (2008), Olson et al. (2016, 2019), Fan et al. (2017), and Massoud et al. (2019a). The BMA method offers an alternative to the selection of a single model from a number of candidate models, by weighting each candidate model according to its statistical evidence, which is proportional to the model’s skill and independence (see Fig. 1). Since the BMA method estimates a distribution of model weights, various model combinations become possible, which implicitly takes care of the model dependence issue. In other words, consider that in the BMA framework there is a hypothetical Model A

FIG. 3. TRMM satellite daily precipitation data for 1998–2016. All of the observed summary metrics (or objective functions) that are used to train the various BMA-weighted model ensemble averages are shown in this figure.

and a Model B that are similar and therefore not independent; Model A may have higher weights in some combinations, and conversely, Model B might have higher weights in other combinations. Consequently, if both models are rewarded in the same set of weights, it is very likely that each model receives a reduced weight due to the fact that both models are providing information to the model average. Therefore, model dependence can play a role in the BMA scheme since both of the dependent models can affect each other's weights, which can be portrayed in the posterior samples. See section 2 in the online supplemental material for additional details on how dependence is implicitly inferred with the BMA method.

To explain how BMA estimates the model weights, consider that at a given location we have the output of multiple models (e.g., Fig. 1). The goal is to weigh the different models such that the weighted estimate is a better predictor of the observed system behavior

than any of the individual models of the ensemble. Thus, the estimated model weights using BMA are as follows:

$$\mathbf{w}_{m,\text{BMA}} = [w(m_1), w(m_2), \dots, w(m_K)], \quad (1)$$

where $w(m_i, i = 1, 2, 3, \dots, K)$ represents the optimized weights of K models after fitting to the observations using a chosen likelihood function. The range of $w(m_i)$ is between 0 and 1, with a weight of 0 for models that do not contribute any information and a weight of 1 for models that fully contribute to the projection. The sum of K model weights $\sum_{i=1}^K w(m_i)$ is equal to 1. The BMA weights are estimated using a Markov chain Monte Carlo (MCMC) algorithm (Vrugt 2016), and the final estimated model weights are K distributions of weights where each distribution is not required to follow any form, e.g., Gaussian, bimodal, etc. The final estimates of the BMA model weights, or $\mathbf{w}_{m,\text{BMA}}$ in Eq. (1), are

TABLE 1. Estimated mean model weights for each model averaging strategy. “Ens mean” refers to the equal weighting strategy that produces the multimodel ensemble mean with no training from the observations. “Mean weight” refers to the average weight each model receives from all five objective functions. “BMA-OFx” refers to the weights produced for optimizing the fit of the model average to each observed precipitation summary metric (i.e., those shown in Fig. 2). The values in this table are depicted in Figs. 3 and 4. These model weights inform and constrain the uncertainties in the projections of end-of-century precipitation.

CMIP5 model	Ens mean	Mean weight	BMA-OF1	BMA-OF2	BMA-OF3	BMA-OF4	BMA-OF5
1) ACCESS1.0	0.08	0.33	0.34	0.67	0.01	0.53	0.10
2) BNU-ESM	0.08	0.03	0.01	0.01	0.10	0.01	0.02
3) CanESM2	0.08	0.18	0.28	0.01	0.03	0.09	0.48
4) CCSM4	0.08	0.04	0.01	0.04	0.12	0.01	0.04
5) CESM1-CAM5	0.08	0.04	0.03	0.04	0.07	0.02	0.03
6) CNRM-CM5	0.08	0.03	0.01	0.02	0.02	0.07	0.01
7) CSIRO Mk3.6.0	0.08	0.09	0.03	0.02	0.31	0.01	0.09
8) GFDL-CM3	0.08	0.01	0.01	0.02	0.02	0.02	0.01
9) HadGEM2-CC	0.08	0.08	0.14	0.01	0.07	0.07	0.11
10) MPI-ESM-LR	0.08	0.11	0.12	0.12	0.13	0.14	0.06
11) MRI-CGCM3	0.08	0.01	0.01	0.01	0.01	0.03	0.01
12) NorESM1-M	0.08	0.04	0.00	0.04	0.12	0.01	0.03

utilized to constrain the spread of uncertainty in the projected end of century climate.

SETTING UP BAYES’ THEOREM

The likelihood function is a critical property of the BMA calculation. According to Bayes’ theorem, the probability of an event is estimated based on prior knowledge of conditions that might be related to the event. In equation form, this looks like

$$P(A|B) = \left(\frac{P(B|A)P(A)}{P(B)} \right). \quad (2)$$

In the BMA application of the current study, event A can be representative of the chosen model combination and event B can be representative of the observed data, $P(A|B)$ is the conditional probability or the likelihood of event A occurring given that event B is true, $P(B|A)$ is a conditional probability of event B occurring given that event A is true, and $P(A)$ and $P(B)$ are the probabilities of observing events A and B independently of each other. For the purposes of this study, we can express $P(A)$ as the prior information of our calculation, which is an equally weighted distribution for each of the model weights and thus can be taken out of the calculation. Probability $P(B)$ is the evidence of the observation or models, and is a normalizing constant and therefore taken out of the equation. This leaves us with

$$P(A|B) \propto P(B|A); \quad P(A|B) \propto L(\mathbf{w}_{m,\text{BMA}}), \quad (3)$$

where $P(A|B)$ is the final distribution of the model weights, or $\mathbf{w}_{m,\text{BMA}}$ in Eq. (1), and $P(B|A)$ is equivalent to the chosen likelihood function $L(\mathbf{w}_{m,\text{BMA}})$ described

in the next paragraph. Therefore, the BMA algorithm searches for model weight combinations $\mathbf{w}_{m,\text{BMA}}$ that will maximize the fit to the observed data, and thus will maximize the value of the likelihood function $L(\mathbf{w}_{m,\text{BMA}})$.

In recent decades, Bayesian inference has emerged as a working paradigm for modern probability theory, parameter and state estimation, model selection, and hypothesis testing (Vrugt and Massoud 2019). According to Bayes’ theorem, the distribution of model weights $P(A|B)$ depends upon the prior distribution $P(A)$, which captures our initial beliefs about the values of the model weights, and a likelihood function $L(\mathbf{w}_{m,\text{BMA}})$, which quantifies the confidence in the model weights $\mathbf{w}_{m,\text{BMA}}$ in light of the observed data $\mathbf{Y}$. The observed data in this case are a spatial map of precipitation characteristics (as shown in Fig. 3), and our goal is to find the optimal set of model weights $\mathbf{w}_{m,\text{BMA}}$ that produces a model combination $\mathbf{X}$, which maximizes the fit, or the likelihood, relative to one or more of the observations. Our likelihood function is set up in the simplest terms as

$$L(\mathbf{w}_{m,\text{BMA}}) = -\frac{1}{2} \sum_{i,j} [\mathbf{Y}_{ij} - \mathbf{X}_{ij}(\mathbf{w}_{m,\text{BMA}})]^2, \quad (4)$$

where i, j refer to the longitudinal and latitudinal indices of grids on the map; $\mathbf{Y}(i, j)$ is the observed precipitation metric at grid i, j obtained from TRMM; and $\mathbf{X}(i, j)$ is the BMA-weighted model ensemble average of the precipitation metric at grid i, j . We apply MCMC sampling on the model weights in an optimization framework until the likelihood function in Eq. (4) is maximized, which allows for the estimation of the optimized model weights, or $\mathbf{w}_{m,\text{BMA}}$ in Eq. (1). These optimized model weights will be used to inform, as well as constrain the spread of

FIG. 4. The distributions for the estimated BMA model weights for each CMIP5 model, represented in the form of box-and-whisker plots. (a) Samples from the prior distribution $P(A)$. (b)–(f) The distributions of each model weight for each precipitation metric (OF1–5).

uncertainty in, the model projections of precipitation change at the end of the century.

Successful use of the MCMC applications depends on many input factors, such as the number of chains, the prior used for the parameters, number of generations to sample, the convergence criteria, among other things. For our application, we used $C = 8$ chains, the prior was a uniform distribution from 0 to 1 for each model weight and each sampled set of weights was normalized so that the sum of weights is equal to 1, the number of generations was set at $G = 5000$ for each metric being fit, and the convergence of the chains relied on the Gelman and Rubin (1992) diagnostic, where we applied the commonly used convergence threshold of $R = 1.2$.

3. Results

a. Model evaluation and BMA weighting

First, equally weighted multimodel average (“ens mean” in Table 1) are produced by averaging the 12 CMIP5 models, then the BMA-weighted model ensemble averages for the five summary metrics are produced based on evaluation of each model against the TRMM observations. Table 1 lists all the model weights for each strategy, including the mean value of the optimized model weight distributions estimated for each BMA strategy. Ens mean refers to the equal weighting strategy; for this, the uncertainty of the projections spans the range of uncertainty from all models considered. In other words, the spread of model simulations from all of the CMIP5 models used in this study produce the

uncertainty for the ens mean strategy. The “mean weight” column of Table 1 lists the average weight each model receives from combining all of the other columns that show the five BMA weights. Last, the “BMA-OFx” columns of Table 1 list the weights produced for optimizing the fit of the model average to each precipitation performance metric (i.e., those depicted in Fig. 2). BMA-OFx refers to the BMA weighting for the model performance metric, or “objective function,” of interest. So, for example, BMA-OF1 refers to the BMA-weighted model ensemble average for the first performance metric, or the long-term mean daily precipitation.

The boxplots in Fig. 4 show the distribution for the estimated BMA model weights for each CMIP5 model. Figure 4a displays samples from the prior distribution $P(A)$. The remaining panels show the distributions of each model weight for each precipitation metric. These distributions are the outcome of maximizing the likelihood function in Eq. (4) for each objective function. Figure 5 shows the model weights listed in Table 1, or the mean weights from the distributions in Fig. 4, but in a colored diagram. In Fig. 5, the red boxes indicate models with a higher weight than the ensemble mean weights, and blue boxes show models with lower weights.

b. Historical simulations

In Fig. S1 in the online supplemental material, we show maps of the first precipitation performance metric, or the historical long-term mean daily precipitation, for each individual CMIP5 model as well as the model averages produced (i.e., the ensemble mean and BMA-OF1). These maps are compared with the

FIG. 5. Mean BMA model weights for each metric presented in a color-coded diagram. This model evaluation informs and constrains the projections of end-of-century precipitation. The color scale is chosen so that a model with the same weight as the equal weights estimate [$w_{m,EqW} = (1/12) = 0.0833$] is shown in white, and weights that are higher than this are in red and any weights that are lower are in blue.

observational reference from TRMM (see Fig. 3a), and we can analyze the bias from each model; these bias maps are shown in Fig. 6. To understand the magnitude of error from each model, Table 2 lists the RMSE of each model compared to TRMM data, for each of the precipitation performance metrics. We can see in Table 2 that the best performing model with the lowest RMSE is the BMA-weighted model ensemble average estimated for each metric. In other words, BMA-OF1 is the best performing model ensemble average for the first precipitation performance metric, or the long-term mean daily precipitation. BMA-OF2 is the best for the second precipitation performance metric, or the annual cycle variability. This is also the case with BMA-OF4 and BMA-OF5, which are the best performing model ensemble averages for the fourth and fifth precipitation performance metrics, respectively. The exception here is the third precipitation performance metric, since the best performing models for this are the MPI-ESM-LR and ACCESS1.0 models, yet the BMA-OF3 model ensemble average is also very skillful for this metric. Thus, the BMA-weighted model ensemble averages are generally some of the most skillful model candidates when compared to the TRMM data.

c. Trade-offs in model weights

Each model combination has a different level of performance skill for each metric that is chosen. For instance, the optimized BMA-OF1-weighted model ensemble average will theoretically perform the best for OF1 (mean precipitation), but this model's performance may degrade for the other metrics (variability or

extremes in precipitation). We can check how each BMA-OFx-weighted model ensemble average performs for the other metrics with the values shown in Table 2. Interestingly, each of the BMA-weighted model ensemble averages have a relatively strong performance skill regardless of the metric of interest. Table 2 shows how each BMA-weighted model ensemble average has a relatively low RMSE for all the metrics. Interestingly, the “ensemble mean” and the “mean skill weights” model averages perform well for mean precipitation; however, much of the variability in the simulation is lost for these model averages and they are not as skillful in simulating metrics related to variability (i.e., OF2 and OF3). Therefore, the ensemble mean and the mean skill weights are the least accurate model averages for simulating OF2 and OF3, or annual and interannual variability.

Figure 7 shows how the performance trade-offs look like for each metric. In Fig. 7, there are plots of the root-mean-square error space for triplet combinations of the BMA-weighted model ensemble averages, along with the ensemble mean model average, the mean weight model average, and various solutions from the prior samples. A perfect model would be located where the 0–0–0 point is, which reflects a model with 0 error for all objectives, and is depicted with a red bull's-eye in the figure. The goal is to move from the location of the prior samples (light blue dots) toward the red bull's-eye, which is noted with a large black arrow in each figure. The prior samples, along with the ensemble mean (red X) and the mean weight (black X) sets of models averages, generally have a higher RMSE than the

Bias in Historical Mean Daily Precipitation (1998-2005)

FIG. 6. Bias plots of mean daily precipitation (OF1) for each individual model as well as the model averages, i.e., ensemble mean and BMA-weighted model ensemble average, relative to TRMM satellite observations.

BMA-weighted model ensemble averages (shown in blue, magenta, and green in the plots). These figures show how the BMA-OF1 model has the best performance for OF1 (mean precipitation), BMA-OF2 performs the best for OF2 (annual cycle variance), etc. Generally, the BMA-OF x models have higher skill for all precipitation summary metrics compared to the prior samples, the ensemble mean, or mean weight model averages, which matches the values shown in Table 2.

One thing that is important to note in Fig. 7 is that these plots indicate what the Pareto trade-offs are for each model average [see Langenbrunner and Neelin (2017) for more information on Pareto fronts]. In general, the larger the distance between the RMSE distributions in

these figures, the larger the trade-offs and thus the differences in the future change estimates are between the model combinations. Alternatively, the closer the RMSE distributions, the more consistent the projected estimates are between the model combinations. This is important to assess the uncertainty of the future projected changes from each model average.

4. Projected changes in precipitation over CONUS

a. Future changes in precipitation

The motivation for this study is that climate models simulate different projections for future changes in

TABLE 2. RMSE for each CMIP5 model and for all the model averages estimated in this study, i.e., BMA-weighted model ensemble averages, ensemble mean, and the model average produced from the mean skill weights. The RMSE for each objective OF1–5 is shown in columns, and the scores for each model are shown in the rows. All values are in millimeters (mm). The best performing model for each metric is highlighted in bold. For all metrics (with the exception of OF3), the best performing model with the lowest RMSE is the BMA-weighted model ensemble average estimated for this metric (i.e., BMA-OF1 is the best performing model for OF1—mean precipitation, BMA-OF2 is the best for OF2—annual variance, etc.).

CMIP5 model	RMSE OF1—mean	RMSE OF2—annual cycle	RMSE OF3—interannual	RMSE OF4—maximum	RMSE OF5—minimum
1) ACCESS1.0	0.61	1.16	0.36	0.74	0.57
2) BNU-ESM	0.94	2.08	0.48	1.21	0.84
3) CanESM2	0.71	1.52	0.41	1.04	0.51
4) CCSM4	0.73	1.79	0.43	1.09	0.56
5) CESM1-CAM5	0.68	1.60	0.45	0.98	0.61
6) CNRM-CM5	0.74	1.41	0.43	0.92	0.70
7) CSIRO Mk3.6.0	0.73	1.63	0.45	1.06	0.57
8) GFDL-CM3	0.92	1.97	0.54	1.17	0.84
9) HadGEM2-CC	0.64	1.37	0.47	0.83	0.61
10) MPI-ESM-LR	0.64	1.48	0.32	0.87	0.55
11) MRI-CGCM3	1.13	2.07	0.54	1.34	1.03
12) NorESM1-M	0.86	1.90	0.45	1.14	0.74
A) BMA-OF1 weights	0.49	1.56	0.73	0.70	0.55
B) BMA-OF2 weights	0.57	1.10	0.43	0.68	0.57
C) BMA-OF3 weights	0.58	1.46	0.38	0.84	0.54
D) BMA-OF4 weights	0.50	1.41	0.68	0.67	0.56
E) BMA-OF5 weights	0.52	1.95	0.82	0.89	0.48
Mean skill weights	0.54	2.22	1.02	0.99	0.81
Ensemble mean	0.65	2.58	1.14	0.85	0.67

precipitation, which results in significant uncertainty for most regions of CONUS. To illustrate this, we show in Fig. S2 the end-of-century (RCP8.5) mean daily precipitation (OF1) for each individual model and for the model averages, i.e., ensemble mean and BMA-OF1 models. In Fig. 8, we show the difference between the future precipitation and the historical estimate.

The maps in Fig. 8 can be used to show the difference between each model's estimate of future precipitation changes. The model averages, such as the ensemble mean and BMA-weighted model ensemble averages, show a smaller magnitude change relative to what the individual models might indicate. However, spatial patterns of change are similar between most of the models, i.e., the East Coast and Northwest will experience increases in precipitation, while the remainder of the country will not experience much change. Some disagreements in the estimated future change between models can be seen, for example the CanESM2 model simulates an increase of precipitation over the majority of the West Coast, including California, whereas other models, such as ACCESS1.0 or HadGEM2-CC show decreases in precipitation for these areas. This results in significant uncertainty in estimated changes of future precipitation for these regions.

Changes in future precipitation may vary depending on the season. A number of studies (e.g., Collins et al. 2013)

have shown that projections may increase in one season and decrease in the other, resulting in no change when looking at just annual precipitation change. For example, Fig. S3 shows that the east coast of CONUS is projected to experience significant increases in precipitation in the winter months (DJF), but not in other seasons. In winter (DJF) and spring (MAM), the northern part of the country is projected to become wetter as the global climate warms, but drier for southwestern CONUS during these seasons. During the summer (JJA), most of the country, with the exception of the northeast, is projected to experience decreases in precipitation. For the fall (SON), many regions of the country will not experience significant changes in average precipitation, except for the northwest, which is projected to become significantly wetter.

b. Verifying results to different observation data and time periods

Climate model evaluation for precipitation indices over CONUS can be highly sensitive to the reference observational products (Gibson et al. 2019). We compare the BMA-weighted model ensemble average trained using TRMM with BMA-weighted model ensemble averages using CPC station data. We apply the BMA model weight estimation to various time periods to minimize the influence of the results to climate variability. In Fig. S4, the first column of figures shows the historical mean precipitation

FIG. 7. 3D RMSE space for triplet combinations of the objective functions used in this study. Each plot shows the RMSE of the BMA-weighted model ensemble averages, the ensemble mean, and the mean skill weights model average. A perfect model would be located where the 0–0–0 point is, which reflects a model with 0 error for all objectives, and is depicted with a red bull's-eye in the figure. Theoretically, the larger the distance between the RMSE distributions, the larger the trade-offs and thus the differences in the future change estimates are between the model combinations. Alternatively, the closer the RMSE distributions, the more consistent the projected estimates are between the model combinations. Units are in mm day^{-1} .

obtained from each set of BMA weights. The middle column shows the estimated BMA weights for the various data products and time periods. The right column of figures shows the projected future change (RCP8.5) in mean precipitation using each of the BMA weights. As shown in Fig. S4, similar BMA weights are produced when constrained by the TRMM satellite or the CPC in situ based datasets, regardless of the time period chosen. Furthermore, the maps of the first precipitation performance metric, or the historical and the future long-term mean daily precipitation, are similar for all the model averages produced.

We would like to point out that there could be concern about using the TRMM precipitation data product over CONUS when there are other longer (arguably more reliable) in situ products available, such as CPC. However, for precipitation over the oceans and for higher temporal resolution, TRMM data can provide some benefits worth mentioning. These two advantages of TRMM will enable future studies that apply other evaluation metrics with BMA for many different parts of the world. Since TRMM is shown to be viable at constraining models despite its short record over CONUS (i.e., Fig. S4), one could use TRMM in other regions of the world, where long-term in situ measurements are sparse (e.g., parts of Africa or South America), to constrain future projections of precipitation from models.

c. Reduction of uncertainty in the BMA model averages

A main goal of producing weighted model averages using the BMA framework is to reduce the spread of

uncertainty in the historical and the future estimates. Figure 9 shows the spread of uncertainty, calculated as the standard deviation of the ensemble spread, of the long-term mean daily precipitation for both the Ensemble Mean (Figs. 9a,b) and the BMA-OF1 (Figs. 9c,d) weighted model averages. What is apparent in this figure is that the BMA method reduces the spread of uncertainty in the historical (Fig. 9c) and the future (Fig. 9d) projections for most regions of CONUS by about a third compared to the original subset of CMIP5 models used (Figs. 9a,b). The biggest reduction in uncertainty seems to be located near the western parts of CONUS, where uncertainty in precipitation is generally very significant. The next section will discuss precipitation changes in the future more thoroughly and will show how the results can change depending on the choice of model average that is used.

d. Sensitivity of future estimates to the choice of performance metric

The choice of ensemble average (based on which performance metric to optimize) can impact the estimates of future precipitation changes considerably. To examine how each ensemble average can affect the projections, we show in Fig. 10 a matrix of maps that displays the sensitivity of each precipitation metric to the choice of model average. The plots in Fig. 10 show that, by the end of the century, the mean rainfall (first column of figures) is projected to increase for most of the East Coast and the Northwest by roughly 10%, might decrease in the High Plains area by about 10%, and very little change is expected for the Southwest region, which

Change in Future Mean Precipitation (2080–2100)

FIG. 8. Changes in end-of-century mean daily precipitation (OF1) relative to historical climatology for each individual model as well as the model averages, i.e., ensemble mean and BMA-weighted model ensemble average.

matches results reported in [Sanderson et al. \(2017\)](#) and [Langenbrunner and Neelin \(2017\)](#). As for the annual cycle (second column of figures), variance is projected to increase for the Northwest and the Southeast by about 10%, but decrease for the High Plains area by roughly 20%. For interannual variability (third column of figures), variance is projected to increase for most of the United States by 5%–10%, especially on the East where it might increase by up to 30%. Finally, for extremes (fourth and fifth columns of figures), it is apparent that the rainfall distributions are getting wider, as the wettest year on record is projected to be 20% more wet for most regions and the driest year on record seems to get drier

by about 10%, except for the Northwest and Southeast regions where the driest year may be slightly more wet in the future.

The uncertainty and sensitivity between the various BMA produced ensemble model averages for each metric is shown in Fig. S5. These plots show the spread in possible outcomes for each precipitation metric, using all the BMA-OFx model ensemble averages. This in essence provides a view of how the choice of constraint (i.e., performance metric) can affect the future projections, and which regions of CONUS have uncertainty in the expected changes in precipitation. We see that for mean precipitation (Fig. S5a), the highest uncertainty in the projected changes is in the West and in the Great

Uncertainty Spread in Mean Precipitation (CMIP5 Ensemble vs BMA)

FIG. 9. Spread of uncertainty in the ensemble of model simulations (standard deviation of the ensemble spread). (a),(b) The uncertainty from the entire CMIP5 ensemble is shown, with the historical simulations on the left and the RCP8.5 simulations on the right. (c),(d) The uncertainty from the BMA-OF1 model ensemble is shown, with the historical simulations on the left and the RCP8.5 simulations on the right. What is apparent in this figure is that the BMA method reduced the spread of uncertainty in the historical and the future projections.

Lakes regions. This metric is shown to be somewhat sensitive to the choice of constraint relative to the other metrics. For the annual cycle (Fig. S5b), the greatest uncertainty in the projections is again in the West. There is not much uncertainty between the various model averages for interannual variability (Fig. S5c). For the metrics related to variability (Fig. S5b,c), there is low sensitivity to the choice of constraint relative to other metrics. Last, for wet and dry extremes (Fig. S5d,e), there are significant uncertainties between the various BMA ensemble model averages, especially in the West. These metrics related to extremes seem to be the most sensitive to the choice of constraint, compared to the other metrics, especially for the wet extremes. This indicates that the projection of the maximum precipitation (OF4) is the most sensitive to the choice of constraint out of the precipitation metrics we chose for this study.

This study shows the use of BMA and its strength for constraining climate projections. It is likely that climate models which produce simulations closer to historical observations have similar performance for the future under a stationary assumption. Yet, growing evidence dictates a more nonstationary future climate. It is questionable that climate models identified as better tools based on the relatively stationary history will be better representative of the future, particularly the future climate extremes. Yet, it has been shown in the literature that trends in precipitation might not be as significant as

trends in other climatic variables, such as temperature. This has been shown in Gibson et al. (2019) and in Figs. S1 and S2, where trends in various precipitation metrics and from various products are shown for CONUS. The amplitude of the trends ($\sim \text{mm yr}^{-1}$) are orders of magnitude smaller than the variability itself or the magnitude of the change in future precipitation metrics ($\sim \text{mm day}^{-1}$). This allows us to believe that the climate models that have higher weights based on historical fit might be better representative of the future of precipitation.

5. Conclusions

This study showcased Bayesian model averaging (BMA) as a framework to achieve more accurate simulations and to constrain the spread of uncertainty in climate change projections. We provided an extensive look at end of century precipitation changes over CONUS (Fig. 10), including estimates of changes in long-term mean daily precipitation (OF1), annual cycle variability (OF2), interannual variability (OF3), and wet (OF4) and dry (OF5) precipitation extremes (as depicted schematically in Fig. 2, and shown in Fig. 3 for the TRMM satellite product). A suite of models from the CMIP5 archive were used for the model averaging, and the BMA weights (Table 1; Figs. 4–5) were trained to reduce the bias between the model ensemble averages and the TRMM satellite observations (Figs. 6 and 7). We found that the

FIG. 10. End-of-century change in precipitation for CONUS (mm day⁻¹ for all panels). The 7 × 5 panel shows the multimodel ensemble mean in the top row, the model average using the mean BMA weights (mean skill weight) in the second row, with all other BMA-weighted model ensemble averages in the rows below. Each column represents a different precipitation metric. The metrics displayed in this figure are mean precipitation (OF1), the annual cycle variation (OF2), the interannual variability (OF3), the mean precipitation during the wettest year (OF4), and the mean precipitation during the driest year (OF5). Red colors indicate a positive change, while blue colors indicate a negative change.

BMA-weighted model ensemble averages used in this study were generally more accurate than the individual CMIP5 models or other model averages, such as the ensemble mean, when compared to the TRMM satellite data (Table 2; Figs. 6 and 7), and had less uncertainty in the simulations (Fig. 9) compared to the original ensemble spread.

We presented a sensitivity experiment to the future precipitation projections, based on the various model weighting strategies that were used in this study (Fig. 10). Our results showed that, by the end of the century, the mean daily rainfall is projected to increase for most of the East Coast and the Northwest, may decrease in the High Plains area, and with little change expected for the Southwest. For mean daily rainfall, the projected changes were consistent between the ensemble mean and the BMA-weighted model ensemble average. The strength of the annual and interannual cycles is expected to increase for most of the United States, with the exception of the High Plains area where annual variability is projected to decrease. For annual and interannual variability, the projected changes in variability were more pronounced for the BMA-weighted model ensemble

average than for the multimodel ensemble mean. As for extremes, the wettest year on record is projected to become wetter for the majority of CONUS and the driest years to become drier. For extremes, the projected changes were consistent between the multimodel ensemble mean and the BMA-weighted model ensemble average. It is important to note that different fitting metrics produce different sets of model weights. One must consider these differences when selecting a set of model weights for potential downscaling or other applications, and this consideration should be dependent on the precipitation metric of interest.

We used BMA as a method to average an ensemble of model projections, with the aim of constraining the spread of uncertainty that is generated with using multimodel ensembles with highly varying climate simulations. We believe that the BMA approach can be used for different climate variables, including temperature. However, given the nonstationarity of temperature trends around the globe, we recommend that one must use an additional fitting metric that represents the trends in temperature in a given grid cell. Similar to how we used five precipitation metrics in our study to represent various

aspects of the precipitation distribution, one could use similar metrics to represent the distribution of other climate variables, and BMA can be applied to fit those metrics.

In our study, the BMA-weighted model reduced the uncertainty in future precipitation projections by about a third (Fig. 9). The BMA framework can be very useful for efforts that estimate changes in climate by the end of the century. With many versions of climate models being produced and developed and with the preparation of models for CMIP6, it is important to answer calls in the literature (e.g., Eyring et al. 2019) to investigate advanced weighting methods that will produce more tightly constrained climate projections (as we showed in Fig. 9) and to more rigorously quantify uncertainty sources. We think that BMA shows promise in being one of these methods, and believe it can be used for climate model evaluation and for constraining the spread of uncertainty in climate model projections.

Acknowledgments. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This research was funded by NCA Award 106232. All rights reserved. All data used in this study are publicly available. The TRMM satellite dataset is available via <https://pmm.nasa.gov/data-access/downloads/trmm>, and the CMIP5 model output data are available via https://cmip.lnl.gov/cmip5/data_portal.html.

REFERENCES

- Abramowitz, G., and C. H. Bishop, 2015: Climate model dependence and the ensemble dependence transformation of CMIP projections. *J. Climate*, **28**, 2332–2348, <https://doi.org/10.1175/JCLI-D-14-00364.1>.
- Alexander, K., and S. M. Easterbrook, 2015: The software architecture of climate models: A graphical comparison of CMIP5 and EMICAR5 configurations. *Geosci. Model Dev.*, **8**, 1221–1232, <https://doi.org/10.5194/gmd-8-1221-2015>.
- Almazroui, M., 2011: Calibration of TRMM rainfall climatology over Saudi Arabia during 1998–2009. *Atmos. Res.*, **99**, 400–414, <https://doi.org/10.1016/j.atmosres.2010.11.006>.
- Annan, J. D., and J. C. Hargreaves, 2011: On the generation and interpretation of probabilistic estimates of climate sensitivity. *Climatic Change*, **104**, 423–436, <https://doi.org/10.1007/s10584-009-9715-y>.
- Bishop, C. H., and K. T. Shanley, 2008: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Wea. Rev.*, **136**, 4641–4652, <https://doi.org/10.1175/2008MWR2565.1>.
- , and G. Abramowitz, 2013: Climate model dependence and the replicate Earth paradigm. *Climate Dyn.*, **41**, 885–900, <https://doi.org/10.1007/s00382-012-1610-y>.
- Collins, M., and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 1029–1136.
- Dettinger, M. D., 2011: Climate change, atmospheric rivers, and floods in California—A multimodel analysis of storm frequency and magnitude changes. *J. Amer. Water Resour. Assoc.*, **47**, 514–523, <https://doi.org/10.1111/j.1752-1688.2011.00546.x>.
- , and D. R. Cayan, 2014: Drought and the California delta—A matter of extremes. *San Francisco Estuary Watershed Sci.*, **12** (2), <https://doi.org/10.15447/SFEWS.2014V12ISS2ART4>.
- , F. M. Ralph, T. Das, P. J. Neiman, and D. R. Cayan, 2011: Atmospheric rivers, floods and the water resources of California. *Water*, **3**, 445–478, <https://doi.org/10.3390/w3020445>.
- Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian, 2007: Multimodel ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.*, **30**, 1371–1386, <https://doi.org/10.1016/j.advwatres.2006.11.014>.
- Easterling, D. R., and Coauthors, 2017: Precipitation change in the United States. *Climate Science Special Report: Fourth National Climate Assessment*, Vol. I, U.S. Global Change Research Program, 207–230, <https://doi.org/10.7930/J0H993CC>.
- Espinoza, V., D. E. Waliser, B. Guan, D. A. Lavers, and F. Martin Ralph, 2018: Global analysis of climate change projection effects on atmospheric rivers. *Geophys. Res. Lett.*, **45**, 4299–4308, <https://doi.org/10.1029/2017GL076968>.
- Eyring, V., and Coauthors, 2019: Taking climate model evaluation to the next level. *Nat. Climate Change*, **9**, 102–110, <https://doi.org/10.1038/S41558-018-0355-Y>.
- Fan, Y., R. Olson, and J. P. Evans, 2017: A Bayesian posterior predictive framework for weighting ensemble regional climate models. *Geosci. Model Dev.*, **10**, 2321–2332, <https://doi.org/10.5194/gmd-10-2321-2017>.
- Feng, Z., L. R. Leung, S. Hagos, R. A. Houze, C. D. Burleyson, and K. Balaguru, 2016: More frequent intense and long-lived storms dominate the springtime trend in central US rainfall. *Nat. Commun.*, **7**, 13429, <https://doi.org/10.1038/ncomms13429>.
- Gao, Y., J. Lu, L. R. Leung, Q. Yang, S. Hagos, and Y. Qian, 2015: Dynamical and thermodynamical modulations on future changes of landfalling atmospheric rivers over western North America. *Geophys. Res. Lett.*, **42**, 7179–7186, <https://doi.org/10.1002/2015GL065435>.
- Gelman, A., and D. B. Rubin, 1992: Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472, <https://doi.org/10.1214/ss/1177011136>.
- Gibson, P. B., D. E. Waliser, H. Lee, B. Tian, and E. Massoud, 2019: Climate model evaluation in the presence of observational uncertainty: Precipitation indices over the contiguous United States. *J. Hydrometeorol.*, **20**, 1339–1357, <https://doi.org/10.1175/JHM-D-18-0230.1>.
- Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310**, 248–249, <https://doi.org/10.1126/science.1115255>.
- Hagos, S. M., L. R. Leung, J.-H. Yoon, J. Lu, and Y. Gao, 2016: A projection of changes in landfalling atmospheric river frequency and extreme precipitation over western North America from the large ensemble CESM simulations. *Geophys. Res. Lett.*, **43**, 1357–1363, <https://doi.org/10.1002/2015GL067392>.
- Herger, N., G. Abramowitz, R. Knutti, O. Angélil, K. Lehmann, and B. M. Sanderson, 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>.

- Hibbard, K. A., G. A. Meehl, P. M. Cox, and P. Friedlingstein, 2007: A strategy for climate change stabilization experiments. *Eos, Trans. Amer. Geophys. Union*, **88**, 217–221, <https://doi.org/10.1029/2007EO200002>.
- Hidalgo, H. G., and E. J. Alfaro, 2015: Skill of CMIP5 climate models in reproducing 20th century basic climate features in Central America. *Int. J. Climatol.*, **35**, 3397–3421, <https://doi.org/10.1002/joc.4216>.
- Higgins, R., W. Shi, E. Yarosh, and R. Joyce, 2000: Improved US precipitation quality control system and analysis. NCEP/Climate Prediction Center ATLAS 7, 40 pp., https://www.cpc.ncep.noaa.gov/research_papers/ncep_cpc_atlas/7/index.html.
- Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.*, **14**, 382–417.
- Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.*, **8**, 38–55, <https://doi.org/10.1175/JHM560.1>.
- Janssen, E., D. J. Wuebbles, K. E. Kunkel, S. C. Olsen, and A. Goodman, 2014: Observational-and model-based trends and projections of extreme precipitation over the contiguous United States. *Earth's Future*, **2**, 99–113, <https://doi.org/10.1002/2013EF000185>.
- , R. L. Sriver, D. J. Wuebbles, and K. E. Kunkel, 2016: Seasonal and regional variations in extreme precipitation event frequency using CMIP5. *Geophys. Res. Lett.*, **43**, 5385–5393, <https://doi.org/10.1002/2016GL069151>.
- Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, <https://doi.org/10.1175/BAMS-D-13-00255.1>.
- Knutti, R., and J. Sedláček, 2013: Robustness and uncertainties in the new CMIP5 climate model projections. *Nat. Climate Change*, **3**, 369–373, <https://doi.org/10.1038/nclimate1716>.
- , R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, <https://doi.org/10.1175/2009JCLI3361.1>.
- , J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, 2017: A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.*, **44**, 1909–1918, <https://doi.org/10.1002/2016GL072012>.
- Kummerow, C., W. Barnes, T. Kozu, J. Shiue, and J. Simpson, 1998: The Tropical Rainfall Measuring Mission (TRMM) sensor package. *J. Atmos. Oceanic Technol.*, **15**, 809–817, [https://doi.org/10.1175/1520-0426\(1998\)015<0809:TTRMMT>2.0.CO;2](https://doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2).
- , and Coauthors, 2000: The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit. *J. Appl. Meteor.*, **39**, 1965–1982, [https://doi.org/10.1175/1520-0450\(2001\)040<1965:TSOTTR>2.0.CO;2](https://doi.org/10.1175/1520-0450(2001)040<1965:TSOTTR>2.0.CO;2).
- Langenbrunner, B., and J. D. Neelin, 2017: Pareto-optimal estimates of California precipitation change. *Geophys. Res. Lett.*, **44**, 12 436–12 446, <https://doi.org/10.1002/2017GL075226>.
- Lee, H., A. Goodman, L. McGibbney, D. E. Waliser, J. Kim, P. C. Loikith, P. B. Gibson, and E. C. Massoud, 2018: Regional climate model evaluation system powered by Apache open climate workbench v1. 3.0: An enabling tool for facilitating regional climate studies. *Geosci. Model Dev.*, **11**, 4435–4449, <https://doi.org/10.5194/gmd-11-4435-2018>.
- Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.*, **38**, L08703, <https://doi.org/10.1029/2011GL046864>.
- Massoud, E. C., A. J. Purdy, M. E. Miro, and J. S. Famiglietti, 2018: Projecting groundwater storage changes in California's Central Valley. *Sci. Rep.*, **8**, 12917, <https://doi.org/10.1038/s41598-018-31210-1>.
- , V. Espinoza, B. Guan, and D. Waliser, 2019a: Global climate model ensemble approaches for future projections of atmospheric rivers. *Earth's Future*, **7**, 1136–1151, <https://doi.org/10.1029/2019EF001249>.
- , and Coauthors, 2019b: Identification of key parameters controlling demographically structured vegetation dynamics in a land surface model: CLM4. 5 (FATES). *Geosci. Model Dev.*, **12**, 4133–4164, <https://doi.org/10.5194/gmd-12-4133-2019>.
- , M. Turmon, J. Reager, J. Hobbs, Z. Liu, and C. H. David, 2020: Cascading dynamics of the hydrologic cycle in California explored through observations and model simulations. *Geosciences*, **10**, 71, <https://doi.org/10.3390/geosciences10020071>.
- Meehl, G. A., and K. Hibbard, 2007: Summary Report: A Strategy for climate change stabilization experiments with AOGCMs and ESMs: Aspen Global Change Institute 2006 Session, Earth System Models: The Next Generation (Aspen, Colorado, July 30–August 5, 2006). WCRP Informal Rep. 3/2007, ICPO Publ. 112, IGBP Rep. 57, 37 pp., https://www.agci.org/sites/default/files/pdfs/lib/publications/06S1_WhitePaper.pdf.
- , and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.*, **90**, 1467–1486, <https://doi.org/10.1175/2009BAMS2778.1>.
- Melillo, J. M., T. C. Richmond, and G. W. Yohe, Eds., 2014: *Climate Change Impacts in the United States: The Third National Climate Assessment*. U.S. Global Change Research Program, 841 pp., <https://doi.org/10.7930/J0Z31WJ2>.
- Olson, R., Y. Fan, and J. P. Evans, 2016: A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures. *Geophys. Res. Lett.*, **43**, 7661–7669, <https://doi.org/10.1002/2016GL069704>.
- , S.-I. An, Y. Fan, and J. P. Evans, 2019: Accounting for skill in trend, variability, and autocorrelation facilitates better multi-model projections: Application to the AMOC and temperature time series. *PLOS ONE*, **14**, e0214535, <https://doi.org/10.1371/journal.pone.0214535>.
- Payne, A. E., and G. Magnusdottir, 2015: An evaluation of atmospheric rivers over the North Pacific in CMIP5 and their response to warming under RCP 8.5. *J. Geophys. Res. Atmos.*, **120**, 11 173–11 190, <https://doi.org/10.1002/2015JD023586>.
- Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*, **24**, 2358–2367, <https://doi.org/10.1175/2010JCLI3814.1>.
- Peterson, T. C., and Coauthors, 2013: Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United States: State of knowledge. *Bull. Amer. Meteor. Soc.*, **94**, 821–834, <https://doi.org/10.1175/BAMS-D-12-00066.1>.
- Pierce, D. W., and Coauthors, 2013: The key role of heavy precipitation events in climate model disagreements of future annual precipitation changes in California. *J. Climate*, **26**, 5879–5896, <https://doi.org/10.1175/JCLI-D-12-00766.1>.
- Radić, V., A. J. Cannon, B. Menounos, and N. Gi, 2015: Future changes in autumn atmospheric river events in British Columbia, Canada, as projected by CMIP5 global climate models. *J. Geophys. Res. Atmos.*, **120**, 9279–9302, <https://doi.org/10.1002/2015JD023279>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast

- ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- Sanderson, B. M., M. Wehner, and R. Knutti, 2017: Skill and independence weighting for multi-model assessments. *Geosci. Model Dev.*, **6**, 2379–2395, <https://doi.org/10.5194/GMD-10-2379-2017>.
- , and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.*, **39**, L16708, <https://doi.org/10.1029/2012GL052665>.
- , —, and P. Caldwell, 2015: Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Climate*, **28**, 5150–5170, <https://doi.org/10.1175/JCLI-D-14-00361.1>.
- Shepherd, T. G., 2014: Atmospheric circulation as a source of uncertainty in climate change projections. *Nat. Geosci.*, **7**, 703–708, <https://doi.org/10.1038/geo2253>.
- Shields, C. A., and J. T. Kiehl, 2016a: Atmospheric river landfall-latitude changes in future climate simulations. *Geophys. Res. Lett.*, **43**, 8775–8782, <https://doi.org/10.1002/2016GL070470>.
- , and —, 2016b: Simulating the Pineapple Express in the half degree Community Climate System Model, CCSM4. *Geophys. Res. Lett.*, **43**, 7767–7773, <https://doi.org/10.1002/2016GL069476>.
- Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, <https://doi.org/10.1175/BAMS-D-11-00094.1>.
- Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **A365**, 2053–2075, <https://doi.org/10.1098/RSTA.2007.2076>.
- USGCRP, 2017: *Climate Science Special Report: Fourth National Climate Assessment (NCA4)*. Vol. I, D. J. Wuebbles et al., Eds., U.S. Global Research Program, 470 pp., <http://doi.org/10.7930/J0J964J6>.
- Vrugt, J. A., 2016: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environ. Modell. Software*, **75**, 273–316, <https://doi.org/10.1016/j.envsoft.2015.08.013>.
- , and B. A. Robinson, 2007: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resour. Res.*, **43**, W01411, <https://doi.org/10.1029/2005WR004838>.
- , and E. C. Massoud, 2019: Uncertainty quantification of complex system models: Bayesian analysis. *Handbook of Hydrometeorological Ensemble Forecasting*, Springer, 563–636.
- , C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson, 2008: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.*, **44**, W00B09, <https://doi.org/10.1029/2007WR006720>.
- Walsh, J., and Coauthors, 2014: Our changing climate. *Climate Change Impacts in the United States: The Third National Climate Assessment*, J. M. Melillo, T. C. Richmond, and G. W. Yohe, Eds., U.S. Global Change Research Program, 19–67, <https://doi.org/10.7930/J0KW5CXT>.
- Wang, C.-C., B.-X. Lin, C.-T. Chen, and S.-H. Lo, 2015: Quantifying the effects of long-term climate change on tropical cyclone rainfall using a cloud-resolving model: Examples of two landfall typhoons in Taiwan. *J. Climate*, **28**, 66–85, <https://doi.org/10.1175/JCLI-D-14-00044.1>.
- Warner, M. D., C. F. Mass, and E. P. Salathé Jr., 2015: Changes in winter atmospheric rivers along the North American west coast in CMIP5 climate models. *J. Hydrometeorol.*, **16**, 118–128, <https://doi.org/10.1175/JHM-D-14-0080.1>.
- Wenzel, S., P. M. Cox, V. Eyring, and P. Friedlingstein, 2014: Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models. *J. Geophys. Res. Biogeosci.*, **119**, 794–807, <https://doi.org/10.1002/2013JG002591>.