https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

Earth System Dynamics
Discussions

To weight or not to weight: assessing sensitivities of climate model weighting to multiple methods, variables, and domains

Adrienne M. Wootten[1], Elias C. Massoud[2], Duane E. Waliser[3], Huikyo Lee[3]

[1]South Central Climate Adaptation Science Center, University of Oklahoma, Norman, OK, 73019, USA
[2]Department of Environmental Science, Policy and Management, University of California Berkeley, Berkeley, CA, 94720, USA
[3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 91109, USA

Correspondence to: Adrienne M. Wootten (amwootte@ou.edu)

Abstract. Given the increasing use of climate projections and multi-model ensemble weighting for a diverse array of applications, this project assesses the sensitivities of climate model weighting, and their resulting ensemble means, to multiple components, such as the weighting schemes, climate variables, or spatial domains of interest. The analysis makes use of global climate models from the Coupled Model Intercomparison Project Phase 5 (CMIP5), and their statistically downscaled counterparts created with the Localized Canonical Analogs (LOCA) method. This work focuses on historical and projected future mean precipitation and daily high temperatures of the south-central United States. Results suggest that model weights and corresponding weighted projections are highly sensitive to the weighting method as well as to the selected variables and spatial domains. For instance, when estimating model weights based on Louisiana precipitation, the weighted projections show a wetter and cooler south-central domain in the future compared to other weighting schemes. Alternatively, for example, when estimating model weights based on New Mexico temperature, the weighted projections show a drier and warmer south-central domain in the future. However, when considering the entire south-central domain in estimating the model weights, the weighted future projections show a compromise in the precipitation and temperature estimates. If future impact assessments utilize weighting schemes, then our findings suggest that how the weighting scheme is derived and applied to the projections may depend on the needs of an impact assessment or adaptation plan. From the results of our analysis, we summarize our recommendations concerning multi-model ensemble weighting as follows:

- Weighted ensemble means should be used not only for national and international assessments but also for regional impacts assessments and planning.
- Multiple strategies for model weighting are employed when feasible, to assure that uncertainties from various sources (e.g., weighting strategy used, domain or variable of interest applied, etc.) are considered.
- That weighting is derived for individual sub-regions (such as the NCA regions) in addition to what is derived for the continental United States.
- That domain-specific weighting be derived using both common (e.g. precipitation) and stakeholder-specific (e.g. streamflow) variables to produce relevant analysis for impact assessments and planning.

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

1 Introduction

The simulation output from climate models has been traditionally used for research into characterizing and understanding the climate system across multiple spatial scales. In recent years, ensembles of climate projections are increasingly used for impact and vulnerability assessments (e.g., Massoud et al., 2018, 2019, 2020ab; Wootten et al., 2020ab). These include large-scale assessments, such as the National Climate Assessment (Wuebbles et al. 2017), and local and regional assessments for individual areas of the United States. Large and local scale assessments can make use of the entire ensemble of climate projections (composed of global climate models [GCMs]), or make use of the ensemble mean, which provides representative information from multiple GCMs. For these assessments, using the ensemble mean provides a useful and convenient way to assess projected changes in a region. Given the coarse resolution of the GCMs (typically > 100km^2), many of these assessments make use of downscaled climate projections to translate larger-scale changes to local scales.

Alongside the use of climate modeling and downscaling for climate research and increased use for impact and vulnerability assessments, there has also been a transition in the last 20 years toward using weighted multi-model means. Model weights based upon skills of historical simulations have been shown to have greater accuracy than an arithmetic multi-model mean in many cases, provided that there is enough information to determine a weight for each model (Knutti et al. 2010; Weigel et al. 2008; Pena and Van den Dool, 2008; Min and Hense, 2006; Robertson et al. 2006). In the last few years, weighting based solely on skill has given way to weighting based upon both skill and independence. This transition has resulted from the recognition that some models can be more skillful for certain variables and regions, but also as common bases of model structure, parameterizations and associated programming code can result in a lack of independence between GCMs (Massoud et al. 2019, 2020a; Sanderson et al. 2015, 2017; Knutti, 2010; Knutti et al. 2017). In acknowledgment of studies indicating that the global climate models are not fully independent, the Fourth National Climate Assessment (NCA4) was the first major climate assessment in the United States to use skill and independence-based model weighting on the ensemble of climate models (Sanderson and Wehner, 2017).

The authors of this paper have extensively investigated the effect of model weighting on the outcome of climate change projections from multi-model ensembles (Massoud et al. 2019, 2020a; Wootten et al. 2020a). For example, in Massoud et al. (2019), the authors utilized information from various model averaging approaches to evaluate 21 global climate models from the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012), and they based their weighting strategies on model independence as well as performance skill of atmospheric rivers globally. In Massoud et al. (2020a), the authors used Bayesian model averaging (BMA) as a framework to constrain the spread of uncertainty in climate projections of precipitation over the contiguous United States (CONUS). In Wootten et al. (2020a), the authors applied various ensemble-weighting schemes to constrain precipitation projections in the south-central United States and applied these strategies to both the 26-model ensemble from the CMIP5 archive and the downscaled version of the models. The latter study is distinct from prior research, because it compared the interactions of ensemble-weighting schemes with GCMs and statistical downscaling to produce multi-model ensemble means.

Some studies have applied model weighting to a certain variable (e.g. precipitation) and went on to investigate climate change impacts for other variables (e.g. temperature or streamflow) (c.f. Knutti et al., 2017; Massoud et al., 2018). The National Climate Assessment had previously considered weighting based only on commonly used climate variables (e.g. precipitation and temperature, Wuebbles et al., 2017), but discussions to use additional variables are currently ongoing. Other studies have applied model weighting to a specific domain (e.g. globally) and went on to apply the developed weights on a different domain (e.g. North America or Europe) (Massoud et al., 2019). However, these studies are rare, as are studies providing comparisons of various weighting schemes (e.g. Shin et al. 2020; Brunner et al., 2020a; Kolosu et al. 2021), and no previous study offers a comprehensive cross-comparison of the effects on the ensemble means from the choices of the domain, variable, weighting scheme, and ensemble. The current study will answer the two following questions regarding model weighting: Should model weights be developed separately when investigating different climate variables? Should model weights be estimated separately when investigating different domains?

Taking these points into consideration, we assess the choice of model weighting strategy by developing and investigating a multi-dimensional sensitivity matrix for applying model averaging for the south-central region of the US - as defined by the NCA. To this end, we look at mean precipitation and high temperatures as our climate variables of interest. Furthermore, we split the entire south-central region into three different domains; Louisiana, New Mexico, and the entire domain. Overall, we created and apply various sets of model weights based on several choices: a) the choice of the ensemble (CMIP5 or downscaled), b) the choice of model weighting scheme, c) the choice of climate variable of interest (precipitation vs temperature), and d) the choice of the domain used to derive weighting (entire south-central region vs smaller sub-domain). Therefore, one example of a strategy that we apply to estimate a set of weights uses the BMA weighting method on the CMIP5 ensemble projections of the precipitation variable for the Louisiana domain. To our knowledge, there has not been a model weighting study that included as many dimensions in the experimental matrix as this study, again these are model ensemble, domain, variable, and importantly, the weighting scheme itself.

Our analysis results in a wide array of possible future outcomes, which comes with high uncertainties on what to expect in the future in this domain. The main question we are after is whether or not some variables or domains have projected climate change signals that have high certainty. Alternatively, we would like to find out whether or not there are climate variables in any of the regions that have highly uncertain climate change projections. We aim to address these uncertainties by applying the multi-dimensional experimental matrix of model weighting strategies and hope to inform the scientific community of these sensitivities for the benefit of future stakeholders, including climate modelers and boundary organizations providing climate services.

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

2 Methods and Data

2.1 Study Domain and Variables

The south-central United States (from about 26°N 108.5°W to 40°N 91°W) has a varied topography with a sharp gradient in mean annual precipitation from the east (humid) to the west (arid), and a generally warm climate. The Mississippi River Valley and the Ozark Mountains in the eastern portion of the region (elevations of 200–800 m), the Rocky Mountains in the west (1500–4400 m), and the Gulf of Mexico in the southeast (near sea level). Precipitation in the southeast portion of the domain can be eight times higher than drier western locations and average high temperatures can reach 40°C (Figure 1).

2.2 Climate Projection Datasets

We use one member each from 26 GCMs in the CMIP5 archive to form the GCM multi-model ensemble. To form the downscaled ensemble, the same 26 GCMs are used from the downscaled projections created with the Localized Constructed Analogs (LOCA) method (Pierce et al. 2014). The LOCA-downscaled projections have been used in other studies, including the NCA4 (USGCRP, 2017) and Wootten et al. (2020a). Table S1 lists the GCMs used for both the GCM ensemble (hereafter CMIP5 ensemble) and downscaled ensemble (hereafter LOCA ensemble). See Wootten et al. (2020a) for more details on the climate projection datasets.

To facilitate analysis, the data for each ensemble member are interpolated from their native resolution to a common 10 km grid using a bi-linear interpolation similar to that described in Wootten et al. (2020b). We examine projected daily precipitation (pr) and daily high temperature (tmax) changes from 1981–2005 to 2070–2099 under the RCP 8.5 scenario, which ramps the anthropogenic radiative forcing to 8.5 W/m^2 by 2100. We chose RCP 8.5 to maximize the change signals and allow us to analyze greater differences between weight schemes and downscaling techniques. The historical period (1981–2005) is used for both the historical simulations and observations to facilitate comparisons with other studies (Wootten et al. 2020b) and because the historical period of the CMIP5 archive ends in 2005 (Taylor et al. 2012).

2.3 Observation Data

Many publicly available downscaled projections (including LOCA) are created using gridded observation-based data for training. Gridded observations are based largely on station data that are adjusted and interpolated to a grid in a manner that attempts to account for biases, temporal/spatial incoherence, and missing station data (Behnke et al. 2016; Wootten et al. 2020b; Karl et al. 1986; Abatzoglou, 2013). In this study, we use Livneh version 1.2 (hereafter Livneh [Livneh et al. 2013]), interpolated to the same 10 km grid using bilinear interpolation, as the gridded observation data used for comparison to the ensembles. Livneh is used in part to facilitate any comparisons between this study and the results of Wootten et al.

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

(2020a). The LOCA ensemble used the Livneh data as the training data, so it is expected that LOCA will be more accurate than the CMIP ensemble when compared to the Livneh dataset. While we recognize that different gridded observations and downscaling techniques influence projections of precipitation variables (e.g. number of days with rain, heavy rain events), the effect is minimal on the mean annual precipitation (Wootten et al. 2020b). Therefore, we find it is appropriate to make use of only one statistical downscaling method and one gridded observation dataset.

2.4 Weighting Schemes

In this analysis, we make use of model weighting schemes detailed in Wootten et al. (2020a) and similar to the weighting schemes applied in Massoud et al. (2020a). The resulting weighting schemes are applied multiple times to complete an experimental matrix allowing for in-depth comparisons of the sensitivity of the ensemble mean to various approaches to deriving and applying the multi-model weights. These weighting methods include the unweighted model mean, the historical skill weighting (hereafter Skill), the historical skill and historical independence weighting (SI-h), the historical skill and future independence weighting (SI-c), and the Bayesian Model Averaging (BMA) method. In essence, the unweighted strategy takes the simple mean of the entire ensemble. The Skill strategy utilizes each model's skill in representing the historical simulations. The SI-h strategy also uses historical skill but considers the independence of each model in the historical simulations. The SI-c strategy uses historical skill and independence of each model found in the climate change signal (i.e. in the future projections) Finally, the BMA strategy employs a probabilistic search algorithm to find an optimal set of model weights that produce a model average that has high skill when compared to the observation and its uncertainty. Refer to Wootten et al. (2020a) and Massoud et al. (2019, 2020a) for more information on how the model weighting schemes are applied.

2.5 Experimental Matrix

Each weighting scheme (Skill, SI-h, SI-c, and BMA) is applied to both ensembles (CMIP5 and LOCA) to fill out an experimental matrix of weights. The weighting schemes are applied to find the best historic fit of two climate variables (tmax and pr). The weighting schemes are also applied to find the best historic fit for three different domains; the full domain (Southern Great Plains), Louisiana only, and New Mexico only. As a result, for each weighting scheme (skill, SI-h, SI-c, and BMA) and ensemble (CMIP5 and LOCA), there are six sets of weights produced (i.e. 3 regions and 2 variables). One example of this would be a BMA weighting strategy used on the CMIP5 ensemble trained on tmax for the entire domain. Another example would be a skill-based weighting strategy used on the LOCA ensemble trained on precipitation in Louisiana. There are a total of 48 such model weighting strategies (ensemble choice x weighting methods choice x variable choice x domain choice = 2 x 2 x 3 x 4 = 48). In addition to the set of 48 weighting strategies, an unweighted ensemble mean is also used. The unweighted strategy effectively has equal weights for all models regardless of variable, domain, or ensemble. As such, including an unweighted ensemble mean represents only one additional modeling strategy, which brings the total to 49 model averaging strategies in our experimental matrix.

The various model weights from each scheme are calculated, and the derived sets of weights are then applied to create ensemble means for the three domains and two variables. In other words, a certain set of weights can be used to determine projected changes in either tmax or pr and can be used for any of the domains, i.e. full domain, Louisiana, or New Mexico. There are a total of 288 such maps that can be created to investigate future climate change. These are 48 model averaging choices described above, applied to 2 different variables in 3 different domains, or 48 x 2 x 3 = 288 combinations of maps. This collection of 288 is in addition to the results from unweighted means of temperature and precipitation. Including these unweighted means, there are 290 combinations of maps from this project. This explains the highly dimensional experimental matrix applied in this study, which provides the total uncertainty that is estimated with our future change projections. See Figure 2 for a schematic describing the various choices made to create each model weighting strategy and the choices made to how each of these model weights can be applied.

3 Results

This section will first consider the sensitivity of the model weighting schemes to the ensembles, variables, and domains used. This section will then focus on the bias and change signal from the resulting combinations of ensemble means.

3.1 Ensemble weights – results from various model weighting strategies

The resulting sets of model weights for the CMIP5 ensemble based on the weighting scheme, variable, and domain, are shown in Figure 3. The 24 sets of model weights for the LOCA ensemble based on the weighting scheme, variable, and domain, are shown in Figure 4. Alongside the best-estimated weight from the BMA weighting scheme, the box-whisker plots in the image show the spread of weights from the 100 iterations of BMA for each ensemble, variable, and domain to which BMA was applied. The grey dots in these figures depict the outliers from the BMA distributions of weights.

One observation seen in these weighting combinations is that the weighting schemes themselves are all sensitive to the ensemble, variable, and domain for which they are derived. This is reflected further when one considers which models from each ensemble are given the strongest weights by each model weighting scheme (Table 1). From Table 1, no model appears in the top three for all model combinations. The model most consistently in the top three is the CanESM2, which is in the top three for 35.4% of the 48 weighting combinations.

Although the weighting schemes are sensitive to ensemble, variable, and domain, the weights produced by Skill, SI-h, and SI-c are similar to each other, while the BMA weighting tends to be different. This is particularly true for precipitation and follows what was shown by Wootten et al. (2020a) and Massoud et al. (2020a). The BMA approach provides a distribution of weights for each model and this distribution of weights overlaps the weights of the Skill, SI-h, and SI-c approaches. This distribution of weights covers a broader region of the model weight space, but the best BMA combination (marked as orange squares in Figures 3 and 4) is significantly different from the other schemes.

Aside from the difference within each combination of ensemble, variable, and domain, there are also notable differences between these combinations. The pattern of the weights, shown in Figures 3 and 4, changes significantly between combinations, particularly among the BMA weights and in the CMIP ensemble. Among the BMA and CMIP5 ensemble combinations (Figure 3), there are no common patterns to the model weights based on domain or variable. However, while the patterns between Skill, SI-h, and SI-c are similar to each other, their magnitude is consistently smaller than BMA. This indicates that when applying different weighting schemes, different models are given higher weights when applying the CMIP5 ensemble for different domains or variables.

When using the LOCA ensemble (Figure 4), there is more consistency in which models are given higher weights, particularly when weights are derived based on high temperature (tmax). For the LOCA ensemble, the distribution of the BMA weights has a similar pattern across all three domains for the tmax derived weights, and the best-weighted models are also somewhat consistent between domains. Similar to the CMIP5 ensemble in Figure 3, the BMA weights tend to be larger for the highest weighted models in the LOCA ensemble compared to those derived with the Skill, SI-h, and SI-c schemes. For weights derived with tmax, the Skill, SI-h, and SI-c have very similar patterns for both the full and New Mexico domains. The Skill and SI-h weighting schemes, which focus entirely on the historical period, created nearly identical weights for the 26 models when weights are derived based on tmax in the full and New Mexico domains. While the weights from Skill and SI-h are not identical when derived using tmax in the Louisiana domain, the weights for the LOCA ensemble generally range from 0.025 to 0.050. The SI-c weights derived using tmax in the LOCA ensemble have a similar pattern between the full and New Mexico domains, but a very different pattern in the Louisiana domain (Figure 4). In addition, the SI-c also tends to have a different pattern from the Skill and SI-h weights when tmax and LOCA are used for derivation. There is much more sensitivity to domains when using precipitation and the LOCA ensemble to derive weights, compared to that of tmax. Regardless of the weighting scheme, there is no common pattern in the weights between domains when the LOCA ensemble and precipitation are used to derive weights. Again, the BMA scheme applies much larger weights to the top models for precipitation-based LOCA weighting compared to the Skill, SI-h, and SI-c weighting schemes.

The LOCA statistical downscaling method, like most statistical downscaling methods, incorporates a bias correction approach, which inherently improves the historical skill. In addition, the Skill, SI-h, and SI-c methods focus primarily on the first moment of the ensemble distribution when deriving weights, which limits the ability to penalize for co-dependence between models in an ensemble. Finally, the BMA considers multiple moments of the ensemble distribution using multiple samples via Markov Chain Monte Carlo (MCMC), rewarding skillful models and penalizing co-dependency. Of the weighting combinations used here, the BMA tends to be the most sensitive to the ensemble, variable, and domain used to determine weights. Given that the BMA focuses on multiple moments of the distribution and is most sensitive to the different choices considered here (ensemble, variable, and domain) it is plausible that the BMA approach responds to and captures the changes in skill and co-dependence among the ensemble members resulting from these various choices.

3.2 Size of the experimental matrix of model weights and how to apply them

One can apply the 48 weighting combinations described above in a similar manner to the way the weighting combinations themselves are created. For example, one could apply the weights derived from the CMIP5 ensemble precipitation for the full domain using BMA to create a weighted ensemble mean of CMIP5 precipitation for Louisiana. As shown in Figure 2, each weighting combination is applied to the variables (high temperature and precipitation) and domains (full, Louisiana, and New Mexico) to produce a set of ensemble means. Altogether, the maximum number of weighted ensemble means produced with these 48 weighting combinations is 48x2x3=288. However, this maximum number of ensemble means resulting from the experiment contains several duplicates. For example, when using the same set of weights, the resulting ensemble mean in a subdomain will be the same as the resulting ensemble mean from the same portion of the full domain. As such, the actual number of ensemble means in this experiment is smaller than 288.

3.3 Historical Bias and Future Projected Changes in unweighted model ensembles

The figures shown in later sections focus on the ensemble means from the 48 weighting combinations applied to the full domain. The discussion surrounding bias and projected changes represented by the ensemble means in the following subsection will be compared to the unweighted ensemble means of high temperature and precipitation from the CMIP5 and LOCA ensembles. For this reason, we first show the historical ranges and the ranges of the future projected changes using the unweighted model ensemble (Figure 5) before reporting on the results using the weighted ensembles. The unweighted CMIP5 ensemble as a whole tends to underestimate high temperatures in the historical period, overestimate precipitation in New Mexico, and underestimate precipitation in Louisiana (top left panel of Figure 5). The LOCA ensemble is much closer to the Livneh observations, which is expected given the bias correction applied in statistical downscaling. Yet, for the unweighted LOCA ensemble, there is a tendency to underestimate precipitation in the whole domain and the New Mexico subdomain and to overestimate temperature in all of the domains (bottom left panel of Figure 5). For the future projected changes in the unweighted CMIP and LOCA ensembles, the projected high temperature changes are consistent between ensembles (bottom right panel of Figure 5), and the projected changes in precipitation are less variable in the LOCA ensemble for the New Mexico domain and more variable for the Louisiana domain (top right panel of Figure 5). Given this baseline information, the following subsections discuss and compare the unweighted and weighted ensemble means for each ensemble (CMIP5 and LOCA).

3.4 Historical Bias and Future Projected Changes using the weighted ensembles

The 48 combinations of model weights are then applied across three domains and two variables to produce 288 ensemble means. The mean projected changes can be sensitive to the weighting scheme, domain, and variable used. The future projected changes from the different ensemble means are summarized in Figure 6, where the boxplots represent the range of the ensemble mean change from the 100 BMA posterior weights. When the weighting is derived using tmax, the resulting CMIP5 mean projected change shows predominantly a decrease in precipitation for all domains (top-left group of panels in Figure 6, top row of figures). For the tmax derived weighting with the LOCA ensemble (top right group of panels in Figure 6, top row of figures), the mean precipitation projections are more variable concerning the domain the weighting is applied.

Using weights derived with precipitation and the CMIP5 ensemble, the mean projected precipitation increases/decreases when Louisiana/New Mexico is used to derive weights across all three applied domains (top-left group of panels in Figure 6, bottom row of figures). Using weights with precipitation in the LOCA ensemble, the mean projected precipitation generally decreases for most weighting schemes (top right group of panels in Figure 6, bottom row of figures), except for the resulting means for Louisiana with the BMA weighting scheme. In contrast to precipitation, the ensemble mean changes for tmax are fairly consistent for both CMIP and LOCA ensembles (bottom groups of panels in Figure 6, all rows of figures), with all model weighting strategies indicating a consistent increase in temperature for all domains.

The following section and corresponding figures compare the results from the various weighting schemes applied in this study. Figure 7 looks at historical biases and Figure 8 shows the projected future change signals in precipitation for the CMIP5 ensemble of models. Figures 9 and 10 look at historical bias and projected future change signals in high temperature for CMIP5. Figure 11 looks at the projected future change signal in precipitation for the LOCA ensemble, and Figure 12 looks at the projected future change signal in high temperature for the LOCA ensemble. For an in-depth analysis of how the model weighting strategies impact the resulting historical bias and climate change signals shown in Figures 7-12, readers are referred to the supplementary section, with a discussion on the main findings reported in the next section. For additional results that complete the analysis, readers are referred to the supplementary section (Figures S1-S6), which includes bias maps from the LOCA ensemble (S1-S2) as well as error distributions from the historical simulations of both ensembles (S3-S6).

4 Discussion

Among climate scientists and the climate modeling community, there is a debate regarding the weighting of multi-model ensembles and, if one does apply weighting, how to do so. This is the first study, to the authors' knowledge, to comprehensively assess the sensitivities of the model weights and resulting ensemble means to the combinations of variables, domains, ensemble types (raw or downscaled), and weighting schemes used. Therefore, this study quantifies multiple weighting sensitivities to inform the larger discussion on multi-model ensemble weighting.

4.1 Sensitivities of the Results to the Experimental Design

The results from individual weighting schemes are sensitive to the choice of domain and variable of interest, regardless of whether the ensemble is downscaled or not. However, one can also note that the BMA weighting scheme tends to be more sensitive than the others. As noted by Wootten et al. (2020a) and Massoud et al. (2019, 2020a), the Skill, SI-h, and SI-c weighting schemes focus on the first moment of the distribution of a variable, while the BMA approach focuses on multiple moments of the distribution of weights. The BMA weighting can therefore produce weights that are significantly different from the other schemes. In addition, the BMA will also be more sensitive to the differences between domains and variables that are provided to derive model weighting. This is particularly the case with regards to the CMIP5 ensemble results for both variables but also is evident in the LOCA ensemble results for precipitation. The ensemble weights are most sensitive to the variable and domain using the CMIP5 ensemble and the weights created with the LOCA ensemble are less sensitive. A statistical downscaling procedure reduces the bias of the ensemble members compared to the raw CMIP5 ensemble, which likely results in there being less sensitivity when the LOCA ensemble is used. This is particularly likely for high temperatures, which is traditionally much less challenging for both global models and downscaling techniques to capture.

We find that, for precipitation, the ensemble mean projected change from a multi-model ensemble is sensitive to the various choices associated with the derivation of model weighting. In contrast, for tmax, the ensemble mean projected change is less sensitive. The larger domain of the south-central region contains multiple climatic regions. The western portion of the domain includes the arid and mountainous New Mexico and Southern Colorado. The eastern portion of the domain is the much wetter and less mountainous area of Louisiana, Arkansas, and southern Missouri. The complexity of the region presents a challenge to GCM representation of precipitation and temperature. Deriving ensemble weights based on Louisiana precipitation favors models which are wetter while deriving ensemble weights based on New Mexico precipitation favors those models which are drier. This effect translates into the projected changes for precipitation in the CMIP5 ensemble that can reverse the change signal in the domain. The sensitivity for precipitation is evident when precipitation is the focus for deriving model weights, but also present to a lesser degree when high temperature is the focus for deriving model weights. The high temperature changes are also sensitive to the domain when precipitation weighting is used because precipitation-based weighting favors wetter or drier models. In contrast, the high temperature change from the CMIP5 ensemble is much less sensitive when calculated with weights derived from high temperatures. The sensitivity present using the CMIP5 ensemble is less apparent for the projected changes with the LOCA ensemble. In particular, LOCA ensemble means derived using the BMA weighting are more sensitive to the variable and domain used to derive weights. The LOCA downscaling, like most statistical downscaling methods, corrects the bias of the CMIP5 ensemble, pushing all models to have similar historical skill. It follows that the BMA weighting is more sensitive to the different choices considered here (ensemble, variable, and domain) and that the BMA weighting responds to and captures changes in skill and co-dependence resulting from the different options of ensemble, variable, and domain.

4.2 Broader Questions

Weighted multi-model means have primarily been focused on GCMs and continental scales. However, the use of climate projections has extended to regional, state, local, and tribal uses for climate impact assessments and adaptation planning. In these regional to local efforts, the raw projection data has been used but also provided to impact models (such as hydrology or crop models). Currently, impact assessments outside the traditional venues of climate modeling tend not to use weighted multi-model means but tend to use unweighted means created using downscaled GCM ensembles. From this study, several questions arise. First, should impact assessments make use of weighted multi-model means? If yes, then a second question is, should multiple weighting schemes and ensemble means be used? Third, for situations where projections are provided to impact models, does this type of study need to be repeated using impact model results? These three questions are also related to the questions mentioned earlier. Should model weights be developed separately when investigating different climate variables? Should model weights be estimated separately when investigating different domains? All such questions could be considered in terms of climate modeling or broader impact assessments and applications.

4.3 To weigh or not to weigh?

At the time of writing, discussion surrounding the use of weighted multi-model ensembles has been limited to climate model developers and the production of national or international climate assessments. The authors know of no impact assessments or adaptation planning exercises where a weighted multi-model ensemble mean is discussed (although the NCA reports address some of these topics), let alone used, by the authors or planners involved or considered by the boundary organizations that serve them. Among climate model developers, Knutti et al. (2017) argue that model weighting is a necessity in part to account for situations where the model spread in the present-day climatology is massive resulting in some models having biases so large that using an unweighted mean is difficult to justify. In other situations, model interdependence becomes increasingly relevant with the increased use of common code bases across institutions causing unweighted means to be overconfident (Brunner et al., 2020b). This concern was also shared by Wootten et al. (2020a) with respect to the common modeling code base applied in the statistical downscaling process. Others have argued that using model weighting tunes the ensemble mean to those models favored during the historical period and allows no flexibility for the change in climate that may be better represented by models that perform poorly in the historical period.

The results from this study demonstrate that the weights and resulting ensemble means are sensitive to the ensemble (CMIP or LOCA), variable, and domain used. However, the concerns of Knutti et al. (2017) and Wootten et al. (2020a) still stand. An unweighted mean will continue to over-favor models with large biases and co-dependencies regardless of the domain or variable of interest for either climate models or impact assessments. For this reason, we recommend the use of weighted ensemble means not only for national and international assessments but also for regional impacts assessments and planning. Additionally, given the sensitivities presented in this study, we recommend not only that model weighting is applied, but that multiple strategies for model weighting are employed when feasible, to assure that uncertainties from various sources (e.g., weighting strategy used, domain or variable of interest applied, etc.) are considered.

4.4 Consideration of weighting scheme, variables of interest, and domain choice

The questions regarding the use of multiple weighting schemes and deriving such schemes with a specific focus on domains or variables of interest are interrelated given the sensitivities of the various weighting schemes to variable and domain. The use of multiple weighting schemes would allow for the sensitivities associated with model weighting to be captured and considered. However, it is important to note that the added value of using multiple weighting schemes may well depend on the domain and variables of interest. Mean projections of temperature are much less sensitive to the weighting scheme used, while mean projections of precipitation are more sensitive, particularly if the domain is very wet or very arid.

Weighting for a specific variable is a more difficult question. In an impact assessment, one might justifiably argue that one should weigh the ensemble on the specific variable of interest for that assessment. Likewise, for national-level assessments and climate modeling, weighting on specific variables could be used to address the large biases and co-dependencies with respect to that variable among the models and produce ensemble means that reflect the appropriate confidence with regards to that variable. However, temperature, precipitation, and multiple other variables have strong physical relationships and thus are not fully independent themselves. As such, creating separate weights for variables independently may break the physical relationships in resulting ensemble means. Nevertheless, the weighting schemes used in this study have the capacity for multivariate ensemble weighting. Future work by the authors will explore multivariate ensemble weighting, in part to assess if multivariate weighting results in robust weighting for the variables used while retaining the physical relationships between the variables of interest. For national assessments, we recommend the use of multiple weighting schemes with multiple variables to assess the sensitivity and ultimately reduce the uncertainty for projected mean changes. For individual impact assessments, the focus on individual variables is likely context-dependent as individual planning decisions and impact assessments are strongly dependent on the region of interest and local climatic changes. A local/regional assessment often focuses on variables uncommon to climate model evaluations that are (or can be) derived from common variables in climate model evaluations. As such, a stakeholder-specific variable (such as growing season length) has a strong relationship with a common climate variable (such as temperature). With this in mind, weighting used in impact assessments should likely be derived using multiple variables incorporating both common and stakeholder-specific variables to produce relevant analysis for impact assessments and planning.

Climate model evaluations and national assessments typically focus on the continental United States or North America. However, the individual National Climate Assessment regions are climatically very different from each other. The individual

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

GCMs in the CMIP ensemble likely do not have the same performance across all regions and an individual downscaling technique can be evaluated in one of these regions but applied to the entire continental United States or North America. In addition, the regions of Alaska, the U.S. Pacific Islands, and the U.S. Caribbean Islands have vastly different climates to the continental United States. The model weighting for each of these regions will likely be vastly different than the weighting for the continental United States as a whole. Given the different climates, it is recommended that weighting is derived for the NCA regions in addition to what is derived for the continental United States. This will allow for larger-scale assessments to account for the ability of the ensemble to reflect the unique climate of these regions while considering the ability of the ensemble to reflect the larger scale patterns which influence the climate in the different subregions.

4.5 Caveats, Challenges, and Future Work

An impact assessment or adaptation planning effort can span a range of spatial scales from municipalities to states or regions. For impact assessments involving larger states or regions, we also recommend a domain-focused weighting, both to capture the needs of the planners or stakeholders involved and to capture the climate in the area of interest. However, for smaller states or local municipalities, we do not recommend deriving model weighting based on these small regions. At small scales, the natural variability of a climate model may result in a model having the local climate correct but the larger climatic patterns represented incorrectly. As such, for impact assessments involving smaller areas, we recommend that model weighting be derived using the larger region that the small domain is situated in to avoid the confounding factor of natural variability in model weighting. One caveat in this study is that the sub-domains of New Mexico and Louisiana are quite small compared to the resolution of the GCMs in CMIP5. This suggests that natural variability may have had some effect on the results. In future work, the authors will repeat this analysis using the larger regions of the United States used in the National Climate Assessment.

The authors recognize that the above recommendations are similar between the community of climate model developers invested in evaluation and assessment generally and the users and stakeholders now using climate projections for local and regional impact assessments. The authors also recognize that implementing such recommendations is more feasible for the former community than the latter. The latter community, users and stakeholders invested in impact assessments and adaptation planning, faces the added challenge that some impact assessments or planning efforts require using climate model projections (or downscaled climate projections) as inputs to additional modeling efforts such as hydrology modeling or crop modeling. While most impact assessments have not incorporated model weighting directly, some are beginning to do so (e.g. Skahill et al., 2021). Knowing this and the sensitivities that this study demonstrates, it is recommended for future efforts to examine the weighting of impacts model outputs from climate model inputs. Would weighting based on climate model inputs produce the same result as weighting based on, for example, streamflow output using an ensemble of climate projections as inputs? Given the sensitivities for weighting schemes, variables, domains, and ensembles, we suspect that the weighting would not be the same and that the translation of error and co-dependencies from climate model projections to impacts models (such as a hydrology model) may result in a higher degree of sensitivity with respect to the resulting ensemble mean of stakeholder specific variables (such as streamflow). While there is less capacity among the users of climate projections to address such questions, the boundary organizations in the United States and internationally are developing the capacity to provide or derive ensemble weights with emphasis on the need of stakeholders. Therefore, the questions of sensitivity of weighting schemes and ensemble means bear increasing relevance as the number of users of climate projection output continues to increase.

5 Conclusions

This study examines the sensitivity of the multi-model ensemble weighting process and resulting ensemble means to the choices of variable, domain, ensemble, and weighting scheme for the south-central region of the US. In general, we see that weighting for Louisiana makes the future wetter and less hot, weighting for New Mexico makes the future drier and hotter, and accounting for the whole domain provides a compromise between the two. In addition, we see that ensemble mean projections for precipitation are more sensitive to the various aspects tested in this study, while ensemble mean projections for high temperature are less sensitive. As such, some domains/variables have uncertain outcomes, regardless of the weighting method. But for other domains/variables, the uncertainty is dramatically reduced, which can be helpful for the assessment of climate models and climate adaptation planning. The sensitivity of precipitation and temperature projections is reduced when LOCA is used, which is likely the result of the bias correction associated with the LOCA downscaling method. In addition, the BMA weighting scheme is more sensitive than the other weighting schemes. BMA's sensitivity is the result of the BMA approach focusing on multiple moments of the distribution to account for model biases and co-dependencies.

Although there is sensitivity associated with the model weighting, a multi-model ensemble of climate projections should incorporate model weighting. Model weighting still accounts for issues of bias and co-dependence that preclude a model democracy approach to crafting multi-model ensemble means. Incorporating multiple weighting schemes allows for assessing and capturing the sensitivity associated with model weighting to the benefit of both climate modeling efforts and climate adaptation efforts. Given the sensitivity associated with weighting for different variables and domains, one may also consider crafting weighting schemes with a focus on the domains or variables of interest to an application. In addition, since some impact assessments or adaptation planning efforts make use of climate projections as inputs to impacts models (such as hydrology or crop models) there is a need to consider similar research to this study with regards to the direct outputs of impacts models using climate projections.

From the results of our analysis, we summarize our recommendations concerning weighting as follows:

- Weighted ensemble means should be used not only for national and international assessments but also for regional impacts assessments and planning.
- Multiple strategies for model weighting are employed when feasible, to assure that uncertainties from various sources (e.g., weighting strategy used, domain or variable of interest applied, etc.) are considered.
- That weighting is derived for individual sub-regions (such as the NCA regions) in addition to what is derived for the continental United States.
- That domain-specific weighting be derived using both common (e.g. precipitation) and stakeholder-specific (e.g. streamflow) variables to produce relevant analysis for impact assessments and planning.

There are a couple of caveats and suggested future research. First, this study makes use of domains that are fairly small compared to the natural variability present in a climate model. Second, this study focused on the south-central United States. Future efforts should consider this analysis using larger regions, such as the continental United States and the NCA sub-regions. Future efforts should also consider examining multivariate weighting to account for the physical relationships between variables. Finally, in the case of impacts models using climate projections, the weighting of the raw ensemble is likely different from weighting that may be applied using the output from impacts models using the ensemble as input. Given the increasing use of climate model ensembles in impacts models, future efforts should consider a similar investigation to this study using an impacts model. Such future efforts will answer multiple questions regarding the appropriate model weighting schemes, but also provide potential guidance to boundary organizations building capacity to assist in regional and local climate adaptation planning and impact assessments.

6 Code Availability

R Code to calculate weights associated with the Skill, SI-h, and SI-c weighting and produce all analysis in this study are available from Dr. Wootten on request. Programming code for BMA calculations is available from Dr. Massoud on request.

7 Data Availability

CMIP5 GCM output are available through the Earth System Grid Federation Portal at Lawrence Livermore National Laboratory (https://esgf-node.llnl.gov/search/cmip5/). The LOCA downscaled climate projections for CMIP5 GCMs are available through numerous portals included the USGS Center for Integrated Data Analytics GeoData Portal (cida.usgs.gov/gdp). The Livneh gridded observations are available from the National Centers for Environmental Information (https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.nodc:0129374;view=html).

8 Author Contribution

Dr. Wootten and Dr. Massoud – Conceptualization, Formal Analysis, Investigation, Methodology, Writing – original draft preparation, Writing – review and editing, Visualization, Validation. Dr. Wootten – Data Curation. Dr. Waliser and Dr. Leet – Supervision, Writing – review and editing.

9 Competing Interest

The authors declare that they have no conflict of interest.

10 Acknowledgements

A part of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). The authors thank the reviewers for their comments and critiques to strengthen this article.

References

Abatzoglou, J.: Development of gridded surface meteorological data for ecological applications and modeling, International Journal of Climatology, 33, 121-131, doi:10.1002/joc.3413, 2013.
Amante, C. and Eakins, B.W.: ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis; NOAA Technical Memorandum NESDIS NGDC-24; National Geophysical Data Center, NOAA: Boulder, CO, USA, 2009.
Behnke, R., Vavrus, S., Allstadt, A., Thogmartin, W., and Radelhoff, V.C.: Evaluation of downscaled gridded climate data for the conterminous United States, Ecological Applications, 26, 1338-1351, doi:10.1002/15-1061, 2016.
Bishop, Craig H., and Shanley, K.T.: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it, Monthly Weather Review, 136, 12, 4641-4652, 2008.
Brunner, L., McSweeney, C., Ballinger, A.P., Befort, D.J., Benassi, M., Booth, B., Coppola, E.: Comparing methods to constrain future European climate projections using a consistent framework, Journal of Climate, 33, 20, 8671-8692, 2020a.
Brunner, L., Pendergrass, A.G., Lehner, F., Merrifield, A.L., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth System Dynamics, 11, 4, 995-1012, 2020b.
Cesana, G., Suselj, K., and Brient, F.: On the Dependence of Cloud Feedbacks on Physical Parameterizations in WRF Aquaplanet Simulations, Geophysical Research Letters, 44, 10,762-10,771. doi:10.1002/2017GL074820, 2017.
Dilling, L, and Berrgren, J. 2014: What do stakeholders need to manage for climate change and variability? A document-based analysis from three mountain states in the Western USA, Regional Environmental Change, 15, 657-667, doi:10.1007/s10113-014-0668-y, 2014.
Duan, Q., Newsha, K., Ajami, X.G., and Sorooshian,S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources ,30, 5, 1371-1386, 2007.
Eyring, V., Bony, S., Meehl, G. A., Senior, C.A., Stevens, B., Stouffer, R.J., and Taylor, K.E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.
Eyring, V., Cox, P.M., Flato, G.M., Gleckler, P.J., Abramowitz, G., Caldwell, P., and Collins, W.D.: Taking climate model evaluation to the next level, Nature Climate Change, 1.

Fan, Y., Olson, R., and Evans, J.P.: A Bayesian posterior predictive framework for weighting ensemble regional climate models, Geoscientific Model Development, 10, 6, 2321-2332, 2017.

Gibson, Peter B., Waliser, D.E., Lee, H., Tian, B., and Massoud, E.: Climate model evaluation in the presence of observational uncertainty: precipitation indices over the Contiguous United States, Journal of Hydrometeorology, 2019, 2019.

Gneiting, T., and Raftery, A.E.: Weather forecasting with ensemble methods, Science, 310, 5746, 248-249, 2005.

GRDC: Major River Basins of the World/Global Runoff Data Centre, GRDC, 2nd ed.; Federal Institute of Hydrology (BfG): Koblenz, Germany, 2020.

Hoeting, J. A., Madigan, D., Raftery, A.E., and Volinsky, C.T.: Bayesian model averaging: a tutorial, Statistical Science, 382-401, 1999.

Karl, T.R., Williams, C.N., Young, P.J., and Wendland, W.M.: A Model to Estimate the Time of Observation Bias Associated with Monthly Mean Maximum, Minimum, and Mean Temperatures for the United States, Journal of Climate and Applied Meteorology, 25, 1986.

Knutti, R: The end of model democracy?, Climatic Change, 102, doi:10.1007/s10584-010-9800-2, 2010.

Knutti, R., Sedlacek, J., Sanderson, B.M., Lorenz, R., Fischer, E.M., and Eyring, V.: A climate model weighting scheme accounting for performance and independence, Geophysical Research Letters, 44, DOI:10.1002/2016GL072012, 2017.

Kolosu, S.R., Siderius, C., Todd, M.C., Bhave, A., Conway, D., James, R., Washington, R., Geressu, R., Harou, J.J. and Kashaigili, J.J.: Sensitivity of projected climate impacts to climate model weighting: multi-sector analysis in eastern Africa, Climatic Change, 164, doi: 10.1007/s10584-021-02991-8, 2021.

Kotamarthi, R., Mearns, L., Hayhoe, K., Castro, C.L., and Wuebbles, D.: Use of Climate Information for Decision-Making and Impacts Research: State of Our Understanding, Prepared for the Department of Defense, Strategic Environmental Research and Development Program, 55pp, 2016.

Lee, H., Goodman, A., McGibbney, L., Waliser, D.E., Kim, J., Loikith, P.C., Gibson, P.B., and Massoud, E.C.: Regional Climate Model Evaluation System powered by Apache Open Climate Workbench v1. 3.0: an enabling tool for facilitating regional climate studies, Geoscientific Model Development, 2018.

Livneh, B., Rosenberg, E.A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K.M., Maurer, E.P., and Lettenmaier, D.P.: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions, Journal of Climate, 26, 9384-9392, 2013.

Massoud, E.C., Purdy, A.J., Miro, M.E., and Famiglietti, J.S.: Projecting groundwater storage changes in California's Central Valley, Scientific Reports, 8, 1, 1-9, 2018.

Massoud, E.C., Espinoza, V., Guan, B., Waliser, D.E.: Global Climate Model Ensemble Approaches for Future Projections of Atmospheric Rivers, Earth's Future, doi:10.1029 / 2019EF001249, 2019.

Massoud, E.C., Lee, H., Gibson, P. B., Loikith, P., and Waliser, D.E.: Bayesian model averaging of climate model projections constrained by precipitation observations over the contiguous United States, Journal of Hydrometeorology, 21, 10, 2020, 2401-2418, 2020a.

Massoud, E.C., Massoud, T., Guan, B., Sengupta, A., Espinoza, V., De Luna, M., Raymond, C., and Waliser, D.E.: Atmospheric rivers and precipitation in the middle east and north Africa (Mena), Water, 12, 10, 2863, 2020b.

Olson, R., Fan, Y., and Evans, J.P.: A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures, Geophysical Research Letters, 43, 14, 7661-7669, 2016.

Olson, R., An, S.-I., Fan, Y., and Evans, J.P.: Accounting for skill in trend, variability, and autocorrelation facilitates better multi-model projections: Application to the AMOC and temperature time series, PloS one, 14, 4, e0214535, 2019.

Parding, K.M., Dobler, A., McSweeney, C., Landgren, O.A., Benestad, R., Erlandsen, H. B., Mezghani, A., Gregow, H., Räty, O., and Viktor, E.: GCMeval - An interactive tool for evaluation and selection of climate model ensembles, Climate Services, 18, doi:10.1016/j.cliser.2020.100167, 2020.

Pierce, D.W., Cayan, D.R., and Thrasher, B.L.,: Statistical downscaling using Localized Constructed Analogs (LOCA), J. Hydrometeorology, 15, doi:10.1175/JHM-D-14-0082.1, 2014.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Monthly Weather Review, 133, 5, 1155-1174, 2005.

Rummukainen M.: State-of-the-art with regional climate models, WIREs Climate Change, 1, doi:10.1002/wcc.008, 2010.

Sanderson, B.M., Knutti, R., and Caldwell, P.: Addressing interdependency in a multimodel ensemble by interpolation of model properties, Journal of Climate, 28, 13, 5150-5170, 2015.
Sanderson, B.M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geoscientific Model Development, 2379-2395, doi:10.5194/gmd-2016-285, 2017.
Sanderson, B.M. and Wehner, M.F.,: Model weighting strategy. In: Climate Science Special Report: Fourth National Climate Assessment, Volume I [Wuebbles, D.J., D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, pp. 436-442, doi: 10.7930/J06T0JS3, 2017.
Schoof, J.T.: Statistical downscaling in climatology, Geography Compass, 7, 249-265, 2013.
Shin, Y., Lee, Y., and Park, J.-S.: A Weighting Scheme in A Multi-Model Ensemble for Bias-Corrected Climate Simulation, Atmosphere, 11, doi:10.3390/atmos11080775, 2020.
Skahill B., Berenguer B., Stoll M.: Ensembles for Viticulture Climate Classifications of the Willamette Valley Wine Region, Climate, 9, 9, 140, doi:10.3390/cli9090140, 2021.
Smith, L. and Stern, N.: Uncertainty in science and its role in climate policy, Philosophical Transactions of the Royal Society A, 369,1-24. doi:10.1098/rsta.2011.0149, 2011.
Tapiador, F.J., Roca, R., Genio, A.D., Dewitte, B., Petersen, W., and Zhang, F.: Is Precipitation a Good Metric for Model Performance? Bulletin of the American Meteorological Society**,** 100, 223-233, doi: 10.1175/bams-d-17-0218.1, 2019.
Taylor, A., Gregory, J.M., Webb, M.J., and Taylor, K.E.: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, Geophysical Research Letters, 39, doi:10.1029/2012GL051607, 2012.
USGCRP: Climate Science Special Report: Fourth National Climate Assessment, Volume I [Wuebbles, D.J., D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 470 pp, doi:10.7930/J0J964J6, 2017.
Vrugt, J.A. and Robinson, B.A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resources Research, 43, 1, 2007.
Vrugt, J.A., and Massoud, E.C.: Uncertainty quantification of complex system models: Bayesian Analysis, Handbook of Hydrometeorological Ensemble Forecasting, Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, HL, Schaake, JC, Eds, 2018.
Vrugt, J.A., Cajo, J.F., Ter Braak, M. P.C., Hyman, J.M., and Robinson, B.A. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resources Research, 44, 12, 2008.
Weart, S.: The development of general circulation models of climate, Studies in History and Philosophy of Science Part B - Studies in History and Philosophy of Modern Physics, 41, 208-217. doi:10.1016/j.shpsb.2010.06.002, 2010.
Wootten, A.M., Massoud, E.C., Sengupta, A., Waliser, D.E., and Lee, H.: The Effect of Statistical Downscaling on the Weighting of Multi-Model Ensembles of Precipitation, Climate*,* 8, 12, 138, 2020a.
Wootten, A.M., Dixon, K.W., Adams-Smith, D.J. and McPherson, R.A. Statistically downscaled precipitation sensitivity to gridded observation data and downscaling technique, Int. J. Climatol., doi:10.1002/joc.6716, 2020b.
Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., DeAngelo, B., Doherty, S., Hayhoe, K., Horton, R., Kossin, J.P., Taylor, P.C., Waple, A.M., and Weaver, C.P.: Executive summary. In: Climate Science Special Report: Fourth National Climate Assessment, Volume I [Wuebbles, D.J., D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, pp. 12-34, doi: 10.7930/J0DJ5CTG, 2017.

Figures

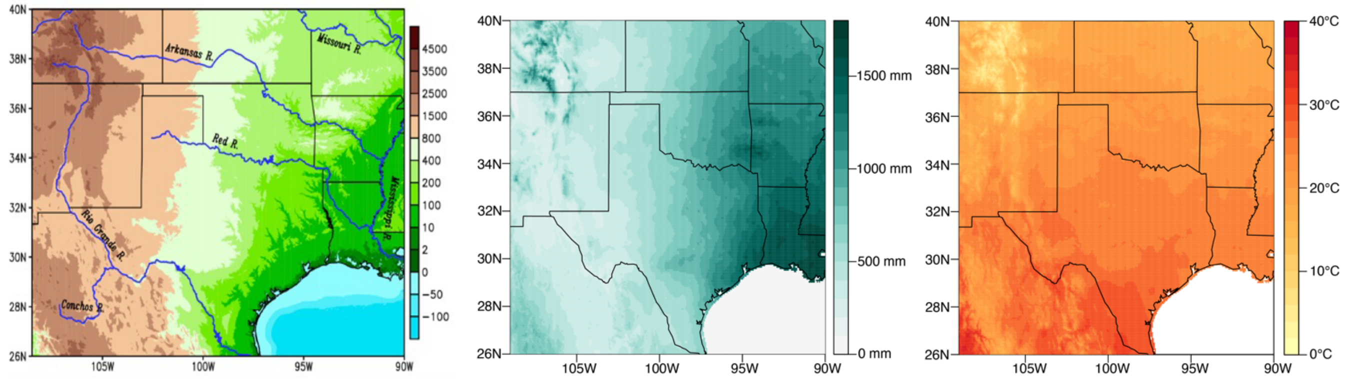


Figure 1: Topographical map for the study domain: The elevation map of the south-central United States with major rivers overlaid on it. Brown/green shading denotes elevation (in units of m), while the rivers are outlined in blue. Topography, bathymetry, and shoreline data are obtained from the National Oceanic and Atmospheric Administration (NOAA) National Geophysical Data Center's ETOPO1 Global Relief Model (Amante and Eakins, 2009). This is a 1 arc-minute model of the Earth's surface developed from diverse global and regional digital datasets and then shifted to a common horizontal and vertical datum. River shapefiles are obtained from the Global Runoff Data Centre's Major River Basins of the World (GRDC 2020). Center — Study domain overlaid with annual average precipitation (mm) from Livneh v. 1.2 (Livneh et al. 2013). Right— Study domain overlaid with annual high temperatures (°C) from Livneh v. 1.2 (Livneh et al. 2013).

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

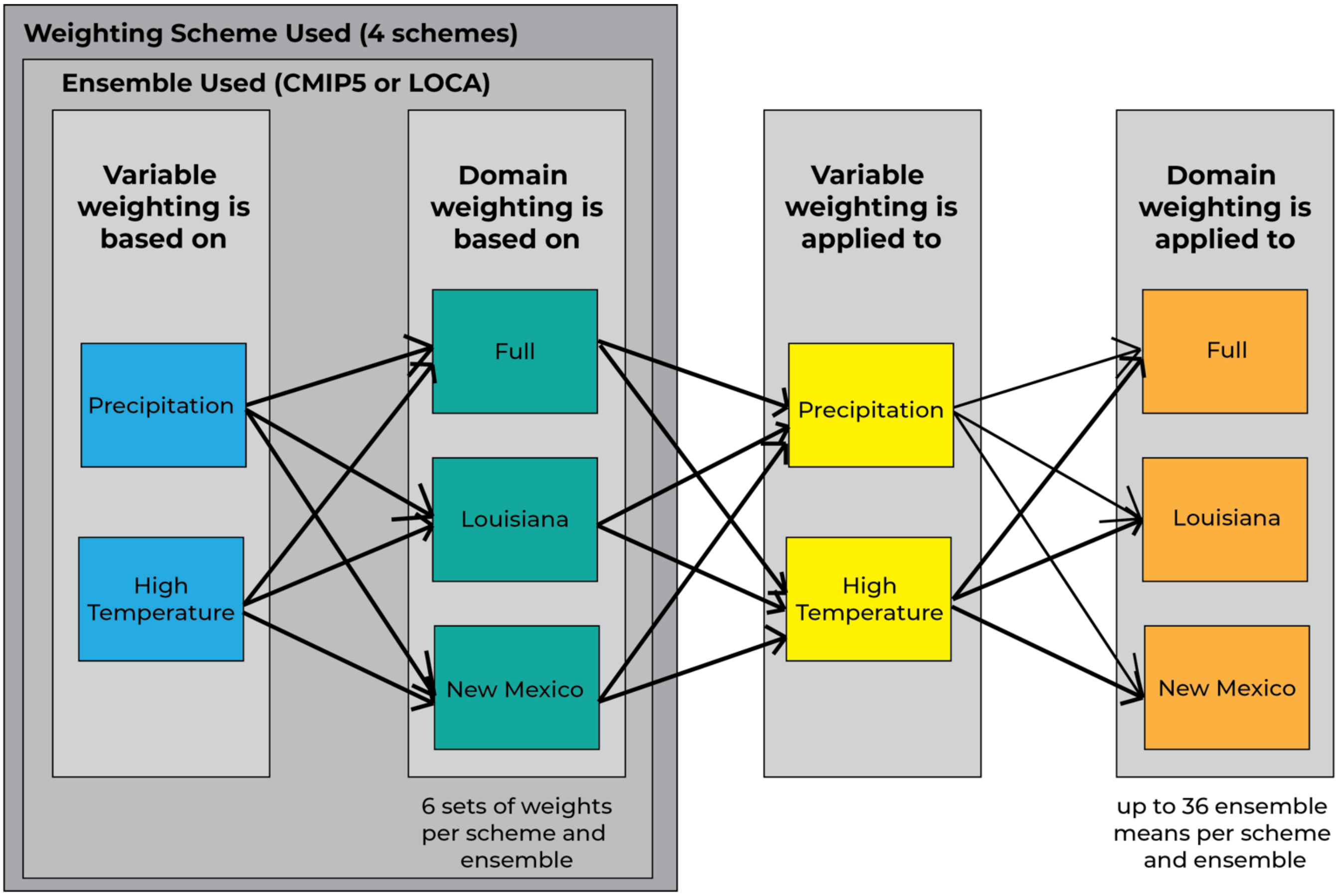

Weighting Scheme Used (4 schemes)
Ensemble Used (CMIP5 or LOCA)
Variable weighting is based on
Precipitation
High Temperature
Domain weighting is based on
Full
Louisiana
New Mexico
6 sets of weights per scheme and ensemble
Variable weighting is applied to
Precipitation
High Temperature
Domain weighting is applied to
Full
Louisiana
New Mexico
up to 36 ensemble means per scheme and ensemble

Figure 2: Flowchart showing the process of analysis with weighting schemes. Each version of the model average is constructed based on several choices: a) the choice of the ensemble (CMIP vs LOCA), b) the choice of model weighting strategy (unweighted, Skill, SI-h, SI-c, or BMA), c) the choice of climate variable of interest (precipitation or temperature), and d) the choice of the domain used for the ensemble averaging (entire south-central region, Louisiana, or New Mexico). These various choices give up to 48, plus the unweighted version, so 49 overall choices of model weighting strategies. Then, once the model average is constructed and trained, there is a choice to be made on which variable and which domain to apply this model average to. Therefore, this results in 48 x 2 x 3 = 288 possible future outcomes in our experimental matrix plus 2 unweighted outcomes, for a total of 290 combinations.

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

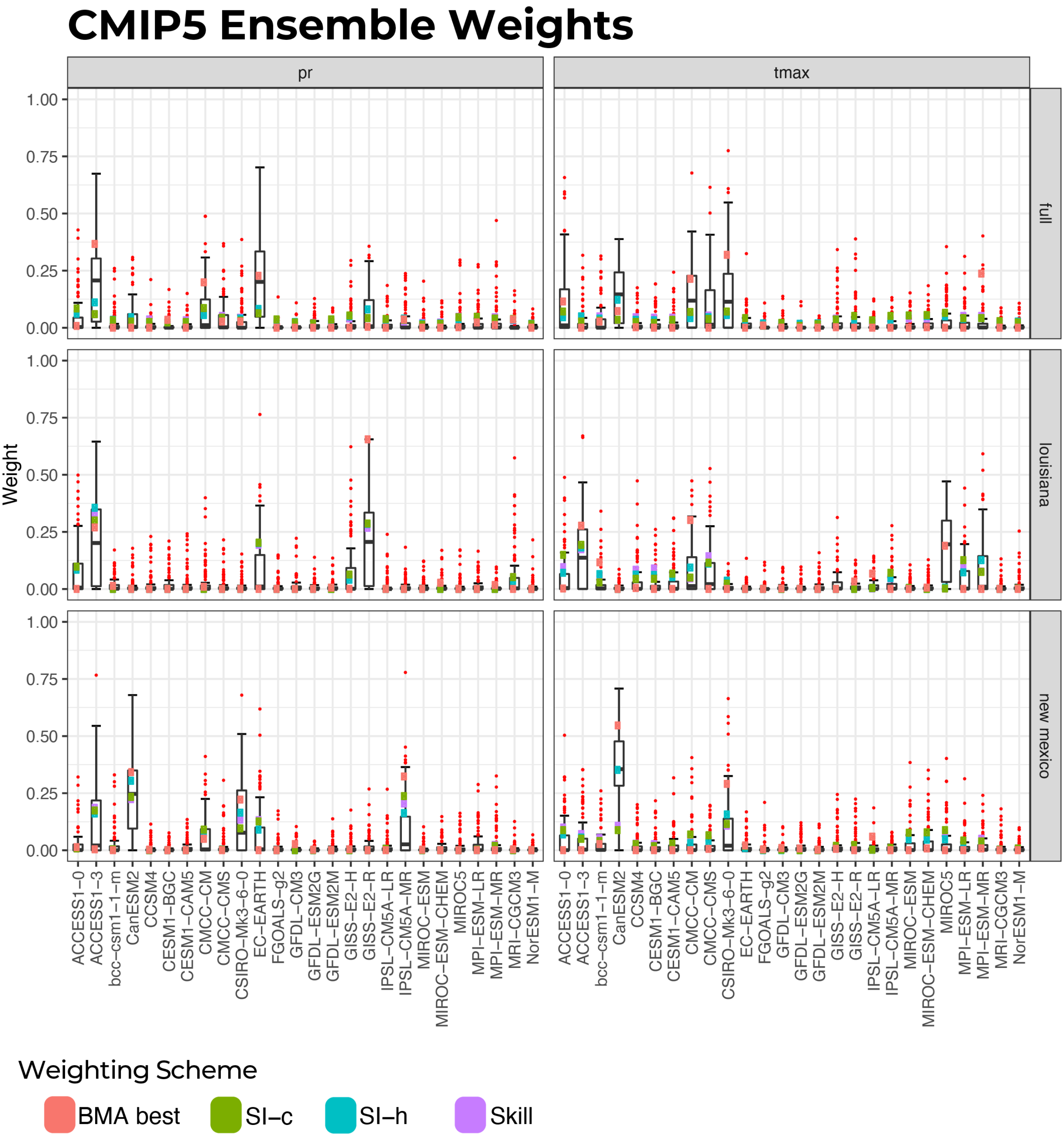


Figure 3: Model Weights for each of the 4 weighting schemes using the CMIP5 ensemble. The left column is weights based on precipitation (pr) alone and the right column is weights based on high temperature (tmax) alone. The top row is weights based on the full domain, the middle row is weights based on Louisiana alone, the bottom row is weights based on New Mexico alone. The boxplots are the spread of weights from the 100 iterations of the BMA weighting scheme. The grey dots in these figures depict the outliers from the BMA distributions of weights.

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

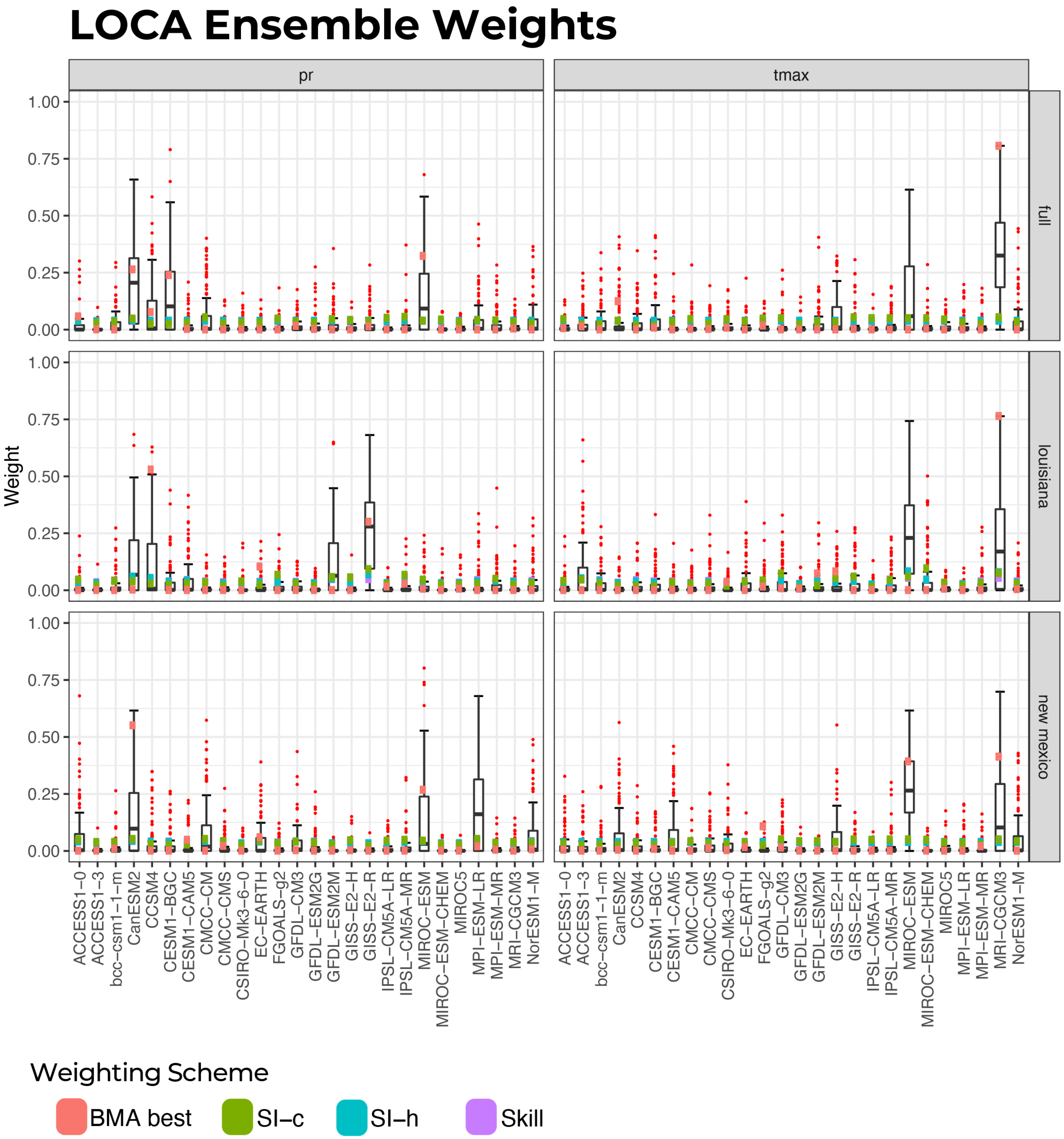


Figure 4: Same a Figure 3, but for the LOCA ensemble.

640

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

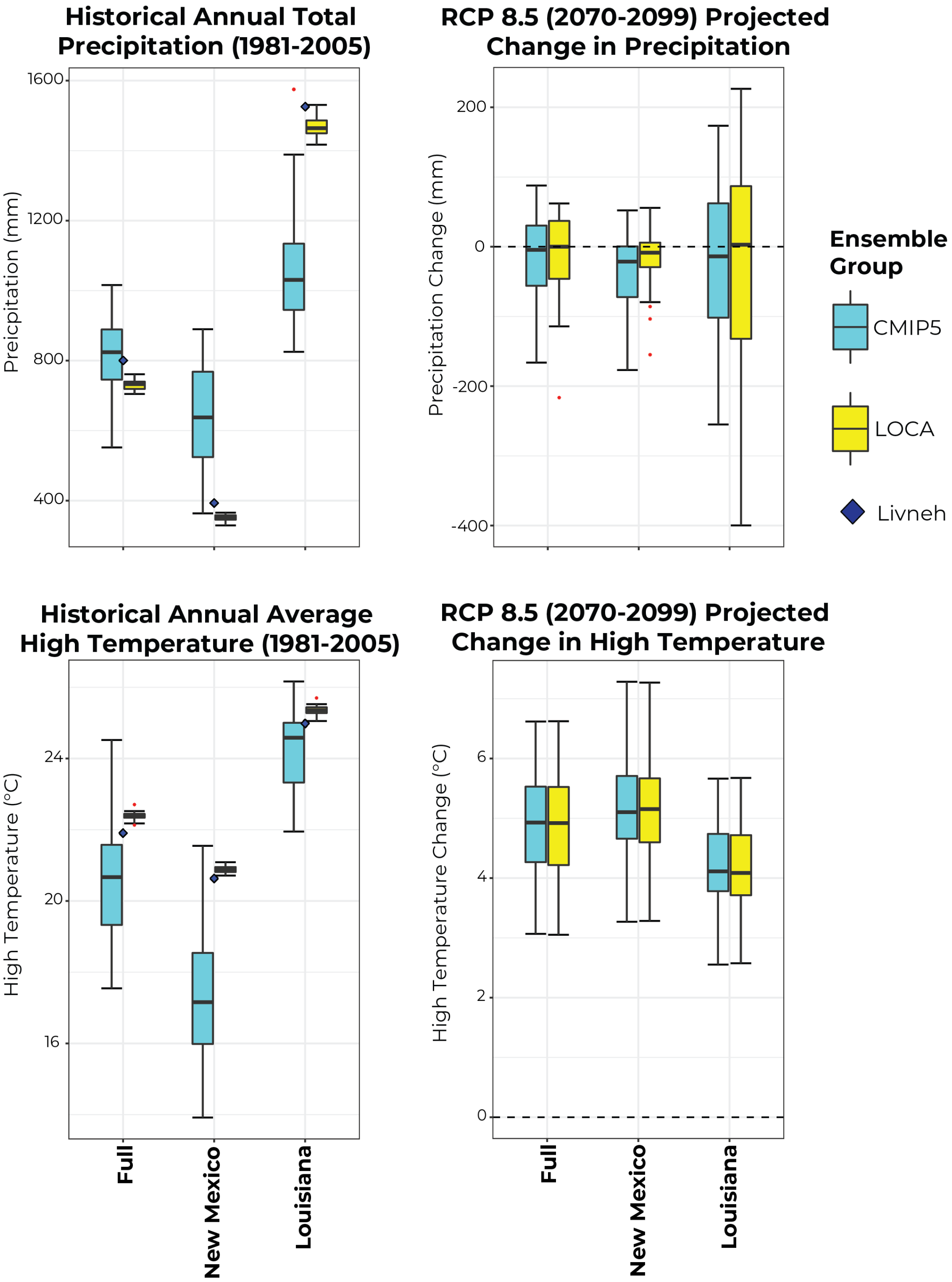


Figure 5: The unweighted model values across each of the three domains. The left column is during the historical period (1981-2005) and the raw ensemble is compared to the same values from the Livneh observations. The right column is the 2070-2099 projected changes under RCP 8.5 from both ensembles. The top row is for precipitation, the bottom row is for high temperature.

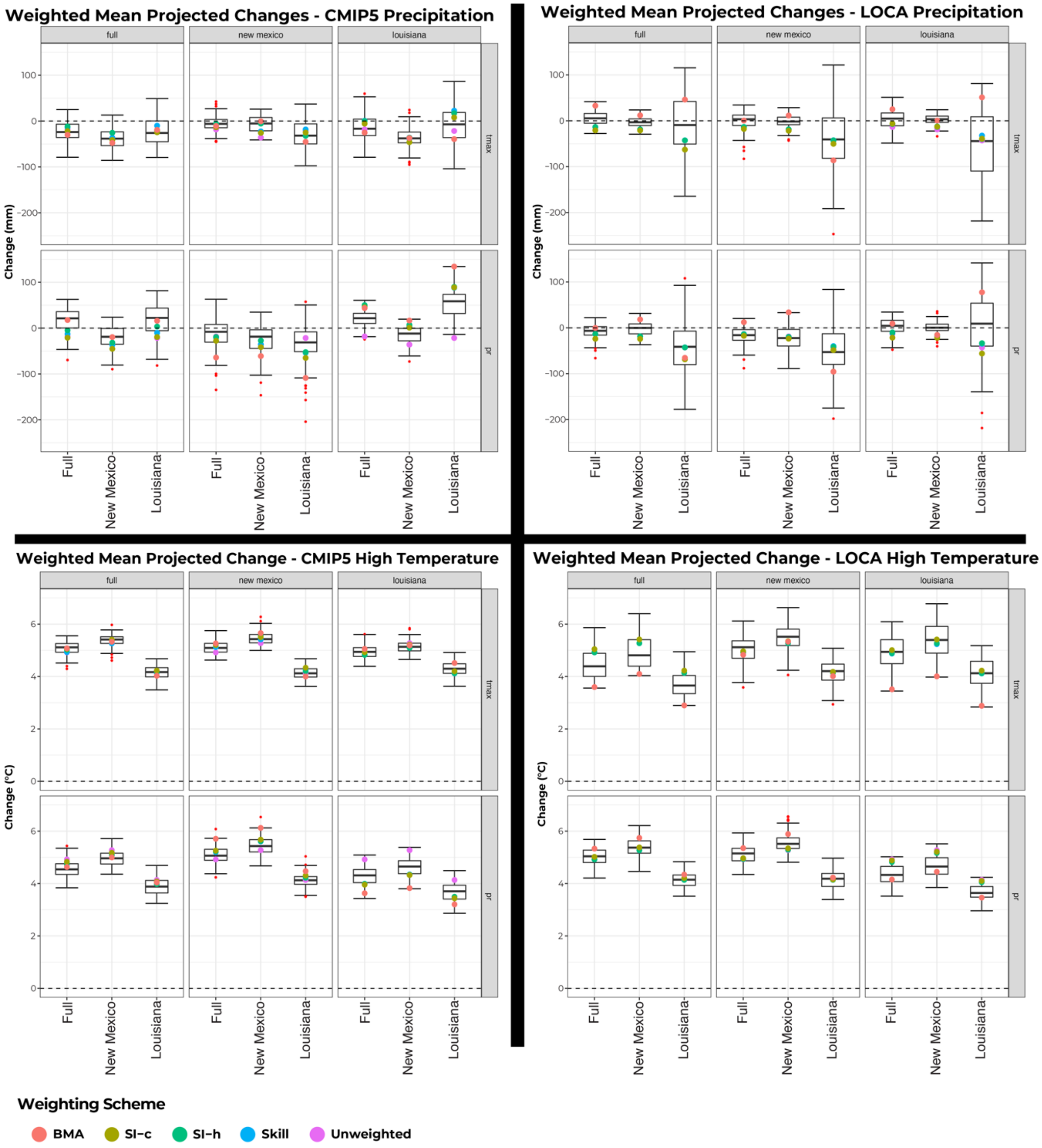
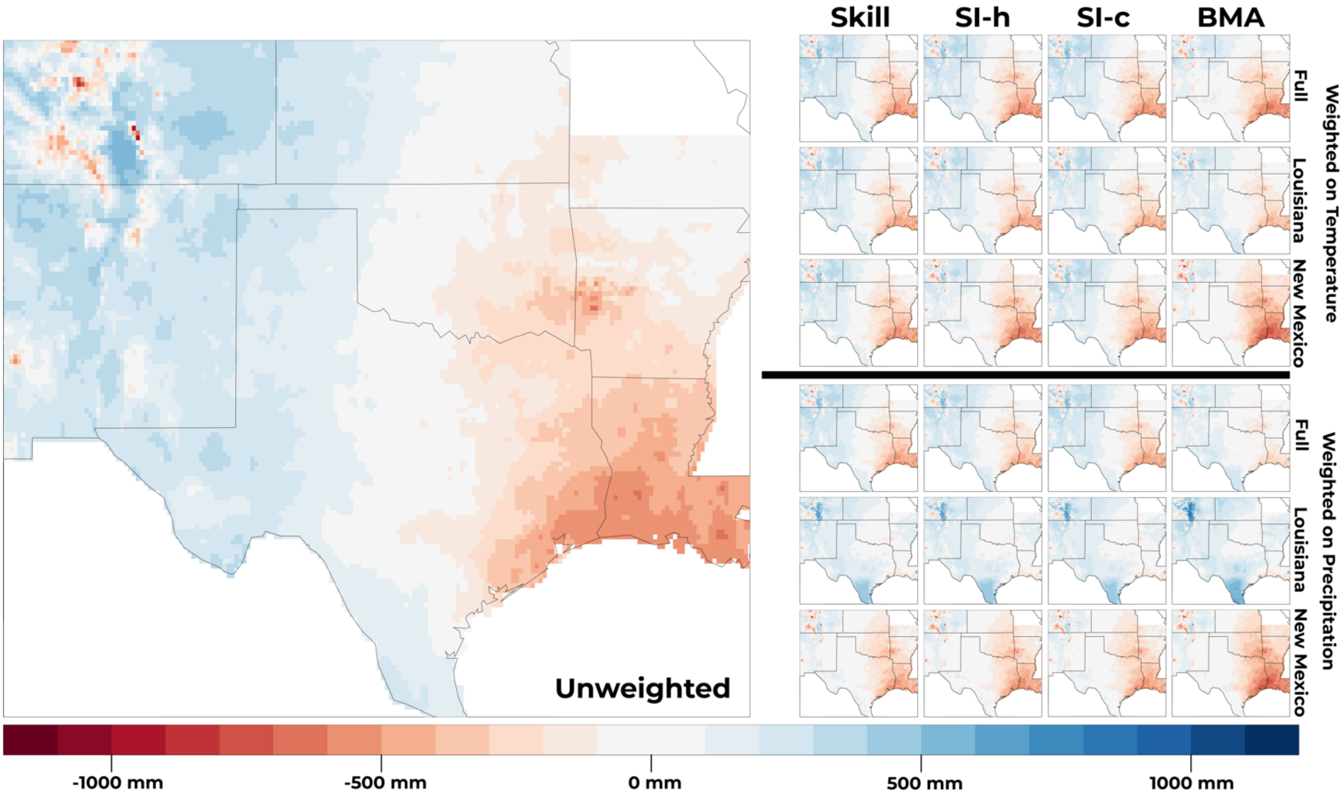
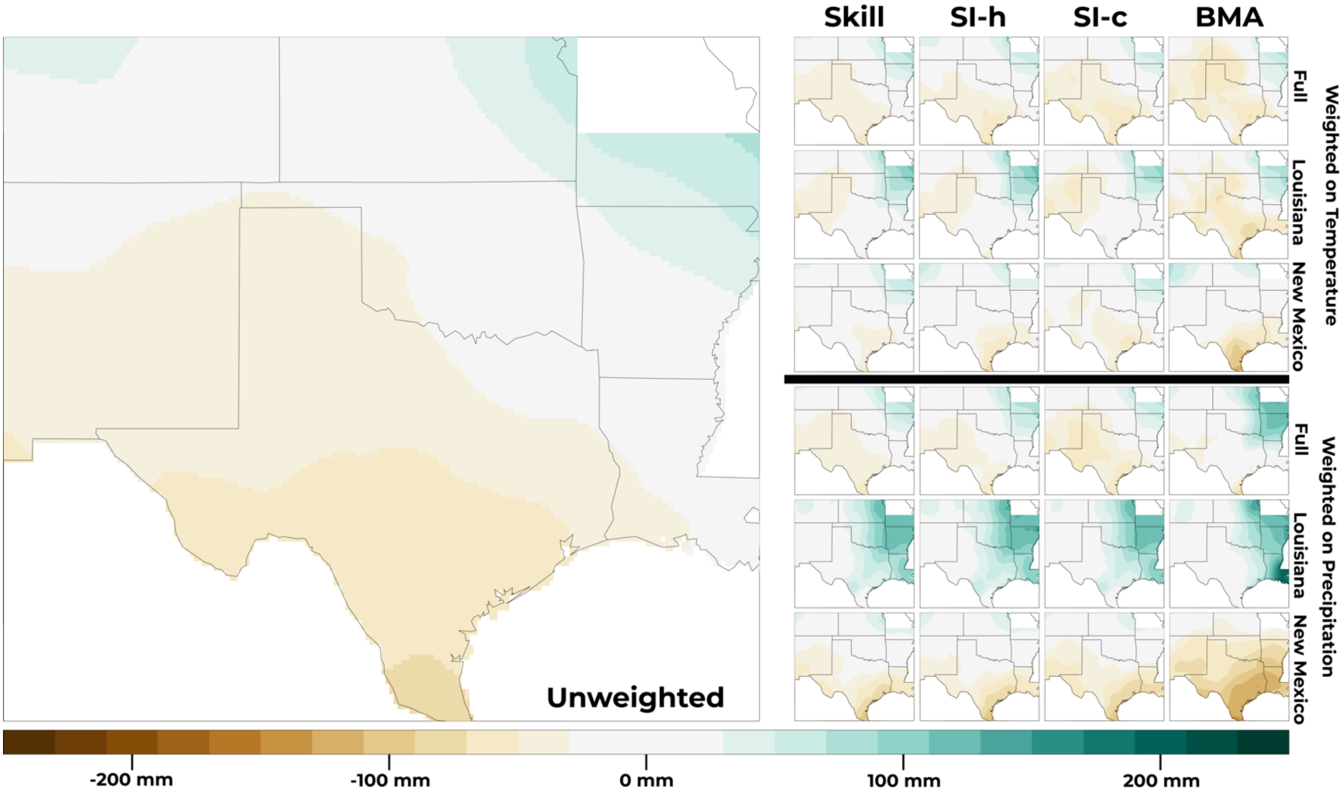


Figure 6: Mean projected changes in temperature and precipitation using all 48 weighting schemes, applied to all three domains and both variables (tmax and pr). The top group focuses on pr, the bottom row focuses on tmax, the left group focuses on the CMIP5 ensemble, and the right group focuses on the LOCA ensemble. In an individual group, the top row is the results from weighting schemes derived with tmax, and the bottom row is the results from weighting schemes derived with pr. In addition, within an individual group, the left column is the results for weighting derived using the full domain, the middle column is the results for weighting derived using the New Mexico domain, and the right column is the results for weighting derived using the Louisiana domain. Within a given domain and variable, the results are shown from left to right for the domain the weights are applied to. The boxplots are the results from the 100 BMA posterior weights.



655 **Figure 7: Bias of CMIP5 ensemble mean precipitation (1981-2005) from the unweighted ensemble (left) and each weighted ensemble mean (right). On the right side, the columns from left to right are for the Skill, SI-h, SI-c, and BMA weighting schemes respectively. On the right side, the top group of twelve plots are the results for weights derived using temperature (tmax) and the bottom group of twelve plots are the results for weights derived using precipitation (pr). Within a group of twelve on the right hand side, the top row is for weights deriving using the full domain, the middle row is for weights derived using the Louisiana**
660 **domain, and the bottom row is for weights derived using the New Mexico domain.**

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022



665 **Figure 8: CMIP5 ensemble mean projected precipitation change (2070-2099, RCP 8.5) from the unweighted ensemble (left) and each weighted ensemble mean (right). On the right side, the columns from left to right are for the Skill, SI-h, SI-c, and BMA weighting schemes respectively. On the right side, the top group of twelve plots are the results for weights derived using temperature (tmax) and the bottom group of twelve plots are the results for weights derived using precipitation (pr). Within a group of twelve on the right hand side, the top row is for weights deriving using the full domain, the middle row is for weights derived using the Louisiana domain, and the bottom row is for weights derived using the New Mexico domain.**

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

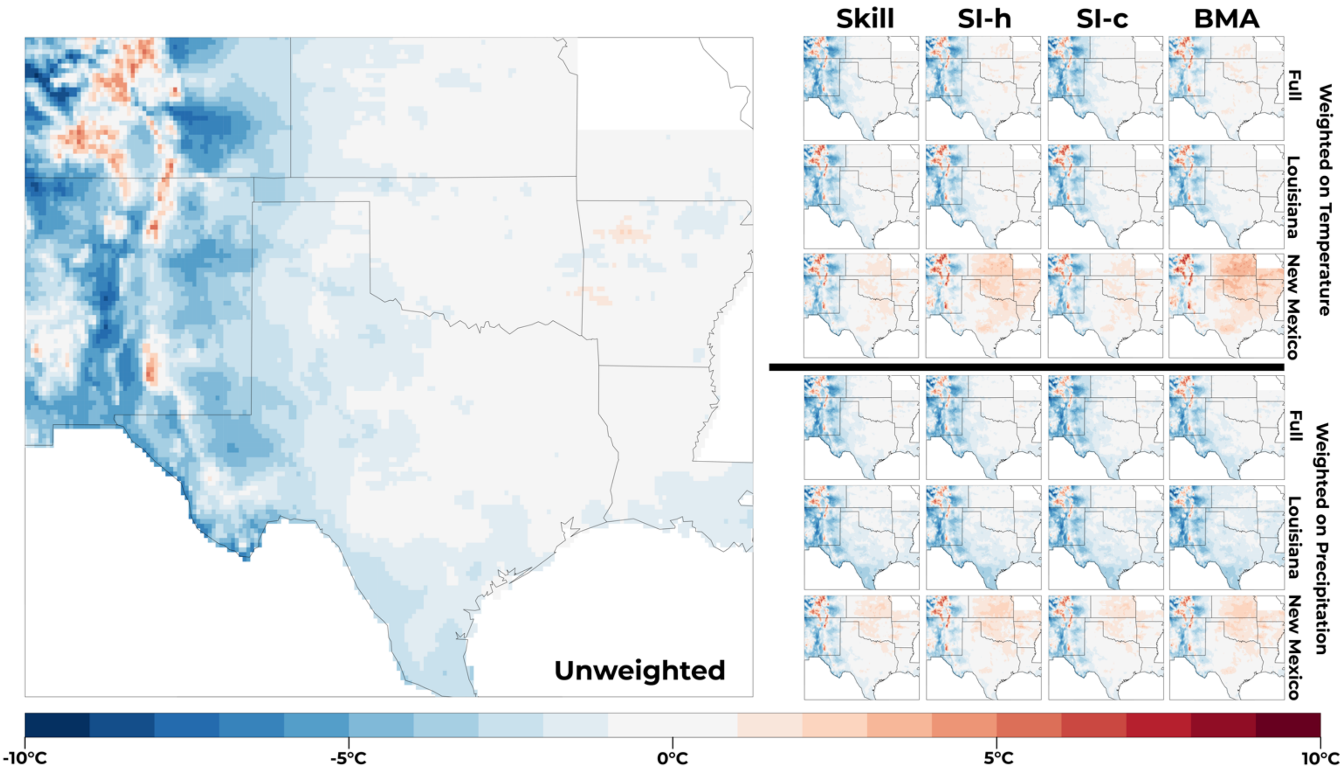


Figure 9: Same as Figure 7, but for the bias of high temperature of the CMIP5 ensemble.

670

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022

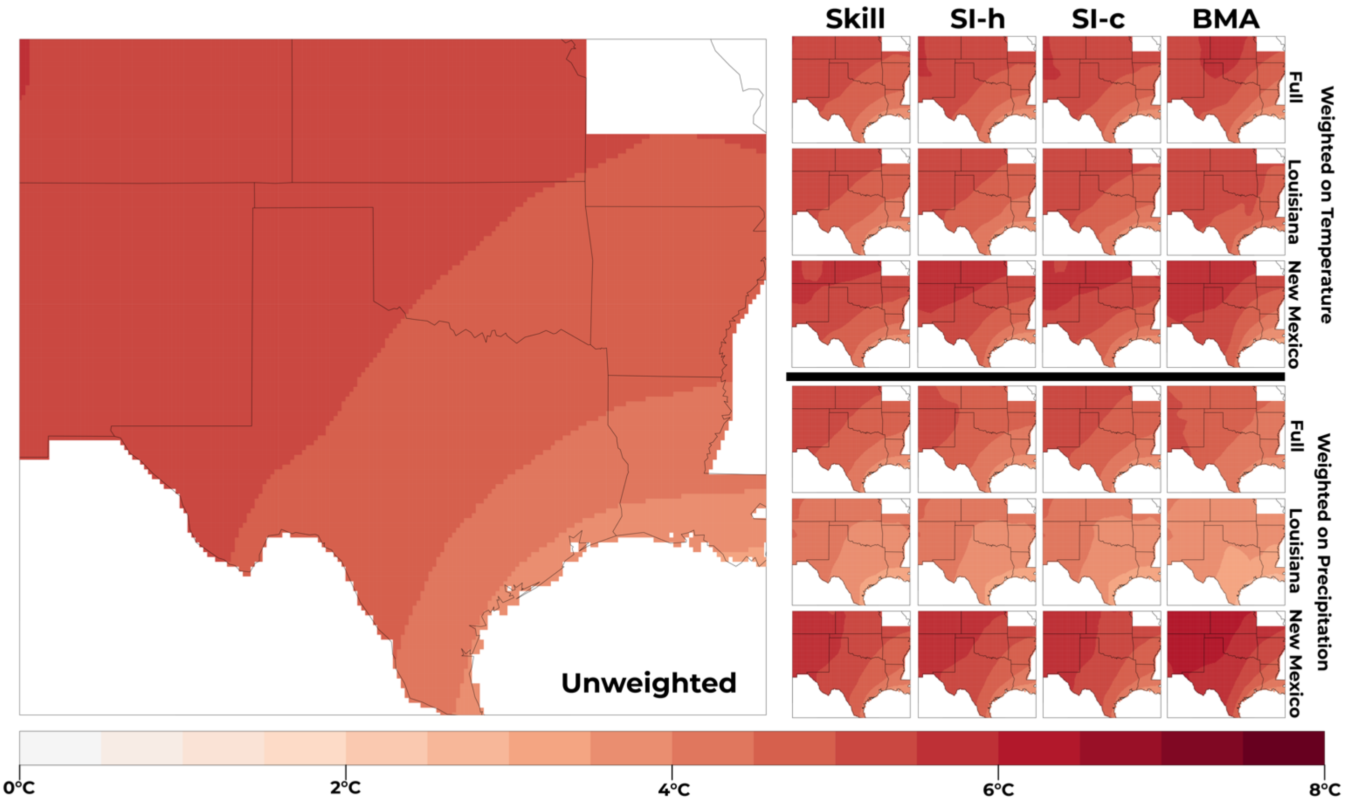
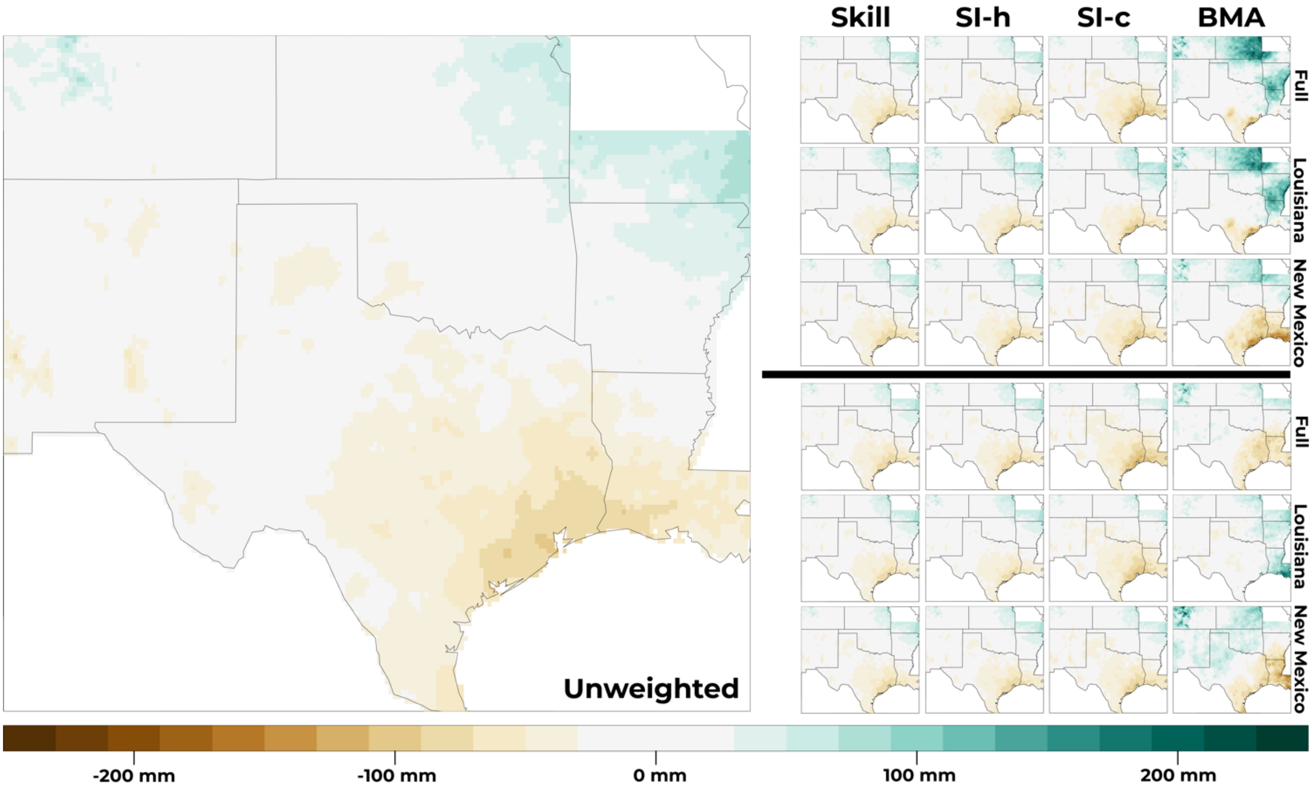


Figure 10: Same as Figure 8, but for the mean projected change of high temperature from the CMIP5 ensemble.

https://doi.org/10.5194/esd-2022-15
Preprint. Discussion started: 22 April 2022



675 **Figure 11: Same as Figure 8, but for the mean projected change of precipitation from the LOCA ensemble.**

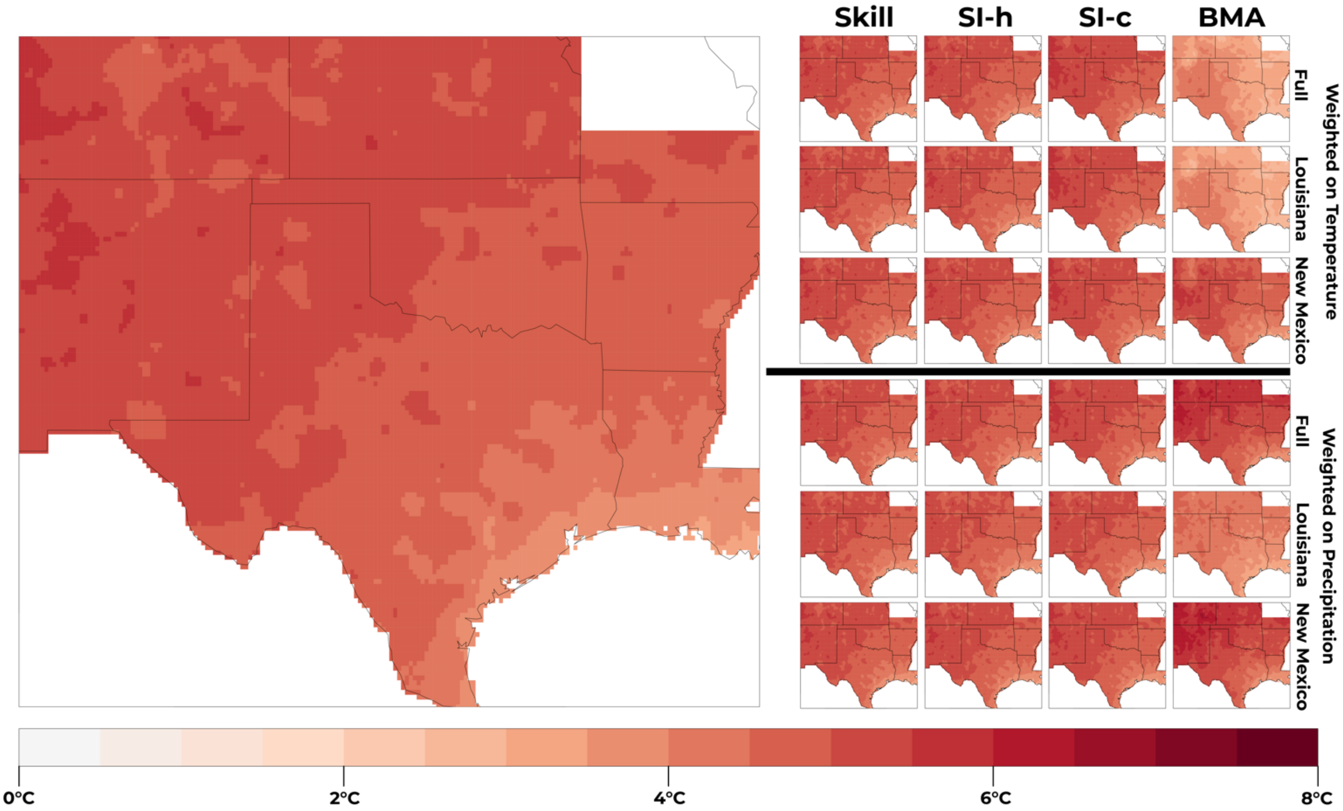


Figure 12: Same as Figure 10, but for the mean projected change of high temperature from the CMIP5 ensemble.

Table 1: Top three highest weighted models from each of the 48 weighting combinations.

Domain Weighting is Based On	Variable Weighting is Based On	Ensemble	Skill	SI-h	SI-c	BMA
Full	**tmax**	CMIP5	ACCESS1-0	CanESM2	CSIRO-Mk3-6-0	CSIRO-Mk3-6-0
			CSIRO-Mk3-6-0	CSIRO-Mk3-6-0	ACCESS1-0	MPI-ESM-MR
			CMCC-CMS	MIROC-ESM	CMCC-CM	CMCC-CM
		LOCA	MRI-CGCM3	MRI-CGCM3	MRI-CGCM3	MRI-CGCM3
			MIROC-ESM	MIROC-ESM	GISS-E2-R	CanESM2
			CESM1-BGC	CESM1-BGC	IPSL-CM5A-MR	FGOALS-g2
	pr	CMIP5	EC-EARTH	ACCESS1-3	CMCC-CM	ACCESS1-3
			CMCC-CM	EC-EARTH	ACCESS1-0	EC-EARTH
			ACCESS1-0	GISS-E2-R	EC-EARTH	CMCC-CM
		LOCA	CESM1-BGC	CanESM2	IPSL-CM5A-MR	MIROC-ESM
			CanESM2	MIROC-ESM	ACCESS1-0	CanESM2
			MIROC-ESM	CESM1-BGC	CMCC-CM	CESM1-BGC
Louisiana	**tmax**	CMIP5	ACCESS1-3	ACCESS1-3	ACCESS1-3	CMCC-CM
			CMCC-CMS	MPI-ESM-MR	ACCESS1-0	ACCESS1-3
			MPI-ESM-LR	CMCC-CMS	MPI-ESM-LR	MIROC5
		LOCA	MRI-CGCM3	MIROC-ESM	MIROC-ESM-CHEM	MRI-CGCM3
			MIROC-ESM	MRI-CGCM3	MRI-CGCM3	GISS-E2-H
			ACCESS1-3	ACCESS1-3	GFDL-CM3	GFDL-ESM2M
	pr	CMIP5	ACCESS1-3	ACCESS1-3	ACCESS1-3	GISS-E2-R
			GISS-E2-R	GISS-E2-R	GISS-E2-R	ACCESS1-3
			EC-EARTH	EC-EARTH	EC-EARTH	MIROC-ESM-CHEM
		LOCA	CCSM4	GISS-E2-R	GISS-E2-R	CCSM4
			GISS-E2-R	CanESM2	IPSL-CM5A-MR	GISS-E2-R
			GFDL-ESM2M	CCSM4	FGOALS-g2	EC-EARTH
New Mexico	**tmax**	CMIP5	CanESM2	CanESM2	CSIRO-Mk3-6-0	CanESM2

			CSIRO-Mk3-6-0	CSIRO-Mk3-6-0	ACCESS1-0	CSIRO-Mk3-6-0
			ACCESS1-0	ACCESS1-0	CanESM2	IPSL-CM5A-LR
		LOCA	MRI-CGCM3	MIROC-ESM	MRI-CGCM3	MRI-CGCM3
			MIROC-ESM	MRI-CGCM3	MIROC-ESM	MIROC-ESM
			GISS-E2-H	CanESM2	GFDL-CM3	FGOALS-g2
	pr	CMIP5	CanESM2	CanESM2	IPSL-CM5A-MR	CanESM2
			IPSL-CM5A-MR	CSIRO-Mk3-6-0	CanESM2	IPSL-CM5A-MR
			ACCESS1-3	IPSL-CM5A-MR	ACCESS1-3	CSIRO-Mk3-6-0
		LOCA	MPI-ESM-LR	MPI-ESM-LR	CanESM2	CanESM2
			CanESM2	CanESM2	MPI-ESM-LR	MIROC-ESM
			MIROC-ESM	MIROC-ESM	CMCC-CM	EC-EARTH