Assessment of the Benefits of Climate Model Weights for Ensemble Analysis in Three Urban Precipitation Frequency Studies

Kevin Grady, Momcilo Markus, Shu Wu, Fuyao Wang, Seid Koric

Prairie Research Institute (Grady, Markus), University of Illinois at Urbana-Champaign, Champaign, Illinois, USA; Nelson Institute for Environmental Studies, (Wu, Wang), University of Wisconsin–Madison, Madison, Wisconsin, USA; National Center for Supercomputing Applications (Koric), University of Illinois at Urbana-Champaign, Champaign, Illinois, USA (Correspondence to Markus: mmarkus@illinois.edu).

KEYWORDS: projected frequency, precipitation frequency, climate models, weighted ensemble, climate change adaptation, urban hydrology, urban adaptation

RESEARCH IMPACT STATEMENT: The hypothesis that the use of weights in climate ensemble studies provides more accurate PF was largely confirmed by the experimental results making the case for continuing to investigate this issue.

ABSTRACT

In hydrology, projected climate change impact assessment studies typically rely on ensembles of downscaled climate model outputs. Due to large modeling uncertainties, the ensembles are often averaged to provide a basis for studying the effects of climate change. A key issue when analyzing averages of a climate model ensemble is whether to weight all models in the ensemble equally, often referred to as the equal-weights or unweighted approach, or to use a weighted approach, where, in general, each model would have a different weight. Many studies have advocated for the latter, based on the assumption that models that are better at simulating

the past, i.e., the models with higher hindcast accuracy, will give more accurate forecasts for the future and thus should receive higher weights. To examine this issue, observed and modeled daily precipitation frequency (PF) estimates for three urban areas in the United States, namely Boston, Massachusetts; Houston, Texas; and Chicago, Illinois, were analyzed. The comparison used the raw output of 24 Coupled Model Intercomparison Project Phase 5 (CMIP5) models. The PFs from these models were compared with the observed PFs for a specific historical training period to determine model weights for each area. The unweighted and weighted averaged model PFs from a more recent testing period were then compared with their corresponding observed PFs to determine if weights improved the estimates. These comparisons indeed showed that the weighted averages were closer to the observed values than the unweighted averages in nearly all cases. The study also demonstrated how weights can help reduce model spread in future climate projections by comparing the unweighted and weighted ensemble standard deviations in these projections. In all studied scenarios, the weights actually reduced the standard deviations compared to the equal-weights approach. Finally, an analysis of the results' sensitivity to the areal reduction factor used to allow comparisons between point station measurements and grid-box averages is provided.

INTRODUCTION

In many cities in the United States, heavy storms have become more frequent and stronger than those used to design the existing urban drainage infrastructure, causing more frequent floods. Moreover, climate studies suggest that heavy storms may become even more frequent and intense in the future (Douglas and Fairbank, 2011; Markus *et al.*, 2012; Wuebbles *et al.*, 2017; Um *et al.*, 2017; Um *et al.*, 2018; Li *et al.*, 2019). Large metropolitan areas are

particularly vulnerable to climate change because of the complex interaction between climate and urban environments (Zhang *et al.*, 2018). To address the problem effectively, ensembles of model-generated data are often used to simulate the variability of modeling outputs for future scenarios and time horizons. When analyzing climate model ensembles, projections from multiple climate models are often aggregated into simple averages (USGCRP, 2018). A key issue, however, is whether to weight all models in an ensemble equally (unweighted approach), an approach often referred to as model democracy (Knutti, 2010), or to use a weighted approach, i.e., to assign different weights to the models when finding these averages. It can be assumed that since all models exhibit differences from each other, such as dynamic cores, parametrization, and model resolutions, some perform better in certain applications than others. Thus, a scheme in which these better performing models are given higher weights than the underperforming ones in theory would be better at predicting future climate than an unweighted average. The recent Fourth National Climate Assessment (NCA4: Wuebbles *et al.*, 2017) addresses model weighting and recommends using a method to determine weights based on model performance (Sanderson and Wehner, 2017). Although in the NCA4 report it is assumed that the models that are better at predicting past climate will also perform better with future climates, complexities and uncertainties of weighted approaches also need to be considered (Sanderson et al. 2015). Since future performances are not known, hindcast accuracies are often used instead as a proxy for model fitness when determining weights. Thus, the models with better hindcast accuracies will receive higher weights in ensembles used for future projections.

Many studies have advocated for using a weighted model ensemble approach. Sanchez *et al.* (2009) and Räisänen and Ylhäisi (2012) both found that using weights improved their results. Masson and Knutti (2011) argued that often not all models in an ensemble are independent.

Some models share a common underlying structure, but with slightly different parameters or resolutions, especially if the models come from the same institution. Model democracy does not account for these dependencies and can risk giving too much emphasis to a particular underlying model structure. Knutti *et al.* (2017) also argued that model democracy can allow poorer models to introduce biases and found that a weighting scheme based on both model performance and independence improved their results. Likewise, the NCA4 also used model skill and independence to determine their weights (Sanderson and Wehner, 2017; USGCRP, 2017). They found that weighting did not strongly influence mean projections, but they still recommended its use to guard against highly replicated but poorly performing models.

At the same time, there are some opposing viewpoints. Wiegel *et al.* (2010) pointed out that in order to use weights, accurate knowledge of each individual model's skill is required, and unpredictable model noise must also be considered. If these uncertainties are not fully taken into account, weights could actually do more harm than good by making the ensemble perform more poorly. Wiegel *et al.* were concerned that this is a real possibility since there is no universal, objective consensus on how to find weights, so they suggested using equal weights to be safe. Christensen *et al.* (2010) experimented with how to determine model weights but in the end suggested that using weights was not beneficial. They argued that the subjective nature of determining weights and the associated uncertainties led to even more uncertainties during the weighting process itself. Hagedorn *et al.* (2005) argue that robust optimal weights are difficult to calculate given the short samples available to train the model.

The purpose of this study is to contribute to this debate by determining the benefits of weights by designing an experiment in which the observed data were divided into training (1961–2000 or 2005) and testing (2006–2018) data sets (Figure 1). The training period was

selected to match the period of observed record used in climate model development, ending in the year 2000 or 2005, depending on the city. The testing period used the observed data not used in climate model development, starting with 2006 and ending with 2018, the last year of the observed record at the time of this study. The selection of the training and testing periods is related to the CMIP5 models settings, where the observed values for the external forcings such as concentrations of greenhouse gasses are used to force the climate models for the period of 1950-2005 (referred to as the training period), while the projected values of external forcings are used to force the models for period of 2006-2100. It is the projection portion of the CMIP5 model output on which the weighted approach needs to be evaluated. Even though the testing period selected for the evaluation of weights is relatively short (2006-2018), the values used should be stable due to averaging large ensembles. Therefore, we leverage the latest observations to give a first test on CMIP5 model projections. This experiment allowed a cross-validation methodology to assess the performance of the weights determined in the training stage by applying them in the testing stage. This paper tests future projections based on a supervised method, i.e., when the projections are actually known. As more observations become available, we will be able to use longer periods for future studies.

In the first step, weights were determined by comparing daily precipitation frequency (PF) estimates based on different climate models with those based on the training set of the observed data (Figure 1). Models with PF estimates closer to the observed ones received higher weights. These weights were then applied to the testing dataset to determine if they provide a more accurate approximation of the observed PF. If the weighted ensemble means were closer to the observed PF than the equal-weights ensemble mean, the weighted approach would be considered beneficial. Additionally, the approach with a smaller standard deviation of the model

results was considered advantageous on account of the smaller variability in the model results. It was hypothesized that the addition of weights will provide a more accurate average PF and smaller model variability.

Figure 1. Schematic of the method to determine the benefits of using weighted averages.

The climate modeling data we used are based on the Climate Model Intercomparison Project Phase 5 (CMIP5) raw model output. Specifically, we employed 24 CMIP5 models to study three small climatically relatively homogeneous urban areas. These models were the same ones used to create the University of Wisconsin Probabilistic Downscaling (UWPD) dataset (Notaro *et al*., 2014; Wu *et al*., 2019). To compare the model-generated grid-based precipitation data with those based on point observations, similar to Markus *et al*. (2018), the model data were multiplied by an inverse areal reduction factor based on the curves in Hershfield (1961).

Inputs into our statistical models are the outputs from many climate models. The computations used high-throughput capabilities of the peta-scale Blue Waters high performance computing (HPC) system at the National Center for Supercomputing Applications (NCSA), where we were able to send each of these computational tasks to a different processing unit of Blue Waters. This extremely efficient computational workflow enabled us to experiment with different climate models and scenarios while tuning the weights of models and comparing numerous simulations.

DATA

For this study, we focused on the areas around three large urban centers in the United States: Boston, MA, Houston, TX, and Chicago, IL. All three are adjacent to relatively large bodies of water (the Atlantic Ocean, the Gulf of Mexico, and Lake Michigan, respectively). Quasi-rectangular areas surrounding each city (Figure 2) were chosen for this study. These areas were selected to be small enough so that the climate would be relatively homogeneous across the entire area, yet big enough to include at least 10 weather stations, to minimize the effects of potential outliers. The areas selected for Boston and Houston are contained within one climate region each as defined by National Oceanic and Atmospheric Administration (NOAA) Atlas 14. The definition of the Chicago area is the same as in Markus *et al*. (2018). Table 1 lists the coordinates and other relevant data for each of these targeted areas.

The training period selected for Boston and Houston was 1960–2005 (46 years), as that was the historical period for the available data. Only those stations within the study target areas with at least 80% observed coverage during this period were selected (Figure 2) (Wu *et al*., 2019), with data from NOAA used in the development of Atlas 14. For Chicago, the selected stations were a subset of those used in Markus *et al*. (2017) from the Global Historical Climatology Network Daily (GHCND). Markus *et al.* (2017) looked at observations from 1961 to 2000 (40 years). That same period was selected here as the training period to take advantage of the already prepared observational datasets from that study. Station data for Chicago were manually checked, and only those with at least 80% completeness for 1961–1980 and at least 70% completeness for 1981–2000 were retained. As a result, 15 of the 30 stations were selected (Figure 2).

Following Markus *et al.* (2017) and the NCA4 (USGCRP, 2017), we employed model data for this period from two Representative Concentration Pathways (RCPs): RCP4.5 and RCP8.5. RCP4.5 represented a low-end emission scenario, while RCP8.5 represented a high-end one, as recommended in the NCA4. While the differences between the two scenarios might not be large for the selected testing period, these scenarios were selected because they have different numbers of available climate models and some differences in the results are expected. Observed data for this period were obtained using the Midwestern Regional Climate Center's Application Tools Environment (cli-MATE) (https://mrcc.purdue.edu/CLIMATE/welcome.jsp, last accessed 10/31/2019). However, not all stations used during the training periods had data readily available for the testing period on cli-MATE, sometimes because observations at a station ended before or during the testing period. Thus, the stations with partial or incomplete records were removed, so some stations ended up being used for training but not for testing (as shown in Figure 2).

For all three cities, the testing period was selected to be 2006-2018 (13 years). It should be noted that the testing period analysis for Houston was performed twice, once with 2017 and once without 2017 in both the models and observations. This is because of Hurricane Harvey (Emanuel, 2017), which affected the Houston area in late August 2017 and produced extremely high daily rainfall amounts well over 10 inches for many stations. Including such an extreme event in such a short time interval strongly influences the observed PFs, especially for longer return periods. This influence is not seen in the climate model data, leading to large disparities between the observed and modeled PFs. The results will show the effects of the year 2017.

Table 1. The regions and time periods selected for this study.

Figure 2. Location of selected training period stations for Boston (top), Houston (center), and Chicago (bottom).

The modeled data used were the raw CMIP5 data, consisting of the same 24 models that were used to create the UWPD dataset (Table 2). These models have different resolutions, with some having only a couple of grid points in each region. The closest model grid point was selected for comparison with each station for each model; however, these grid points were not always in the observed regions in Table 1. Thus, slightly larger regions (the second row in Table 1) were used for the models to ensure that the closest model grid point to each station in the observed region was included in the analysis. Finally, two models (CMCC-CESM and MRI-ESM1) did not have RCP4.5 data available, so two sets of weights for each city needed to be found: one without those models for RCP4.5 and one with them for RCP8.5. This is necessary as the weights for each scenario should average to 1, so removing models may affect the weights of the remaining ones.

For all grid cells, due to their large (greater than 1,000 square kilometers) and highly variable size (Table 2), a constant reduction factor of 0.90 was assumed based on Hershfield (1961) and the Technical Paper 29 (U.S. Weather Bureau, 1957), producing an inverse areal reduction factor (ARF) equal to 1.11. However, several later studies (Sivapalan and Blöschl, 1998; Allen and DeGaetano, 2005) indicated that the ARFs in Hershfield (1961) were too high. The reduction factors in the Technical Paper 29 were even higher, probably due to the assumption of constant ARF for large areas. To accommodate the recommendations from Sivapalan and Blöschl (1998) and Allen and DeGaetano (2005), two other lower ARF values, 0.80 and 0.67, were also calculated to assess the sensitivity of results based on the uncertain areal reduction factors resulting from variable grid-cell size for different models.

Table 2. The 24 CMIP5 models used in this study along with their grid-cell sizes (latitude × longitude). No RCP4.5 data were available for models with an (*).

METHODOLOGY

The initial step in the methodology of this study (Figure 3) is adopted based on NOAA Atlas 14 (Perica et al. 2018; 2019), where PFs were estimated based on L-moments (Hosking and Wallis, 1997) applied to the annual maximum series (AMS) and corrected by the Langbein (1949) formula. The AMS for a location is the series of the single largest 24-hour precipitation values for each year (the annual maximum) in the period of interest at that location. In the training period for each city, for each weather station, the observed data were used to find the AMS. Likewise, the AMS (corrected by inverse ARF) were found at each of the model grid points in the modeled region from the daily model-generated precipitation data for each of the models in Table 2. From these AMS, PFs were found at each location using L-moments to fit the data to generalized extreme value (GEV) distributions. The return periods used were 2, 5, 10, 25, 50, and 100 years, which, after applying Langbein's (1949) formula for AMS, became 2.54, 5.52, 10.51, 25.50, 50.50, and 100 years, respectively. The model weights were then determined based on the similarity between the observed PFs at each station and the modeled PFs at the closest model grid point to each station for each model. Similar to Markus et al. (2018), a percent difference was found at each station for each model/return period combination:

$$\%\ Difference = \frac{(Model\ PF - Obs\ PF)}{Obs\ PF}. \quad (1)$$

Once these differences were found, they were averaged for each model across all stations within a city domain and return periods, resulting in a single value, d_i, for each of the 24 models.

The model weights can be derived from this set of d_i, but here we deviated a bit from Markus *et al*. (2018) by using a variation of Tukey's (1977) formula:

$$w_i = \begin{cases} (1 - \left|\frac{d_i}{h}\right|^3)^3, if\ |d_i| \le h \\ 0, \qquad\qquad if\ |d_i| > h \end{cases} \tag{2}$$

Whereas h is commonly the standard deviation of the set of d_i, here we instead took it to be their standard deviation from 0, not their mean:

$$h = \sqrt{\frac{\sum_i {d_i}^2}{n}} \tag{3}$$

where n is the number of models used (here 24 for RCP8.5, 22 for RCP4.5). This was done because most of the d_i for each of the cities were large and negative, meaning that many of their absolute values were greater than their standard deviation and, thus, the weights for those models would be 0. This is different from the more typical case for Tukey's formula where the values are more evenly distributed around 0. After finding the weights for each model using Eq. 2, they were then normalized so that their mean was 1.

Next, observed and modeled PFs for the testing period were found in the same way as in the training period. For each station and return period combination, the PFs of the closest grid point for each model were averaged across all 24 models (or 22 for RCP4.5), once using the weights found in the training period and once using equal weights. These averages could then be compared with their corresponding observed PFs to see whether using weights made a significant improvement overall over using equal weights.

Figure 3. Steps in determining the weights (training) and evaluating the methodology (testing)

It was also necessary to investigate both the weighted and unweighted standard deviations of the model PFs for a period. To do so we find the PFs for that period and emission scenario at the grid points within the modeled regions in Table 1 for the appropriate city, and then average them spatially. The result for each city–time–emission scenario–return period combination is one average PF for each model, x_i. The unweighted standard deviation of these 24 (22 for RCP4.5) x_i, σ_u, is found in the typical way. From Rimoldini (2014), the equation for the weighted standard deviation of the x_i is

$$\sigma_w = \sqrt{\frac{V_1^2}{V_1^2 - V_2} k_2} \tag{4}$$

where

$$V_p = \sum_{i=1}^{n} w_i^p, \quad p = 1, 2; \qquad k_2 = \frac{1}{V_1} \sum_{i=1}^{n} w_i (x_i - \mu)^2; \qquad \mu = \frac{1}{V_1} \sum_{i=1}^{n} w_i x_i.$$

Like before, w_i are the weights, and n is the number of models.

RESULTS

Table 3 lists the weights found for each model for each city for ARF=0.90, ARF=0.80, and ARF=0.67, respectively. The presented weights were averaged for the two climate scenarios (RCP8.5 and RCP4.5), as the difference between those scenarios on model weights was found to be insignificant. For the two models with only RCP8.5 data (CMCC-CESM and MRI-ESM1),

the weight for the RCP8.5 scenario is presented instead. Also, because the two scenarios are averaged together, the weights in a column of the table may not necessarily average to exactly 1. All models generally showed a relatively high degree of consistency across the three regions and for all three ARF values. Models showing high (greater than 1.5) average weights (Figure 4) for the three sites, three ARF values, and two climate scenarios, included ACCESS1-0, ACCESS1-3, CMCC-CM, HadGEM2-CC, IPSL-CM5A-MR, IPSL-CM5B-LR, and MRI-ESM1. On the other hand, several models had zero weights in all cases (e.g., CMCC-CESM, inmcm4, MIROC-ESM, MIROC-ESM-CHEM and NorESM1-M), possibly explained by the large spread of the d_i and the large dry biases of some of the models.

Table 4 shows the weighted and unweighted percent differences (Eq. 1) for each ARF averaged across all return periods and stations for Boston, Houston with the complete record, Houston without the hurricane year (2017), and Chicago. All results in Table 4 are presented for two assumed climate scenarios (RCP4.5 and RCP8.5) and applied to the testing period (2006–2018). The negative values in Table 4 indicate that the models on average underestimate the observed frequency estimates (dry bias). Small absolute values of the differences indicate that the models are accurate and vice versa. In 23 of the 24 cases in Table 4 (except for Boston for RCP4.5 and ARF=0.80, shown in bold numbers), the percent differences for the weighted approach were smaller in absolute value than those of the equal-weights approach, indicating that the weighted approach produced more accurate results than that with equal weights. Houston and Chicago have dry biases in all cases, while that is not the case for Boston, where the signs of the errors are positive in some cases. The weighted percent differences for the testing period averaged for all cases, as shown in the bottom row in Table 4, are about a half of the equal-weights case for both scenarios, highlighting the benefits of weights.

Table 3. Climate model weights averaged for climate scenarios RCP4.5 and RCP 8.5 for ARF=0.90, 0.80, and 0.67.

Figure 4. Weights for each climate model averaged for the three geographic locations, three ARFs, and two climate scenarios.

Table 4. Weighted and unweighted percent differences of PFs (Eq. 1) averaged across all return periods and stations in the testing period (2006–2018) for ARF=0.90, 0.80, and 0.67.

In addition to the steps shown in the flowchart (Figure 3), another testing of the method has been performed to assess the effects of weights by comparing weighted and unweighted ensemble spreads. The weights were found to reduce spread in the training stage because the weighting scheme removes very inaccurate models (outliers). As a result of removing outliers, the weighted standard deviation in the training stage was typically smaller than that of the equal-weights case. To test the consistency of weights for different periods, it was examined if the same will hold for the testing stage. It was examined whether the models deemed "good" in the training stage, will also remain grouped together around the mean (or relatively close to the mean) in the testing stage, and similarly if the bad models in the training stage will remain poor-performing models (outliers) in the testing stage. The standard deviations of the model data for two periods in the future, both with and without weights, were compared to investigate if the weights can reduce ensemble spread in future climate projections. It can be speculated that the consistency in model weights in training and testing periods, i.e., reduced weighted ensemble spread compared to the equal weights option, would support the assumption that past performance is an indication of future accuracy. Only relative effects of weights were examined,

while other sources of variability in model performances, e.g. natural variability over short time periods and variability of the future climate, were not analyzed in this research.

Similar to Wu et al (2019), to get an approximate estimate of the changes in the 21st century, we split the available future model data (2006–2100) into two approximately equal periods, 2006–2053 and 2054–2100. We find both the weighted and unweighted standard deviations, σ_w and σ_u, of the PFs for each period using the technique described at the end of the Methodology section and Eq. 4. The statistical quantity selected for this test was the percent reduction (PR), expressed as

$$PR\ (\%) = (\frac{\sigma_w - \sigma_u}{\sigma_u}) \times 100. \quad (5)$$

A negative PR indicates that the weighted approach resulted in a smaller standard deviation; zero means that the standard deviations are equal; and a positive PR means that the weights increased the standard deviation. The results for ARF=0.90, ARF=0.80, and ARF=0.67 are presented in Table 5. The results show that the PR was negative, ranging between -61.6% and -6.7%, for all cases with the average of -37.0%, meaning that the weighted standard deviations are smaller than the unweighted ones and that the weights successfully reduced the projected model spread.

The relative increases in precipitation frequency estimates based on this study are illustrated in Figure 5, where the locations and climate scenarios are included in the title for each subplot. Names in the legend start with "U" (unweighted method) or "W" (weighted method) and end with "mid" (2006-2053) or "late" (2054-2100), representing the first and the second halves of the 21st century respectively, and the changes with respect to "present" time (1960-2005). For example, "Umid" denotes the results for the mid-21st century using the unweighted

ensemble approach. The expected frequency estimates increase with time, i.e., late-century frequency estimates are larger than those of mid-century, meaning that heavy storms will continue to increase throughout the century. The values for RCP8.5 are also typically higher than those of RCP4.5. The projected increases are generally larger for the weighted approach. Large differences between the weighted and equal-weights signify the importance of the decision to either select the weighted approach or adopt equal weights. This difference is small for Chicago RCP8.5 mid- and late-century statistics, but it is very significant for Houston RCP4.5 late-century statistics, where the 100-year event is about 50% higher for some return periods when variable weights are used.

Table 5. Percent reduction (PR) values comparing weighted vs. unweighted ensembles spread according to Eq. 5.

Figure 5. Projected relative percent increases in heavy storm events at three selected sites.

CONCLUDING REMARKS

The experiment described in this study was based on several key assumptions. The range of results based on uncertain future emission scenarios was represented by using RCP4.5 and RCP8.5 as in many other studies (Wuebbles *et al.*, 2017), but structural (model) and data uncertainties were not explored herein. There are potential issues in ensemble analyses when the models are correlated (Knutti *et al.*, 2010; Steinschneider *et al.*, 2015); nonetheless, in this study it was assumed that the model results are independent. It appears that this assumption would be less problematic for extreme precipitation, for which the correlations among different models

were weak, even for very similar models. Additional insights on this issue were offered by Wuebbles *et al*. (2017) and Shortridge and Zaitchik (2018).

The study is limited given the small number of years in both the training and testing historical periods producing large variability in the results (which may have been alleviated by averaging large ensembles). A relatively small number of stations and model grid points were used as well. More of all of these would help improve and strengthen the conclusions, especially more model grid points, so that each grid point maps to only one station. Having multiple sets of raw data would also allow more models to have positive weights by finding the weights for each set and then averaging them. Finally, the results are specific to one selected method for determining weights. Different methods would certainly result in different weights, which is an aspect that requires further exploration.

A critical factor contributing to the results of this study is the assumed ARF. The initial ARF adopted for this study matched that in the recent NOAA Atlas 14 publications (e.g. Perica *et al*., 2018). The value of 0.90 used in this study was obtained by extrapolating the original Hershfield (1961) curve to the size of the grid cells of the climate models. Many researchers found these curves to be very uncertain (Pavlovic *et al*., 2016) and deemed too high (Sivapalan and Blöschl, 1998; Allen and DeGaetano, 2005), prompting the addition of two lower ARF curves that might be more appropriate for this problem. ARF=0.80 and ARF=0.67 were assumed conveniently to produce inverse values of 1.25 and 1.50, respectively. Although different ARFs produced variable results, the application of weights resulted in a significant reduction in percent difference between observed and model-based precipitation frequency estimates, regardless of the adopted ARF (Table 4). Weighted results consistently outperformed the unweighted ones, indicating that the past performance can be an indication of future accuracy. Models that were

more accurate in the training stage were typically more accurate in the testing stage and vice versa. Additionally, application of weights resulted in reduced standard deviation among the projected ensemble members, thereby potentially reducing the confidence limits of the future projections. Moreover, these models tended to have similar performances at all three geographic locations.

Despite the simplifications, limitations, and uncertainties of the proposed approach, the hypothesis that the addition of weights will provide a more accurate average PF and smaller model variability was largely confirmed by the experimental results. This study makes the case for continuing to investigate the use of weights and demonstrates that they can have value and could prove to be useful tools when studying future climate.

ACKNOWLEDGMENTS

The paper is supported by NOAA/UCAR Contract # SUBAWD000255. The authors would like to thank the National Center for Supercomputing Applications (NCSA) Industry Program at the University of Illinois and its Blue Waters project, funded by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois for hardware and application support. Lisa Sheppard (Illinois State Water Survey) provided editorial assistance.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the lead and corresponding authors upon reasonable request.

LITERATURE CITED

Allen, R. J. and A. T. DeGaetano. 2005. "Areal Reduction Factors for Two Eastern United States Regions with High Rain-Gauge Density." *Journal of Hydraulic Engineering* 10 (4): 327-335.

Christensen, J. H., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen. 2010. "Weight Assignment in Regional Climate Models." *Clim. Res.* 44 (2-3): 179-194.

cli-MATE: Midwestern Regional Climate Center Application Tools Environment. (https://mrcc.purdue.edu/CLIMATE/welcome.jsp, last accessed 09/30/2019).

Douglas, E. M. and C. A. Fairbank. 2011. Is Precipitation in Northern New England Becoming More Extreme? Statistical Analysis of Extreme Rainfall in Massachusetts, New Hampshire, and Maine and Updated Estimates of the 100-Year Storm." *Journal of Hydrologic Engineering* 16 (3): 203-217.

Emanuel, K. 2017. "Assessing the present and future probability of Hurricane Harvey's rainfall." Proc. Natl Acad. Sci. USA 114, 12681–12684.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005. "The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept." *Tellus,* 57A, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hershfield, D. M. 1961. "Rainfall Frequency Atlas of the United States." Technical paper 40.

Hosking, J. R. M., and J.R., Wallis. 1997. "Regional frequency analysis: An approach based on L-moments," Cambridge University Press, Cambridge, U.K.

Knutti, R. 2010. "The End of Model Democracy?" *Climatic Change* 102 (3-4): 395-404. doi:10.1007/s10584-010-9800-2

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl. 2010. "Challenges in Combining Projections from Multiple Climate Models." *J. Climate* 23: 2739-2758.

Knutti, R., 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.

Langbein, W. B. 1949. "Annual Floods and the Partial-duration Flood Series." *Trans. Am. Geophys. Union* 30 (6): 379.

Li, Z., X. Li, Y. Wang, and S. M. Quiring. 2019. "Impact of Climate Change on Precipitation Patterns in Houston, Texas, USA." *Anthropocene* 25, ISSN 2213-3054. doi:10.1016/j.ancene.2019.100193

Markus, M., J. R. Angel, K. Wang, G. Byard, S. McConkey, and Z. Zaloudek. 2017. *Impacts of Potential Future Climate Change on the Expected Frequency of Extreme Rainfall Events in Cook, DuPage, Lake, and Will Counties in Northeastern Illinois.* Champaign, IL: Illinois State Water Survey.

Markus, M., J. R. Angel, G. Byard, S. McConkey, C. Zhang, X. Cai, M. Notaro, and M. Ashfaq. 2018. "Communicating the Impacts of Projected Climate Change on Heavy Rainfall Using a Weighted Ensemble Approach." *J. Hydrol. Eng.* 23 (4): 04018804.

Markus, M.; Wuebbles, D.J.; Liang, X.Z.; Hayhoe, K.; Kristovich, D.A. (2012). Diagnostic analysis of future climate scenarios applied to urban flooding in the Chicago metropolitan area. Clim. Chang. 111 (3-4): 879–902.

Masson, D. and R. Knutti. 2011. "Climate Model Genealogy." *Geophys. Res. Lett.* 38 (8): L08703.

Notaro, M., D. J. Lorenz, C. Hoving, and M. Schummer. 2014. "Twenty-first-century Projections of Snowfall and Winter Severity across Central-eastern North America." *J. Clim.* 27 (17): 6526-6550.

Pavlovic, S., S. Perica, M. St Laurent, and A. Mejía. 2016. "Intercomparison of Selected Fixed-area Areal Reduction Factor Methods." *J. Hydrol.* 537 (38): 419-430. https://doi.org/10.1016/j.jhydrol.2016.03.027

Perica, S., S. Pavlovic, M. S. Laurent, C. Trypaluk, D. Unruh, and O. Wilhite. 2018. *Precipitation-frequency Atlas of the United States. Version 2.0: Texas*. Silver Spring, MD: National Weather Service.

Perica, S., S. Pavlovic, M. S. Laurent, C. Trypaluk, D. Unruh, D. Martin, and O. Wilhite. 2019. *Precipitation-frequency atlas of the United States. Version 3.0: Northeastern States*. Silver Spring, MD: National Weather Service.

Räisänen, J., and J. S. Ylhäisi. 2012. "Can Model Weighting Improve Probabilistic Projections of Climate Change?" *Clim. Dyn.* 39 (7-8): 1981-1998.

Rimoldini, L. 2014. "Weighted Skewness and Kurtosis Unbiased by Sample Size and Gaussian Uncertainties." *Astron. Comput.*, 5: 1-8.

Sánchez, E., R. Romera, M. A. Gaertner, C. Gallardo, and M. Castro. 2009. "A Weighting Proposal for an Ensemble of Regional Climate Models over Europe Driven by 1961-2000 ERA40 Based on Monthly Precipitation Probability Density Functions." *Atmos. Sci. Lett.* 10 (4): 241-248.

Sanderson, B.M., R. Knutti, and P. Caldwell, 2015. "A representative democracy to reduce interdependency in a multimodel ensemble." *Journal of Climate*, 28, 51715194. http://dx.doi.org/10.1175/jcli-d-14-00362.1

Sanderson, B. M. and M. F. Wehner. 2017. "Model Weighting Strategy." In: *Climate Science Special Report: Fourth National Climate Assessment, Volume I*, edited by D. J. Wuebbles, D. W. Fahey, K. A. Hibbard, D. J. Dokken, B. C. Stewart, and T. K. Maycock, 436-442. Washington, DC: U.S. Global Change Research Program. doi: 10.7930/J06T0JS3

Shortridge, J. E. and B. F. Zaitchik. 2018. "Characterizing Climate Change Risks by Linking Robust Decision Frameworks and Uncertain Probabilistic Projections. *Climatic Change* 151: 525-539.

Sivapalan, M. and G. Blöschl. 1998. "Transformation of Point Rainfall to Areal Rainfall: Intensity-Duration-Frequency Curves." *Journal of Hydrology* 204: 150-167.

Steinschneider, S., R. McCrary, L. O. Mearns, and C. Brown. 2015. "The Effects of Climate Model Similarity on Probabilistic Climate Projections and the Implications for Local, Risk-based Adaptation Planning: Intermodel Correlation and Risk." *Geophys Res Lett* 42: 5014-5044.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, MA: Addison Wesley.

Um, M.-J., Y. Kim, M. Markus, and D. J. Wuebbles. 2017. "Modeling Nonstationary Extreme Value Distributions with Nonlinear Functions: An Application Using Multiple CMIP5 Precipitation Projections for U.S. Cities." *Journal of Hydrology* 552: 396-406.

Um, M.-J., J.-H. Heo, M. Markus, and D. J. Wuebbles. 2018. "Performance Evaluation of Four Statistical Tests for Trend and Non-stationarity and Assessment of Observed and Projected Annual Maximum Precipitation Series in Major United States Cities." *Water Resources Management* 32 (3): 913-933.

USGCRP, 2017: *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.

USGCRP, 2018: Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II [Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 1515 pp. doi: 10.7930/NCA4.2018.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller. 2010. "Risks of model weighting in multimodel climate projections." *J. Clim*. 23 (15): 4175-4191.

Wu, S., M. Markus, D. J. Lorenz, J. R. Angel, and K. Grady. 2019. "A Comparative Analysis of the Historical Accuracy of the Point Precipitation Frequency Estimates of Four Data Sets and their Projections for the Northeastern United States." *Water* 11: 1279.

Wuebbles, D. J., D. W. Fahey, K. A. Hibbard, B. DeAngelo, S. Doherty, K. Hayhoe, R. Horton, J. P. Kossin, P. C. Taylor, A. M. Waple, and C. P. Weaver. 2017. Executive Summary. In: *Climate Science Special Report: Fourth National Climate Assessment, Volume I*, edited by D. J. Wuebbles, D. W. Fahey, K. A. Hibbard, D. J. Dokken, B. C. Stewart, and T. K. Maycock. 12-34. Washington, DC: U.S. Global Change Research Program. doi: 10.7930/J0DJ5CTG

Zhang, W., G. Villarini, G. A. Vecchi, and J. A. Smith. 2018. "Urbanization Exacerbated the Rainfall and Flooding Caused by Hurricane Harvey in Houston. *Nature* 563: 384-388.

Table 1. The regions and time periods selected for this study.

City	Boston, MA	Houston, TX	Chicago, IL
Coordinates of Observed Region	42-43°N × 70.5-72°W	28.5-30.5°N × 94-96°W	41-43°N × 87-88.5°W
Coordinates of Modeled Region	41-44°N × 70-73.2°W	27.8-31.6°N × 92.8-97.6°W	40-44°N × 86-90°W
# of Stations in Training Period	35	40	15
# of Stations in Testing Period	22	23	11
Training Period	1960-2005 (46 years)	1960-2005 (46 years)	1961-2000 (40 years)
Testing Period	2006-2018 (13 years)	2006-2018 (13 years); 2006-2016, 2018 (12 years)	2006-2018 (13 years)

Table 2. The 24 CMIP5 models used in this study along with their grid-cell sizes (latitude × longitude). No RCP4.5 data were available for models with an (*).

Model	Grid-Cell Size	Model	Grid-Cell Size
ACCESS1-0	1.25° × 1.875°	inmcm4	1.5° × 2°
ACCESS1-3	1.25° × 1.875°	IPSL-CM5A-LR	1.8948° × 3.75°
CanESM2	2.7904° × 2.8125°	IPSL-CM5A-MR	1.2676° × 2.5°
CMCC-CESM*	3.75° × 3.75°	IPSL-CM5B-LR	1.8948° × 3.75°
CMCC-CM	0.75° × 0.75°	MIROC5	1.4007° × 1.4063°
CMCC-CMS	1.8651° × 1.875°	MIROC-ESM	2.7904° × 2.8125°
CNRM-CM5	1.4007° × 1.4063°	MIROC-ESM-CHEM	2.7904° × 2.8125°
CSIRO-Mk3-6-0	1.8651° × 1.875°	MPI-ESM-LR	1.8651° × 1.875°
GFDL-CM3	2° × 2.5°	MPI-ESM-MR	1.8651° × 1.875°
GFDL-ESM2G	2.0225° × 2.5°	MRI-CGCM3	1.1215° × 1.125°
GFDL-ESM2M	2.0225° × 2.5°	MRI-ESM1*	1.1215° × 1.125°
HadGEM2-CC	1.25° × 1.875°	NorESM1-M	1.8948° × 2.5°

Table 3. Climate model weights averaged for climate scenarios RCP4.5 and RCP 8.5 for ARF=0.90, 0.80, and 0.67.

Climate Model	ARF=0.90			ARF=0.80			ARF=0.67		
	Boston	Houston	Chicago	Boston	Houston	Chicago	Boston	Houston	Chicago
ACCESS1-0	2.98	5.23	2.73	1.58	5.01	2.87	0.00	4.32	2.34
ACCESS1-3	2.03	6.72	2.06	2.20	5.99	2.38	0.44	4.48	2.34
CanESM2	2.27	0.00	1.34	2.22	0.00	1.76	0.08	0.00	2.29
CMCC-CESM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CMCC-CM	2.93	2.35	3.96	2.06	2.60	3.51	0.00	3.12	2.10
CMCC-CMS	0.73	0.00	0.84	1.52	0.00	1.23	3.28	0.00	2.15
CNRM-CM5	2.28	0.28	0.02	2.22	0.39	0.07	0.06	0.76	0.71
CSIRO-Mk3-6-0	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.00
GFDL-CM3	0.00	0.00	0.00	0.00	0.00	0.00	3.02	0.00	0.00
GFDL-ESM2G	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.00
GFDL-ESM2M	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
HadGEM2-CC	1.73	3.27	0.43	2.13	3.45	0.72	1.16	3.67	1.83
inmcm4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IPSL-CM5A-LR	0.00	0.00	0.00	0.00	0.00	0.01	2.77	0.00	0.41
IPSL-CM5A-MR	1.90	0.09	3.59	2.18	0.13	3.35	0.73	0.32	2.23
IPSL-CM5B-LR	2.96	3.86	5.67	1.89	3.95	3.83	0.00	3.94	0.07
MIROC5	1.57	0.05	1.74	2.08	0.07	2.12	1.60	0.20	2.33
MIROC-ESM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MIROC-ESM-CHEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MPI-ESM-LR	0.02	0.00	0.03	0.26	0.00	0.09	3.64	0.00	0.78
MPI-ESM-MR	0.00	0.00	0.15	0.00	0.00	0.32	0.68	0.00	1.33
MRI-CGCM3	0.91	0.44	0.08	1.68	0.58	0.19	3.04	1.04	1.08
MRI-ESM1	1.44	1.46	0.81	2.00	1.70	1.19	1.99	2.32	2.11
NorESM1-M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

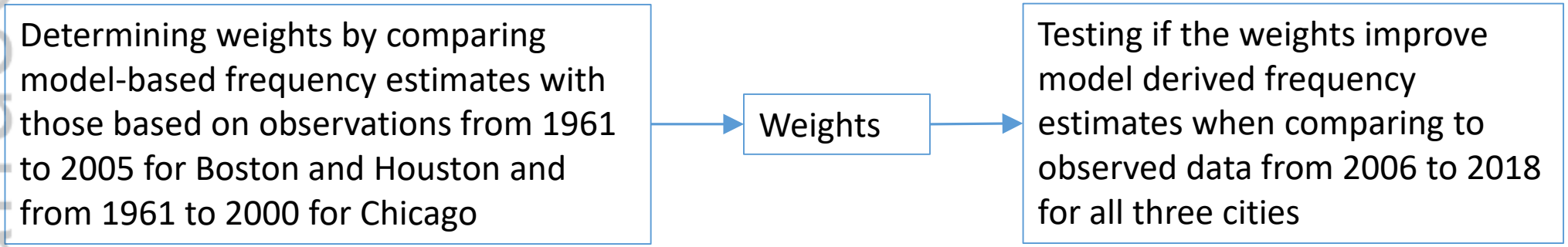
Table 4. Weighted and unweighted percent differences of PFs (Eq. 1) averaged across all return periods and stations in the testing period (2006–2018) for ARF=0.90, 0.80, and 0.67.

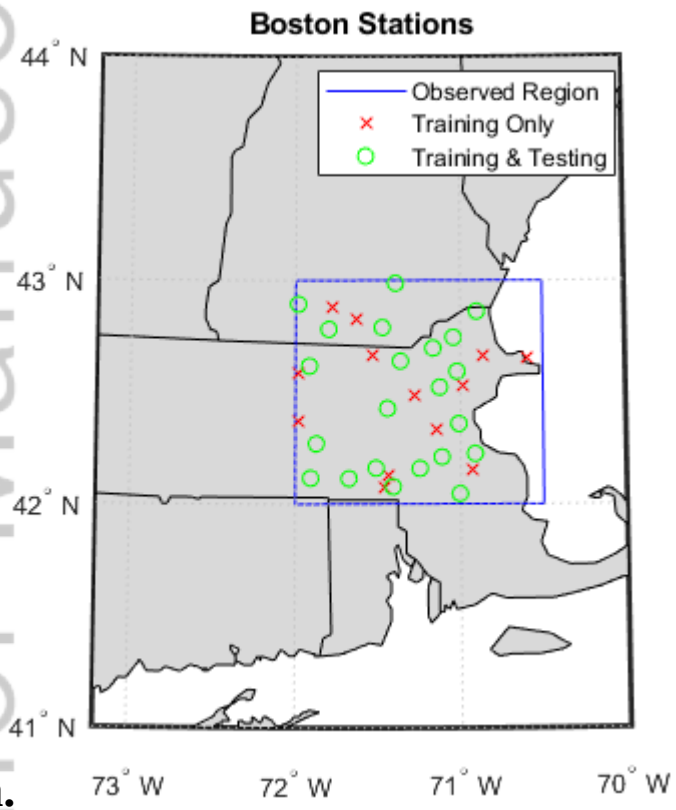
		RCP4.5		RCP8.5	
		Weighted	Unweighted	Weighted	Unweighted
ARF=0.90	Boston	5.86%	-17.60%	-8.94%	-24.84%
	Houston with 2017	-43.85%	-62.28%	-39.97%	-61.58%
	Houston without 2017	-33.21%	-55.20%	-29.28%	-54.98%
	Chicago	-20.11%	-38.46%	-19.60%	-35.41%
ARF=0.80	Boston	**13.66%**	**-7.21%**	-0.27%	-15.36%
	Houston with 2017	-37.28%	-57.52%	-33.23%	-56.73%
	Houston without 2017	-25.38%	-49.55%	-21.32%	-49.30%
	Chicago	-12.71%	-30.69%	-12.16%	-27.27%
ARF=0.67	Boston	5.96%	11.35%	0.11%	1.57%
	Houston with 2017	-26.42%	-49.02%	-22.50%	-48.08%
	Houston without 2017	-12.46%	-39.46%	-8.73%	-39.16%
	Chicago	-4.15%	-16.83%	-2.30%	-12.72%
Average	Average	-15.84%	-34.37%	-16.52%	-35.32%

Table 5. Percent reduction (PR) values comparing weighted vs. unweighted ensembles spread according to Eq. 5.

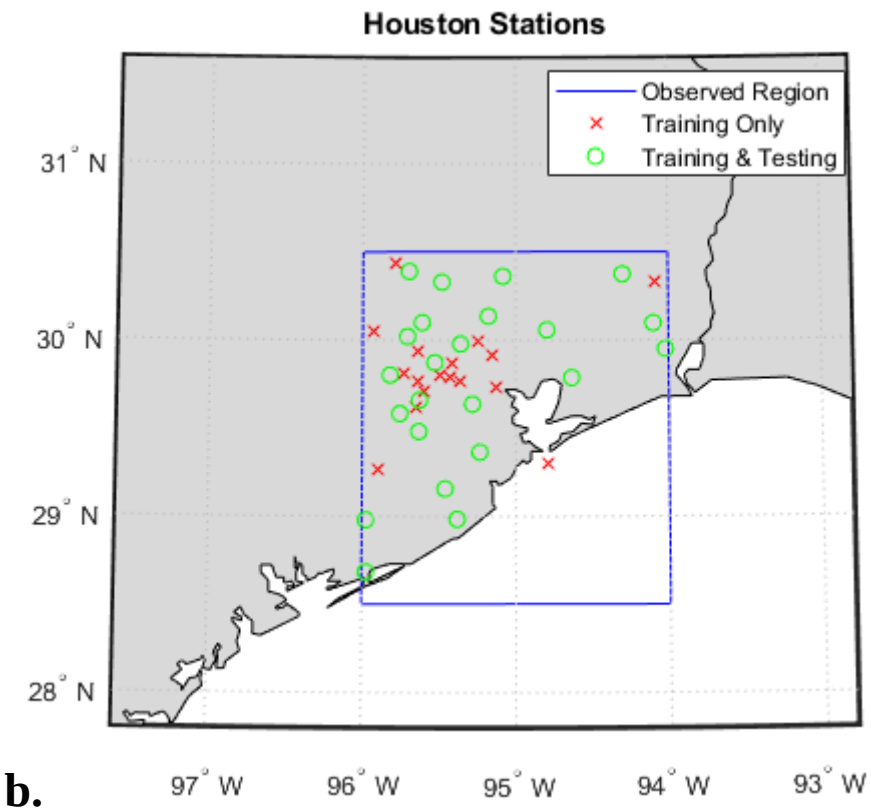
Climate Scenario		Boston				Houston				Chicago			
		2006-2053		2054-2100		2006-2053		2054-2100		2006-2053		2054-2100	
		RCP4.5	RCP8.5	RCP4.5	RCP8.5	RCP4.5	RCP8.5	RCP4.5	RCP8.5	RCP4.5	RCP8.5	RCP4.5	RCP8.5
ARF=0.90 Return Period (years)	2	-41.0	-29.4	-37.7	-39.8	-54.1	-28.2	-41.8	-47.9	-21.2	-16.2	-25.6	-8.2
	5	-40.9	-32.6	-38.8	-33.6	-59.8	-35.2	-49.5	-55.8	-24.6	-20.2	-28.3	-15.2
	10	-39.5	-31.9	-38.1	-27.6	-61.6	-41.3	-53.5	-59.1	-27.0	-23.9	-29.5	-20.2
	25	-36.5	-28.4	-35.7	-19.1	-60.4	-48.6	-56.8	-57.9	-30.0	-29.1	-30.0	-26.8
	50	-33.6	-24.8	-32.9	-12.7	-57.2	-52.2	-57.9	-53.4	-32.1	-33.1	-29.6	-31.7
	100	-30.4	-20.8	-29.8	-6.7	-52.8	-53.6	-57.9	-47.0	-33.9	-37.0	-28.6	-36.4
ARF=0.80 Return Period (years)	2	-41.6	-33.6	-37.8	-40.6	-51.3	-27.6	-39.9	-45.3	-24.1	-19.5	-26.2	-9.8
	5	-40.2	-34.8	-37.3	-35.0	-56.8	-34.3	-47.3	-52.9	-26.5	-22.7	-28.1	-15.7
	10	-38.3	-32.9	-36.1	-30.0	-58.5	-40.2	-51.1	-56.2	-28.1	-25.8	-28.5	-20.3
	25	-34.9	-28.3	-33.6	-23.1	-57.4	-47.0	-54.4	-55.6	-30.1	-30.3	-28.0	-26.5
	50	-31.9	-24.0	-31.1	-18.1	-54.5	-50.4	-55.6	-51.5	-31.4	-33.8	-26.9	-31.2
	100	-28.9	-19.5	-28.4	-13.5	-50.3	-51.7	-55.8	-45.6	-32.6	-37.1	-25.3	-35.6
ARF=0.67 Return Period (years)	2	-21.7	-43.2	-22.1	-46.0	-42.0	-24.7	-32.6	-36.8	-43.3	-40.1	-36.7	-22.6
	5	-22.9	-44.6	-25.4	-47.5	-47.4	-30.9	-39.2	-44.0	-42.6	-41.1	-36.8	-25.9
	10	-23.9	-45.3	-27.8	-48.6	-49.4	-36.4	-43.0	-47.8	-41.6	-41.7	-35.6	-28.9
	25	-24.8	-45.0	-30.6	-48.9	-49.1	-42.6	-46.6	-48.7	-40.1	-42.6	-32.9	-33.2
	50	-25.1	-43.8	-32.4	-47.8	-47.0	-45.6	-48.3	-46.3	-38.9	-43.1	-30.3	-36.3
	100	-25.0	-41.9	-33.8	-45.8	-43.7	-46.7	-49.1	-41.9	-37.7	-43.5	-27.5	-38.9

List of Figures:

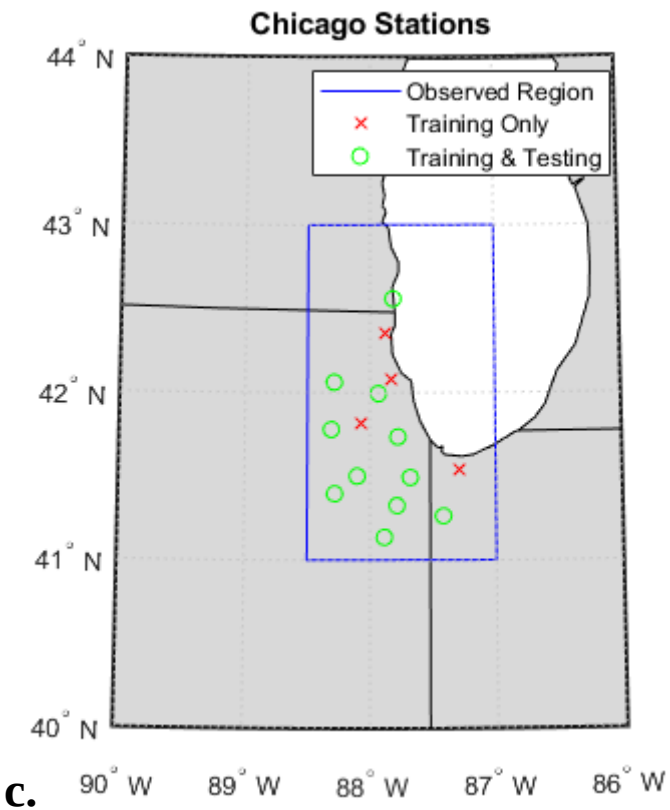




a.



b.



c.