Statistical Methods for Defining Climate-Similar Regions around Weather Stations Using NLDAS-2 Forcing Data

John Hathaway[1], Trenton Pulsipher, Jeremiah Rounds, and Jim Dirks
Applied Statistics and Computational Modeling, Pacific Northwest National Laboratory

1 Introduction

The North American Land Data Assimilation System Phase 2 (NLDAS-2) (Mitchel 2004) is a well-studied and well-regarded climate data set that provides reliable hourly estimates of multiple variables on a ⅛-degree grid spatial resolution for the years 1979 to the present.

Of interest is apportioning the ⅛-degree grid cells used by the NLDAS-2 study into climate-similar regions (spatial partitions) with a constraint that each of the climate-similar regions must be associated with a reliable weather station (e.g. reference grid cell). We provide a new method for this apportioning of ⅛-degree grid cells into climate-similar regions and/or identifying the most representative reference grid cell within a fixed region.

This work is embedded in a larger research project: Platform for Regional Integrated Modeling and Analysis, which is a unique modeling framework developed at Pacific Northwest National Laboratory (PNNL) to simulate the interactions among natural and human systems at scales relevant to regional decision-making. The Building Energy Demand Model is used to estimate building energy consumption across large geographic areas with thousands of different building representations. Climate-similar regions are identified such that the simulated building's heating and cooling energy demand profiles are appropriately representative of an area much larger than the single ⅛-degree grid cell in which a weather station is located.

It is not atypical in building energy demand literature to divide a spatial region of interest into climate regions around weather stations (as in Kalamees et al. 2012 and ASHRAE 90.1), or to complete this process algorithmically. As far back as 1986, Andersson and co-workers from Lawrence Berkeley Laboratory published work detailing a computer program they called GLOM, which was designed to interactively use population weighted Typical Meteorological Year (TMY) data to divide the United States into climate-similar regions (1986). The new challenge is incorporating an increased level of detail about weather and climate previously not available to researchers. This work introduces new algorithms built for the much larger scale and detailed NLDAS-2 data that allocate grid cells into climate-similar regions.

While the climate comparison model can be leveraged for many purposes, our current objectives are oriented toward building energy demand. We contrast our results with the International Energy Conservation Code (IECC) climate partitioning of the United States (Briggs et al. 2003). The IECC divides the contiguous United States into seven climate zones. Each of these IECC climate zones is partitioned according to climate type (marine, moist, dry) with the southeast United States designated a Warm-Humid region. When these factor levels are considered together, the IECC creates 14 regions for the contiguous United States (Figure 1) and recommends varying building standards according to region.

[1] John.Hathaway@pnnl.gov

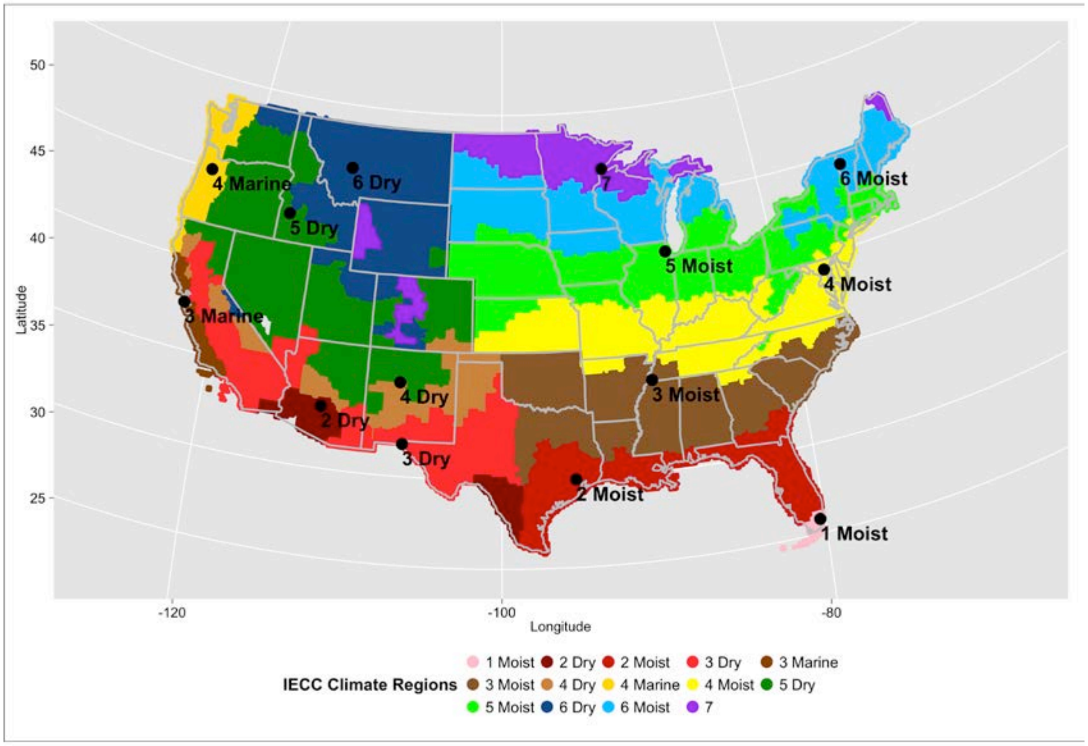

Figure 1. ⅛-degree grid cells colored by IECC climate zone and type. White regions correspond to areas of the United States with almost zero population that did not impact our work. Briggs and co-workers (2003) recently suggested the reference cities marked with dots as representative for each climate region. Note, 14 reference cities and climate regions are shown.

The Typical Meteorological Year (TMY3) data set from Wilcox and Martin (2008) is often used in building energy demand literature and was influential in implementing our methodology. In Wilcox and Martin (2008), for each month, the most representative (or most typical) month of weather observations are selected from among over 30 years of historical data at specified fixed weather stations. This is done by comparing differences in cumulative distribution functions (CDFs) for daily observations by month and by variable across all observed years. In our work, the goal is to find representative grid cells and climate-similar regions by comparing CDFs between grid cells, and the method of Wilcox and Martin (2008) is adjusted accordingly. While we do not use the TMY3 data set for any purpose, the Sandia Method (Hall et al. 1978), Finkelstein's Goodness-of-Fit procedure (Finkelstein 1971), and the National Solar Radiation Data Base (NSRDB) class for weather stations as used by Wilcox and Marion (2008) influenced our work.

In building energy demand simulations, such as that done by EnergyPlus©, a common procedure is to model a building's important heating, cooling, and insulating characteristics and simulate how these characteristics interact with observed weather station data. We identified **⅛**-degree grid cells in the NLDAS-2 data set that contained NSRDB Class I through III weather stations, and labeled them reference grid cells so that each reference grid cell corresponds to an available reference weather station.

The concept of a reference weather station is not new. Indeed, the IECC climate zones and types are all associated with reference cities. Recently, Briggs and co-workers suggested the reference cities marked in Figure 1 as having representative weather for each climate region (Briggs et al. 2003). We consider the usefulness of those representative cities in this paper.

2 Data

Daily Summary Data

The NLDAS-2 data set, as used in this work, provides smoothed hourly observations from 53,636 ⅛-degree grid cells in the contiguous United States from 1971 to 2009.

We construct a daily summary data set from NLDAS-2 hourly data of nine variables: Dry-Bulb Temperature (daily minimum, maximum, mean), Dew Point (daily minimum, maximum, mean), Wind Speed (daily maximum, mean), and Total Solar Radiation (daily total). Selecting these variables follows Hall's Sandia method (Hall et al. 1978) and subsequent related meteorological literature. The variables in NLDAS-2 used to construct these derived variables were Pressure, Vapor Pressure, Specific Humidity, Downward Shortwave Radiation Flux, Dry-Bulb Temperature, Dew Point Temperature, Horizontal Wind Speed, and Vertical Wind Speed.

Total Solar Radiation (Daily) is the sum of Downward Shortwave Radiation Flux for that day.

Dew Point (T_d) is in Celsius and calculated in the following manner for each hourly observation:

$$T_d = \frac{243.5 \log(v)}{17.62 - \log(v)},$$

where $v = \frac{Specific\ Humidity \times Pressure \times 6.112}{0.622}$.

Generally, daily minimum, maximum, and mean values are drawn from hourly observations in NLDAS-2 with the same name.

3 Methods

We are most interested in two operations related to this data. 1) Apportioning grid cells into sets associated with a representative reference grid cell. This operation requires both the apportioning of grid cells and the assignment of partitions to reference grid cells that contain a weather station location. 2) Optimally subset the reference grid cells into a smaller reference set (rather than using all 1,290 reference grid cells in building energy demand simulations) to lose as little accuracy as possible. Our methods divide these operations into three sub-problems:

- Calculating climate similarities between grid cells and reference grid cells.
- Assigning grids cells to the most climate-similar reference grid cell and scoring performance.
- Selecting a most representative subset of reference grid cells.

Throughout this work, a triplet (P,$\boldsymbol{r}$,G) will help keep the work tractable and organized. G is the set of ⅛-degree grid cells (e.g., the contiguous United States). P is a partitioning of G such that every element is assigned to one partition. For example, naively, P could be the 48 contiguous states, then $\boldsymbol{r}$ would be a set of reference grid cells such that there is one for every partition (i.e., state in this example). P and $\boldsymbol{r}$ are considered to be such that elements of $\boldsymbol{r}$ (reference grid cells) are uniquely assigned to elements of P (partitions). Continuing the naïve example, the 48 states all have 48 capitals, so $\boldsymbol{r}$ could arbitrarily be the cell that contains the capital of each state. We could label these terms ($P_{\text{State}}, \boldsymbol{r}_{\text{State}}, G_{\text{US}}$), and in so doing so, we would indicate that every grid cell of the United States has been assigned a partition and a reference. A more practical example is to consider the IECC data in Figure 1. Every grid cell has an IECC climate zone and every IECC climate zone has a reference city. Thus, to refer to the IECC organization of the United States, we would state that we were working with partitions and references ($P_{\text{IECC}}, \boldsymbol{r}_{\text{IECC}}, G_{\text{US}}$).

In this work we consider constructing $\boldsymbol{r}$ for a given partitioning P from an available pool of references (R). The cardinality of R need not be the same as the cardinality of P. The vector notation is used to indicate that now there is a smaller set of references ordered to correspond to P such that elements of $r_{\text{i}} \in \boldsymbol{r}$ are assigned to specific elements $p_i \in P$. Continuing the naïve United States example with partitions by States (P_{State}), R could be the NSRDB Class I through III weather stations for the United States, and now $\boldsymbol{r}$ needs to be determined for P_{State}.

Given *G*, *R*, and the daily summary NLDAS-2 data, we illustrate our methods for comparing and grouping climate-similar regions. We describe the climate similarity score (CSS) and the methods for its derivation at each grid cell. We then discuss the different applications of assigning grid cells to optimal reference grid cells.

Climate Similarity Scores

A score measuring climate similarity between any two grid cells in the NLDAS-2 data set can be constructed by comparing the quantiles of observed variables in the daily summary data set. Let g and r be two grid cells from NLDAS-2. Let v be a variable from the derived daily summary data, and let m be a month from the derived data as an element of 1 to 12. We treat each variable from the same month and grid cell crossing as having its own random distribution over the climate period from 1979 to 2009. Let X_i be a daily observation for v in a month over all years. Empirical CDFs exist for g ($\hat{F}_{\text{gmv}}$) and r ($\hat{F}_{\text{rmv}}$) that summarize the observed distribution of X_i.

Let $\mathcal{X}_{mv} = (x_1, \dots, x_N)$ be a set of appropriate evenly spaced quantities. For any two grid cells g and r, the Finkelstein-Schafer statistic for variable v in month m is:

$$FS_{mv}(g,r) = \frac{1}{N} \sum_{x_i \in \mathcal{X}_{mv}} \left| F_{gmv}(\mathrm{x_i}) - F_{rmv}(x_i) \right|.$$

Let $\mathcal{W} = (w_1, \dots, w_9)$ be a set of weights for variables that sum to 1. Let $\mathcal{U} = (u_1, \dots, u_{12})$ be a set of weights for months that sum to 1. A CSS for grid cells g and r is a weighted average as:

$$CSS(g,r) = \sum_{m=1}^{12} \sum_{v=1}^{9} u_m w_v FS_{mv}(g,r).$$

The decision on the values to use for $\mathcal{W}$ and $\mathcal{U}$ are project driven. In our current applications, we set each u_m to 1/12 and used the International Weather Year for Energy Calculations (IWEC) weights shown in Table 1 (Thevnard, 2001). For comparison purposes, we show the Sandia Method and TMY3 weights for $\mathcal{U}$. We note that g and r can be the same grid cell, which would result in $CSS(g,r) = 0$. As CSS values approach zero, g and r have similar weather across months and variables.

Table 1. Variable weights $\mathcal{U}$.

Variable (units)	IWEC	Sandia Method	TMY3
Maximum Dry-Bulb Temperature (K)	0.050	0.0417	0.05
Minimum Dry-Bulb Temperature (K)	0.050	0.0417	0.05
Mean Dry-Bulb Temperature (K)	0.300	0.0833	0.10
Maximum Dew Point Temperature (K)	0.025	0.0417	0.05
Minimum Dew Point Temperature (K)	0.025	0.0417	0.05
Mean Dew Point Temperature (K)	0.050	0.0833	0.10
Maximum Wind Speed (m/s)	0.050	0.0833	0.05
Mean Wind Speed (m/s)	0.050	0.0833	0.05
Total Solar (or Global Horizontal) Radiation (W/m^2)	0.400	0.5000	0.25
Direct Radiation (W/m^2)	NA	NA	0.25

Climate-Similar Partitions

The first type of problem that can be solved with CSS is assigning grid cells to reference grid cells or weather stations according to climates similarity. In this process, we assign each grid cell of a region to a reference grid cell with the lowest CSS. Let $\boldsymbol{r}$ be an ordered subset of reference grid cells that contain NSRDB Class I through III weather stations (R). Let G be a subset of grid cells for the contiguous United States or other region of interest within NLDAS-2. For every $g \in G$, define a partition assignment function:

$$P(g, \boldsymbol{r}) = \underset{r_i \in \boldsymbol{r}}{argmin}\, CSS(g, r_i).$$

Thus, P maps G into $\boldsymbol{r}$. The inverse $P^{-1}(r_i)$ is the set of all grid cells assigned to NSRDB Class I-III weather station *r*. The set of $P(g, \boldsymbol{r})$ for all $g \in G$ is a climate partitioning of G. We note that P is dependent on available reference grid cells (***r***); thus, **r** is an argument to *P*, but outside of the context where the possible reference grid cells might be changing, it is simpler to omit that argument. Effectively, *P* should be thought of as having divided the map of the United States into regions around reference grid cells.

For any $g \in G$, the minimal CSS to an $r_i \in \boldsymbol{r}$ is $CSS\big(g, P(g, \boldsymbol{r})\big)$, a fact that is used in the next section. We label this min $CSS_{\boldsymbol{r},P}(g)$ and left it is as implied that a partition around a reference was created. This notion of minimal CSS becomes the basis for saying how well represented a region (set of grid cells) is by a set of reference grid cells. It also forms the basis of selecting ***r*** from R such that all the grid cells are optimally represented.

Reference Cell Optimization

Having defined CSS and minimum CSS for any grid cell given an ***r***, there are practical questions that can be asked around minimizing a score of summed CSS.

- Which 200 of 1290 NSRDB Class I through III weather stations would best represent the spatial region of interest? Repeated in terms of our defined equations, which **r** subset of R (200 of 1100 weather stations) would minimize the total CSS over G (spatial region of interest)?
- For all the **⅛**-degree grid cells in a specific IECC climate region which single weather station is the most climate-similar option (i.e., provides the minimum total CSS across all grid cells in the region)? In this case, ***r*** is a single station out of many, and a G is all the grid cells in a single IECC climate region.

For these applications, population in a grid cell must be considered to find representative weather as it affects buildings or population. We use a grid cell weighting $\frac{p_g}{p_.}$ of the population of the grid cell g divided by the population of an entire spatial region (*p*). This gives population centers more weight in determining which reference grid cells are used to create climate partitions (i.e., climate-similar regions). The minimum CSS is an absolute value difference. As such it has analogous quantities to mean absolute error and root mean square error. The mean score (MS) for the climate partitions *P(g,**r**)* is:

$$MS(P,\boldsymbol{r},G) = \sum_{g\in G} \frac{p_g}{p_.}\left[\min CSS_{\boldsymbol{r},P}(g)\right].$$

Where p_g is the population and MS measures goodness-of-fit of a climate partition for grid cells weighted by population given a reference set $\boldsymbol{r}$. In the next section, we omit *P* because it is clear from context that the partition is created around one particular set reference grid cells. We use this score when comparing our results to IECC climate zone data, because those results are developed with an emphasis in optimizing climate similarity where buildings are located. For example, Briggs and co-workers write, "… a representative city should, to the extent possible, favor weather conditions where buildings are located" (Briggs et al. 2003). If we did not weight by population, we would not be considering and relating to existing work fairly. For contrast, we do consider results in which no population weighting occurs, and only the minimum CSS(g) dictates how the reference and partition is scored.

Best Reference Cells within Assigned Partitions (Two-Stage Process)

Sometimes the partitioning *P* is assumed to be fixed and immutable because of outside constraints. We demonstrate an example of this with P_{IECC} (the IECC climate regions); within $p \in P_{IECC}$ (a select IECC climate zone) we compare the optimal reference grid cell to the published guidance (Briggs et al. 2003) using the detailed CSS data derived as described above using the NLDAS-2 data. This becomes a two-stage process in which partitioning is done first and then reference grid cell assignments are made. As we will illustrate later, performing this operation in two parts has drawbacks.

Best Representative Subset

Among all possible picks for the reference sets ***r***, there is a best set that minimizes *MS(**r**,G)*. As before, let ***r*** be a subset of reference grid cells with NSRDB Class I through III weather stations (R). Let *G* be a subset of grid cells for the contiguous United States or other region-of-interest within the NLDAS-2 grid. The power set $\mathcal{P}(R)$ is all possible subsets of R. For convenience, let $\mathcal{P}_K(R)$ be all possible subsets of size *K* from R.

The reference subset of R with the minimum possible population-weighted climate scores over the best possible CSS for each grid cell is:

$$\mathcal{B}_K(R,G) = \underset{\boldsymbol{r}\,\in\,\mathcal{P}_K(R)}{argmin} MS(r,G)$$

$\mathcal{B}_K$ is the reference subset that provides the lowest weighted climate similarity given that it must be of size *K*. Finding $\mathcal{B}_K$ is computationally infeasible except when *K* is near 1 or near *R*-1 (e.g., when selecting the best reference for an IECC climate zone or removing the worst). The cardinality of $\mathcal{P}_K(R)$ is $\binom{|R|}{K}$. For example, in the case of selecting the best 14 reference grid cells for the contiguous entire United States, as we do in the results section, there are 1290 weather stations that identified our reference grid cells, and we select subsets of size *K*=14, so $|\mathcal{P}_{14}(1290)| \approx 3.776625e+32$.

4 Approximating Best Representative Subsets

Because it is often computationally infeasible for us to find the best representative subset by exhaustive search, we consider two approaches to find optimal representative subsets. The first method is a two-stage process that leverages K-means clustering, and the second method is a novel algorithm we have named Monte-Carlo Sifting (MC-Sifting).

K-Means Clustering

For the K-means clustering application, we use a three-step process to define the climate-similar regions and reference grid cell. In the first step, we cluster grid cells ($g \in G$) using the CSS of each grid cell to all reference grid cells ($r \in R$). Grid cells from approximately the same spatial locations have vectors of CSS to reference grid cells that are nearly the same. In this way, each **⅛**-degree grid cell has its weather described not by direct observation, but by similarity to all the reference grid cells (i.e., weather stations) in the United States. Using this clustering, we construct meaningful groupings of grid cells that all have similar relationship to the weather observed at reference grid cells. We used K-means, with a Euclidean distance, as the clustering algorithm. K clusters of grid cells are the result of this step. These clusters form a partition for which no reference grid cells have yet been selected, and it is not the final P in (P, **r,** G). We can refer to each cluster of cells within G as $c_k$with k in 1 to K.

Next, we consider each cluster alone, and in turn to, find the reference r in R such that $MS(r, c_k)$ is minimized. Individually these minimum correspond to c_k, and collectively in order they are $\boldsymbol{r}$ of (P, **r**, G). Thus, the clustered organization of G is leveraged to select reference grid cells. We can label these $\boldsymbol{r_{clustering}}$ to be very specific.

Finally, we construct P, the partitioning of G that will minimize the CSS given the references $\boldsymbol{r_{clustering}}$. For each $g \in G$, we calculate P(g,$\boldsymbol{r}$) as in the "Climate-Similar Partitions" subsection. Each grid cell is assigned to the reference to which is has the minimum CSS. That is to say, for each $g \in G$ we find:

$$P(g, \boldsymbol{r}) = \underset{r_i \in \boldsymbol{r_{clustering}}}{argmin} CSS(g, r_i).$$

We can label the resulting partition $P_{clustering}$, and the result of this algorithm is a complete specified ($P_{clustering}, \boldsymbol{r_{clustering}}, G$). This approach does not leverage the population weighting and there is no evidence provided that this set of references is approximately the best representative subset.

Monte Carlo Sifting

We now introduce a novel algorithm called MC-Sifting to find reasonable best representative subsets $\mathcal{B}_K$. This algorithm leverages structure and concepts of a greedy search algorithm (Cormen et al. 2001) and has some similarity to the multi-armed bandit problem (Gittins 1979). We consider diagnostics for the performance of this algorithm in the results section. MC-Sifting can be applied to any scenario in which there is a scoring function and a power set that must be searched to find an optimally scoring subset.

MC-Sifting is the iterative process of sifting through a set of Monte-Carlo runs to identify a candidate and dropout reference grid cell until an optimal subset is selected. In general, the algorithm repeatedly draws random samples from available subsets of reference grid cells in an iterative process. At each iterative step, the Monte-Carlo selection of reference grid cells is evaluated to select and remove a reference grid cell based on the contribution of each grid cell in returning optimal scoring subsets. The candidate reference grid cell is fixed in the remaining iterative steps, and the dropout reference grid cell is removed from further sifting (Monte-Carlo runs). As the MC-Sifting process completes the needed steps to select the necessary number of reference grid cells (K), the algorithm is more selective about the set of reference grid cells drawn during the Monte-Carlo runs. At the end of the MC-Sifting process, the algorithm returns the reference grid cell set that has the optimal score.

Algorithm 1 documents the MC-Sifting process previously described. The Monte-Carlo-Sifting process relies on two supporting processes labeled Monte-Carlo-Runs and Rank-Evaluate for selecting candidate and dropout reference grid cells. Monte-Carlo-Sifting and Monte-Carlo-Runs are straightforward and follow the previous description of MC-Sifting. The Rank-Evaluate process is specific to our application and would vary depending on the application of MC-Sifting. Rank-Evaluate is applied over the set of Monte-Carlo runs at each iterative step. In general the reference grid cell that most often appears in 200 best performing $MS(\mathbf{r},G)$ values becomes the candidate reference grid cell. Similarly, the reference grid cell that most often appears in 200 worst performing $MS(\mathbf{r},G)$ values becomes the dropout grid cell.

```
Monte-Carlo-Sifting(R, K):
        C = R
        B_K = ∅
        min_set, min_score = Inf
        while (|C| + |B_K| ≠ K And |B_K| ≠ K):
                M := Monte-Carlo-Runs(C, B_K, K − |B_K|)
                best, worst := Rank-Evaluate(C, M)
                drop best and worst from C
                add best to B_K
                If a score in M is lower than min_score it becomes the min_set with the min_score
        return min_set

Monte-Carlo-Runs(C, B, K):
        set = []
        score = []
        for i in 1 to N:  #N is the number of monte carlo runs usually quite large
                select at random r ∈ P_K(C)
                u = r ∪ B
                set[i] = r
                score[i] = Score(u)  #Score in our application is MS
        return set, score

Rank-Evaluate(C, M):
        # Rank-Evaluate is very application specific.
        # Returns best and worst of C given Monte Carlo scores M
        set, score := M
        sort set by ascending score
        for t in 1 to 200:
                for e in C:
                        p_1[e] = proportion of time e is in a top t set
                        p_2[e] = proportion of time e is in a bottom t set
                        keep_score[e] = p_1(1 − p_2)
                        drop_score[e] = p_2(1 − p_1)
                various_worst[t] = which e has highest drop_score
                various_best[t] = which e has highest keep_score
        worst = which e occurs most often in various_worst
        best = which e occurs most often in various_best
        return best, worst
```

Algorithm 1. Monte-Carlo sifting for best representative subset. For each iterative step, MC-Sifting decides which reference grid cells have performed the best and worst in the random selections. The best performing reference grid cell is fixed in the solution set B_k, and the worst performing reference grid cell is removed from further consideration (sifting).

5 Results

Considering the CSS of IECC Climate Regions

IECC uses the climate regions shown in Figure 1 to develop building construction guidelines. In this section, we consider results in which the IECC climate regions are considered immutable partitions of the United States. Inherently, climate regions and reference weather becomes a two-stage process under this guidance. First, IECC-related research documentation suggests climate regions, and then references are selected for these climate regions. Essentially, in terms of the previous methodology, this means the partition (P_{IECC}) and assigned reference grid cells (R_{IECC}) for the United States (G) are dictated by IECC methods as documented. We evaluate CSS performance using (R_{IECC}, P_{IECC}, G), where G is the set of grid cell CSS values documented in this paper.

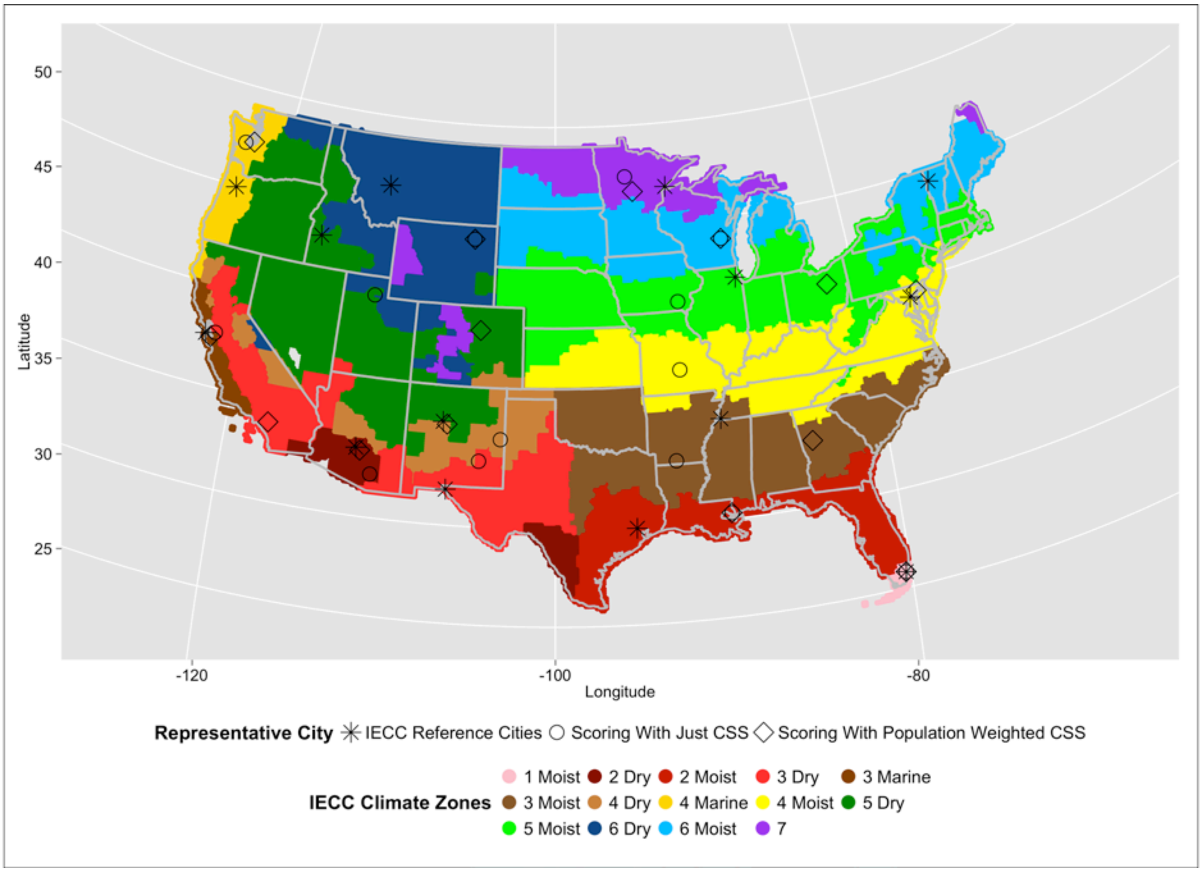

Figure 2 IECC climate zones with optimal reference grid cells identified using IECC literature, minimum CSS and population-weighted minimum CSS.

In general, climate-only scoring tends to disagree with population-weighted scoring. For example, all the eastern United States reference grid cells selected by just climate are toward the geographic center of the climate region; whereas, with population weighting, the eastern seaboard population has a distinct pull. Also, the climate-only scoring selects Tucson as the ideal reference grid cell for climate region 2 Dry (Arizona Desert), while both IECC literature and the population-weighted climate scoring agree that Phoenix area is best.

A limitation of the two-stage procedure also is evident in this result; for example, consider the climate scoring result for regions 4 Dry and 3 Dry. Because within each region, a reference grid cell is selected without any adjustment for what reference grid cell is selected in the other climate regions, the climate only scoring selects two reference grid cells for two regions that are very close to each other geographically and have very similar climates. This does not happen when regions are selected with the reference cities, which is what we recommend and demonstrate how to do in the following section.

Figure 3 focuses on the comparisons to the CSS performance between the IECC reference grid cells and population-weighted reference grid cells. The box-and-whisker plots (Figure 3a) show the CSS performance for all the grid cells in each IECC climate region when using the two different reference grid cells. The diamonds show the population-weighted average CSS score for the climate region using the two different reference grid cells. Figure 3b shows each grid cell and its respective change in CSS when moving from the IECC reference grid cells and the population-weighted reference grid cells where grid cells are still separated by IECC climate regions. Many of the climate regions have similar CSS performances, as the reference grid cells selected are generally the same. Climate region 5 Moist is unique in that some distance separates the two reference grid cells, but the CSS values are generally the same. Climate region 3 Marine has the exact opposite feature. The two reference grid cells are within 50 kilometers of each other, but the change provides a dramatic improvement in the CSS values over the IECC reference grid cells located in South San Francisco. Finally, climate region 3 Dry shows a big improvement in the population-weighted CSS performance when moving from the IECC reference grid cell. This likely is due to the large population in the California portion of the 3 Dry climate region.

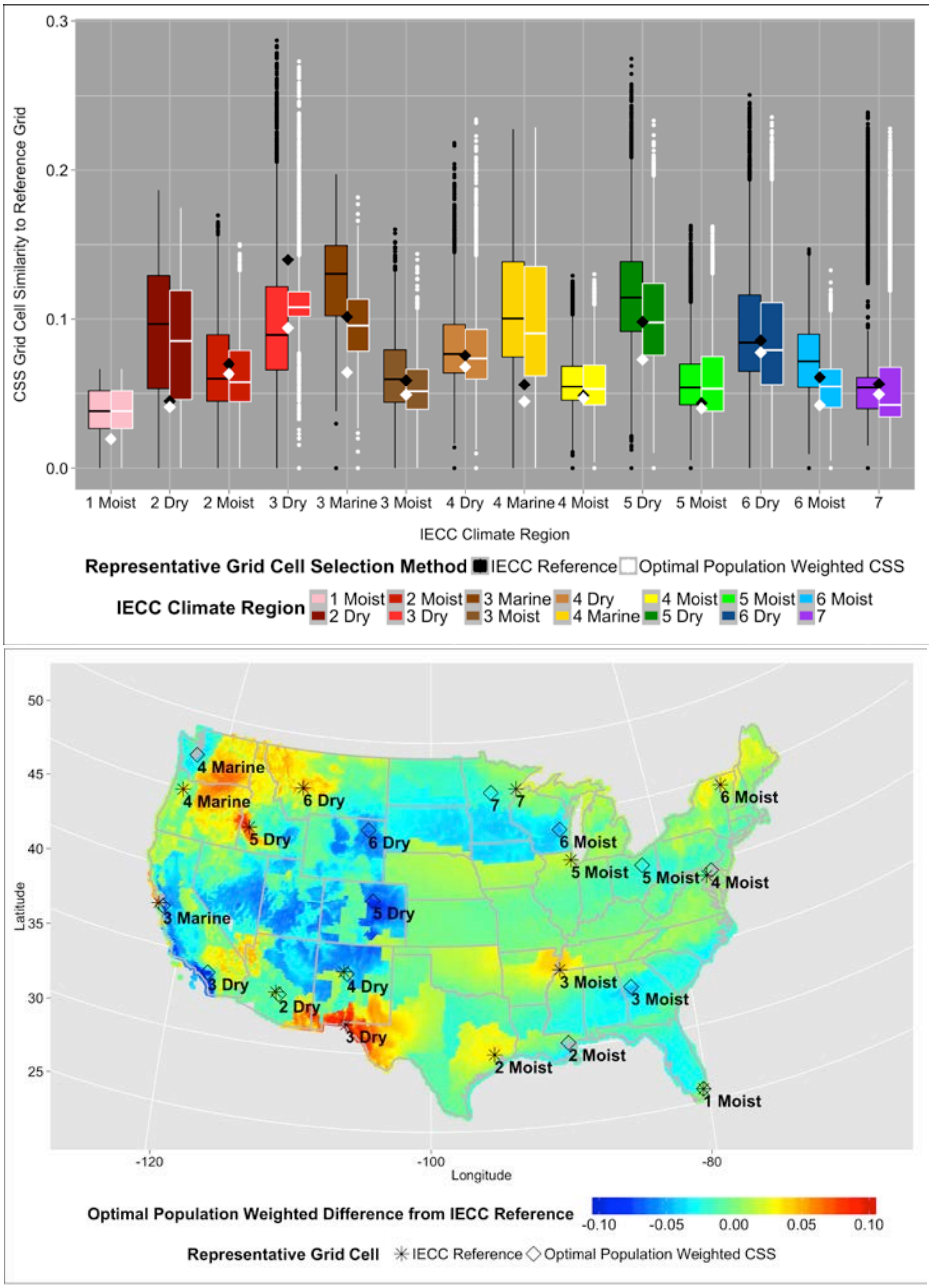

Figure 3. Comparisons of the CSS performance when using the IECC reference grid cell and the optimal population-weighted CSS grid cell. A) Each climate region and the distribution of CSS scores for each reference grid using a box-and-whisker plot. The diamonds locate the population-weighted average CSS for the climate region. B) Spatial map of the differences in CSS to different reference grid cells by climate region – Optimal population weighted CSS reference – IECC grid cell reference performance.

Defining U.S. Climate Regions with MC-Sifting

Much of our emphasis with the MC-Sifting process has focused on picking an optimal subset for the eastern portion of the United States from a smaller set of reference grid cells (~300) or selecting an optimal subset of reference grid cells for a given state. In this application, we demonstrate the use of MC-Sifting to find an optimal set of 14 reference grid cells using the full set of reference grid cells across the contiguous United States. This process will provide results that can be compared to the CSS performance using the 14 IECC climate regions discussed above. For comparison, we discuss the results when using K-means clustering to establish 14 unique groupings across the contiguous United States.

The map in Figure 4 plots all 1290 reference grid cells as the white circles. The CSS performance of each grid cell when using all reference grid cells to create 1290 unique climate regions is shown as well. This map is the basis for the results that are shown in the following sections and has a population-weighted CSS performance of 0.01051.

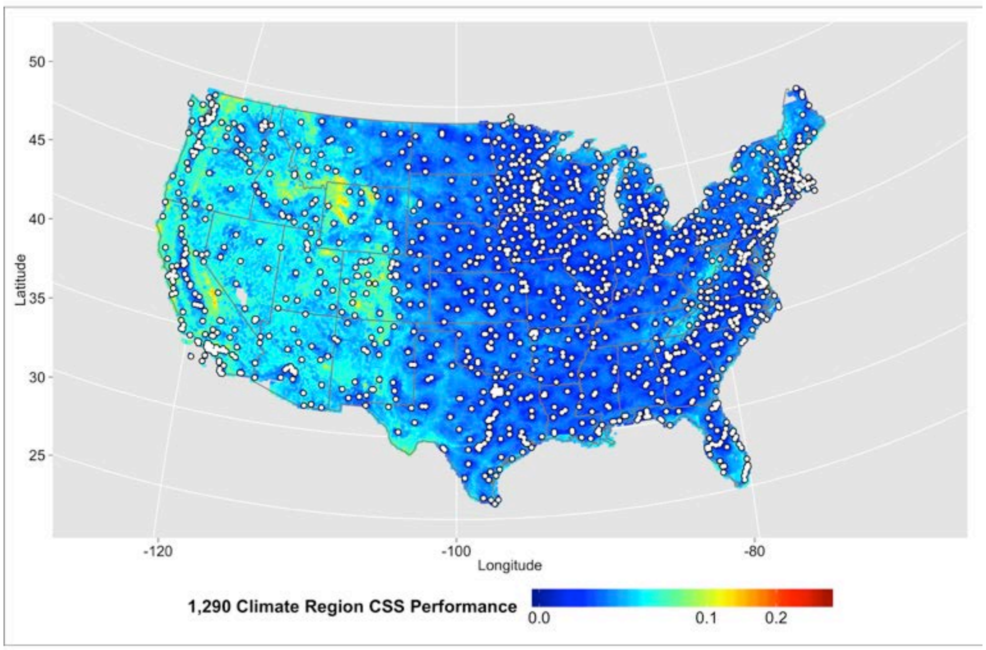

Figure 4. CSS performance when using the complete set of 1290 reference grid cells to represent the climate of the contiguous United States.

Climate Region Comparisons

Figure 5 displays a stacked histogram of the MC-Sifting approach with vertical lines demarcating the performance of the four different methods evaluated to select 14 reference grid cells and regions to represent the climate for the contiguous United States. Each color grouping shows the 100,000 Monte-Carlo runs stacked in order of progression through the first run of the MC-Sifting.

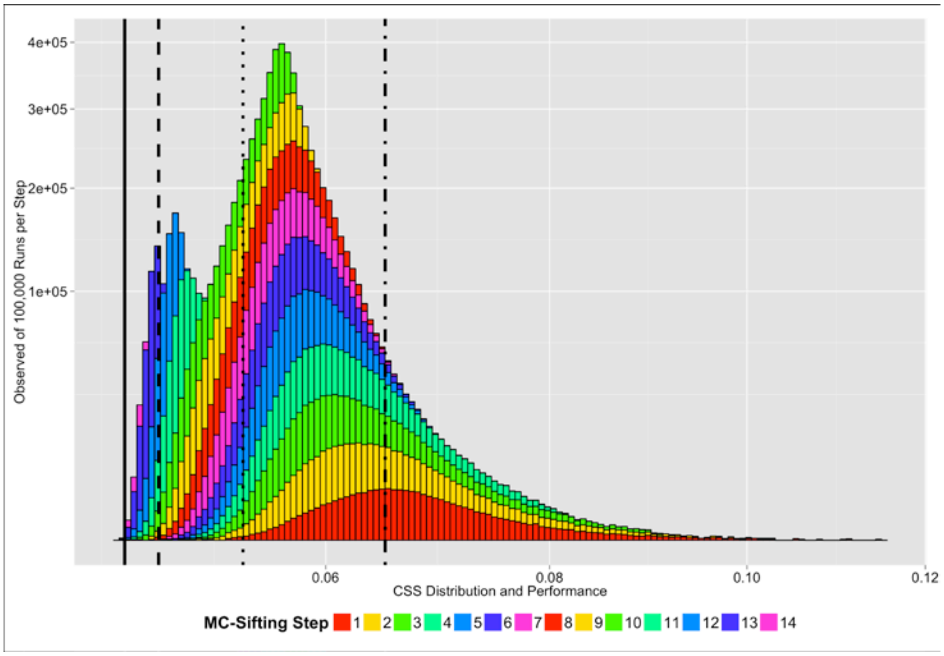

Figure 5. Stacked histogram showing the population-weighted CSS performance for the 100,000 Monte-Carlo selections of 14 reference grid cells at each MC-Sifting step. At each step, a candidate and dropout reference grid cell are selected. The steady shift to the left and reduction in variability at each step demonstrate proper performance of MC-Sifting. The four black lines show the overall population-weighted CSS performance for the four methods discussed, from the left: 1) solid line marks the performance for the MC-Sifting routine (0.04445), 2) dashed line marks the K-means clustering performance (0.4695), 3) dotted line marks the IECC climate zone when using the optimal reference grid cells (0.05334), and 4) dotted dashed line marks the IECC clime zone when using the IECC reference grid cells (0.06503).

The maps in Figures 6 through 8 show the climate region assignments for the K-means regions and two unique runs of the MC-Sifting routine to find the optimal set of 14 reference grid cells. Both runs of the MC-Sifting routine returned approximately the same population-weighted CSS to five decimal points, but had some different regional assignment boundaries. Much of the boundary differences are in regions of the United States where climate differences are low in relative terms. Figure 9 shows the CSS difference comparisons between the two MC-Sifting runs and the relatively low impact in performance among the grid cells that switched climate-similar regions among climate regions 9, 11, 12, and 13. The differences between MC-Sifting run “a” compared to the K-means climate region grouping are shown Figure 10. Comparing the results in Figure 9 to Figure 10 shows that the difference between the two MC-Sifting runs is quite small compared to the difference between the K-means method and the first MC-Sifting run. The K-means method misses the Dallas and Miami reference grid cells and picks two reference grid cells near each other in northern California. The map of the first MC-Sifting run, “a”, CSS performance, is shown in Figure 11 with the optimal reference grid cells shown as white dots. This map highlights the final CSS performance for the optimal set of 14 reference grid cells as selected using MC-Sifting.

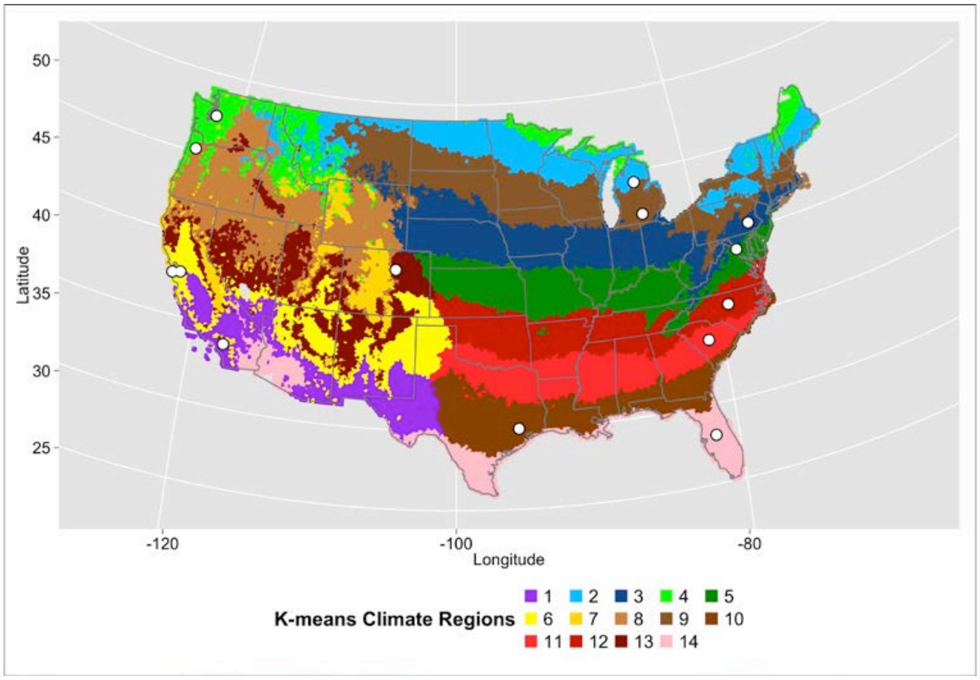

Figure 6. Fourteen climate-similar regions defined when using K-means clustering to create the regions. The U.S. population-weighted CSS performance is 0.04844.

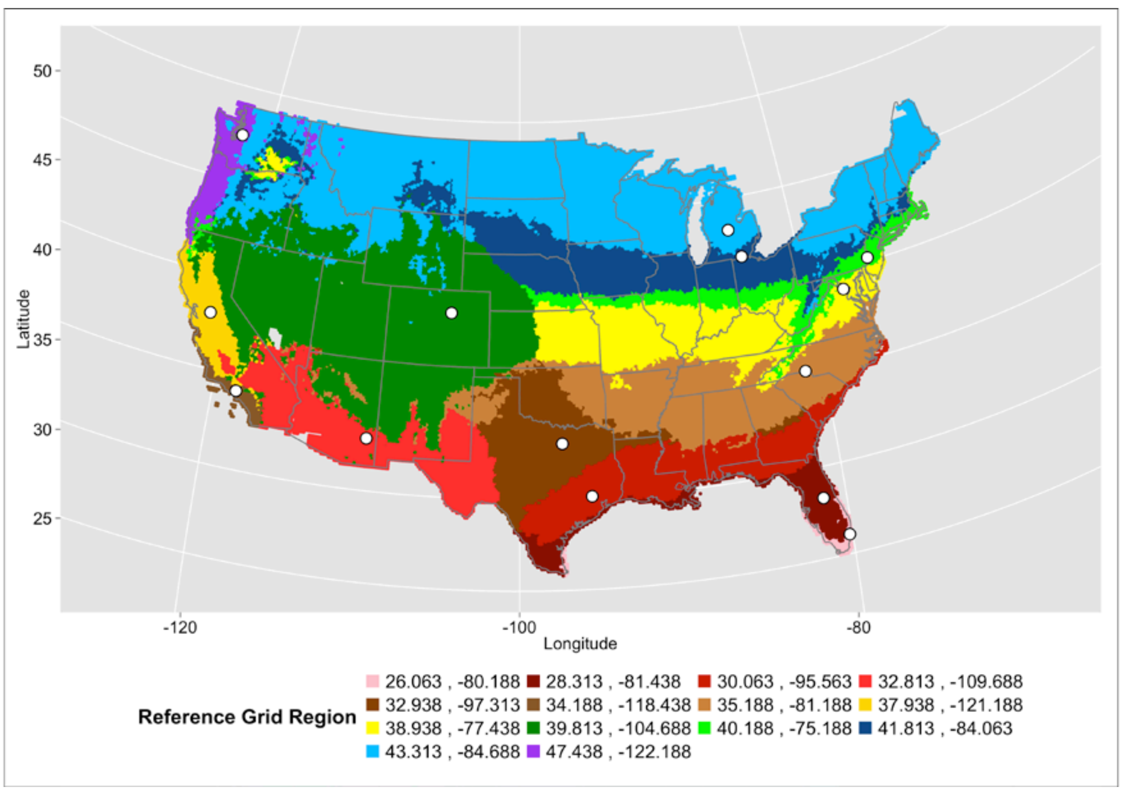

Figure 7. Fourteen climate-similar regions using the first of two MC-Sifting runs (a). The U.S. population weighted CSS performance is 0.044494.

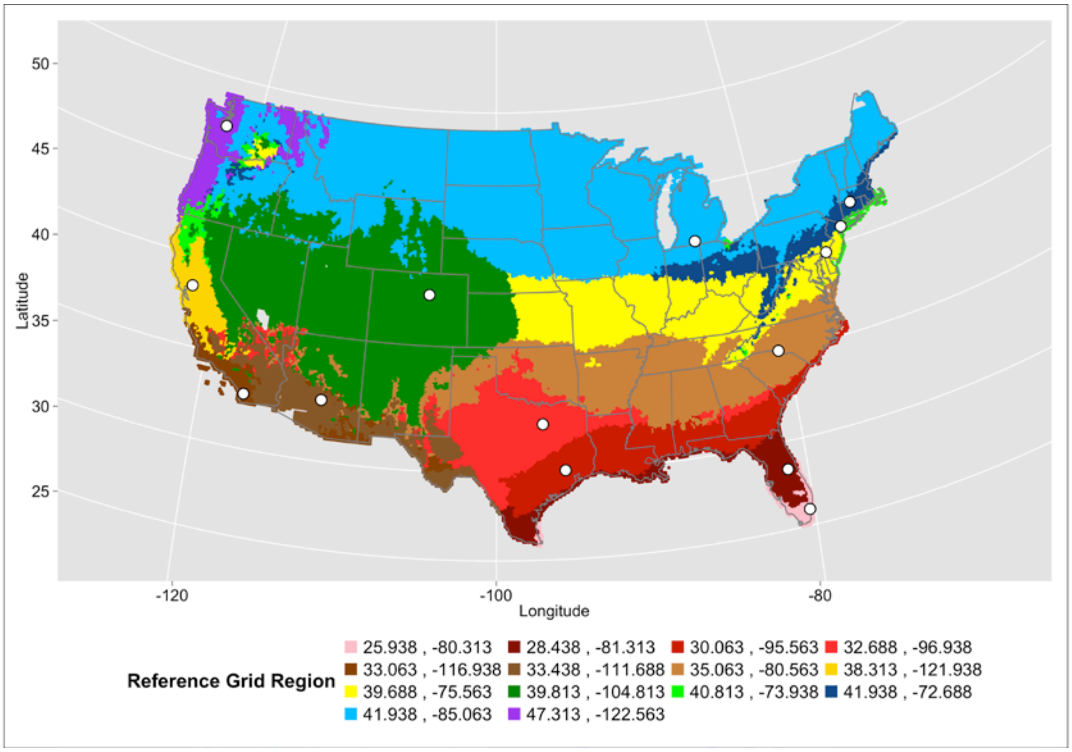

Figure 8. Fourteen climate-similar regions using the second of two MC-Sifting runs (b). The U.S. population weighted CSS performance is 0.04451.

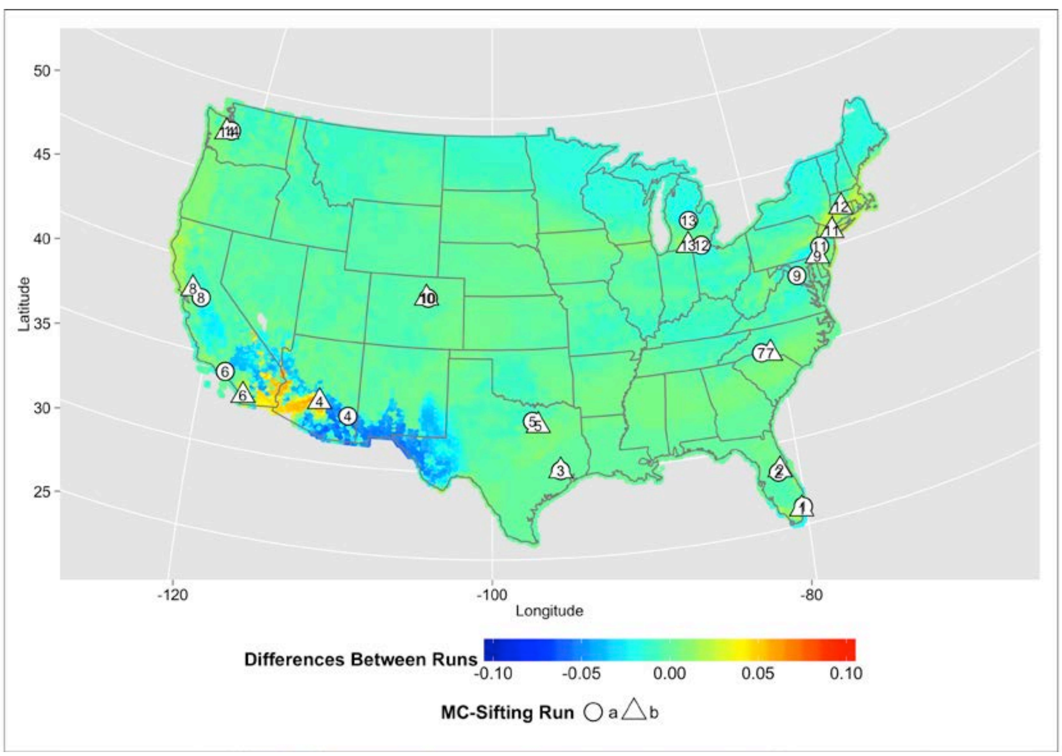

Figure 9. Difference comparison between the two MC-Sifting runs shown above. The reference grid cells selected from each run are displayed. Notice that there is little difference between the two applications for climate groups 9 through13 that show the most drastic changes in the climate region assignments in Figure 7 and Figure 8.

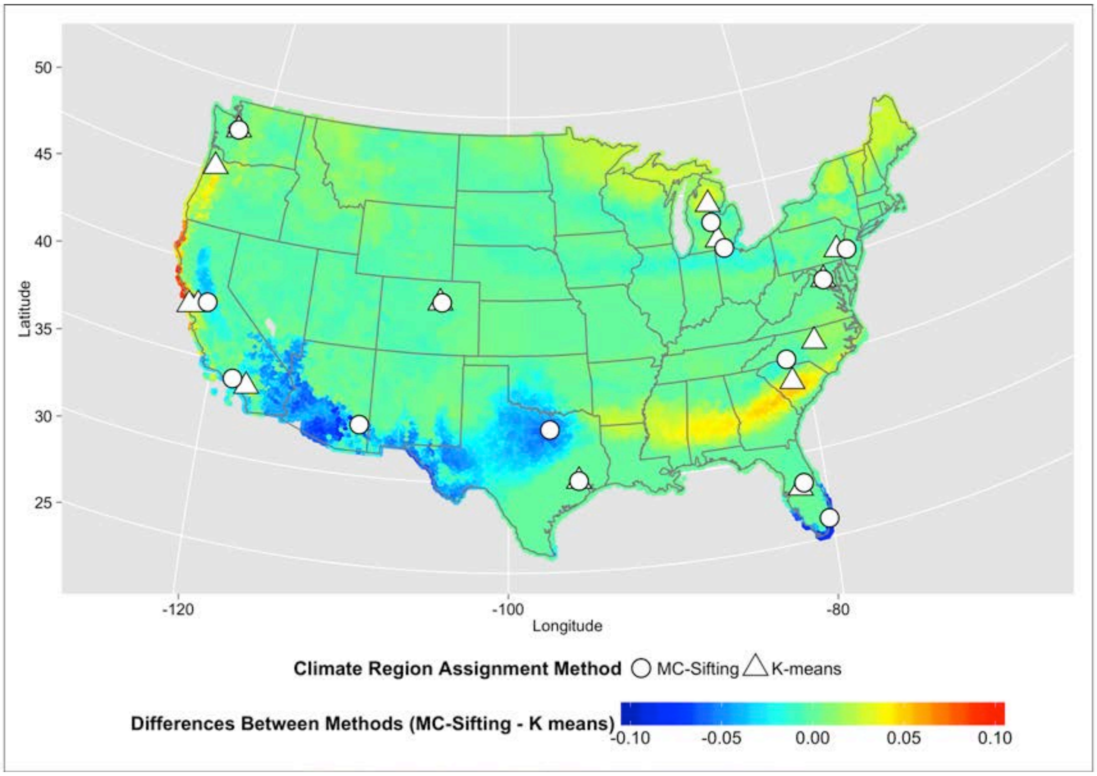

Figure 10. Difference comparison between MC-Sifting run "a" and the K-means method. The reference grid cells selected from each run are displayed. The two large differences, with significant population, in support of MC-Sifting are shown in Dallas and Miami.

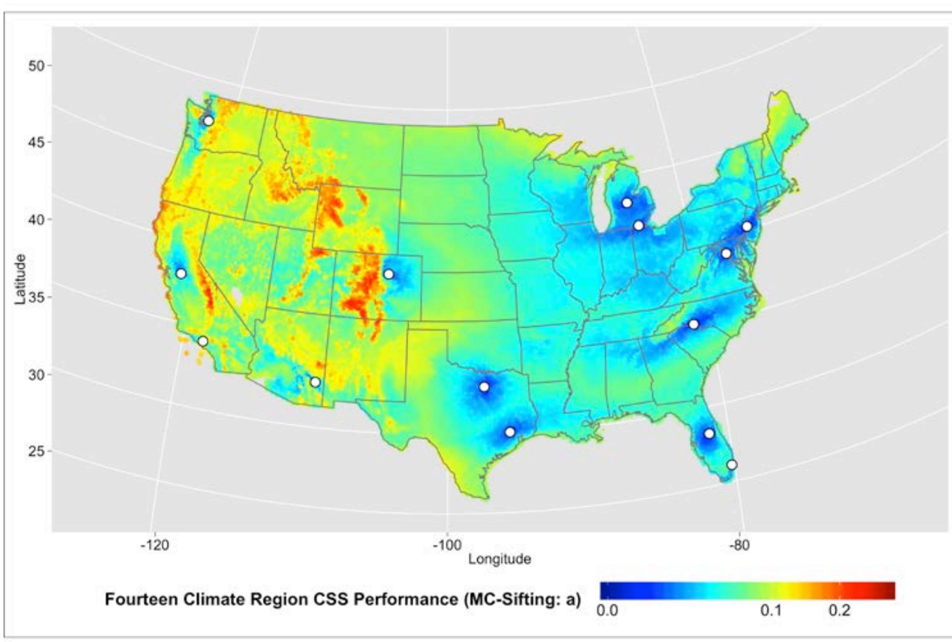

Figure 11. CSS performance when using the optimal set of 14 reference grid cells from run "a" of the MC-Sifting process. The color scale is identical to that in Figure 4.

Conclusions

A CCS is a novel method of combining multiple climate variables of varied measurement types into one unique comparison index. We leveraged previously developed ideas from the work used to create TMY data for use in building energy modeling. Our work integrated nine climate variables across the 12 months of the year, but the methods could easily be applied to just one of the variables and/or some of the months. We also developed climate-similar regions with the assumption that each of the 12 months in a year was equally important, but expect that different applications would weight certain months more heavily in defining climate-similar regions (e.g., interest in summer heat waves).

The 14 climate regions defined by the IECC were developed using heating and cooling degree-day data. The IECC also forces a separation between the humid and dry portions of the United States, which our current variable weighting did not emphasize (see Table 1). The comparison of the IECC reference grid cells to the population-weighted CSS reference grid cells within each respective climate region would not be affected by this difference. Thus, we propose that this improved set of IECC reference cities be used for future energy modeling. However, moisture importance could account for some of the differences in the comparisons between the MC-Sifting selections of 14 regions as compared to the IECC climate regions. The variable weighting used in our analysis (see Table 1) gave humidity (wet bulb) a 10 percent relative impact; if humidity is as important as is suggested by the forced humidity separation represented by the IECC climate types (marine, moist, dry), one could easily rerun the analysis with an increased emphasis associated with the wet bulb variables.

The MC-Sifting routine provides a new approach to finding an optimal set of reference grid cells. In many applications the routine will find the best set without the need of evaluating all possible combinations. For the extreme cases shown in the paper, the MC-Sifting routine outperforms K-means clustering (see Figures 5 and 10) and shows similar CSS performance results across different runs. For future work, we believe the selection of the number of Monte-Carlo runs at each step could be optimized to improve the opportunity of finding the minimum set. Some improvements to the candidate and dropout selection algorithm could be examined for improved performance. One of the best options, which we leverage in other work, is to reduce the set of candidate reference grid cells. Removing many of the reference grid cells that are collocated and have similar climate patterns (possibly by K-means clustering) also could be a viable approach.

Acknowledgements

With the advancements in computing and data storage, we can tackle much larger problems than previous work in this area could achieve. The process to calculate the CSS values and run the MC-Sifting routine does require access to high-performance computing, and we thank the PNNL institutional computing team for their support. The research described in this paper was conducted in support of the Platform for Regional Integrated Modeling Assessment, which is a Laboratory Directed Research and Development initiative at PNNL. PNNL is a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

References

Andersson, Brandt, William L. Carroll, Marlo R. Martin, 1986: Aggregation of U.S. Population Centers Using Climate Parameters Related to Building Energy Use. J. Climate Appl. Meteor., 25, 596–614. doi: http://dx.doi.org/10.1175/1520-0450(1986)025<0596:AOUPCU>2.0.CO;2

ASHRAE. Energy Standard for Buildings Except Low-Rise Residential Buildings. 2007.

Briggs, R. S., R. G.Lucas, and T. Taylor. 2003. “Climate Classification for Building Energy Codes and Standards: Part 2-Zone Definitions, Maps, and Comparisons, Technical and Symposium

Papers." ASHRAE Winter Meeting, Chicago, Illinois, January 2003. Atlanta, Georgia: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Cormen, T. H., C. E. Leiserson, R.L. Rivest, and C. Stein. 2001. *Introduction to Algorithms*. Second Edition, MIT Press and McGraw-Hill, Cambridge, Massachusetts. ISBN 0-262-03293-7.

Finklestein, J. M., and R.E. Schafer. 1971. "Improved Goodness-Of-Fit Tests." *Biometrika* 58(3):641.

Gittins, J C. 1979. "Bandit Processes and Dynamic Allocation Indices." *Journal of the Royal Statistical Society Series B (Methodological)* 41(2): 148–177.

Hall, I.J.; Prairie, R.R.; Anderson, H.E.; Boes, E.C 1978. *Generation of Typical Meteorological Year for 26 SOLMET Stations.* SAND 78-1601, Sandia National Laboratory, Albuquerque, New Mexico.

Kalamees, T, K Jylhä, H Tietäväinen, and J Jokisalo. 2012. Development of "Weighting Factors for Climate Variables for Selecting the Energy Reference Year According to the EN ISO 15927-4 Standard." *Energy and Buildings* 47:53-60.

Mitchell, K. E., D. Lohmann, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, B. A. Cosgrove, J. Sheffield, Q. Duan, L. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, D. P. Lettenmaier, C. H. Marshall, J. K. Entin, M. Pan, W. Shi, V. Koren, J. Meng, B. H. Ramsay, and A. A. Bailey. 2004. "The Multi-Institutional North American Land Data Assimilation System (NLDAS): Utilizing Multiple GCIP Products and Partners in a Continental Distributed Hydrological Modeling System." *Journal of Geophysical Research* 109: D07S90, doi:10.1029/2003JD003823.

PRIMA. http://prima.pnnl.gov/

Thevenard, D., and A. Brunger. 2001. *Typical Weather Years for International Locations*. ASHRAE Research Project, 1015-RP, Ontario, Canda.

Wilcox, S., and W. Marion. 2008. *User Manual for TMY3 Data Sets.* NREL/TP-581, National Renewable Energy Laboratory, Golden, Colorado.