World Climate Search and Classification Using a Dynamic Time Warping Similarity Function

Pawel Netzel and Tomasz F. Stepinski

Abstract We present a data-mining approach to climate classification and analysis. Local climates are represented as time series of climatic variables. A similarity between two local climates is calculated using the dynamic time warping (DTW) function that allows for scaling and shifting of the time axis to model the similarity more appropriately than a Euclidean function. A global grid of climatic data is clustered into 5 and 13 climatic classes, and the resultant world-wide map of climate types is compared to the empirical Köppen–Geiger classification. We also present a concept of climate search—an interactive, Internet-based application that allows retrieval and mapping of world-wide locations having climates similar to a user-selected location query.

Keywords Climate classification · Dynamic time warping · Climate search · Clustering · Similarity

1 Introduction

Global climate classification (CC) schemes discretize a multitude of local climates (LCs) across the earth's land surface in order to identify several typical climate types and to map their geographical extents. Thus, a CC scheme simplifies spatial variability of climates and makes analysis of their temporal change easier. CCs are used to study relationships between climate and other elements of environment, including the biota, hydrology, and agriculture. They also are used to provide visualization of global climate datasets and to illustrate climate change.

Classical CC schemes, such as, the popular Köppen–Geiger classification (KGC) (Köppen 1936; Kottek et al. 2006), rely on heuristic decision rules reflecting a body

P. Netzel · T.F. Stepinski (✉)
Space Informatics Lab, University of Cincinnati, Cincinnati 45221-0131, USA
e-mail: stepintz@uc.edu

P. Netzel
e-mail: netzelpl@ucmail.uc.edu

D.A. Griffith et al. (eds.), *Advances in Geocomputation*,
Advances in Geographic Information Science,
DOI 10.1007/978-3-319-22786-3_17

of environmental and geographical research to identify climate types and to delineate boundaries between them. The KGC has been widely used, and has become a de facto standard for global CC. However, from a modern perspective, the KGC appears to be rather arbitrary and not consistent with today's preference for data-oriented approaches to classification problems.

A data-oriented approach is feasible because of the availability of large, worldwide climate datasets. A climate dataset is a large collection of geographical locations for which climate variables are given. The coordinates of these locations constitute a geographic space; they are not used directly to find climate types. Climate variables constitute a climate space, and are used to find climate types through a clustering process. Clustering divides climate space into mutually exclusive and collectively exhaustive clusters in a way that maximizes intra-cluster homogeneity of LCs and their inter-cluster heterogeneity between cluster exemplars. Cluster exemplars are identified with climate types. Once clusters are found, their members are reassembled in geographic space, resulting in a map of climate types.

Previous studies about global classification of climates via clustering (Zscheischler et al. 2012; Metzger et al. 2012; Zhang and Yan 2014) relied on a generic, not a climate domain-specific, set of clustering techniques that includes: representing LCs by vectors of climate variables, and using Euclidean distance to calculate dissimilarity between LCs. However, a LC is an annually repeating pattern of weather conditions, and thus is more naturally represented by a time series rather than by a vector. A time series representation of a LC takes into consideration month-to-month sequencing information that vector representation does not provide. Given a time series representation of LCs, the dynamic time warping (DTW) distance (Berndt and Clifford 1994), rather than Euclidean distance, is the most logical choice for measuring dissimilarities between two LCs.

In this chapter, we present a new, data-oriented approach to a global CC, one that is based on a time series representation of LCs and on using the DTW as a distance function. We also introduce a new analytical tool—climate search—as a means of studying similarities of climates across the world. Climate search (CS) is an interactive, internet-based application that retrieves locations having climates similar to a user-identified query. It works using a paradigm of similarity search. This paradigm was previously used in a spatial domain for finding patterns of land cover similar to a query pattern (Stepinski et al. 2014), but here we apply it to a temporal domain for finding LCs (time series) similar to a query climate.

2 Data and Methods

In the following sections, we present input data and their relevant pre-processing steps, the method of calculating the similarity of LC and CC, and the method used to compare CC results.

2.1 Data Source

We use a gridded climatic dataset available from the WorldClim project (Hijmans et al. 2005). The following monthly variables are used: air temperature (T), maximum air temperature (T_{max}), minimum air temperature (T_{min}), and total precipitation (P). These are the only WorldClim variables available as monthly variables. All data are long-term averages calculated from measurements taken between 1950 and 2000. The 30 arc-second grid, having spatial extent of (180° W, 60° S) – (180° E, 90° N), is obtained by interpolating measurements from world-wide networks of climate stations. The geographic distribution of locations of these stations can be found in Hijmans et al. (2005). The original reference system of this dataset is WGS84 (EPSG:4326).

2.2 Data Preprocessing

WorldClim data are preprocessed differently, depending on application to either global CC or CS.

For CC, the data first were reprojected to the Mollweide projection (EPSG:54009). This allows clustering of cells having near equal areas (Usery and Seong 2001). We need equal area cells for evaluating similarity between different classifications (Cannon 2012). To reduce the size of the grid, we resampled the Mollweide grid to spatial resolution 75 km × 75 km, resulting in 213 × 482 grid cells, of which 23,979 represent land surface and the rest represent water (nodata).

For CS, the data first were reprojected to the spherical Mercator projection (EPSG:3857). By using this projection, the data and the results can be visualized in a web browser using Google or Bing maps as background for convenient reference. Reprojected data were resampled to a spatial resolution of 4 km × 4 km. This resolution is fine enough to distinguish between LCs, but, at the same time, coarse enough so that the time of search across the entire world is only about 40 seconds per query.

2.3 Variables and Their Normalization

A LC is defined as a 12-month-long time series constructed from climate data at a particular grid cell. The data first are corrected to remove a phase shift caused by sun position change during the year. With this correction, the clustering algorithm can find climate types that occur simultaneously in both northern and southern hemispheres. We use the following three climate variables:

- air temperature, T, a measure of average thermal conditions;
- precipitation, P, a measure of humidity of the climate and
- air temperature range, $dT = T_{max} - T_{min}$, a measure of thermal conditions, variability.

CC and CS can be conducted using either two variables (T, P) or three variables (T, P, dT). The two-variable classification (T, P) adheres to the KGC protocol of describing climate types using variables derivable only from the values of monthly averages of temperature and precipitation. By adding the third variable, the classification also takes into account an in-month variability of thermal conditions.

The ranges of the three selected variables are as follows:

- air temperature: (−50, 40) [°C],
- air temperature range: (0, 25) [°C] and
- precipitation: (0, 1550) [mm/month].

For the variables to contribute equally to the value of dissimilarity between two LCs, they need to be normalized to a common range of [0,1]. This is a standard way to normalize climate variables (see, e.g., Zhang and Yan 2014); we refer to such normalization as "global" and denote it by the letter g. Another problem is the distribution of the values of these variables. In particular, P has a distribution that is highly skewed toward large values. This means that an overwhelming number of normalized values of precipitation are very small, resulting in diminishing influence of P on the overall value of dissimilarity. To prevent this from happening, we introduce a "modified" normalization of P (denoted by the letter l):

$$P \leftarrow \begin{cases} \frac{P}{350}, & \text{if } P \leq 350 \\ 1, & \text{if } P > 350 \end{cases} \tag{1}$$

With this definition, the top 1 % of the highest values of P (≥350 mm/month) are normalized to one, the remaining 99 % have a better behaving distribution of their values, and the influence of precipitation on the value of dissimilarity is restored.

A time series representing a LC is a multivariate time series. Each cell (location) is described by a time series of 12 two-dimensional (or three-dimensional) vectors V_i, i, 1, … , 12:

$$LC = (V_1, V_2, ..., V_{12}),$$

where

$$V_i = \begin{cases} (T_i, P_i), & \text{in the case of two variables} \\ (T_i, P_i, dT_i) & \text{in the case of three variables} \end{cases} \tag{2}$$

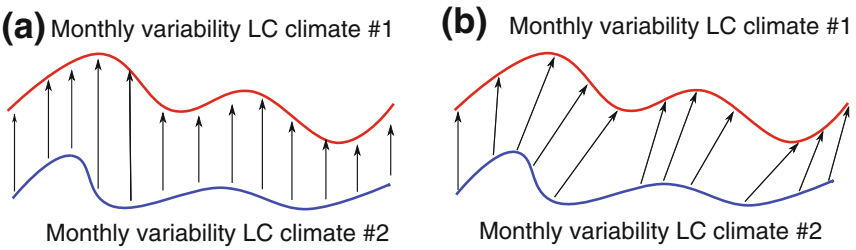

Fig. 1 Two ways of calculating distance between time series: **a** Euclidean distance and **b** DTW

2.4 Dissimilarity Measure

We use the DTW algorithm (Rabiner and Juand 1993) to calculate a dissimilarity (also referred to as a "distance") between any two LCs. DTW is widely used in calculating distances between time series.

The main idea of the DTW is to find an optimal alignment of two time series before calculating a distance between them. This alignment minimizes the sum of distances between the elements of the two series. The difference between Euclidean distance and the DTW distance (Fig. 1) is that DTW synchronizes two time series, thus finding two LCs similar even if there are some time shifts in their variability.

The matrix of distances between two series consists of Euclidean distances between each pair of their constituent vectors. For example, the $\{2, 3\}$ element of such a matrix is the Euclidean distance between a vector of climatic variables characterizing the month of February in the first series, and a vector of climatic variables characterizing the month of March in the second series. The general description of the DTW is as follows: for two time series v_1, v_2 with length N, the DTW algorithm finds alignments $s_1, s_2, ..., s_M$ and $t_1, t_2, ..., t_M$, such that

$$\begin{cases} s_1 = t_1 = 1, & \\ s_M = t_M = N, & \\ 0 \le s_{k+1} - s_k \le 1 & \text{for } k = 1, 2, ..., M-1, \\ 0 \le t_{k+1} - t_k \le 1 & \text{for } k = 1, 2, ..., M-1. \end{cases} \tag{3}$$

and

$$DTW(v_1, v_2) = \min_{s,t} \sum_{i=1}^{M} \|v_{1_{s_i}} - v_{2_{t_i}}\| \tag{4}$$

where $\|.\|$ is the Euclidean distance between the two vectors. Note that the length of the alignment sequence $M \ge N$ because the alignments do not have to be one-to-one. Thus, DTW finds a minimum value path crossing the matrix of distances from the

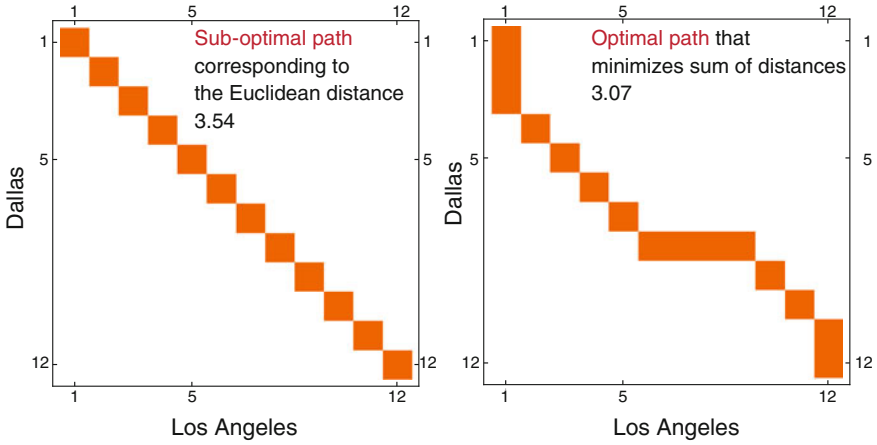

Fig. 2 Paths of summation of distances in a distance matrix: Euclidean distance (*left*) and the DTW distance (*right*)

upper left-hand corner to the lower right-hand corner of the matrix. This is illustrated in Fig. 2, with the left panel showing a path corresponding to the Euclidean distance (which can, but does not have to, be minimal), and the right panel showing the minimal path corresponding to the DTW distance. Finding an optimal time warping path is calculated efficiently using dynamic programming (Rabiner and Juand 1993).

Figure 3 shows an example of two locations for which a value of dissimilarity between their climates is very different, depending on whether it is calculated using Euclidean distance or the DTW. Two-dimensional time series are shown as an illustration. Each time series is depicted as a three-dimensional curve showing dependence of temperature and precipitation on time (month). The curves in both panels are identical, but the dotted lines, which show pairings of months used to calculate

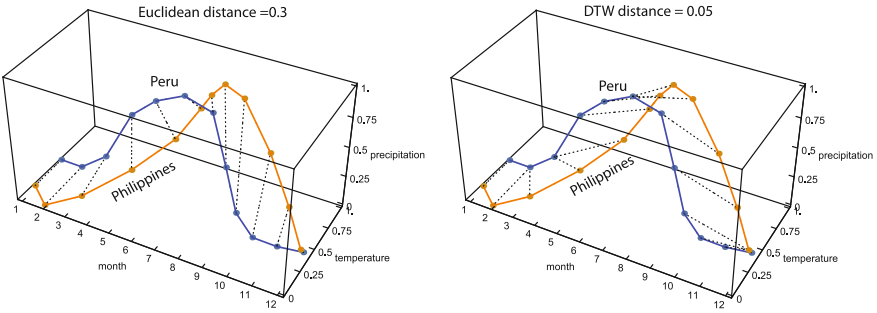

Fig. 3 Example of climates at two locations for which the value of dissimilarity calculated using the DTW measure is significantly different from the value of dissimilarity calculated using Euclidean distance

a dissimilarity value, are different. In the case of Euclidean distance, corresponding months are compared, for example, January to January, whereas in the case of the DTW, non corresponding months can be compared. As a result, the DTW dissimilarity is six times smaller than the Euclidean dissimilarity. According to Euclidean distance, the two climates are not very similar because a month-to-month comparison of temperature and precipitation shows significant differences. However, according to the DTW, the two climates are similar because their climatograms (blue and orange curves in Fig. 3 have similar topologies despite differences in a phase and a stretch. Therefore, human perception of climate is expected to be similar in the two locations.

2.5 Clustering Methods and CCs Comparisons

To explore how CC depends on the number of variables, normalization protocol, a dissimilarity function, clustering method, and a number of clusters, we calculate a large number of CCs corresponding to the combination of these parameters. There are two sets of variables ([T, P] and [T, P, dT]), two protocols of normalizations (global and modified), two dissimilarity functions (Euclidean and DTW), two clustering methods (see the following paragraph), and two values for the number of clusters (see ensuing discussion). These result in 32 possible classifications plus the two KGC for two different numbers of clusters.

We used two different clustering methods: hierarchical clustering with Ward linkage (Ward 1963) and partition around medoid (PAM) (Kaufman and Rousseeuw 1987). Both methods use only the values of dissimilarities between LCs; thus, the first step in using them is to calculate a $23{,}979 \times 23{,}979$ matrix of dissimilarities between climates in all locations. We do not attempt to find an optimal number of clusters; rather, we assume either 5 or 13 clusters, which corresponds to the number of clusters at the first two levels of the KGC.

To quantify a degree to which two classifications partition the land surface into similar or dissimilar climatic zones, we follow Zscheischler et al. (2012) in calculating an information theoretic index called the V-measure (Rosenberg and Hirschberg 2007), which measures the degree of association between two different ways of delineating climatic zones. Briefly, the V-measure evaluates spatial association between two sets of climatic zones using two criteria: homogeneity and completeness. An association satisfies the homogeneity criterion if for all climatic zones in the first classification, each zone contains only locations that have a single climatic label in the second classification. An association satisfies the completeness criterion if for all climatic zones in the second classification, each zone contains only locations that have a single climatic label in the first classification. Only two identical classifications satisfy completely both criteria. In the case of different classifications, the V-measure calculates a degree to which these two criteria are satisfied. The V-measure

is given by computing the harmonic mean of homogeneity and completeness measures. The range of V is [0,1], with 1 indicating a perfect correspondence between two classifications. We use a value (1-V) to quantify dissimilarity between two classifications (not to be confused with dissimilarity between two LCs).

3 Climate Classifications

We calculate a value of (1-V) between each pair of 34 classifications (32 classifications based on our clustering method and the two KGCs), resulting in a 34×34 matrix of distances between classifications. Figure 4 shows the heat map (Wilkinson and Friendly 2009)—a graphical representation of this matrix with rows and columns rearranged so the classifications most similar to each other are next to each other in the matrix. The black-to-white color gradient indicates values of (1-V) from small (similar classifications) to large (dissimilar classifications). Classifications are numbered from 1 to 34, the first two being the KGCs with 13 and 5 climate types,

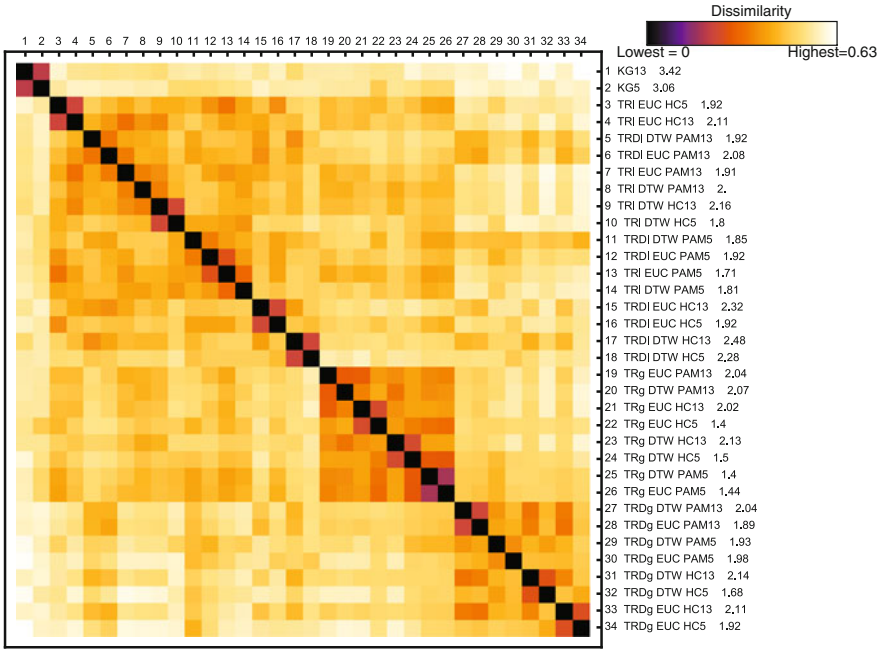

Fig. 4 A heat map illustrating a V-measure-based comparison between 34 different CCs. The color gradient indicates dissimilarities between pairs of classifications, from small to large. The numbers after classification names are the values of the Davies-Bouldin index (see Sect. 3 details)

respectively. The remaining classifications are labeled to indicate the choice of free parameters used; for example, TPdTl DTW PAM5 indicates a classification obtained using variables T, P, dT, modified normalization (l), the DTW dissimilarity function, the PAM clustering algorithm, and 13 clusters. Similar classifications are identified in the heat map as square reddish blocks located on the diagonal.

First, we note that for each protocol (the number of variables, normalization and dissimilarity function) hierarchical clusterings (including KGCs) into 5 and 13 classes are similar. Hierarchical clusterings subdivide more broadly defined climate types (those resulting from dividing a dataset into 5 clusters) into constituent, more narrowly-defined climate types (those resulting from dividing a dataset into 13 clusters). The result of this hierarchy is the high spatial correspondence between partitionings.

Second, we note that, for a given number of variables, classifications obtained using global (g) normalization are similar. This is especially true for classifications based on only two variables. The g normalization reduces the influence of P on the clustering of LCs, resulting in de facto delineation of thermal zones. No grouping is observed for classifications that use "modified" (l) normalization.

Objective determination of the "best" classification is not possible due to lack of validation data. We certainly cannot treat the KGC as the ground truth. Each classification delineates a climate dataset differently, based on the method used. The classification maps reflect definitions of what constitute a climate and how climate similarity is described. Various classifications can be consider the "best" depending on a given point of view or a specific application. From the point of view of homogeneity of climates within climate types, all clusterings are similarly homogeneous and more homogeneous than the KGCs. The degree of homogeneity of climates within a given clustering is measured using the Davies–Bouldin (DB) index (Davies and Bouldin 1979). The values of the DB index are given in Fig. 4 following the names of classifications. The smaller the value of the DB index, the better the homogeneity of climate types in a classification.

All 32 maps of climate types resulting from our classifications are not possible to show here. We have selected only four maps for side-by-side comparison. Figure 5 shows a comparison between the TPdTl DTW PAM5 and the KG5. The KGC shows the five well-established climate types: tropical A, arid B, temperate C, continental D, and polar E. The TPdTl DTW PAM5 shows five unnamed climate types obtained via clustering. A visual comparison of the two maps indicates that climatic types of these two classifications can be matched to each other, although spatial extents of matched types vary.

Figure 6 (top panel) shows a comparison between the TPdTl DTW PAM13 and the KG13. The KGC shows the 13 well-established climate types: tropical rainforest Af, tropical monsoon Am, tropical savanna Aw, arid desert BW, arid steppe BS, temperate dry summer Cs, temperate dry winter Cw, temperate without dry season Cf, continental dry summer Ds, continental dry winter Dw, continental without dry season Df, polar tundra ET, and polar frost EF. The TPdTl DTW PAM13 shows

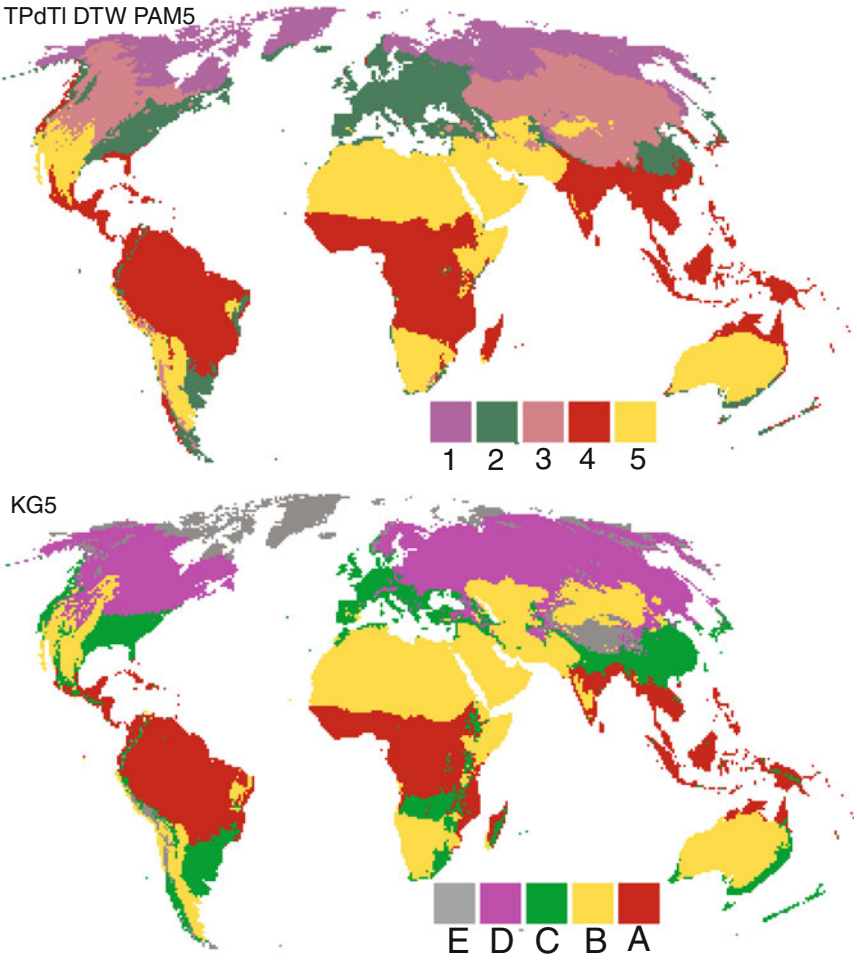

Fig. 5 A comparison of the clustering-based classification (TPdT1 DTW PAM5), having 5 climate types, with the KG5 classification. Each classification has its own legend to stress that cluster-derived climate types may have different meanings from KG climate type

13 unnamed climate types obtained via clustering. Of the 13 types resulting from clustering, 6 can be matched to the KGC climate types: 4→Af, 13→BW, 2→ET, 2→DF, 5→Dw, and 11→BS. The remaining clustering types cannot be matched to a KGC climate type.

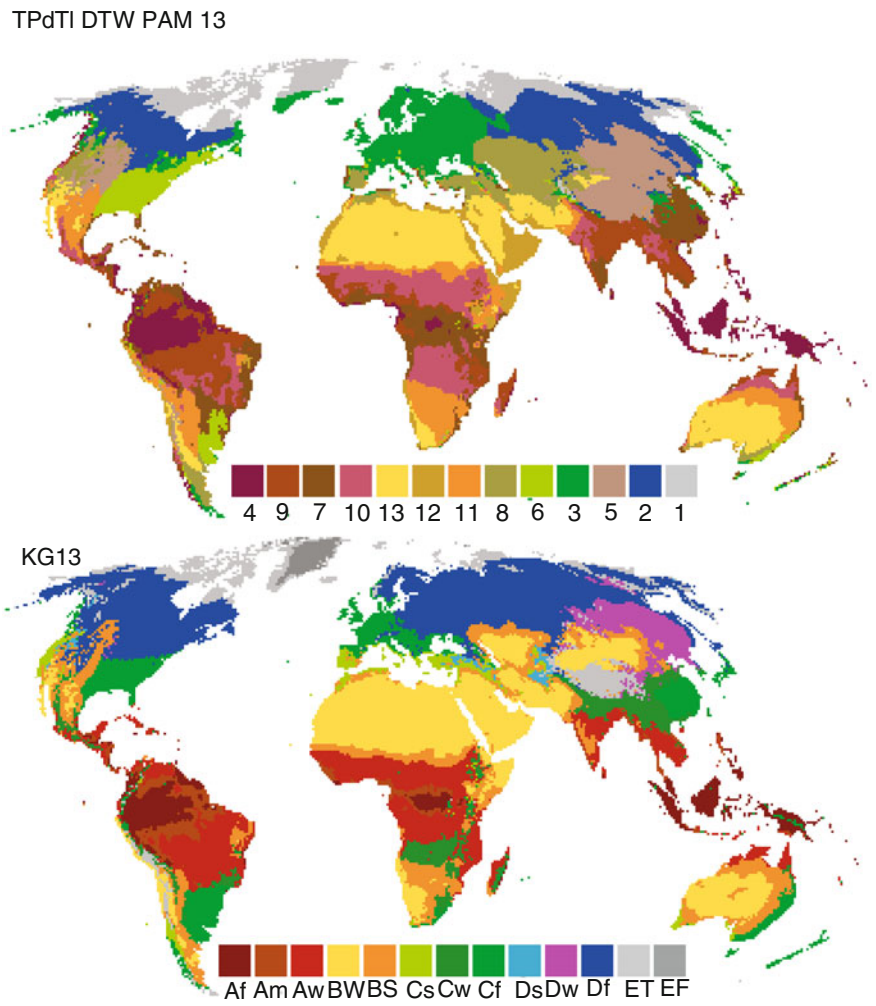

Fig. 6 A comparison of the clustering-based classification (TPdT1 DTW PAM5), having 13 climate types, with the KG13 classification. Each classification has its own legend to stress that cluster-derived climate types may have different meanings than KG climate types

4 Climate Search

CS is an interactive, online visualization tool for finding locations around the world having climate similar to a user-selected query. The main idea of a CS is to select a location, obtain a LC (represented by a time series) for this location, and ask an algorithm to find all locations around the world that have similar climates (on the

basis of the DTW dissimilarity function). Thus, CS works on the principle of query-by-example. A CS uses three climate variables—T, P, dT—and the modified (l) normalization.

We have developed a web application called ClimateEx (Climate Explorer) to enable world-wide CS. ClimateEx is available at (http://sil.uc.edu). The core calculation engine of this application is written in the C language and uses parallel processing. The user interface (Fig. 7) runs in the web browser environment, and is powered by the web-mapping library OpenLayers 3 (Santiago 2015). ClimateEx allows a user to browse a background map, select a location of interest, and calculate a climate similarity map. It also allows a user to download the resulting similarity map in the GeoTIFF format. A single search takes about 40 s.

When working with ClimateEx, a typical work flow is as follows. A user browses a background map (pan and zoom operations) to locate a place of interest and clicks on the map to point to a precise a location that starts the calculations. Once the calculations are finished, a generated similarity map is displayed in the web browser that can be explored using a Bing map in the background as a reference. ClimateEx

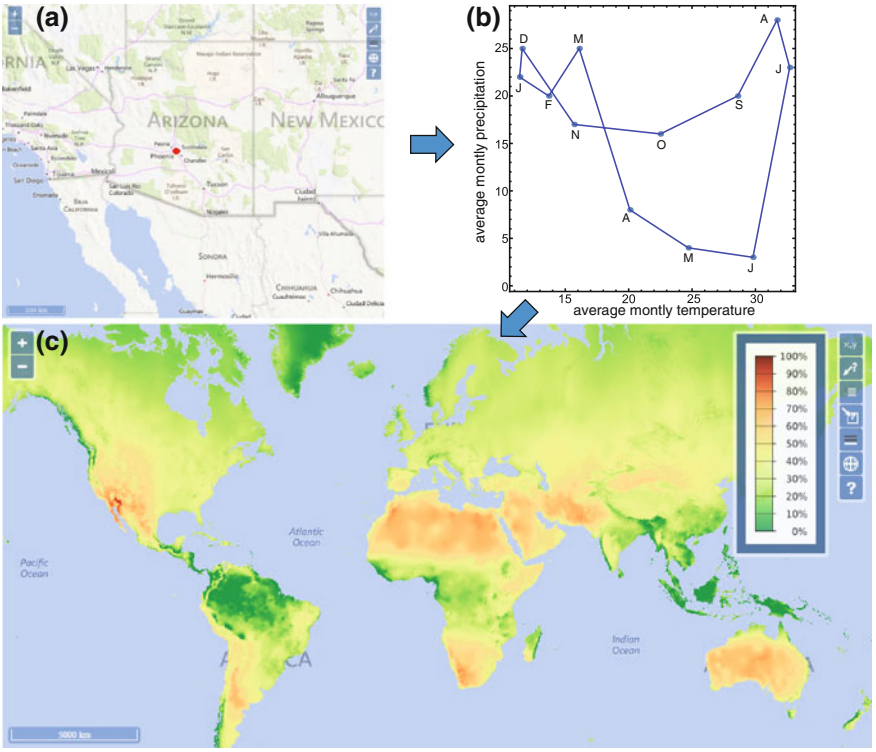

Fig. 7 A search for places with climate similar to that in Phoenix, Arizona. **a** Location of the query climate. **b** Circular climatogram of the query climate; months are labeled by their first letters. **c** A similarity map

can hold more than one similarity map at the same time to compare between different searches.

We demonstrate the utility of CS with two examples. The first example is the location of Phoenix, Arizona (Fig. 7), as a query produced a map of the world showing other locations having climates similar to that in the Phoenix area. A character of the LC in Phoenix (a query) is illustrated by a circular climatogram—a parametric graph with T on the x-axis and P on the y-axis. The closed polyline represents an annual cycle of these two parameters. The climatogram describes Phoenix climate as warm and dry, with the peak P in the summer months (monsoon). A red-to-green color gradient in the similarity map shows the degree of similarity to the query climate that goes large to small. The similarity map reveals that climate in Phoenix is similar (to various degrees) to LCs of North Africa, Australia's interior, the Arabian Peninsula, the Gobi desert, and the Kalahari desert, as well as a LC in Patagonia. These areas also are arid, and their annual cycle of T and P are similar to that in Phoenix except, for a possible time shift.

The second example is the location of Dallas, Texas (Fig. 8). The climatogram indicates that the LC in Dallas is warm and its P is, on average, about five times

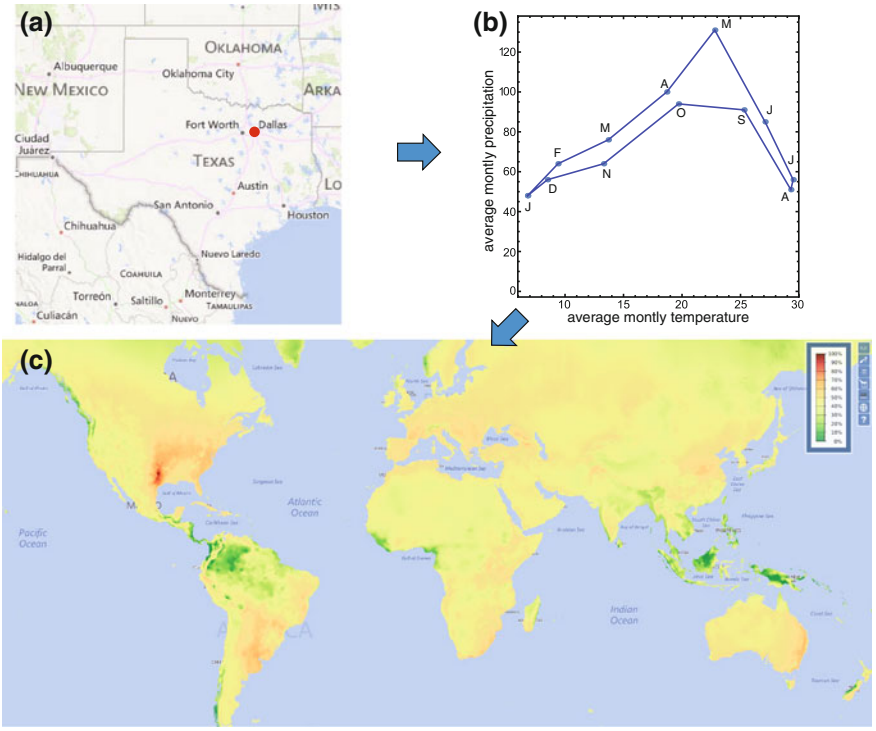

Fig. 8 A search for places with climate similar to that in Dallas, Texas. **a** Location of the query climate. **b** Circular climatogram of the query climate; months are labeled by their first letters. **c** A similarity map

higher then in Phoenix. A significant difference exists in the shape of the climatogram, indicating a different annual cycle of climate variables. The similarity map shows that climates similar to that in Dallas are found across the United States in the region that roughly coincides with "Tornado Alley" and also are found across South America in the region called "El Pasillo de los Tornados". Other regions in the world where similarity is 50 % (denoted by red in the map) also coincide with the occurrence of tornadoes. Users can compare the similarity map with the map of tornado regions at https://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology.

5 Conclusions

Our proposed data-oriented method of CC is a step forward in the effort to study spatial regionalization of world climates. There is no a single "best" classification of climate continuum. An advantage of our method over the KGC is that it is based on a data-driven methodology, and it produces classifications that are customized to a particular task. For example, the KGC uses vegetation as a proxy to discriminate between different climates, whereas our classifications are based purely on climatic data with no ties to vegetation. These differences in approach explain differences in maps of climate types. The different maps being similar is a testimony that various environmental factors (such as, for example, vegetation) are quite correlated with climate.

Within a clustering methodology for global classification of climates, we introduce two innovations: representing LCs as time series, and using the DTW as a dissimilarity function. We submit that such an approach is better suited to compare climates than a standard approach consisting of vector representation and a Euclidean dissimilarity function.

Finally, a similarity search is an original, data-oriented approach to explore and visualize the spatial distribution of climates. It complements classification as a tool to study the spatial distribution of climates. Whereas CC yields an overall but highly simplified view of the spatial arrangement of different climates, the CS provides more narrowly focused, but also more specific information. CS can only be achieved using the notion of climate similarity, and hence cannot be constructed on the basis of the KGC. Note that CS does not just retrieve a few best matches; it produces the entire similarity map showing both similar as well as dissimilar regions. CS can be used for visualization of climate change with a query taken from today's LC and applied to a grid of future climates calculated by a climate prediction algorithm. It also can be used to evaluate the degree of climate homogeneity necessary to maintain a given environment.

Acknowledgements This work was supported by the University of Cincinnati Space Exploration Institute, and by the National Aeronautics and Space Administration through grant NNX15AJ47G.

References

Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. KDD Workshop 10(16):359–370

Cannon AJ (2012) Köppen versus the computer: comparing Köppen-Geiger and multivariate regression tree climate classifications in terms of climate homogeneity. Hydrol Earth Syst Sci 16(1):217–229

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227

Hijmans RJ, Cameron SE, Parra JL, Jones P, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15):1965–1978

Kaufman L, Rousseeuw P (1987) Clustering by means of medoids. In: Dodge Y (ed) Statistical data analysis based on the L1 norm and related methods. North-Holland, pp 405–416

Köppen W (1936) Das geographische system der klimate. In: Köppen W, Geiger R (eds) Handbuch der klimatologie. Gebrueder Borntraeger, Berlin, pp 1–44

Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift 15(3):259–263

Metzger MJ, Bunce RGH, Jongman RHG, Sayre R, Trabucco A, Zomer R (2012) A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. Glob Ecol Biogeogr 22(5):630–638

Rabiner L, Juand B (1993) Fundamentals of speech recognition. Prentice-Hall International Inc

Rosenberg A, Hirschberg J (2007) V-Measure: a conditional entropy-based external cluster evaluation measure. In: Joint conference on empirical methods in natural language processing and computational natural language learning, pp 410–420

Santiago A (2015) The book of OpenLayers 3. Theory and Practice, Leanpub, Victoria, BC

Stepinski T, Netzel P, Jasiewicz J (2014) Landex—a geoweb tool for query and retrieval of spatial patterns in land cover datasets. IEEE J Sel Top Appl Earth Obs Remote Sens 7(1):257–266

Usery EL, Seong J (2001) All equal-area map projections are created equal, but some are more equal than others. Cartogr Geogr Inf Sci 28(3):183–193

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Wilkinson L, Friendly M (2009) The history of the cluster heat map. Am Stat 63(2):179–184

Zhang X, Yan X (2014) Spatiotemporal change in geographical distribution of global climate types in the context of climate warming. Clim Dyn 43(3–4):595–605

Zscheischler J, Mahecha MD, Harmeling S (2012) Climate classifications: the value of unsupervised clustering. Procedia Comput Sci 9:897–906