Delineation of Phenoregions in Geographically Diverse Regions Using *k*-means++ Clustering: A Case Study in the Upper Colorado River Basin

Yuan Zhang,[1] George F. Hepner, and Philip E. Dennison
Department of Geography, University of Utah,
260 S. Central Campus Dr., Rm. 270,
Salt Lake City, Utah 84112-9155

Abstract: Phenological monitoring and modeling over a geographically diverse area has long been a problem due to the spatially variable nature of phenological forcing. Phenoregions, which are phenologically and climatically self-similar clusters, have many potential benefits as a geographic unit for monitoring and modeling of vegetation dynamics. This research develops an improved method to delineate regions of similar phenological forcing in geographically diverse regions, using the Upper Colorado River Basin (UCRB) as a case study. Principal component analysis plus *k*-means++ clustering are adopted to delineate phenoregions in the UCRB, using variables related to elevation, temperature, precipitation, soil, and vegetation history. Raster data at 1 km spatial resolution are used to extract these variables. A series of hierarchical, non-nestable phenoregion maps is generated. The optimal phenoregion map is selected based on spatial homogeneity and spatial concordance with other phenoregion maps generated using different numbers of clusters. This series of phenoregion maps can be considered as a framework for phenological modeling and monitoring, as well as for useful potential vegetation delineation, natural resource conservation, and policymaking.

INTRODUCTION

The delineation of subregions based on their climatic, ecologic and geographic characteristics has been increasingly used for planning, policymaking, natural resource conservation, and management by government agencies and conservation groups (Thompson et al., 2004). Partitioning of regions into functional subregions is dependent on the purpose of the application. Therefore, subregions can take a variety of forms based on the classification logic.

Ecoregions are one type of subregion delineation. The Küchler, Bailey, and Omernik systems are three well-known and widely used ecoregion classification systems generated using different data and classification methods (McMahon et al., 2001; Thompson et al., 2004). The concept of potential natural vegetation (PNV) was introduced by Tüxen (1956) as the vegetation that would exist today if human impacts were removed. The Küchler system (Küchler, 1964) is a potential natural vegetation map

[1]Corresponding author; email: zhangyuan76@gmail.com

GIScience & Remote Sensing, 2012, **49**, No. 2, p. 163–181. http://dx.doi.org/10.2747/1548-1603.49.2.163

of the conterminous United States. The Bailey system (Bailey, 1983) adopted maps of climate, topography, and vegetation to generate ecoregion maps at nine levels of division, while each level is based primarily on one particular map (Omernik, 1987). The Omernik ecoregion system (Omernik, 1987) is based on a combination of four maps: land use, land surface form, potential natural vegetation, and soils. These well-known and widely-used ecoregions are delineated by qualitatively analyzing the homogeneity and generality of each adopted map based on the knowledge and experiences of experts.

Extending on the ecoregion concept, the delineation of subregions has become more function-specific. Agroecoregions (a.k.a. agroecozones, crop growth zones, or soil productivity zones) are generated by delineating subregions of similar expected crop performance. They are used for crop suitability analysis and agricultural policy-making (Williams et al., 2008). The soil map units used in the Natural Resources Conservation Service (NRCS), U.S. Department of Agriculture soil survey are mapped by differentiating the properties of natural bodies of soils and serve as the basic map unit for the widely used STATSGO (State Soil Geographic) and SSURGO (Soil Survey Geographic) databases (NRCS, 2011).

Recently, natural area subregion delineation has focused on phenological processes to yield functional phenoregions. Phenological processes are the relationship between periodic biological phenomena and climatic conditions—how organisms grow and behave in response to environmental conditions (Hodges, 1991). The term "phenoregion" was first defined by White et al. (2005) as phenologically and climatically self-similar clusters. The phenoregion system derived by White et al. (2005) served as a global framework for monitoring phenological responses to climate change.

Geographical diversity, in this research, refers to the inherent heterogeneity of topography, climate, vegetation, and the resulting phenology. Geographical diversity can cause difficulties in monitoring and modeling phenology, as well as in making effective plans and policies in such regions. Therefore, delineation of phenoregions in geographically diverse regions is especially important and necessary.

This research seeks to develop an improved classification of phenoregions for the Upper Colorado River Basin (UCRB)—a geographically diverse region, with a vast range and various patterns of topography, climate, vegetation, and phenology. Thus, a methodology developed for the UCRB could be adapted to areas of complex topography and climate interactions in other regions of the world. Multivariate clustering generates clusters as quantitative subregions based on multiple variables such as temperature, precipitation, and elevation, depending on the function of the clusters (Hargrove and Hoffman, 2004; White et al., 2005). Multivariate clustering has been demonstrated to be effective for subregion delineation at different spatial scales. In this research, principal component analysis (PCA) combined with improved *k*-means clustering (*k*-means++ clustering) is used to generate phenoregion maps of the UCRB. The number of clusters is varied, and the results are compared using a set of evaluative criteria to determine the optimal classification from this series of phenoregion maps. This research is expected to contribute to subregion delineation by demonstrating quantitative regional classifications within a geographically diverse area, and to phenological study by exploring the clustering of phenological forcing variables, including climate, topography, and soils.

METHODOLOGY

The geographical diversity of the UCRB necessitates the delineation of phenology-based subregions. This research identified and analyzed geographic variables related to climate, elevation, topography, and vegetative history to generate phenoregions having similar phenological forcing. Because the UCRB is a highly diverse region, a quantitative approach (PCA plus *k*-means++ clustering) and high spatial resolution data were used to yield increased homogeneity within each defined phenoregion.

In this section, the study area and its geographical diversity are introduced. Variables used for phenoregion classification and data that support these variables are identified. Lastly, a set of evaluative criteria is discussed to determine the optimal phenoregion map from a series of hierarchical phenoregion maps with different numbers of clusters.

Study Area

The UCRB is within portions of the states of Arizona, Colorado, New Mexico, Utah, and Wyoming, which contain the watersheds draining into the Colorado River system above Lee's Ferry (Fig. 1). The UCRB has a drainage area of approximately 280,000 km^2 (Kenney et al., 2009). The largest land uses in the UCRB are rangeland and forest. The elevation in the UCRB varies from east to west with an approximate range of 1000 to 4000 meters. The UCRB has very complex topography and geomorphology. Topographic differences lead to a diversity of climate in this region, including both alpine and semiarid/arid conditions. Precipitation ranges from more than 1000 mm per year in the Rocky Mountains on the east side of the basin to less than 250 mm per year in the west. Different parts of the region have different precipitation seasonality—for example, mountainous areas in the east receive most of their precipitation as snow in winter whereas the high plateaus in the west receive most of their precipitation during monsoon rainfall in August. Mean annual temperatures within the UCRB range from 0.4°C to 12.3°C (USGS, 2006).

Variables and Data for Phenoregion Delineation

Several important features are required for data used for phenoregion delineation. Data should have the spatial resolution suitable for the UCRB region, specifically, fine enough to differentiate phenological forcing and coarse enough to avoid high computational costs. Taking into consideration the area of the UCRB—280,000 km^2—raster data at 1 km spatial resolution were used. Second, data should be from published sources and should have had their accuracy assessed, if possible. Lastly, data should be accessible and their geographic extent should cover the entire study area.

Optimal delineation results require the identification of variables having primary influence on vegetation phenology. Table 1 lists 11 variables included in phenoregion delineation related to temperature, precipitation, elevation, soil fertility, and vegetation. Plant phenology and species distribution patterns have been demonstrated by multiple studies to be strongly affected by elevation (Campbell, 1974; Schuster et al., 1989). Distribution of different species can easily be observed across elevation gradients. The UCRB has an elevation range of about 3000 m, making elevation an especially

Fig. 1. The UCRB study area and the location of UCRB within the conterminous United States.

important input variable in this research. The vegetation cover in the UCRB changes gradually from barren systems above 3500 m, to conifer- and deciduous-dominated forest at around 2500–3000 m, to sagebrush-dominated shrubland below 2000 m. The elevation variable is derived from the Digital Terrain Elevation Data (DTED) level 0 at 30 arc second (~1 km) resolution compiled by the National Geospatial Intelligence Agency (NGA, 1996) in 2001.

Temperature has long been observed to directly influence phenological phases. A large number of papers have scrutinized the effects of temperature on the phenological timings of plants (Fitter et al., 1995; Sparks and Carey, 1995; Sparks et al., 2000; Badeck et al., 2004). In general, higher temperature accelerates plant development,

Table 1. Summary of Variables Used in the UCRB Phenoregion Delineation

Variables	Descriptions	Data sources	Original spatial resolution	Accuracy
Elevation	Mean elevation above sea level	DTED Level 0 by NGA	30 arc seconds	Horizontal accuracy: <60 meters; vertical accuracy: <46 meters
Mean maximum temperature	Annual maximum temperature averaged for 1971–2000	PRISM dataset by PRISM Climate Group at Oregon State University	0.00833 decimal degrees	130 meters
Mean minimum temperature	Annual minimum temperature averaged for 1971–2000			
Mean maximum temperature during growing season	Maximum temperature between May and Oct., averaged for 1971–2000			
Mean minimum temperature during growing season	Minimum temperature between May and Oct., averaged for 1971–2000			
Standard deviation of monthly temperature	Intra-annual variability of temperature			
Mean precipitation	Total annual precipitation averaged for 1971–2000			
Mean precipitation during growing season	Total precipitation between May and Oct., averaged for 1971–2000			
Standard deviation of monthly precipitation	Intra-annual variability of precipitation			
Soil Variability Index	The variable indicating soil variability and was derived by applying PCA on 10 soil characteristics	STATSGO soil characteristics for the conterminous United States by USGS	1 km	N/A
Mean NDVI	Mean annual NDVI averaged for 1990–2005	AVHRR-NDVI by USGS	1 km	<1 km root mean square error

leading to earlier onset of phenological events. For example, two locations with a mean annual temperature difference of about 5°C can cause the onset of green-up to differ by as much as a month in the UCRB. Five temperature variables calculated from the PRISM (Parameter-elevation Regressions on Independent Slopes Model; Daly et al., 1994) dataset were included as input variables: mean annual maximum temperature, mean annual minimum temperature, and standard deviation of monthly temperature as well as mean maximum and minimum temperature during the growing season, defined as from May to October based on the range of first and last freeze/frost occurrence dates at different stations spread over the UCRB (Koss et al., 1988). PRISM data have a spatial resolution of 0.00833 decimal degrees (~925 m) and have a monthly temporal resolution covering 1971 to 2000 (PRISM Climate Group, 2010). The mean annual maximum and minimum temperature represent the general temperature range. Standard deviation of monthly temperature, and mean maximum and minimum temperature during the growing season were adopted to account for the intra-annual variation of temperature.

Precipitation is another major factor having great effects on vegetation. It affects the timings of different phenophases and accounts for a significant amount of the phenological variation, especially in moisture-limited regions like the UCRB (Reed et al., 1994; Peñuelas et al., 2004; Tadesse et al., 2010). Three variables—annual mean precipitation, standard deviation of monthly precipitation, and mean precipitation during the growing season—were included. These three precipitation variables were also extracted from the PRISM dataset (PRISM Climate Group, 2010).

Although not as significant as precipitation, soil fertility has close relationship with species distribution pattern (Swaine, 1996). Soil fertility can greatly influence vegetation abundance and species richness (Gentry and Emmons, 1987; Swaine, 1996). Many soil attributes are directly correlated with fertility—such as pH value, organic matter, and cation exchange capacity (Troeh and Thompson, 2005). The principal component (PC) of several soil attributes is used as an index of soil fertility to preserve maximum variability in soil attributes and allow better discrimination of phenoregions (see the following section of this paper). The SSURGO and STATSGO databases are considered to be reliable data sources to derive soil fertility by providing a series of soil attributes directly related to soil fertility. However, the NRCS soil survey project to populate the SSURGO and STATSGO database is still under way in the western U.S., so this data source is currently unavailable. Instead, the USGS-compiled 1 km data set of STATSGO soil characteristics for the conterminous United States (USGS, 1997) was used, with a full coverage of the study area and limited soil attributes less directly related to soil fertility, yet still greatly influencing vegetation growth. This data set contains 10 soil parameters including the high and low values of the range of organic matter, permeability, available water capacity, bulk density, and depth to seasonally high water table. The first PC was selected as the soil variability index, accounting for 99.6% of the total variance. This PC was named the soil variability index and was used as an input variable.

A vegetation Index (VI) is an indicator of vegetation abundance based on differences in the spectral reflectance at two different wavelengths within the electromagnetic spectrum. VIs have been successfully used for observing vegetation phenology (Reed et al., 1994; Zhang et al., 2001; White and Nemani, 2006; Masialeti et al., 2010). The Normalized Difference Vegetation Index (NDVI) is the most commonly used VI

for monitoring vegetation phenology. The mean annual NDVI is provided by the 1 km dataset from the USGS by averaging the AVHRR-NDVI from 1990 to 2005, to signify the average vegetation growth and vigor.

Eleven variables (Table 1) were thus selected and extracted from corresponding data sources. They were all resampled using bilinear interpolation to a common 1 km resolution, and then subset to the UCRB.

Phenoregions customized by this set of variables can capture particular phenological forcing patterns that may be missing in other types of subregion classification (Hargrove and Hoffman, 2004). For example, by considering the annual mean precipitation and the mean precipitation during the growing season, delineated phenoregions can provide additional information in discriminating places with desert monsoon and with alpine climates. Also, the areas with a higher mean temperature in the growing season resulting in an earlier onset of green-up can be distinguished using these variables. Delineation of more general purpose subregions without these variables might fail to capture these important differences. In this sense, the set of variables can decompose the geographical diversity of topography, climatic conditions and vegetation, and finely differentiate the spatial variation in phenological forcing among different geographic locations in the UCRB.

Principal Component Analysis and *k*-means++ Clustering

Principal component analysis plus iterative *k*-means clustering has been demonstrated to be an effective approach for the delineation of subregions in previous studies (Hargrove and Hoffman, 2004; White et al., 2005). PCA plus *k*-means clustering, as a quantitative method, does not rely on geographic knowledge or familiarity with the data, thus making the delineation of phenoregions more objective. Although this approach is computationally intensive compared with other quantitative methods, its essence of hierarchical non-nestable clustering can lead to independent phenoregion maps with different numbers of clusters, providing a better opportunity to develop an improved classification of phenoregions (Hargrove and Hoffman, 2004).

However, due to the local optima problem associated with the ordinary *k*-means algorithm, this research uses *k*-means++ (Arthur and Vassilvitskii, 2007), rather than the ordinary *k*-means algorithm, to acquire optimal clustering. The *k*-means++ algorithm is an augmentation of ordinary *k*-means by replacing the random seeding with a careful seeding process. The *k*-means++ clustering retains all the advantages of ordinary *k*-means and solves the local optima problem. The combination of PCA and *k*-means++ clustering is thus used in this research to delineate phenoregions by generating clusters based on PCs of the 11 variables in Table 1.

Principal component analysis is an essential prerequisite for *k*-means clustering due to potentially strong correlations among input variables. For example, precipitation is correlated with many soil attributes affecting soil fertility. Soil fertility and vegetation abundance indicated by NDVI are closely related to higher precipitation (Swaine, 1996), and total annual precipitation and its variability are influenced by elevation (Prins and Loth, 1988). PCA can effectively reduce the strong correlations between variables by converting the original set of variables into several PCs that are orthogonal in the data space (Hargrove and Hoffman, 2004). All of the variables were normalized because PCs are sensitive to scaling. Normalized variables have a mean

of zero and variance of one. PCA was applied to normalized variables and the first several PCs were selected based on the variance for which they accounted.

After PCs were selected to transform the two-dimensional geographic map space to a data space formed by these PCs, the *k*-means++ algorithm was used to cluster pixels that are close in the data space (i.e., similar values in elevation, temperature, precipitation, soil fertility and vegetation history). Ordinary *k*-means clustering has an intrinsic flaw associated with the random seeding: the performance of *k*-means clustering is dependent on the initial selection of cluster centroids. Many researchers have been aware of this problem, i.e., that the *k*-means algorithm may terminate at a local optimum instead of a global optimum depending on the initial centroids (Steinley, 2003). The *k*-means++ algorithm (Arthur and Vassilvitskii, 2007) addresses this problem by introducing a careful seeding process in lieu of random seeding in the ordinary *k*-means algorithm, to ensure the initial centroids are as far away from each other as possible in the multivariate data space. The algorithm proceeds as follows:

1. Select at random the first centroid from all data points.
2. Calculate a probability statistic using the following equation for each data point,

   $P = \frac{D(i)^2}{\sum_i D(i)^2}$, where $D(i)$ is the shortest distance (in the data space) from a data point i to its closet centroid that has already been selected.
3. Select the data point with the largest probability (P) as the next centroid.
4. Repeat step 2 and 3 until all k centroids are selected.
5. Proceed with this set of initial centroids as in the ordinary *k*-means algorithm.

Improved seeding can help ensure maximum dissimilarity between phenological forcing clusters and maximum homogeneity within each cluster. Adopting the *k*-means++ seeding process can effectively reduce the uncertainty and avoid the sometimes poor clustering arbitrarily resulting from the ordinary *k*-means algorithm. In summary, the *k*-means++ clustering improves both speed and accuracy compared to the ordinary *k*-means algorithm (Arthur and Vassilvitskii, 2007).

The number of clusters k, is an *a priori* parameter for each execution of *k*-means++ clustering. An inappropriate selection of k could result in a poor classification of phenoregions. Therefore, the clustering was tested with k clusters, with k ranging from 5 to 26. Two types of comparisons were then used to select a phenoregion map with higher homogeneity and spatial concordance (a measure of spatial coincidence and spatial overlap; Hargrove et al., 2006) with other phenoregion maps with different numbers of clusters.

Comparisons of Phenoregion Maps

The optimal phenoregion map should have the following characteristics:

1. The map should be as homogeneous within each phenoregion as possible—i.e., pixels within the same phenoregion should have as low variability or

dispersion as possible in terms of elevation, temperature, precipitation, soil fertility, and vegetation history.

2. Since the phenoregion maps with different numbers of clusters are all generated by the same process, maps with higher spatial concordance with other phenoregion maps may indicate consistently more stable phenoregions.

Two methods were adopted to compare phenoregion maps based on these two criteria. The first comparison method calculated the mean standard deviation of the Euclidean distance in data space formed by PCs for different phenoregion maps, referred to as "absolute comparison." The second method quantified the spatial concordance between pairs of phenoregion maps, referred to as "relative comparison."

The absolute comparison first calculates for each pixel the Euclidean distance from itself to the centroid of the cluster it belongs to in the data space. This indicates the similarity between a pixel and the mean of the phenoregion (the final centroid after running *k*-means++) in terms of phenological forcing variables. Then the standard deviation of a phenoregion is calculated as:

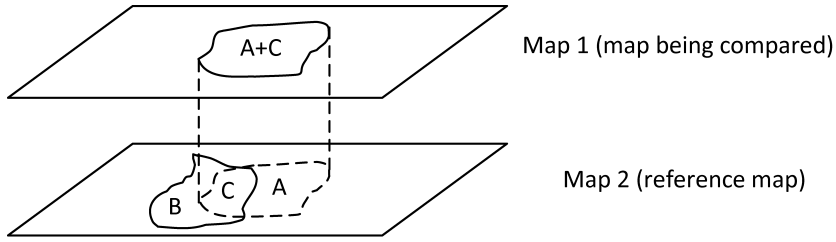
$$STD = \sqrt{\frac{\sum_{i=1}^{N} D^2}{N}},$$

where STD is the standard deviation of a particular phenoregion, N is the number of pixels in this phenoregion, and D is the Euclidean distance between a pixel and the centroid in the data space. The standard deviation of a phenoregion indicates the homogeneity of that phenoregion. The standard deviations of all phenoregions on a specific map are averaged to derive the mean standard deviation, indicating the general degree of homogeneity of that map.

The relative comparison uses Mapcurves goodness-of-fit (GOF) scores (Hargrove et al., 2006) to quantify the degree of spatial concordance between two maps. Each cluster on the map has a GOF score calculated as:

$$GOF = \sum\left(\frac{C}{B+C} \times \frac{C}{A+C}\right),$$

where GOF is the goodness-of-fit score of this cluster, C is the amount of overlapping region, $B + C$ is the total area of the intersected cluster on the reference map, and $A + C$ is the total area of the intersected cluster on the map being compared (Fig. 2). A mapcurve, which is a form of cumulative frequency distribution, was plotted for each comparison direction of phenoregion maps (e.g., the map with five clusters compared to the map with six clusters as a reference). This plot contains an x-axis indicating the GOF score and y-axis indicating the percentage of clusters with a GOF score exceeding the correspondent GOF threshold (Hargrove et al., 2006). Two GOF scores were derived by calculating areas underneath mapcurves of both comparison directions. The higher score indicates favorable direction of comparison and was selected as the GOF score between these two maps. The Mapcurves GOF scores range from 0 to 1, with the higher value indicating better fit.



$$GOF = \sum \left(\frac{C}{B+C} \times \frac{C}{A+C} \right)$$

Fig. 2. Algorithm used to calculate GOF score of cluster $A + C$ on Map 1 (Hargrove et al., 2006)

The average Mapcurves GOF score of a phenoregion map was calculated as the sum of scores between this map and each of 5- to 26-phenoregion maps divided by 22 (the number of phenoregion maps). This research used the average Mapcurves GOF score as a measure of the average spatial concordance of a phenoregion map with the whole series of maps with different numbers of clusters.

RESULTS AND DISCUSSION

Selection of Principal Components

Eleven components were generated from the variables listed in Table 1 using PCA. The first five components accounted for 96% of the total variance (Table 2) and were consequently selected as the PCs used in further analysis. The component scores imply the dominant variables for each of the five selected PCs as well as the correlations between PCs and input variables (Table 2). The first PC has the highest score for elevation and the climatic variables. It is negatively correlated with the elevation and the precipitation variables and positively correlated with the temperature variables. It explained about 63% of the total variance by itself. The second PC is a supplement to the first, representing several of the climatic variables including the mean minimum temperature, the mean minimum temperature during the growing season, the standard deviation of monthly temperature, and the standard deviation of monthly precipitation. The soil variability index was highly positively correlated with the third PC, while having comparably low correlation coefficients with other PCs, making the third PC a dominant representation of the soil fertility. The fourth PC is highly correlated with the mean NDVI. The fifth PC further explains the intra-annual climatic variation by scoring highest for the standard deviations of temperature and precipitation.

Phenoregion Maps Generated by *k*-means++ Clustering

Phenoregion maps with different numbers of clusters (5 to 26) were generated separately following the procedure described above. Figure 3 illustrates the 5-, 12-, 19-, and 26-phenoregion maps. Visually, the phenoregion maps in Figure 3 have

Table 2. Component Scores of and Variance Accounted for by the Principal Components[a]

	PC1	PC2	PC3	PC4	PC5
Elevation	–0.9147	–0.1003	0.0002	–0.0420	–0.1894
Tmax	0.9346	0.2432	0.0014	–0.0709	0.0990
Tmin	0.8348	0.5337	–0.0028	0.0644	–0.1130
Tmax_GS	0.9717	0.1581	0.0015	–0.0517	0.1271
Tmin_GS	0.8629	0.4805	–0.0027	0.0720	–0.0825
Temp_STD	0.7148	–0.5151	0.0058	–0.1288	0.4381
Precip	–0.8888	0.3342	–0.0022	0.1006	0.1360
Precip_GS	–0.9166	0.2478	–0.0010	0.0271	0.0928
Precip_STD	–0.7184	0.4104	–0.0071	0.2795	0.4174
SVI	0.0005	0.0014	0.9998	0.0231	–0.0017
NDVI	–0.4468	0.3259	0.0236	–0.8262	0.0697
Variance explained	0.6321	0.1203	0.0909	0.0734	0.0436
Cumulative variance explained	0.6321	0.7524	0.8433	0.9167	0.96034

[a]Abbreviations: Tmax = mean maximum temperature; Tmin = mean minimum temperature; Tmax_GS = mean maximum temperature during growing season; Tmin_GS = mean minimum temperature during growing season; Temp_STD = standard deviation of monthly temperature; Precip = mean precipitation; Precip_GS = mean precipitation during growing season; Precip_STD = standard deviation of monthly precipitation; SVI = soil variability index; PC1–PC5 = the first to the fifth PCs.

similar structures. For example, all four maps delineate the Northern Wyoming basin (cross sign), Northern Canyonlands (diamond sign), parks and ranges in northern Utah (triangle sign), and the western White River National Forest in Colorado (donut sign), yet with moderate differences. Phenoregions in one map are not simply the subsets of those in another map with smaller number of phenoregions because they are not generated as nestable hierarchical clusters. Instead, each pixel was reassigned to a cluster each time the phenoregion map was generated. Mountainous areas tend to be patchier than lower elevations and have more linear shapes following the direction of the elevation contours. A larger number of phenoregions are associated with mountainous areas such as the Southern Rocky Mountains in Colorado (yellow circles in Fig. 3) than with flat areas such as the Wyoming Basin (red circles in Fig. 3). The patches become smaller and sparser with increased distance from the core area of phenoregions and along the boundaries. This trend is considered to be the representation of gradual instead of abrupt change of phenological forcing in transition areas. These transition areas are called "phenopauses" (first coined by Hargrove and Hoffman (2004) as "ecopauses") (Hargrove and Hoffman, 2004; Williams et al., 2008). Pixels belonging to the same phenoregions are not necessarily contiguous; instead, they can be distributed in either large or small patches far away from each other. This results from similar phenological forcing occurring in different locations within the UCRB.

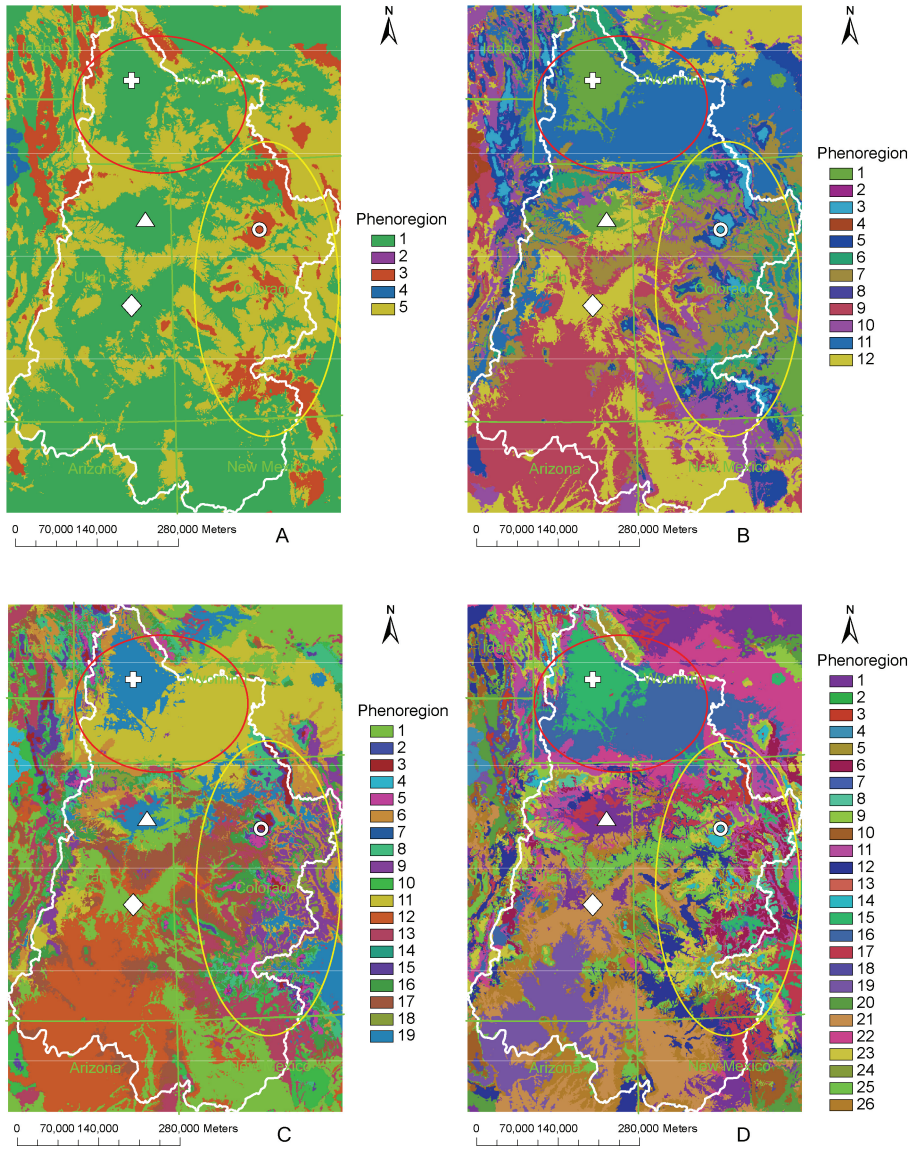


Fig. 3. 5- (A), 12- (B), 19- (C) and 26-phenoregion (D) maps.

Selection of the Optimal Phenoregion Map

The optimal phenoregion map was selected by absolute and relative comparisons as the map with higher homogeneity and spatial concordance with other phenoregion maps. The mean standard deviation of each phenoregion map using both ordinary *k*-means and *k*-means++ clustering is shown in Figure 4. The mean standard deviation using *k*-means++ clustering monotonically decreases—the average homogeneity within clusters always becomes higher as the number of phenoregions increases. An

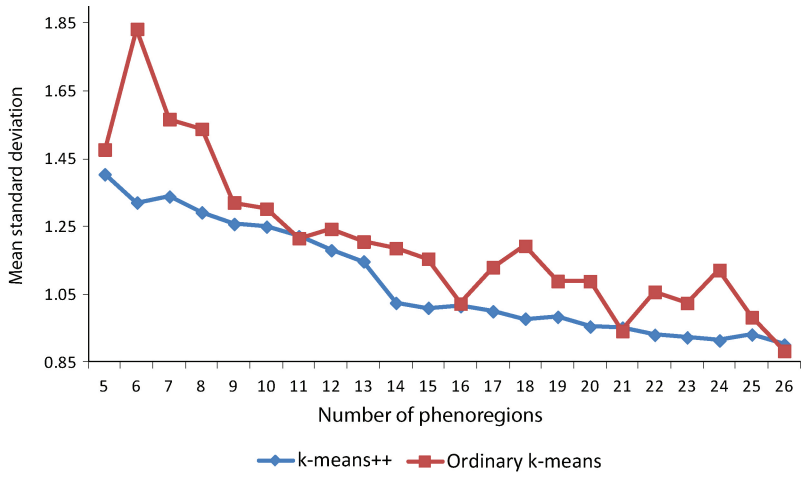


Fig. 4. Mean standard deviations of phenoregion maps using ordinary *k*-means and *k*-means++ clustering

abrupt change of slope can be observed at the 14-phenoregion map. The mean standard deviation decreases much faster from the 5-phenoregion to the 14-phenoregion map (by 0.38), and decreases more slowly from the 14-phenoregion to the 26-phenoregion map (by only 0.12). Therefore, 14- to 26- phenoregion maps—with lower mean standard deviations, i.e., higher average homogeneity—are considered better choices for the optimal phenoregion map.

The mean standard deviation using ordinary *k*-means does not strictly decrease as the number of phenoregions increases. The *k*-means++ clustering has lower mean standard deviation (increases in homogeneity) for almost all phenoregion maps compared with ordinary *k*-means clustering. Among all 22 maps, only four using ordinary *k*-means (11-, 16-, 21-, and 26-phenoregion maps) achieve similar clustering results as *k*-means++. This is because the careful seeding process ensures that the *k*-means++ algorithm can almost always achieve the global optimum and lead to the optimal clustering. Optimal clustering implies maximum similarity within each phenoregion and dissimilarity between different phenoregions in terms of phenological forcing variables. Figure 5A is the matrix of Mapcurves GOF scores between all pairs of phenoregion maps, represented by grayscale values. Brighter tones indicate higher GOF scores. Mapcurves GOF scores between each phenoregion map and itself are always equal to 1, producing the white diagonal from the upper left (5,5) to the lower right (26,26) corner. Except for these perfect fits, most of the phenoregion maps have GOF scores below 0.9, because of the hierarchical yet non-nestable nature of this series of phenoregion maps. However, the 6- and 7-phenoregion map has a GOF score of 1, suggesting that the 7-phenoregion map is a subdivision of the 6-phenoregion map. The GOF score between the 13- and 14-phenoregion map is 0.99992, indicating very high spatial concordance between these two maps. A phenoregion map tends to have a higher GOF score when compared to another map with a similar number of clusters. For example, the Mapcurves GOF score curve of the 17-phenoregion map peaks at 17 clusters and declines on either side of the peak (Fig. 5B). Another trend observed

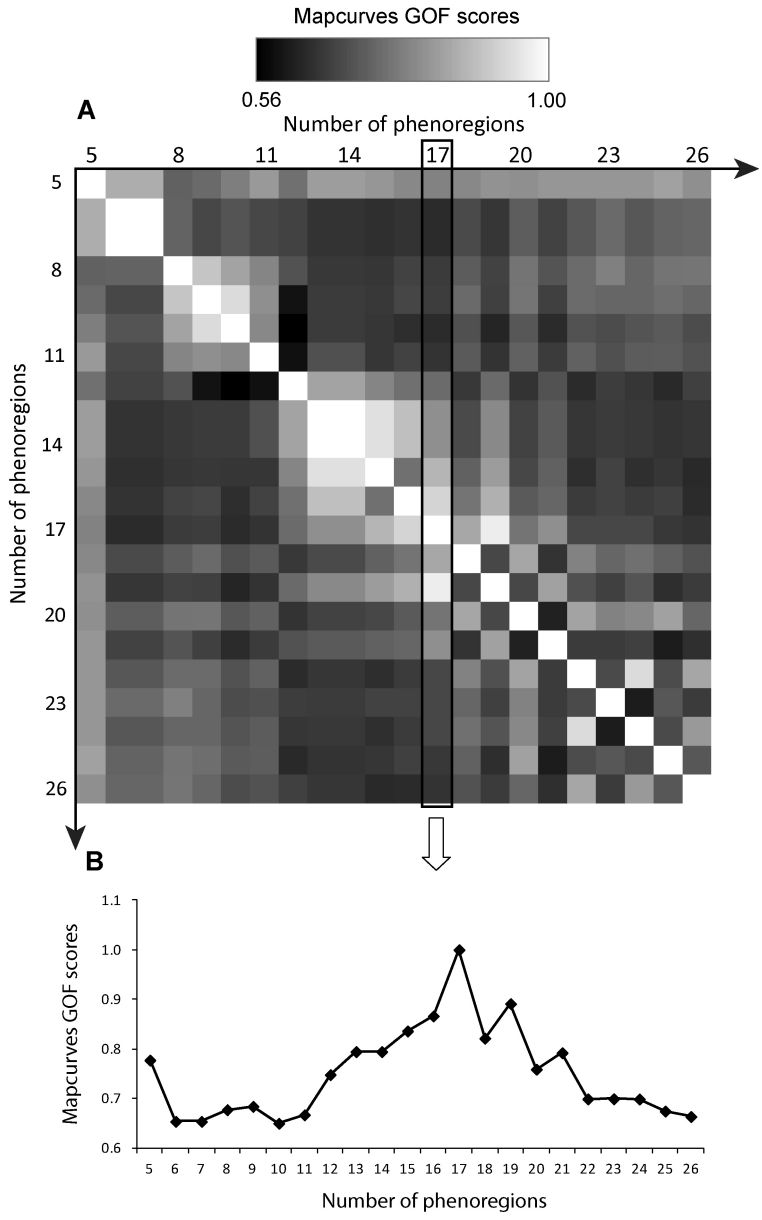


Fig. 5. A. Matrix of Mapcurves GOF scores represented by linearly scaled grayscale values, with black indicating a score of 0.56 and white indicating 1.00. B. Mapcurves GOF scores of the 17-phenoregion map.

from Figure 5A is that phenoregion maps with a larger number of clusters have slightly higher spatial concordance with maps with a smaller number of clusters (the brighter areas in the upper right and lower left corner). There are brighter squares along the

Table 3. Rank of Phenoregion Maps by Average Mapcurves GOF Score

Phenoregion number	Average Mapcurves GOF score	Phenoregion number	Average Mapcurves GOF score
5	0.8016	15	0.7378
20	0.7520	24	0.7350
17	0.7499	6	0.7335
14	0.7494	7	0.7335
13	0.7494	11	0.7277
8	0.7474	26	0.7274
18	0.7452	23	0.7250
16	0.7429	25	0.7245
19	0.7418	10	0.7206
22	0.7407	21	0.7173
9	0.7396	12	0.7110

(5,5) to (26,26) diagonal, showing that 5- to 7-, 8- to 11-, and 12- to 17-phenoregion maps have better fits with the maps within these respective ranges.

Among all phenoregion maps, the 5-phenoregion map has the highest average GOF score (0.8016, Table 3), and the 12-phenoregion map the lowest average score (0.7110). The 5-, 20-, 17-, 14-, and 13-phenoregion maps (in decreasing order) have a high degree of concordance with other maps (Table 3), and thus are considered to be superior choices for the final phenoregion map.

DISCUSSION

The use of this methodology and the phenoregion maps created show promise as useful analytical and policy tools for geographically diverse regions. However, there are limitations that should be taken into consideration when adopting this approach. The first issue is that for a regional phenoregion classification, finer resolution data are often preferred to achieve more accurate results. In this research, all data were either originally 1 km or near 1 km resolution (such as 30 arc seconds or 900 m) that were resampled to 1 km. One kilometer spatial resolution is adequate for a 280,000 km^2 UCRB area, but studies of smaller areas may require higher spatial resolution data. Another issue is associated with the pixels along the boundaries or the transition areas between phenoregions, termed phenopauses. Two adjacent pixels belonging to different phenoregions may have only minor differences in phenological forcing due to the gradual transitions over space. Phenological features of pixels within phenopauses could be considered as the mixture of the features in adjoining phenoregions or pixels. Using phenological modeling as an example, the prediction result for a pixel within a phenopause could utilize an inverse distance weighted average that takes adjacent pixels into account. Thirdly, the climatic features may not be stable over time. This classification is a simplified representation of recent, current, and near-future phenological

forcings. However, phenological forcings such as temperature and precipitation are likely to be altered by climate change. Changes in temperature and precipitation would result in changes to phenoregions derived from these variables. Lastly, due to the same dynamic nature of climatic variables, the different time spans of the PRISM and NDVI datasets can influence the phenoregion maps that are output. However, the influence is limited because the time span difference is within 20 years and the averaged values further attenuate the influence.

CONCLUSIONS

This research demonstrates an improved approach to phenological analysis and modeling in several ways. It has identified improved variables and data sources, and the overall methodological approach, with *a priori* defined evaluative criteria and a repeatable technical procedure using PCA plus *k*-means++ clustering, creates a robust means to delineate phenoregions. In particular, the absolute comparison shows that 14- to 26-phenoregion maps are more homogeneous. The relative comparison shows that 5-, 20-, 17-, 14-, and 13-phenoregion maps have higher spatial concordance with other maps. The optimal phenoregion map can be selected based on the intended use. For example, the phenoregion map used for further research will be employed as the basic map unit for a predictive phenological model. Time- and labor-consuming ground truth and phenological model validation require the number of phenoregions to be small. Therefore, the 14-phenoregion map is selected to serve as the optimal one and the future work of phenological modeling will be conducted for each phenoregion classified in the 14-phenoregion map.

This research developed a framework that could generate improved delineation of phenoregions in geographically diverse regions, using the UCRB as a case study. A unique set of variables was identified with the objective of decomposing the variant phenological forcing in the UCRB. This research is among the first to introduce *k*-means++ clustering to the delineation of phenoregions and natural area subregions. The adoption of the *k*-means++ algorithm in lieu of the ordinary *k*-means algorithm ensures that global optima can almost always be achieved. The results demonstrate the ability of PCA plus *k*-means++ to reduce uncertainty and lead to the optimal set of phenoregions with a given number of clusters *k*. Two evaluative criteria are proposed to compare variable cluster phenoregion maps: the homogeneity within each phenoregion and the spatial concordance with other phenoregion maps. It is believed that this methodology for phenoregion map delineation and evaluation can be employed to enhance potential vegetation delineation, phenological monitoring and modeling, and natural resource conservation and management, especially in geographically diverse regions.

REFERENCES

Arthur, D. and S. Vassilvitskii, 2007, "*k*-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA: Society for Industrial and Applied Mathematics, 1027–1035.

Badeck, F.-W., A. Bondeau, K. Böttcher, D. Doktor, W. Lucht, J. Schaber, and S. Sitch, 2004, "Responses of Spring Phenology to Climate Change," *New Phytologist,* 162(2):295–309.

Bailey, R. G., 1983, "Delineation of Ecosystem Regions," *Environmental Management,* 7(4):365–373.

Campbell, R. K., 1974, "Use of Phenology for Examining Provenance Transfers in Reforestation of Douglas-Fir," *Journal of Applied Ecology,* 11(3):1069–1080.

Daly, C., Neilson, R. P., and D. L. Phillips, 1994, "A Statistical–Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain," *Journal of Applied Meteorology,* 33:140–158.

Fitter, A. H., R. S. R. Fitter, I. T. B. Harris, and M. H. Williamson, 1995, "Relationships between First Flowering Date and Temperature in the Flora of a Locality in Central England," *Functional Ecology,* 9(1):55–60.

Gentry, A. H. and L. H. Emmons, 1987, "Geographical Variation in Fertility, Phenology, and Composition of the Understory of Neotropical Forests," *Biotropica,* 19(3):216–227.

Hargrove, W. and F. Hoffman, 2004, "Potential of Multivariate Quantitative Methods for Delineation and Visualization of Ecoregions," *Environmental Management,* 34:S39–S60.

Hargrove, W., Hoffman, F., and P. Hessburg, 2006, "Mapcurves: A Quantitative Method for Comparing Categorical Maps," *Journal of Geographical Systems,* 8(2):187–208.

Hodges, T, 1991, *Predicting Crop Phenology*, Boca Raton, FL: CRC Press.

Kenney, T. A., Gerner, S. J., Buto, S. G., and L. E. Spangler, 2009, *Spatially Referenced Statistical Assessment of Dissolved-Solids Load Sources and Transport in Streams of the Upper Colorado River Basin*, Reston, VA: U.S. Geological Survey Scientific Investigations Report 2009–5007, 50 p.

Koss, W. J., Owenby, J. R., Steurer, P. M., and D. S. Ezell, 1988, "Freeze/Frost Data," in *Climatography of the U.S.*, Asheville, NC: National Oceanic and Atmospheric Administration.

Küchler, A. W., 1964, *Potential Natural Vegetation of the Conterminous United States (with separate map at scale 1:3,168,000)*, Washington, DC: American Geographical Society, Special Publication No. 36.

Masialeti, I., Egbert, S., and B. D. Wardlow, 2010, "A Comparative Analysis of Phenological Curves for Major Crops in Kansas," *GIScience & Remote Sensing*, 47(2):241–259.

McMahon, G., Gregonis, S. M., Waltman, S. W., Omernik, J. M., Thorson, T. D., Freeouf, J. A., Rorick, A. H., and J. E. Keys, 2001, "Developing a Spatial Framework of Common Ecological Regions for the Conterminous United States," *Environmental Management,* 28(3):293–316.

NGA (National Geospatial Intelligence Agency, formerly National Imagery and Mapping Agency), 1996, Digital Terrain Elevation Data Level 0.

NRCS (Natural Resources Conservation Service, U.S. Department of Agricultural), 2011, "Soil Survey Manual—Chapter One," [http://soils.usda.gov/technical/manual/contents/chapter1.html], accessed January 10, 2011.

Omernik, J. M., 1987, "Map Supplement: Ecoregions of the Conterminous United States," *Annals of the Association of American Geographers,* 77(1):118–125.

Peñuelas, J., Filella, I., Zhang, X., Llorens, L., Ogaya, R., Lloret, F., Comas, P., Estiarte, M., and J. Terradas, 2004, "Complex Spatiotemporal Phenological Shifts as a Response to Rainfall Changes," *New Phytologist,* 161(3):837–846.

Prins, H. H. T. and P. E. Loth, 1988, "Rainfall Patterns as Background to Plant Phenology in Northern Tanzania," *Journal of Biogeography,* 15(3):451–463.

PRISM Climate Group, 2010, Oregon State University, October 2010 [http://prism.oregonstate.edu].

Reed, B. C., Brown, J. F., VanderZee, D., Loveland, T. R., Merchant, J. W., and D. O. Ohlen, 1994, "Measuring Phenological Variability from Satellite Imagery," *Journal of Vegetation Science,* 5(5):703–714.

Schuster, W. S., Alles, D. L., and J. B. Mitton, 1989, "Gene Flow in Limber Pine: Evidence from Pollination Phenology and Genetic Differentiation along an Elevational Transect," *American Journal of Botany,* 76(9):1395–1403.

Sparks, T. H. and P. D. Carey, 1995, "The Responses of Species to Climate Over Two Centuries: An Analysis of the Marsham Phenological Record, 1736–1947," *Journal of Ecology,* 83(2):321–329.

Sparks, T. H., Jeffree, E. P., and C. E. Jeffree, 2000, "An Examination of the Relationship between Flowering Times and Temperature at the National Scale Using Long-Term Phenological Records from the UK," *International Journal of Biometeorology,* 44(2):82–87.

Steinley, D., 2003, "Local Optima in *k*-means Clustering: What You Don't Know May Hurt You," *Psychological Methods*, 8(3):294–304.

Swaine, M. D., 1996, "Rainfall and Soil Fertility as Factors Limiting Forest Species Distributions in Ghana," *Journal of Ecology,* 84(3):419–428.

Tadesse, T., Wardlow, B. D., Hayes, M. J., Svoboda, M. D., and J. F. Brown, 2010, "The Vegetation Outlook (VegOut): A New Method for Predicting Vegetation Seasonal Greenness," *GIScience & Remote Sensing*, 47(1):25–52.

Thompson, R., Shafer, S., Anderson, K., Strickland, L., Pelltier, R., Bartlein, P., and M. Kerwin, 2004, "Topographic, Bioclimatic, and Vegetation Characteristics of Three Ecoregion Classification Systems in North America: Comparisons along Continent-wide Transects," *Environmental Management,* 34:S125–S148.

Troeh, F. R. and L. M. Thompson, 2005, *Soils and Soil Fertility*, Ames, IA: Blackwell.

Tüxen R., 1956, "Die heutige potentielle naturliche Vegetation als Gegenstand der Vegetationskartierung (Contemporary Potential Natural Vegetation as an Object of Vegetation Mapping)," *Angewandte Pflanzensoziologie*, 13:5–42.

USGS (U.S. Geological Survey), 1997, "STATSGO Soil Characteristics for the Conterminous United States," January 1997 (created by Wolock, D. M.) [http://water.usgs.gov/lookup/getspatial?muid].

USGS (U.S. Geological Survey), 2006, "National Water-quality Assessment (NAWQA) Program: Upper Colorado River Basin Study Unit" [http://co.water.usgs.gov/nawqa/ucol/], accessed March 1, 2009].

White, M. A., Hoffman, F., Hargrove, W. W., and R. R. Nemani, 2005, "A Global Framework for Monitoring Phenological Responses to Climate Change," *Geophysical Research Letters*, 32(4):L04705.

White, M. A. and R. R. Nemani, 2006, "Real-Time Monitoring and Short-term Forecasting of Land Surface Phenology," *Remote Sensing of Environment,* 104(1): 43–49.

Williams, C. L., Hargrove, W. W., Liebman, M., and D. E. James, 2008, "Agro-ecoregionalization of Iowa Using Multivariate Geographical Clustering," *Agriculture, Ecosystems & Environment,* 123(1–3):161–174.

Zhang, X., Hodges, J. C. F., Schaaf, C. B., Friedl, M. A., Strahler, A. H., and G. Feng, 2001, "Global Vegetation Phenology from AVHRR and MODIS Data," in *Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International*, 5:2262–2264.