Research Article Environmetrics

Received: 3 August 2015, Revised: 12 October 2015, Accepted: 13 October 2015, Published online in Wiley Online Library: 26 November 2015

(wileyonlinelibrary.com) DOI: 10.1002/env.2370

Bivariate functional data clustering: grouping streams based on a varying coefficient model of the stream water and air temperature relationship

H. Li[a], X. Deng[a], C. A. Dolloff[b] and E. P. Smith[a]*

A novel clustering method for bivariate functional data is proposed to group streams based on their water–air temperature relationship. A distance measure is developed for bivariate curves by using a time-varying coefficient model and a weighting scheme. This distance is also adjusted by spatial correlation of streams via the variogram. Therefore, the proposed distance not only measures the difference among the streams with respect to their water–air temperature relationship but also accounts for spatial correlation among the streams. The proposed clustering method is applied to 62 streams in Southeast US that have paired air–water temperature measured over a ten-month period. The results show that streams in the same cluster reflect common characteristics such as solar radiation, percent forest and elevation. Copyright © 2015 John Wiley & Sons, Ltd

Keywords: bivariate functional data; variogram; varying coefficient model; water–air relationship; weighted distance

1. INTRODUCTION

Temperature is a critical factor determining the distribution, abundance, and growth of fishes in streams (Keleher and Rahel, 1996; Beitinger *et al.*, 2000; Flebbe *et al.*, 2006; Meisner, 1990; Caissie, 2006; Minns *et al.*, 1995; Mohseni *et al.*, 1998; Sinokrot and Stefan, 1993; Stefan *et al.*, 2001). Although much recent research has focused on predicting future stream temperatures, many important issues remain including detection of patterns of water temperature over time and identifying the sensitivity of water temperature to factors that vary on similar spatio-temporal scales (Webb, 1996; Li *et al.*, 2014). The availability of inexpensive temperature sensors has lead to the ability to measure air and water temperatures for periods up to a year at a fine temperature scale (e.g., 30 min) (Trumbo *et al.*, 2014). A number of studies show that air temperature is strongly correlated with water temperature and the modeled relationship of air and water temperatures has been used to predict water temperature with reasonable accuracy across a variety of scales (Mohseni *et al.*, 1998; Webb *et al.*, 2003; Li *et al.*, 2014). These models can be used by managers to help identify streams that are likely to either lose habitat or to provide thermal refugia for obligate coldwater species such as trout under various scenarios of climate change (Flebbe *et al.*, 2006; Trumbo *et al.*, 2014).

The process of identifying and grouping streams for potential management will be more effective if streams can be organized into relatively homogeneous clusters. A key question to address is whether streams can be clustered based on similar water and air temperature relationships. Because streams within clusters likely share similar risk profiles, managers may tailor investment by streams within groups based on watershed specific influences (Mayer, 2012). For example, many climate and landscape factors such as solar radiation and percent forest that are known to affect the stream water and air temperature relationship (Chen *et al.*, 1998) are distinct for each stream but show similarities when compared within a derived cluster.

Although various clustering methods have been used for functional environmental data (Haggarty *et al.*, 2012; Ignaccolo *et al.*, 2013; Ignaccolo *et al.*, 2008), there are two major challenges for clustering streams based on the water–air temperature relationship. Here, the stream sites are not part of a stream network. The modeling and analysis of stream network data are described in Ver Hoef *et al.* (2014). The first challenge is to work with the bivariate curves associated with the air and water temperatures, and the second is to incorporate spatial information associated with the streams into the clustering method. Because each site has two variables, air temperature and water temperature, the observations are not expressed as vectors but in matrices. Hence, creating an interpretable distance measure is not trivial. Ieva *et al.* (2013) adopt a classical distance approach by summing distances between curves for electrocardiography signals. In a case study of air temperature and precipitation in Canada, Jacques and Preda (2014) represent multivariate curves through principal components, which

* *Correspondence to: E. P. Smith, Department of Statistics, Virginia Tech, 406A Hutcheson Hall, Blacksburg, VA, USA. E-mail: epsmith@vt.edu*

a *Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA*

b *USDA Forest Service, Southern Research Station, Department of Fish and Wildlife Conservation, Blacksburg, VA 24060, USA*

are defined based on the projections of functional data on eigenfunctions with equal weight on each curve. These methods may not be appropriate when one of the curves is of more interest or has greater significance to the clustering procedure. The second challenge is to incorporate spatial information into the clustering method. Attributes associated with streams located within a fixed geographical region are likely to be correlated, hence incorporation of spatial correlation into the clustering method will likely enhance biological interpretations of clusters. Several researchers have worked on spatially adjusted distance and/or the clustering algorithm based on spatial correlation. Oliver and Webster (1989) modify curve distances through the variogram (Cressie, 1993). Giraldo *et al.* (2012) extend Oliver and Webster's method and apply hierarchical clustering to spatially correlated data. Ben-Dor *et al.* (1999) propose a cluster affinity search technique for spatially constrained clustering. Brenden *et al.* (2008) modify the cluster affinity search technique method and propose a valley segment affinity search technique to cluster streams and sites within streams. However, these cluster algorithms might not be easily extended to bivariate functional data without defining an appropriate distance measure.

In this work, we propose a weighted distance measure for clustering based on the water–air temperature relationship to deal with the first challenge. Specifically, we cluster streams based on two curves: the time varying average daily maximum water temperature and the sensitivity of daily maximum water temperature to changes in the daily maximum air temperature. The daily maximum temperatures are used because they are critical to the survival of cold water fishes such as trout, which cannot tolerate prolonged exposure to water temperature in excess of $21^{\circ}C$. We used the varying coefficient linear model (VCM) (Li *et al.*, 2014) to quantify the water–air temperature relationship for individual streams. The motivation for using the VCM is to express the coefficients in the linear model as a function of time, which has a similar spirit to functional regression (Ramsay and Silverman, 2005). However, the VCM here focuses on the prediction of response at every time point, not over the whole time period as in functional regression. The intercept and slope functional curves obtained from the VCM correspond to the smoothed average water temperature and the sensitivity of water temperature to air temperature over time. We then develop a weighted Canberra distance (Lance and Williams, 1967) to measure the distance between streams based on the intercept and slope curves. The weighted distance provides flexibility to enable the resultant clustering to balance the focus between the two curves. That is, different weights lead to different interpretations; greater weight on the intercept curves result in clusters of streams separated mainly by smoothed water temperature. In contrast, more weight on the slope curves yield clusters of streams separated mainly by the sensitivity of water temperature to air temperature. To deal with the second challenge, incorporation of spatial information into the clustering, we adopt the method in Giraldo *et al.* (2012) to adjust the weighted distance by a scale of spatial distance using the variogram (Cressie, 1993). The resulting distance measure is the product of the weighted distance between the two curves in the VCM and the variogram. This model-based spatially adjusted distance makes clusters both statistically interpretable and spatially (ecologically) meaningful. The weighted distance focuses on the water–air temperature relationship, and the spatial component leads to streams having similar climate and landscape features within clusters.

The interpretability of cluster analysis result often depends on the choice of distance measure and the clustering algorithm. In this paper, the flexibility in the weighted and spatially adjusted distance measure leads to meaningful interpretations of grouped streams. The weight parameter is used to vary the emphasis from water temperature to the water–air temperature relationship. This flexibility reveals common characteristics for stream groups, such as the importance of solar radiation and percent forest coverage under various weighted distance measures. The parameters from the variogram control how much spatial correlation is incorporated into the analysis. Our method balances statistical (water–air temperature) and spatial (location of streams) variation and provides a valuable tool to natural resource managers and decision makers.

2. DATA

Data for this study were selected from paired air and stream water temperatures collected within 204 randomly selected subwatersheds located in the Southern Appalachian mountains, USA. These 204 are among the nearly 12,000 subwaterheds identified and classified as capable of supporting populations of Eastern Brook trout *Salvelinus fontinalis* in states from Maine to Georgia (EBTJV, 2006). Water and air temperature monitors were placed at the pour-point (downstream termination point) of each subwatershed. We used roughly 10 months of data with the same starting (1 January 2011) and ending dates (15 October 2011) from 62 of the 204 subwatersheds (Figure 1). For each site, we extracted the daily maximum temperature as these values represent the upper temperatures that fish are exposed to during the period; data input consists of $T = 288$ paired daily maximum air and water temperatures.

For stream site $i = 1, 2, \ldots, n$ $(n = 62)$, we denote daily maximum air and water temperatures as $(\mathbf{A}_i(t), \mathbf{W}_i(t))$, where $\mathbf{A}_i(t)$ is the air temperature at time t, $\mathbf{W}_i(t)$ is the water temperature at time t and $t = 1, 2, \ldots, T$. Without loss of generality, we used the centered air temperature. For each site, five site characteristics that may influence the water–air temperature relationship also were measured including a climatic variable, solar radiation, and four landscape variables including latitude, longitude, elevation, and the percentage of forest in the near-stream area (Mayer, 2012). For each site i, $s_i = (u_i, v_i)$ denotes the location variable, where u is latitude and v is longitude. For the initial clustering, only latitude and longitude are used in the clustering algorithm along with water and air temperatures. The other three covariates were used to validate and evaluate the clustering results.

3. PROPOSED CLUSTERING METHOD

We first applied the VCM to the original water and air temperature data and derived the bivariate functional curves to describe the water–air relationship. We then calculated a spatially-adjusted weighted distance measure, which incorporates the latitude and longitude through the variogram. We lastly applied the K-medoids method (Hastie *et al.*, 2009) to implement the clustering algorithm.

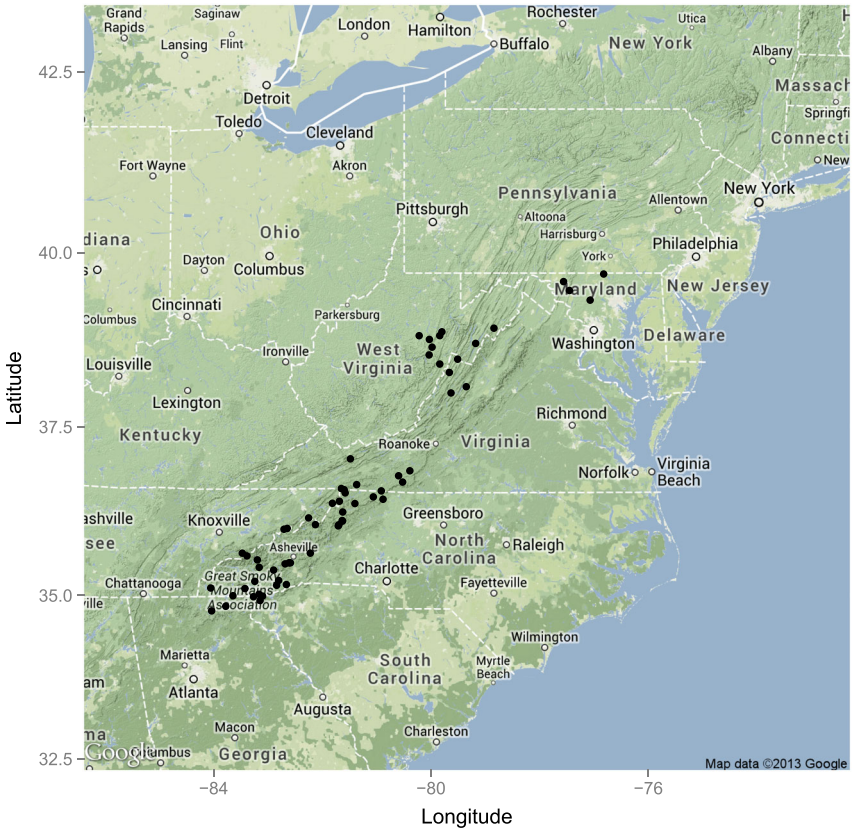

Figure 1. Location of 62 streams in Southeast United States (black dots are the locations)

3.1. Varying coefficient linear model for water–air relationship

It is common to smooth temporal data before clustering such that the curves observed at discrete time points will have a continuous functional form (Ramsay and Silverman, 2005). Popular smoothing techniques include the Karhunen–Loève expansion (Chiou and Li, 2007), the orthonormalized Gaussian basis (Kayano *et al.*, 2010), and the B-spline (Abraham *et al.*, 2003) approaches. In this work, because of the focus of the study, we need a smoothing method to model the sensitivity of water temperature to changes in air temperature. Because the VCM provides a model with a meaningful interpretation of the water and air relationship, it has advantages over other approaches that address the the study goals. In this work, we use the VCM as the smoothing method. Specifically, the intercept profile obtained from the VCM provides information about the smoothed maximum water temperature, and the slope profile reflects the sensitivity of water temperature to changes in air temperature over time. By using varying coefficient profiles as functional data, the clustering results can be expected to produce groups with similar water and air temperature relationships.

For site i, we describe the varying coefficient model (Li *et al.*, 2014) for the air–water temperature relationship as

$$W_i(t) = \theta_{0i}(t) + A_i(t)\theta_{1i}(t) + \epsilon_i(t), \tag{1}$$

where $\theta_{0i}(t) = \sum_{j=1}^{K} \alpha_{ij} b_{ij}(t)$ and $\theta_{1i}(t) = \sum_{j=1}^{K} \beta_{ij} b_{ij}(t)$ are varying intercept and slope coefficients. The $\epsilon_i(t)$ is the error term in the model with $E(\epsilon_i(t)) = 0$ and $\text{var}(\epsilon_i(t)) = \sigma_i^2(t)$. Here, $\{b_{i1}(t), \ldots, b_{iK}(t)\}$ is a set of K specified basis functions and α_{ij}, β_{ij}, $j = 1, 2, \ldots, K$, are parameters in the VCM in (1). To estimate these parameters, we adopt the regression spline methods used in Hoover *et al.* (1998) and Li *et al.* (2014). For each site i, the estimated coefficients $\mathbf{X}_i(t) = (\hat{\theta}_{0i}(t), \hat{\theta}_{1i}(t))$ are the profiles of the bivariate curves used in the following analysis.

3.2. Distance measure

In this section, we develop a weighted distance measure for bivariate functional data and incorporate spatial correlation between streams into this distance. The general form of the proposed distance between site i and site j can be written as

$$d_s(i, j) = r(h)d(\mathbf{X}_i(t), \mathbf{X}_j(t)), \tag{2}$$

where $h = \|s_i - s_j\|$ is the Euclidean distance between sites i and j. Here, $d(\mathbf{X}_i(t), \mathbf{X}_j(t))$ is used to quantify the distance between two sites based on the bivariate functional profile (i.e., the relationship between water–air temperatures), and $r(h)$ is a function of the spatial distance between two sites based on longitude and latitude.

Note that the functional profile for each site includes an intercept curve and a slope curve. Therefore, the calculation of $d(\mathbf{X}_i(t), \mathbf{X}_j(t))$ would involve two distances, the distance between the intercept curves and the distance between the slope curves. To combine the two distances, we consider a weighting scheme to create a single measure, where the weight can be chosen to give different emphasis on slope and intercept. Specifically, we define

$$d(\mathbf{X}_i(t), \mathbf{X}_j(t)) = w d_c(\theta_{0i}(t), \theta_{0j}(t)) + (1-w) d_c(\theta_{1i}(t), \theta_{1j}(t)), \tag{3}$$

where $0 \leq w \leq 1$ is the weight, and $d_c(\cdot,\cdot)$ is the Canberra distance (Lance and Williams, 1967) defined as

$$d_c(\theta_{ki}(t), \theta_{kj}(t)) = \sum_{t=1}^{T} \frac{|\theta_{ki}(t) - \theta_{kj}(t)|}{|\theta_{ki}(t)| + |\theta_{kj}(t)|}, \tag{4}$$

for two curves $\theta_{ki}(t)$ and $\theta_{kj}(t)$, $k = 0, 1$. As a standardized distance, the use of Canberra distance converts the two distances into a common scale to avoid the problem of different magnitudes of the intercept and the slope curves. The weight w here provides the flexibility for clustering streams based on different characteristics. For example, when w is close to 0, the distance in Equation (3) contains more information on the sensitivity of water temperature to air temperature (slope curve). Then the resultant clusters tend to separate streams based on water-to-air temperature sensitivity. In contrast, when w is close to 1, the distance in Equation (3) contains more information on smoothed maximum water temperature (intercept curve). Then the smoothed maximum water temperature statistic will become a key factor in clustering. Therefore, we can alter the emphasis of the cluster results based on different weights.

For the spatial distance $r(h)$ in (2), we consider using the variogram to model it, which is discussed in Oliver and Webster (1989), Cressie (1993), and Giraldo *et al.* (2012). Specifically, we adopt the distance measure in Giraldo *et al.* (2012) and extend it to the bivariate curves situation. As there are two curves for each site, $\mathbf{X}_i(t) - \mathbf{X}_j(t)$ still results in two vectors. Therefore, the distance measure cannot be directly defined by subtraction and taking absolute values. It can be defined more appropriately by using the distance measure (3) in the variogram for bivariate functional data. For the form of the variogram, we use the method suggested by Oliver and Webster (1989) and express the variogram as

$$r(h) = c_0 + c_1 \left(1 - exp\left\{-\frac{|h|}{\rho}\right\}\right). \tag{5}$$

To extend the exponential function in Oliver and Webster (1989), we added parameters c_0 and c_1, where c_0 is nugget effect and $c_0 + c_1$ is the sill (Le and Zidek, 2006). By plugging Equation (5) into (2) and dividing by $c_0 + c_1$, we obtain the following expression:

$$d^*(i,j) = \frac{c_0}{c_0 + c_1} d(\mathbf{X}_i(t), \mathbf{X}_j(t)) + \frac{c_1}{c_0 + c_1}\left(1 - exp\left\{-\frac{|h|}{\rho}\right\}\right) d(\mathbf{X}_i(t), \mathbf{X}_j(t)). \tag{6}$$

The first part of Equation (6) is proportional to the distance (3), and the second part can be modified according to the spatial correlation. The distance measure (6) is the distance we will use in the clustering algorithm.

To estimate the parameters in the variogram in distance measure (6), we adopt a classical formula from Cressie (1993):

$$2\hat{\gamma}(h) \equiv \frac{1}{N_h(N_h - 1)} \sum_i \sum_j ((\mathbf{X}_i(t) - \mathbf{X}_j(t))^2 : (i,j) \in N(h); h \in T(h(l))), \tag{7}$$

where

$$N(h) \equiv \{(i,j) : \|s_i - s_j\| \in (h - \epsilon, h + \epsilon); i, j = 1, 2, \ldots, n\}$$

with N_h being the number of sites in $N(h)$, $T(h(l))$ is the l^{th} tolerance region, $l = 1, 2, \ldots, m$, and m is the number of tolerance regions. In selecting tolerance regions, we keep sizes of regions similar and make the number of distinct pairs to be at least 30 (Journel and Huijbregts, 1978). For bivariate data, we defined the empirical estimation for the variogram as

$$2\hat{\gamma}(h(l)) \equiv \frac{1}{N_h(N_h - 1)} \sum_i \sum_j (d(\mathbf{X}_i(t), \mathbf{X}_j(t))^2 : (i,j) \in N(h); h \in T(h(l))), \tag{8}$$

To estimate c_0, c_1, and ρ in (5), we use the least squares estimates from nonlinear regression between the empirical variogram in (8) and the parametric variogram in (5).

3.3. Clustering algorithm

There are various methods used in functional clustering such as K-means and hierarchical clustering (Hastie *et al.*, 2009). Sangalli *et al.* (2010) consider the misalignment of functional data and define similarity between curves in order to apply K-means. Tokushige *et al.* (2007) extend the K-means method for functional data and propose crisp and fuzzy K-means algorithms. Tarpey and Kinateder (2003) propose a

K-means method based on principal points. Giraldo *et al.* (2012) apply hierarchical clustering method to spatially correlated data in Canada. Classical K-means requires squared Euclidean distance, which is not applicable to our proposed distance. Alternatively, K-medoids clustering (Hastie *et al.*, 2009) can be more flexible for different distance measures.

The choice of the number of clusters k often depends on prior information or the goal of study. In this work, we expect the number of clusters to be relatively small for meaningful interpretation. A naive method to determine the number of clusters k is through visualization of the stream temperature. Such an approach is subjective but helpful when there is a clear pattern of separate groups. In our study, in which we use a spatially weighted distance measure, visualizing curves on a graph can be difficult. Alternatively, we use an analytical method to determine k based on the calculation of within cluster dissimilarity. Such an idea is also used in Tibshirani *et al.* (2001). They compare logarithms of within cluster dissimilarity of the original data to uniformly generated data and estimate the optimal k by the gap between the two. We adopt a similar but simpler method by selecting k based on the silhouette width (Rousseeuw, 1987). Specifically, for a cluster solution with k clusters $C_1, \ldots, C_k$, the silhouette width for stream i is defined as

$$s(i|C_1, \ldots, C_k) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{9}$$

where $a(i) = \frac{1}{|C(i)|} \sum_{j \neq i, j \in C(i)} d_s(i, j)$ is the average distance of stream i to all other streams within the same cluster. Here, $C(i)$ is defined as the cluster that stream i belongs to. Also $b(i) = \min_{i \notin C_h} \frac{1}{|C_h|} \sum_{j \in C(h)} d(i, j)$ is the minimum average distance from stream i to all points in another cluster. Then the optimal number of clusters k is determined by

$$k = \arg\max_m \sum_i s(i|C_1, \ldots, C_m). \tag{10}$$

4. RESULTS

In this work, 62 streams with complete ten-month water and air temperature data are used in the analysis. The data were first summarized by computing the daily maximum air and water temperature. We fit the VCM to the paired temperature data for each stream and obtain the intercept and slope curves. The proposed clustering method is applied to bivariate curves of intercept and slope. For illustration, the performance of the proposed method is evaluated by choosing various values of weight w in Equation (3), that is, $w = 100\%, 75\%, 50\%, 25\%$, and 0%. In the following sections, the clustering results will be viewed from both statistical and geographical perspectives.

4.1. Parameter estimation

The VCM method in Section 3 is used to fit the water and air temperature data of each stream to quantify the water–air temperature relationship. The fitted coefficient curves are shown in Figure 2. For the intercept curves in Figure 2, there is no visible separation between the smoothed maximum water temperature trend patterns for the streams. The magnitude of the smoothed maximum water temperature varies considerably: for each time point, the gap between the maximum intercept and the minimum intercept is about $7^\circ C$. For the slope curves in Figure 2, the variation is relatively large, especially in the summer time (around the period of $t = 200$). Recall that the intercept and slope curves represent the smoothed maximum water temperatures and sensitivities of water temperature to air temperature, respectively. It is thus helpful to use the bivariate intercept and slope curves obtained from the VCM for clustering streams and assessing the risk of high water temperature.

We used the pre-specified formulation in Equation (5) to estimate the variogram (Figure 3). We consistently observe that the value of the variogram increases as the distance between two streams increases, which verifies that streams located in close proximity to each other have a similar water and air temperature relationship. Therefore, it is reasonable to incorporate the spatial information into the clustering method.

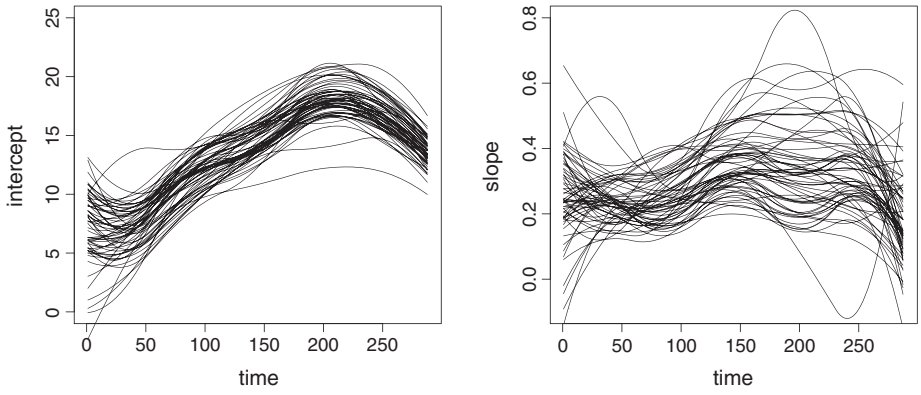

Figure 2. Intercept and slope curves from the varying coefficient linear model for 62 sites

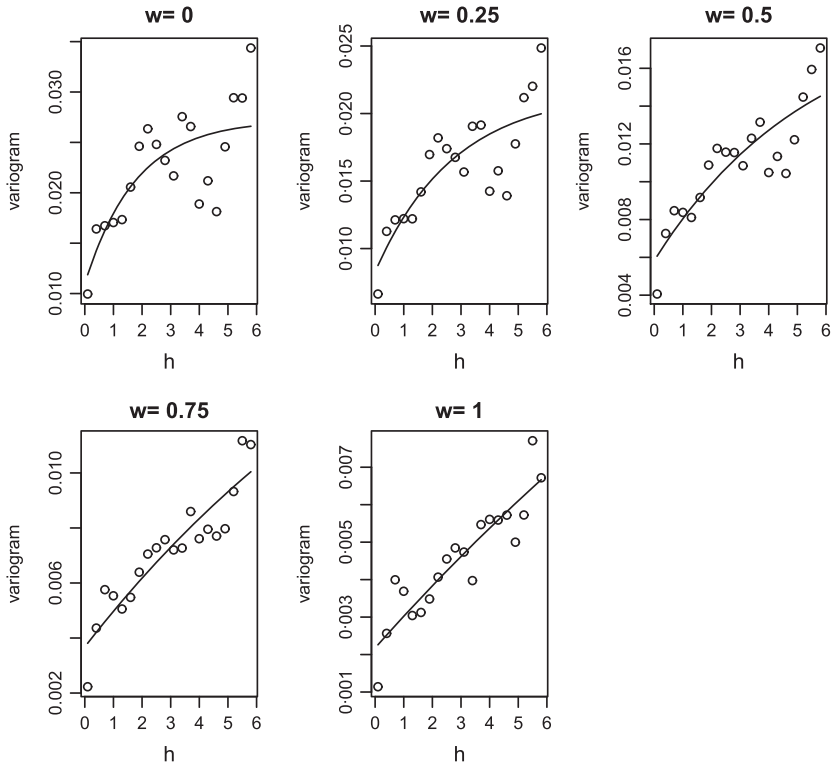

Figure 3. Variogram estimates and the exponential curves for different weights on the intercept curve: circles are empirical estimates of the variogram, and curves are for the estimated exponential variogram

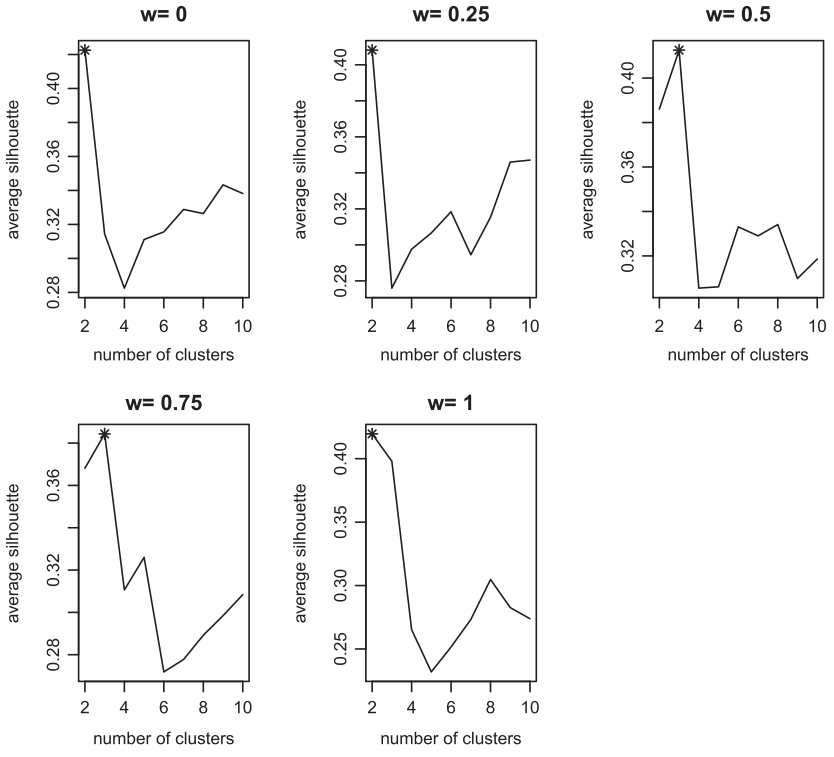

Figure 4. Silhouette statistics for different numbers of clusters under five different weights. The star marks indicate the optimal number of clusters and the corresponding silhouette statistics

4.2. Clustering results

The silhouette statistics for choosing the number of clusters are summarized in Figure 4. From the results in Figure 4, it clearly shows that two or three clusters can be the optimal number of clusters, producing the highest silhouette statistics for all scenarios of five weights. Figures 5–9 report the performance of the proposed clustering method with respect to the weight $w = 100\%, 75\%, 50\%, 25\%$, and 0% on the intercept. For each figure, part (a)–(f) display the location of streams, intercept and slope curves and boxplots for solar radiation, percent forest, and elevation, respectively. For the streams in each cluster, we use their solar radiation, percent forest and elevation to form separate boxplots. By changing the weight w in distance measure (3), we get three interesting findings.

First, the weight can determine the impact of spatial correlation on the cluster results. From Figure 5(a) where the weight on intercept is large, we observe that the proposed clustering method clearly results in two groups of northern and southern streams on the map. For the other figures (part (a) of Figures 6–9) in which the weight on intercept decreases, the northern and southern clusters pattern is not that visible. Such a finding indicates that the major variation in intercept curves is related to geographical variation as the clustering results mainly depend on the location of streams. Because the intercept curve from the VCM is interpreted as the smoothed maximum water temperature, it is expected that streams with smaller mean intercepts are located in the north because streams with higher latitude often have lower stream temperature. In contrast, Figure 9(a) reveals that when the weight on the slope curve is large (w is small), the geographical pattern of stream groups becomes weaker. As shown in Figure 9(a), when $w = 0$ (distance is calculated based on slope only), the stream clusters are not aligned by location on the map. This finding indicates that the spatial correlation among the slope curves is weak.

Second, the intercept and slope curves have connections with the landscape and climate variables in the clustering results. From Figure 5, we observe that when more weight is given to the intercept in the distance measure (w is large); the resultant stream clusters are grouped in a consistent manner with stream elevations. From Figure 5(f), it is seen that the two clusters result in strong separation with respect to the elevation covariate. This result is expected because streams with high elevation usually have lower summer water temperature (Figure 5(b) red curves), which is directly related to the intercept curves. Note that the elevation information is not used in the proposed clustering algorithm. This finding further confirms that the proposed distance measure defined in Equation (2) is meaningful and adequate for studying the connection between the landscape variables and the clustering results. Moreover, we can see from Figure 9 that when the slope is emphasized in calculating the distance measure (w is small), streams in different clusters tend to have distinct magnitude for variables such as solar radiation and percent forest. For example, the boxplots in Figure 9(d) and (e) show that the means of solar radiation and percent forest are separated for different clusters. Because the slope is interpreted as measuring sensitivity, our findings could imply that the sensitivity of the water temperature to the air temperature is related to solar radiation and percent forest. From Figure 9(d), it is observed that for sites having less solar radiation (lower boxplot), water is less sensitive to air temperature (red curves in Figure 9(c)). Furthermore, from

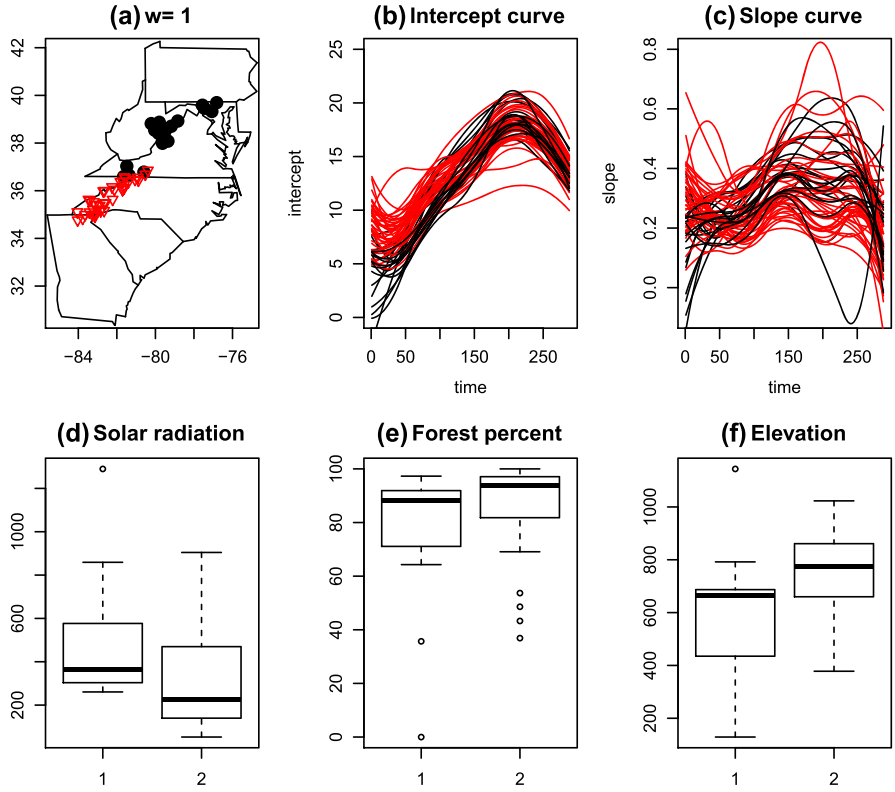

Figure 5. Cluster results for weight = 100% on intercept. Numbers of sites in each cluster are: 21 for cluster 1 (black) and 41 for cluster 2 (red). (a) location of streams, (b) intercept curves, (c) slope curves, (d) boxplots for solar radiation for different clusters, (e) boxplots for percent forest for different clusters, and (f) boxplots for elevation for different clusters

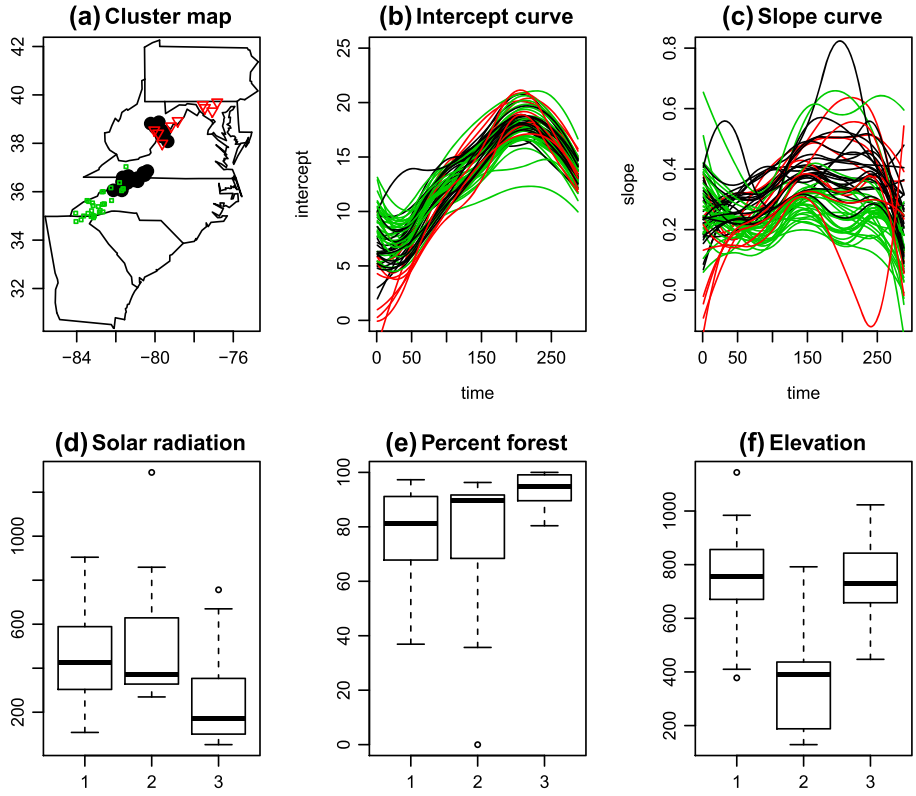

Figure 6. Cluster results for weight = 75% on intercept. Numbers of sites in each cluster are: 23 for cluster 1 (black), 9 for cluster 2 (red), and 30 for cluster 3 (green). (a) location of streams, (b) intercept curves, (c) slope curves, (d) boxplots for solar radiation for different clusters, (e) boxplots for percent forest for different clusters, and (f) boxplots for elevation for different clusters

Figure 9(e), we can see that for the sites with higher percent forest (upper boxplot), water temperature is less sensitive to air temperature (red curves in Figure 9(c)) as expected. Therefore, the clustering analysis infers that solar radiation and percent forest affect the sensitivity of the water temperature to the air temperature.

Note that the slope curves in the second cluster (red curves) are flat during summer time (day 174 through 230) in Figure 9(c). Days 174 and 230 are actually two knots in our regression splines. Therefore we can take advantage of the VCM and test if the slope curves in cluster 2 are constant. Specifically, we test if combinations of $\beta's$ in model (1) are zero. We obtained those combinations from the 29 sites in cluster 2 and conducted Hotelling's T^2 test (Rencher and Christensen, 2012) for constant slope. The test result shows that the slope curves in cluster 2 are constant for the summer period ($p = 0.44$). The result indicates that a linear regression model, with common slope but possibly different intercepts, is reasonable to use over the period. In cluster 2, the percent forest is high. The constant slopes suggest that the percentage forest can help buffer water temperature from the influence of high air temperature.

Third, the proposed clustering methods with distance measure (3) have more advantages if using a considerable amount of weight, that is, $w = 50\%$, on both intercept and slope. Figure 7 reports the clustering results for the weight $w = 50\%$ (equal weight on the intercept and slope). From the boxplots in parts (d), (e), and (f) of Figure 7, elevation, solar radiation, and percent forest exhibit significant differences across clusters. Because equal weight ($w = 50\%$) on the intercept and slope curves is used to calculate the distance measure (3), the clustering result therefore reflects information from all three variables, elevation, solar radiation, and percent forest in the clusters. Recall that those three variables are not used in calculating the proposed distance measure. Such an observation confirms that the proposed distance measure is flexible for investigating which underlying variables affect the water–air relationship, and we can construct cluster results based on different climate and landscape variables.

In Figures 5–9 where the weight on intercept deceases from 100% to 0%, we find that elevation differences between clusters is less relevant whereas solar radiation and percent forest differences become more important among clusters. To verify this, we calculated the F statistic (Rencher and Schaalje, 2008) to evaluate, for each of those three variables, whether the means are the same among clusters (Table 1). Solar radiation and percent forest were highly significant ($p < 0.01$) except when the value of weight $w = 100\%$. This is evidence that the separation on the two variables among clusters is consistent with the weight on the slope curve (sensitivity). Note in the last column in Table 1, the p-value for elevation is larger when there is more weight on the intercept (w is large). This is evidence that the elevation difference among clusters is consistent with the weight on intercept curve (maximum water temperature). This plausible outcome was revealed only by changing the weight in distance measure (3). Depending on the research interest, one could increase the weight on the intercept to emphasize variation in elevation between clusters or increase the weight on slope to emphasize the role of solar radiation and percent forest variation between clusters.

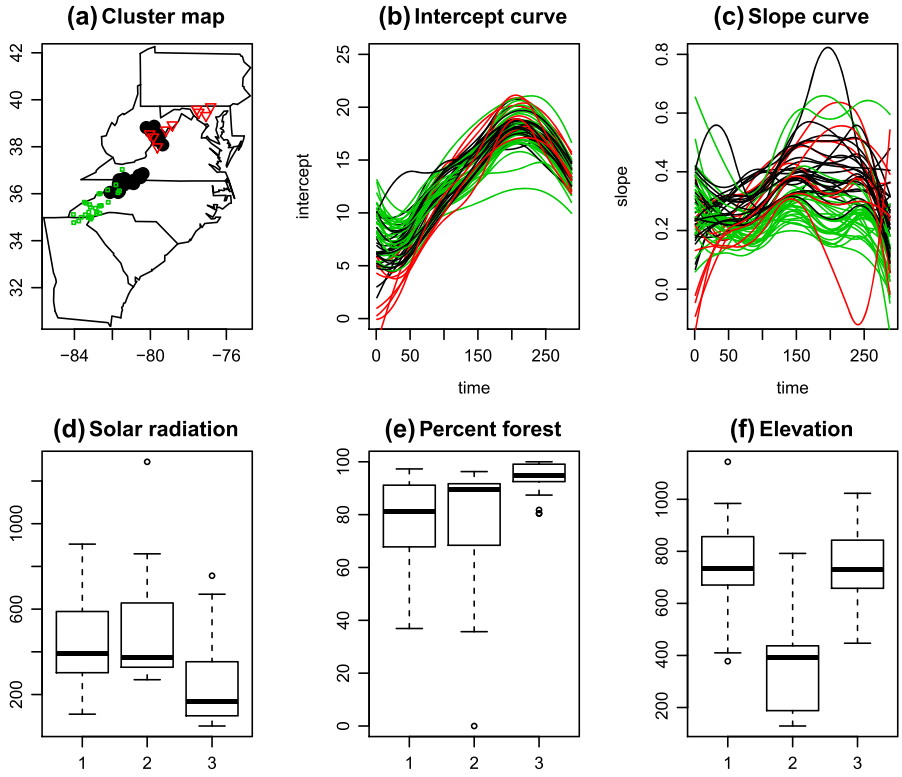

Figure 7. Cluster results for weight = 50% on intercept. Numbers of sites in each cluster are: 24 for cluster 1 (black), 9 for cluster 2 (red), and 29 for cluster 3 (green). (a) location of streams, (b) intercept curves, (c) slope curves, (d) boxplots for solar radiation for different clusters, (e) boxplots for percent forest for different clusters, and (f) boxplots for elevation for different clusters

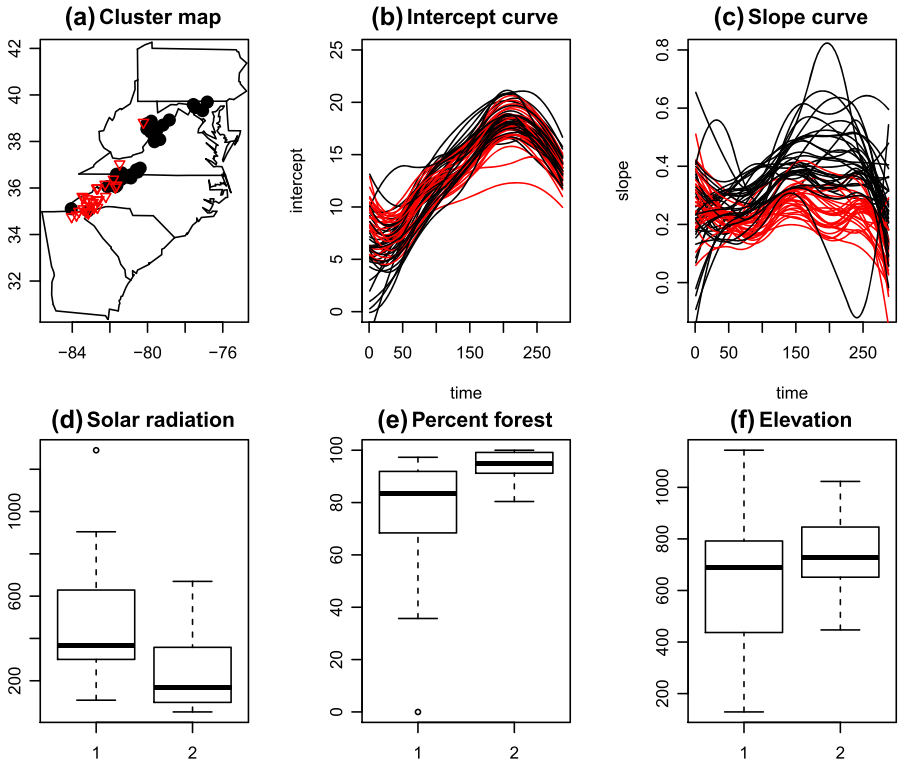

Figure 8. Cluster results for weight = 25% on intercept. Numbers of sites in each cluster are: 34 for cluster 1 (black) and 28 for cluster 2 (red). (a) location of streams, (b) intercept curves, (c) slope curves, (d) boxplots for solar radiation for different clusters, (e) boxplots for percent forest for different clusters, and (f) boxplots for elevation for different clusters

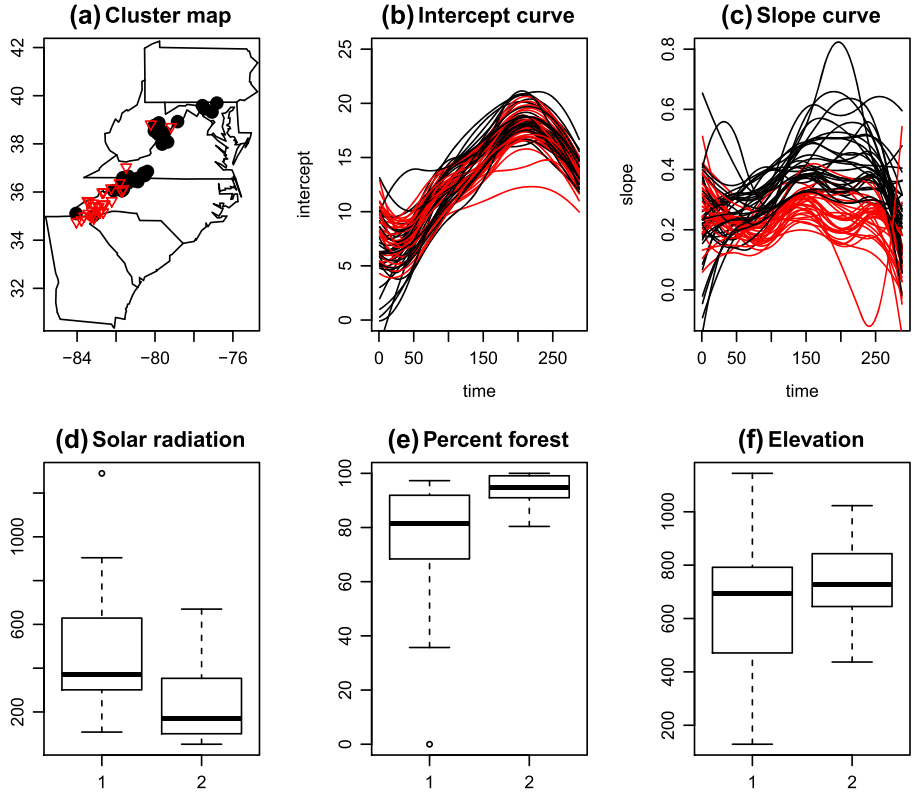

Figure 9. Cluster results for weight = 0% on intercept. Numbers of sites in each cluster are: 33 for cluster 1 (black) and 29 for cluster 2 (red). (a) location of streams, (b) intercept curves, (c) slope curves, (d) boxplots for solar radiation for different clusters, (e) boxplots for percent forest for different clusters, and (f) boxplots for elevation for different clusters

Table 1. F statistics and *p*-values (in parenthesis) from F-test for cluster differences

Weight on intercept	Solar radiation	Percent forest	Elevation
0	17.6 (<0.01)	18.5 (<0.01)	1.84 (0.17)
0.25	17.2 (<0.01)	17.5 (<0.01)	2.82 (0.09)
0.5	8.47 (<0.01)	9.98 (<0.01)	13.9 (<0.01)
0.75	9.29 (<0.01)	10.0 (<0.01)	13.9 (<0.01)
1	4.23 (0.04)	2.98 (0.08)	12.3(<0.01)

5. DISCUSSION

We developed a bivariate functional clustering procedure to group streams based on a varying coefficient model of the water and air temperature relationship. We defined a weighted distance measure adjusted by spatial information and applied the K-medoids clustering method using this distance. Our analysis demonstrates that streams in the same cluster share similar values of solar radiation, percent forest, and elevation. It is worth pointing out that the weighted distance in this work can be extended to multivariate functional data by using multiple weights. One application is to study the relationship between water temperature and other covariates using the VCM, which would generate more than one slope curve. By assigning different weights to multivariate curves in the distance measure, the resultant clusters can reflect the sensitivity of water temperature to different covariates.

Classical methods such as hierarchical clustering (Giraldo *et al.*, 2012) are often used to cluster environmental functional data. In this work, both K-medoids and hierarchical clustering methods result in similar grouping of streams. We chose the K-mediods method for two reasons: first, it is an iterative optimization procedure such that it can gradually improve the clustering quality and second, the K-medoids algorithm is effective in detecting compact spherical-shaped clusters and is easy to use in practice (Aggarwal and Reddy, 2013).

An important decision for this analysis is that the clustering method is based on the water–air temperature relationships using daily maximum water and air temperatures. Our choice was made based on biological considerations; there are clearly other possible choices for summarizing the series. In addition, the analysis did not involve a lag for the air temperature. Although lagged models were considered,

they were only slightly better in a few cases and also found to vary with different days and seasons. For example, if during the previous day rainfall occurred, this sometimes resulted in a poor value of air temperature relative to the current day. In addition, by treating the data with concurrent maximum water and air temperatures, the slope and intercept curves of the VCM can have meaningful interpretations. We will investigate how to incorporate a varying coefficient lag into the VCM framework.

Another property of the proposed method is the ability to deal with simple missing value patterns. The varying coefficients in Equation (1) are based on a linear combination of spline bases and thus have a small number of parameters. Therefore, a complete data series is not required for fitting the VCM and obtaining the bivariate curves. The VCM can be applied to a data set with missing values at random time points. However, when the data set contains missing values over extended periods of time, the VCM cannot work properly because of the lack of sufficient information. We applied the proposed method to the 62 streams with complete data rather than all 204 streams. For the other streams, the missing data often occurs over a period of time, resulting from equipment failure, or irregularities in the installation of the device. In future work, we plan to develop an imputation method for large gaps of missing values using the VCM. Finally, the use of silhouette width for choosing the number of clusters may not be suitable when there is only one cluster. To verify that one cluster was not appropriate, we evaluated the clustering through variables not used in the cluster analysis. The stream clustering in this work clearly shows that there are at least two clusters with meaningful interpretations.

6. CONCLUSION

The challenge of clustering bivariate functional data is that there are two groups of observations associated with each unit (i.e., streams). While the approach of Ieva *et al.* (2013) could be used, the assumption of additivity results in an equal weighting of curves. Although this approach reduces the number of curves, it lacks flexibility in interpreting the effect of each curve on the clustering results. Our assignment of different weighted distances for two curves provides this flexibility. In this study of water and air temperatures, larger weight on the intercept of the VCM results in groups of streams that are clustered by smoothed maximum water temperature. In contrast, larger weight on the slope results in streams grouped by the water and air slope relationship.

Weighted distance can also be used in the variogram definition for bivariate functional data. Incorporating spatial correlation by variogram enables sites in the same cluster to have both similar water–air relationships and geographic characteristics. This is important in that the cluster results provide information on how land management might be tailored to the water–air relationship. For example, we found that the sensitivity of water temperature to air temperature is associated with percent forest (Figures 8 and 9). This suggests that the set of streams with high percent forest is less sensitivity to changes in air temperature, meaning that the relationship between air and water temperature is relatively constant, especially in the summer. Sites with less forest tend to be more sensitive in the summer months.

Acknowledgements

We thank Joe Cline for data collection and management, and the USFS Southern Research Station for support under JV 2010 11330140-137. Helpful comments from Hongxiao Zhu on the manuscript were appreciated.

REFERENCES

Abraham C, Cornillon P, Matzner-Lober E, Molinari N. 2003. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics. Theory and Applications* **30**:581–595.
Aggarwal CC, Reddy CK. 2013. *Data Clustering: Algorithms and Applications*. CRC Press: Boca Raton, FL.
Beitinger TL, Bennett WA, McCauley RW. 2000. Temperature tolerances of North American freshwater fishes exposed to dynamic changes in temperature. *Environmental Biology of Fishes* **58**:237–275.
Ben-Dor A, Shamir R, Yahkini Z. 1999. Clustering gene expression patterns. *Journal of Computational Biology* **6**:281–297.
Brenden T, Wang L, Seelbach P, Clark RJ, Wiley M, Sparks-Jackson B. 2008. A spatially constrained clustering program for river valley segment delineation from GIS digital river networks. *Environmental Modelling & Software* **23**:638–649.
Caissie D. 2006. The thermal regime of rivers: a review. *Freshwater Biology* **51**:1389–1406.
Chen YD, Carsel RF, McCutcheon SC, Nutter WL. 1998. Stream temperature simulation of forested riparian areas: 1. Watershed-scale model development. *Journal of Environmental Engineering* **124**:304–315.
Chiou JM, Li PL. 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **69**:679–699.
Cressie N. 1993. *Statistics for Spatial Data*. John Wiley & Sons: New York.
EBTJV. 2006. Eastern brook trout joint venture.
Flebbe PA, Roghair LD, Bruggink JL. 2006. Spatial modeling to project southern Appalachian trout distribution in a warmer climate. *Transactions of the American Fisheries Society* **165**:1371–1382.
Giraldo R, Delicado P, Mateu J. 2012. Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica* **66**:403–421.
Haggarty RA, Miller CA, Scott EM, Wyllie F, Smith M. 2012. Functional clustering of water quality data in Scotland. *Environmetrics* **23**:685–695.
Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York, NY.
Hoover DR, Rich JA, Wu CO, Yang LP. 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**:809–822.
Ieva F, Paganoni A, Pigoli D, Vitelli V. 2013. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**:401–418.
Ignaccolo R, Ghigo S, Bande S. 2013. Functional zoning for air quality. *Environmental and Ecological Statistics* **20**:109–127.
Ignaccolo R, Ghigo S, Giovenali E. 2008. Analysis of air quality monitoring networks by functional clustering. *Environmetrics* **19**:672–686.
Jacques J, Preda C. 2014. Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis* **71**:92–106.

Journel AG, Huijbregts CJ. 1978. *Mining Geostatistics*. Academic Press: London.
Kayano M, Dozono K, Konishi S. 2010. Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *Journal of Classification* **27**:211–230.
Keleher CJ, Rahel FJ. 1996. Thermal limits to salmonid distributions in the Rocky Mountain region and potential habitat loss due to global warming: a geographic information systems (GIS) approach. *Transactions of the American Fisheries Society* **125**:1–13.
Lance GN, Williams WT. 1967. Mixed-data classification programs. I. Agglomerative systems. *Australian Computer Journal* **1**:15–20.
Le ND, Zidek JV. 2006. *Statistical Analysis of Environmental Space-Time Process*. Springer: New York.
Li H, Deng X, Kim DY, Smith EP. 2014. Modeling maximum daily temperature using a varying coefficient regression model. *Water Resources Research* **50**:3073–3087.
Mayer TD. 2012. Controls of summer stream temperature in the Pacific Northwest. *Journal of Hydrology* **475**:323–335.
Meisner JD. 1990. Effect of climate warming on the southern margins of the native range of brook trout. *Salvelinus Fontinalis. Canadian Journal of Fisheries and Aquatic Science* **47**:1065–1070.
Minns C, Randall R, Chadwick E, Moore J, Green R. 1995. Potential impact of climate change on the habitat and production dynamics of juvenile Atlantic salmon (*Salmo salar*) in eastern Canada. In *Climate Change and Northern Fish Population*, Beamish R (ed.). NRC Research Press: Ottawa, Canada, 699–708.
Mohseni O, Stefan HG, Erickson TR. 1998. A nonlinear regression model for weekday stream temperatures. *Water Resources Research* **34**:2685–2692.
Oliver M, Webster R. 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* **21**:15–35.
Ramsay J, Silverman B. 2005. *Functional Data Analysis* Second, Springer Series in Statistics. Springer: New York.
Rencher AC, Christensen WF. 2012. *Methods of Multivariate Analysis*. Wiley: New York.
Rencher AC, Schaalje GB. 2008. *Linear Models in Statistics*. John Wiley & Sons, Inc.: Hoboken, NJ.
Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics* **20**:53–65.
Sangalli L, Secchi P, Vantini S, Vitelli V. 2010. K-means alignment for curve clustering. *Computational Statistics and Data Analysis* **54**:1219–1233.
Sinokrot BA, Stefan HG. 1993. Stream temperature dynamics: measurements and modeling. *Water Resources Research* **29**:2299–2312.
Stefan H, Fang X, Eaton J. 2001. Simulated fish habitat changes in North American lakes in response to projected climate warming. *Transactions of the American Fisheries Society* **130**:459–477.
Tarpey T, Kinateder K. 2003. Clustering functional data. *Journal of Classification* **20**:93–114.
Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society B* **63**:411–423.
Tokushige S, Yadohisa H, Inada K. 2007. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics* **22**:1–16.
Trumbo B, Nislow K, Stallings J, Hudy M, Smith E, Kim D, Wiggins B, Dolloff C. 2014. Ranking site vulnerability to increasing temperatures in southern appalachian brook trout streams in virginia: an exposure-sensitivity approach. *Transactions of the American Fisheries Society* **143**:173–187.
Ver Hoef JM, Peterson EE, Clifford D, Shah R. 2014. Ssn: an r package for spatial statistical modeling on stream networks. *Journal of Statistical Software* **56**(3):1–45.
Webb B. 1996. Trends in stream and river temperature. *Hydrological Processes* **10**:205–226.
Webb BW, Clack PD, Walling DE. 2003. Water air temperature relationships in a Devon River system and the role of flow. *Hydrological Processes* **17**: 3069–3084.

