Journal of Data Science **9**(2011), 399-426

Multi-scale Clustering of Functional Data with Application to Hydraulic Gradients in Wetlands

Mark C. Greenwood[1], Richard S. Sojda[2], Julia L. Sharp[3],
Rory G. Peck[1] and Donald O. Rosenberry[2]
[1]*Montana State University,* [2]*U.S. Geological Survey*
and [3]*Clemson University*

Abstract: A new set of methods are developed to perform cluster analysis of functions, motivated by a data set consisting of hydraulic gradients at several locations distributed across a wetland complex. The methods build on previous work on clustering of functions, such as Tarpey and Kinateder (2003) and Hitchcock et al. (2007), but explore functions generated from an additive model decomposition (Wood, 2006) of the original time series. Our decomposition targets two aspects of the series, using an adaptive smoother for the trend and circular spline for the diurnal variation in the series. Different measures for comparing locations are discussed, including a method for efficiently clustering time series that are of different lengths using a functional data approach. The complicated nature of these wetlands are highlighted by the shifting group memberships depending on which scale of variation and year of the study are considered.

Key words: Cluster analysis, functional data analysis, generalized additive model, wetlands, hydrology, groundwater.

1. Introduction

Functional data analysis (Ramsay and Silverman, 2005) techniques have recently become popular as a way of analyzing data that are collected as multiple time series, where each series is observed frequently in time. Instead of viewing each time series as a set of n observations over time per series, each series is converted into a continuous, functional representation, often using spline-based techniques. Tarpey and Kinateder (2003) have discussed methods for cluster analysis of functional data using a version of k-means cluster analysis, exploiting the connections between the coefficients of the spline functions and the values that their respective continuous functions attain. Hitchcock et al. (2006) and Hitchcock et al. (2007) employ the k-medoids algorithm of Kaufman and Rousseeuw (1990)

for clustering functional data, assessing the role of different types of smoothing in the clustering of functions. Greenwood (2004) also explored clustering of functions using hierarchical clustering methods. Those methods dealt with functions where the interest was to compare the original observations to detect similar or different trajectories over time. Tarpey and Kinateder (2003) also explored clustering based on first derivatives of functions to remove locally linear portions of the signal and focus the cluster analysis on the variation around that lowest frequency component of the original time series. Greenwood (2004) considered registered curvature functions to compare the shape of different curves. These latter approaches suggest that functional clustering can be adapted to focus on certain aspects of multiple time series, which we extend in this work.

This study focuses on eliciting information on grouping of the thirty-two sampling locations that have similar patterns in their vertical hydraulic gradient time series, a metric related to the groundwater pressure at each site that determines whether a location is in "recharge" - water is moving out of the standing water and into the groundwater, or the location is in "discharge" - water is moving out of the groundwater and into the wetland. Grouping the locations that were studied based on either their long-term trends or diurnal variation in this measure provides insight into (1) whether there are locations that behave similarly on either scale (long term trends or diurnal variation) and (2) consistency (or lack thereof) of the groups of behaviors across different scales and the two years of the study. Establishing groups of locations with similar patterns, along with the form of their patterns, provides a first step in understanding the complex dynamics of the wetland and the role of each location in the underlying hydrologic system. Some conclusions about processes that might be driving variation on the different scales are possible through a detailed comparison of the different clustering results. Prior studies have not measured the variation in gradients at the fine time resolution used in this study and thus were not able to observe detailed changes in the trends or study the average diurnal variation in the gradients.

The data set contains some unique aspects that require novel methods to control the aspects of the time series that were compared and to maximize the use of available information for comparing different locations. First, the series have strong, and possibly non-stationary, low frequency components that tend to dwarf the scale of the high frequency oscillations. The low frequency components are of interest as they relate to whether the location is functioning in recharge or discharge at a particular time of year. The diurnal variations are high frequency effects likely to be related to evapotranspiration; they provide the opportunity to study whether this component of variation is consistent in timing of peaks and amplitude across sites, even if they have different trends. A nonparametric method is proposed to decompose each time series into two continuous compo-

nents, with functional clustering methods applied to each component. Second, data are available for each series for different lengths of time (i.e., different locations begin and end recording on different dates in each year). A new method for clustering functional data recorded over different time intervals is proposed. Our goal is to perform the cluster analysis in a way that fairly compares the series, but also maximizes the use of the available information over the course of the two summer field seasons in 2006 and 2007 and provides information on grouping of locations for the two components of variation.

1.1 Study area

Red Rock Lakes National Wildlife Refuge (NWR) is part of the high mountain Centennial Valley in Southwestern Montana, USA. The valley bottom where we collected the wetland hydrologic data is at an elevation of approximately 2,014 meters (m), with the nearby Centennial Range exceeding elevations of 3,000 m only 5 to 10 km to the south. The most outstanding ecological feature of the refuge is that it is a wetland complex over 10,000 hectares (ha) in size, and includes Upper and Lower Red Rock Lakes and Swan Lake. The complex is comprised of several types of wetlands, with permanent and semipermanent lacustrine aquatic beds and semipermanent and seasonal palustrine emergent wetlands being the most prominent. Much of Red Rock Lakes NWR is a federally designated wilderness with restrictions on mechanized access. Therefore, detailed climatic, hydrologic, and soil data are rare.

The study area's climate is continental, with cold summers and a short growing season (National Oceanic and Atmospheric Administration Western Regional Climate Center, 2005). At the weather station in nearby Lakeview, MT, the average of the yearly means from 1943 to 2008 of the daily maximum temperatures is 9.6°C (2006 and 2007 were amongst the warmest 25% of the years on this measure) and the average of the daily minimum temperatures is -5.9°C (2006 and 2007 were amongst the warmest 85% of the years). The temperatures were similar for the two years and relatively warm for the area, but not the most extreme on record, at least in terms of these measures. The average annual precipitation in the study area ranges from approximately 127 cm in the Centennial Range to 38 cm on the valley floor (U. S. Soil Conservation Service, 1977; Amend et al., 1986; Montana Natural Resources Information System, 2005). At Lakeview, the average yearly total precipitation is 52.8 cm, with 2006 at the 20th percentile and 2007 at the 42nd percentile, suggesting that 2007 was wetter than 2006 but both years were somewhat dry. The differences in yearly precipitation are important in understanding the results below.

A wetland is a complex system, involving linkages among the abiotic factors that drive wetland characteristics and their spatial and temporal patterns.

Wetland management is commonly hampered by a lack of knowledge about such ecological processes, although it is generally understood that hydrology is pivotal and groundwater often is assumed to play a key role. Although relationships among hydrological patterns, soils, and vegetation will be the final target of this research, the current study focuses on the hydrology observed in a network of wells and piezometers, with special interest in the interactions between surface water and groundwater. We documented site-specific groundwater discharge in the wetlands, and this study is an attempt to learn more about hydrologic patterns over time. Specifically, we examine possible groups of locations with similar long-term or diurnal variations in recharge/discharge behavior that suggest a similar function of those locations in the wetland.

Thirty-two locations were monitored in the wetland, half from the shallow lake area and the rest from palustrine (marsh origin) pond habitats, as displayed in Figure 1. To sample ponds, we selected palustrine emergent semipermanent wetlands (PEMF) classified by the National Wetland Inventory (U. S. Fish and Wildlife Service, 2004) that met three criteria: (1) ponds had to be $\geq$ 2,500 m^2 in area; (2) They could not be contiguous with the open water area of Lower Red Rock Lake, itself; and (3) No two ponds could be $\leq$ 200 m apart. These criteria were chosen to accommodate a companion research objective related to waterfowl use of the sites that is not reported here, resulting in pond and lake sites that were planimetrically similar to the lake sites. Eighty-five ponds met these criteria and 16 were chosen randomly for study. We selected sites within the open water habitat of Lower Red Rock Lake by selecting randomly from previously established random points (Paullin, 1973). Using these points as the northeast corner of the site, 50 m by 50 m plots were established that met two criteria: (1) the plots could not include emergent vegetation and (2) no two plots could be $\leq$ 200 m apart. Thirty-four sites met these criteria, and 16 were selected randomly. Sites from both groups were selected to appear visually similar and to present similar opportunities for waterfowl to utilize the sites from a vertical and horizontal habitat structure perspective. Aspects of hydrology were not used in the criteria for selecting the locations. By the nature of such wetlands, water levels varied from being dry at some times in most ponds to nearly 2 m deep at a few lake sites.

Comparisons across sites as to whether a site was dominated by groundwater discharge (i.e., hydraulic head in the subsurface being greater than surface-water stage, causing water to move from the shallow aquifer to the wetland) or by groundwater recharge (i.e., hydraulic head in the subsurface being less than surface-water stage, causing water to move into the shallow aquifer from the wetland) was of key interest to this study (Figure 2). A piezometer and stilling well pair using 3.81 cm diameter polyvinyl chloride (PVC) pipe was placed

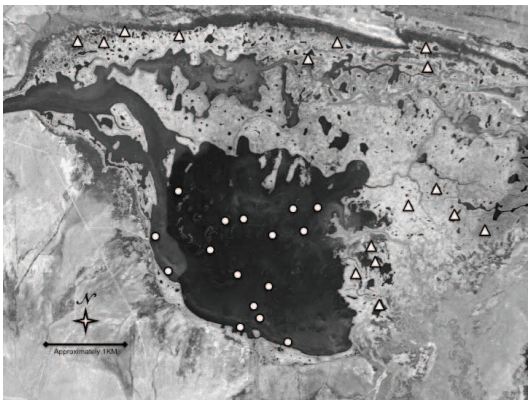

Figure 1: Map of Red Rocks Lake study area with pond (triangle) and lake (circle) sites.

at the northeast corner of each of the lake sites, and each was equipped with an automated, capacitance rod data logger that took a measurement every 30 minutes. Similarly, a piezometer and well pair was installed in each pond at an arbitrary location five meters from the edge of the emergent vegetation. Water levels in the wetlands were monitored with stilling wells whenever surface water was present. Whenever the wetland was dry, shallow ground water levels were monitored. In this system, water levels in the ponds reflect the level of the shallow ground water. Figure 2 shows this arrangement. The vertical hydraulic gradient, the response variable of interest, was calculated by finding the difference between water levels in a well and adjacent piezometer, and dividing that difference by the distance from the piezometer screen to the sediment-water interface of the wetland.

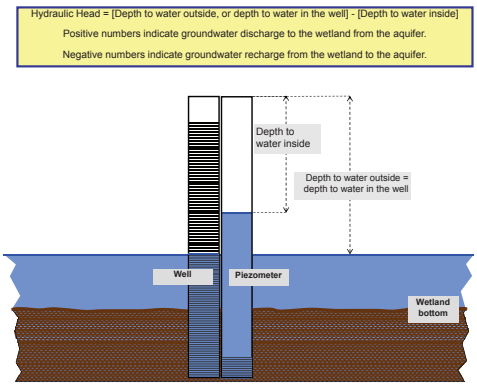

Figure 2: Schematic of piezometer and well measurement system. Surface water in the wetland is equivalent to the water table. This example illustrates a ground water discharge site.

Two years of vertical hydraulic gradient data from the network of 32 sites are considered here, with the number of locations varying slightly between the years due to instrument failure. In 2006, n=27 locations provided useful measurements for comparing trends among locations, with an average of 5,573 observations per location, taken every 30 minutes, leading to an average of 116 days of measurements per site. In 2007, n=29 locations provided useful measurements, with an average of 4,442 observations per location; and, this translates to approximately 93 days of observations for that year. A few outliers (23) were removed from the entire data set of around 284,000 observations, with these outliers typically caused by machine errors or instrument maintenance. Because of the small number of outlying observations relative to the total number of observations, their removal was not sufficient to impact the analysis.

2. Statistical Methods

The first step in the analysis is to model the time series at each location using a generalized additive model (GAM, Hastie and Tibshirani, 1990), with the GAM structured to nonparametrically decompose the time series into a long-term trend and an estimate of diurnal variation. Each component of the decomposition is considered a functional observation for each location for the subsequent cluster analysis. The second step involves defining metrics to assess differences in the functions on these two scales of variation to generate a distance matrix for use in cluster analyses related to each component of variation. Finally, analysis techniques that are well suited to clustering functional data such as these are discussed.

2.1 Nonparametric time series decomposition using GAMs

These hydraulic time series contain information on a range of different frequencies of oscillation. These oscillations range from long-term, low-frequency trends across the entire field season of approximately four months to high-frequency variation both at the diurnal level and higher frequencies. Using an additive modeling approach (Wood, 2006), we decompose each time series into the two components of variation that are of interest: the long-term trends and diurnal variation. By defining specific spline-based components in the GAM, we can decompose the signal into two estimated model components. The model for the vertical hydraulic gradient time-series, y_{kt}, at location k for time t between a and b, is $y_{kt} = \alpha_k + s(t)_k + s_D(t^*)_k + \epsilon_{kt}$. In our notation $s()$ indicates a smooth, nonparametric, penalized regression cubic B-spline component in the model, with the two spline bases defined to extract the desired frequency of the signal, decomposing the time series into two continuous components, trend, $s(t)$, and diurnal

variation, $s_D(t^*)$. The notation t^* indicates that only the fractional component of time is used, $t^* = t - \lfloor t \rfloor$ where $\lfloor\ \rfloor$ denotes the floor function. The time series for all n locations are modeled independently to estimate the model components for use in functional clustering discussed below.

To estimate a spline-based model component, a set of m basis functions, $\mathbf{B}$, such as those presented in Figure 3, are defined for each additive nonparametric model component using the spline formulas for either cubic spline or circular cubic spline bases (e.g., Wood, 2006, p. 150). The main difference between cubic splines and circular cubic splines is that for the circular splines, the mth basis function matches the first basis function, so only $m-1$ basis functions are used for the function. After defining the required basis functions, a smooth function, $s(t)$, is defined as the sum of the product of the m basis functions and their spline coefficients, $\boldsymbol{\beta}$: $s(x) = \sum_{i=1}^{m} \mathbf{B}_i(x)\boldsymbol{\beta}_i$. For penalized cubic regression splines, the spline coefficients are estimated as $\hat{\boldsymbol{\beta}} = (\mathbf{B}^\mathsf{T}\mathbf{B} + \lambda\mathbf{S})^{-1}\mathbf{B}^\mathsf{T}\mathbf{y}$, where $\mathbf{y}$ is a vector of the responses and $\mathbf{S}$ is chosen so that $\int_a^b [s''(t)]^2 dt = \beta^\mathsf{T}\mathbf{S}\beta$. This choice of $\mathbf{S}$ corresponds to working with penalized cubic regression splines and penalizing the variation of the second derivative of the function, which controls the roughness in $s(t)$. A smoothing parameter, λ, is then selected to minimize $||\boldsymbol{y} - \hat{\boldsymbol{y}}|| + \lambda\mathbf{B}^\mathsf{T}\mathbf{S}\mathbf{B}$ by cross-validation or, more often, generalized cross-validation (GCV). Specifically, the smoothing parameter, λ, is selected to minimize $GCV = T\sum_t (y_{kt} - \hat{y}_{kt})^2/(T - \gamma DoF)^2$, where T is the sample size for the time series, $\hat{y}_{kt}$ is the estimate from the model at time t, DoF are the total model degrees of freedom, and γ is a multiplier suggested by Kim and Gu (2004) to improve the selection of the smoothness of the functions. The suggested multiplier of 1.4 provides slightly smoother functions than would otherwise be selected. The effective degrees of freedom or *edf* for the different smoothers in the model can be calculated from the diagonal of the influence matrix implied by the form of the estimator for $\hat{\beta}$; *edf* describe the amount of information used to estimate each model component. DoF are the sum of the *edf* of the nonparametric components and the *df* of the parametric components (just the intercept in these models). For models with j smooth components, the $\lambda\mathbf{S}$ is replaced with $\sum_j \lambda_j\mathbf{S}$ to allow different penalties for the different smooths.

Adaptive smoothing is used for the long-term trend model component since the variation in the rate of change in the low frequency component may vary depending on the time of year. In many series, the difference between the regular regression spline and the adaptive smoother in the estimated effects are negligible except in times of more rapid change in the mean vertical hydraulic gradients. In these sections of the series, using an adaptive smoother provides better-fitting results while retaining smoothness in the trends when there is little variation in the trend. Adaptive smoothing involves choosing a different smoothing param-

eter, $\lambda(i)$ for each of the m spline coefficients, with the selection of the optimal

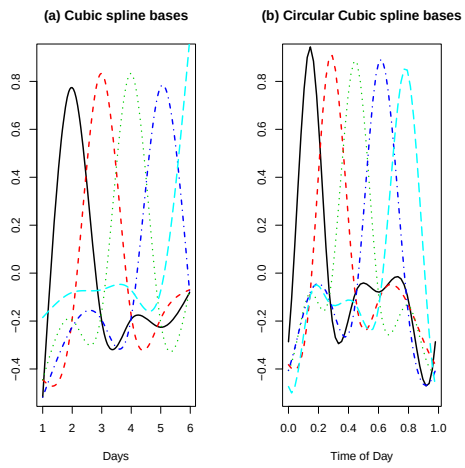

Figure 3: Plot of a simplified version of the cubic spline bases (a) used for the trend component in the model and the circular cubic spline bases (b) used in the diurnal component. More basis functions are used in the actual models but these are presented to illustrate the form of the splines being used for each model component.

values of the parameters using the same generalized cross-validation criterion; full details are available in the `mgcv` documentation (Wood, 2006). The component *edf* can be calculated as discussed above.

The two smooth model components are parameterized to have mean zero. Combining the intercept with the smooth function of time, $\alpha_k + s(t)_k$, provides the trend as a function of the day and time of observation for location k and retains information on the mean vertical hydraulic gradient for comparisons between sites. An example of the type of bases used for the trend component is displayed in Figure 3a, showing an example with five cubic spline bases. One less basis than the number of days with observations is used to model the trend component using adaptive smoothing methods. In contrast, for the diurnal component, $s_D(t^*)_k$, circular bases such as those in Figure 3b are defined so that the estimate at the end of a day, midnight or 1.0 in the plot, matches the estimate at the beginning of the next day, 0. Since 48 measurements were taken every day, 47 basis functions were used for the circular effect. By enforcing the periodicity of one day for this effect, the frequency of variation that this component estimates is fixed. More details on the definitions of the spline bases can be found in Wood (2006).

To understand the decomposition of the time series, the results for a single location across the two years of the study are presented in Figure 4. These results are typical of observations for the two years, with many locations exhibiting a swing from recharge (negative vertical hydraulic gradients) to discharge (positive

vertical hydraulic gradients) during the summer but the variation in both the diurnal components and the trends were greater in 2007 (the wetter year) than in 2006. Almost all diurnal components were significant in the models and the models explained a large amount of the variation in the time series. The four different model components, 2006 trend, 2006 diurnal component, 2007 trend and 2007 diurnal components are extracted for use in separate functional cluster analyses along with a cluster analysis of trends that goes across both years.

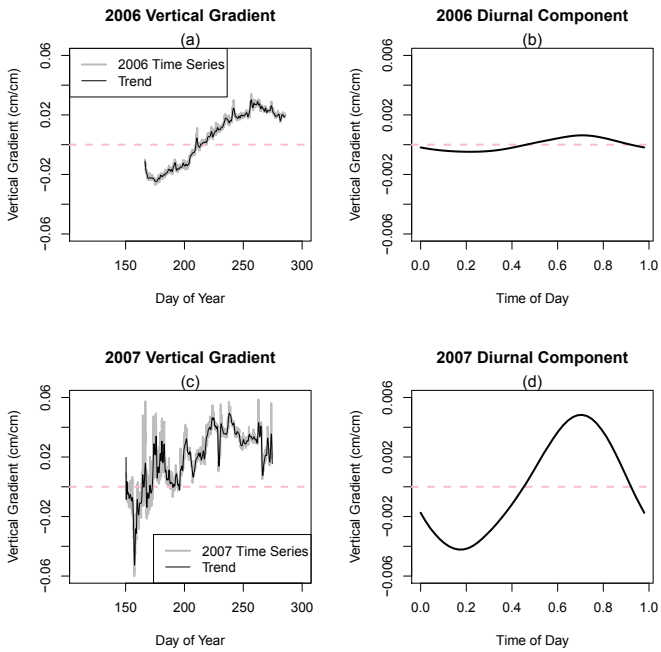

Figure 4: Plot of the estimated vertical hydraulic gradient time series and nonparametric trend function ($\hat{\alpha} + \hat{s}(t)$) for a location in 2006 (a) and 2007 (c) and the estimated diurnal component ($\hat{s}_D(t)$) for the 2006 (b) and 2007 (d) models.

Selecting the amount of smoothing for each model component via GCV is a quick and often useful technique. However, it can select models that under-smooth occasionally and it has been recommended by Kim and Gu (2004) to slightly increase the effective degrees of freedom in the calculation of the GCV criterion, multiplying by 1.4 to provide better results. We employ this recommendation in all the GAMs estimated here. It also is possible to attempt to account for the correlation between neighboring observations using generalized additive mixed model methods, possibly using a continuous autoregressive error structure inherent in the nature of the measurements. With typically around 5,000 observations per series in these data and a rich set of basis functions involved in the fixed effects in each model, this approach was not computationally feasible. In

testing this approach on reduced data sets, the trend estimates did not track the observations closely, being heavily smoothed when autocorrelation was incorporated into the model. So the previous degree of freedom inflation factor was used. Our interests are in extraction of the features, and these simpler models provide estimates of those features that track the actual observations closely, with models typically explaining most of the variation in the series (R^2_{adj} was above 0.95).

2.2 Functional semi-metrics

To perform a cluster analysis, or many other analyses, a metric for comparing observations must be chosen. A Euclidean distance measure is the most commonly used metric, with ties to typical linear models as well as a useful interpretation in terms of physical distances. A dissimilarity measure, $D(x, y)$, must obey four conditions to define a metric space: non-negativity, identity of indiscernables, symmetry, and the triangle inequality (Kaufman and Rousseeuw, 1990). Our primary metric is based on Euclidean distance, but integrates the squared difference of two continuous functions, $x_i(t)$ and $x_j(t)$, over the intersection of times that both functions are observed, from a to b: $D^2_{ij} = \int_{[a,b]}(x_i(t) - x_j(t))^2 dt$. This is similar to taking the sum of the squared pointwise differences of all observations between times a and b, especially if the functions are measured frequently in time. If B-splines are used to estimate the functions and a common set of basis functions is used, then the spline coefficients can also be compared with Euclidean distance to provide equivalent results to the integrated measure we employ. When matched B-spline bases are used, the result of the numerical integration is exact but no faster than when computed using an iterative numerical integration based on the trapezoidal rule (Ramsay et al., 2009).

In some situations, it is useful to relax the conditions for a full metric space to meet the needs of a particular analysis, leading to semi-metrics that maintain many of the properties of a distance metric but allow for comparisons of specific traits of data sets. Ferraty and Vieu (2006) explicitly discuss the possibility of relaxing the identity of the indiscernibles part of the metric definition to allow measures where having a distance of zero does not imply that the original functions are identical. One simple functional data semi-metric space involves comparing first derivatives of functions using Euclidean distance between the first derivatives of the functions. Two functions could differ but have matching first derivatives. In any situation where some pre-processing or smoothing is performed, this relaxed semi-metric definition may be useful. In our initial use of GAMs to decompose each series into two components, we have induced a similar semi-metric for each estimated component as we move from the original observations to focus on different aspects of the series.

For the smoothed trends in each year, $\hat{\alpha}_k + \hat{s}(t)_k$, different intervals of time

of measurement between the series must be considered. If we were to restrict our measure only to the intersection of time of all the locations as in prior functional clustering research, our resulting data set for comparisons would include comparisons over a very short time interval. Even searching for a window of time with generally high levels of recordings leads to a significant loss in time and as well as a few study locations. We instead define our semi-metric for the trend that compares each pair based on the time that both records overlap. In this way, we can maximize the information used for each pairwise comparison. This induces two complications: (1) the triangle inequality might be violated because the distance between two functions may not exceed the sum of the distances between the two functions and third function as different sections of the functions are compared in each distance calculation and (2) each comparison is based on windows of time of different lengths (essentially different numbers of the original discrete observations).

Relaxing the triangle inequality means that this measure loses the geometrical interpretation of distances that the triangle inequality provides, and methods that do not rely on full distance metrics must be used. This is only problematic for some cluster analysis techniques, but many methods of forming groups do not rely on the triangle inequality property. Other researchers (for example see Quackenbush, 2001) actually use the term semi-metric for dissimilarity measures that do not satisfy the triangle inequality. We prefer to refer to this more relaxed notion of differences as "dissimilarities", as suggested in Kaufman and Rousseeuw (1990). If the time of observation was the same for any set of three functions, then our dissimilarity measure meets the conditions of the semi-metric described above; otherwise, it fits with our conditions for a dissimilarity measure. Specifically, the measure is always greater than 0, is zero for the dissimilarity between an observation and itself, is symmetric, and, most importantly, it increases as two functions become more dissimilar. In fact, despite its limitations, the dissimilarity measure provides a physically interpretable difference between the vertical hydraulic gradient functions, with units of a difference in hydraulic gradient per day.

For the comparisons of the cyclic components, $s_D(t^*)_k$, with $t^* \in [0, 1)$, the Euclidean definition of distance between two functions integrates over the course of a day from 0 to 1. This provides a distance semi-metric, $D_{ij}^2 = \int_0^1 (s_D(t^*)_i - s_D(t^*)_j)^2 dt^*$. Because the integration is always between 0 and 1 for the diurnal component, the triangle inequality is satisfied for the clustering of the diurnal functions. However, since two functions could have identical diurnal components but have different original series, especially different trends, this is still a semi-metric measure.

In order to correct for different lengths of time in different comparisons of

trend components, $x_k(t) = \hat{\alpha}_k + \hat{s}(t)_k$, and make the comparisons over different lengths of time equivalent, we standardize the dissimilarity measure based on the length of time of the comparison. Let $L_{ij} = [a, b]$ be the intersection of the times over which both $x_i(t)$ and $x_j(t)$ are observed and let $\delta = b - a$ be the length of this time interval. The standardized dissimilarity measure for L_{ij} is defined as $D_{ij}^2 = \int_a^b (x_i(t) - x_j(t))^2 dt/\delta^2$. By dividing by the squared time of comparison for the pair, D_{ij} is standardized to be a measure of the average difference between the functions *per day*. This type of standardization follows from recommendations made by Kaufman and Rousseeuw (1990) to deal with missing values and a mix of categorical and quantitative variables used in a cluster analysis. A limitation of this method is that the intersection of times of observation, L_{ij}, must not be empty, otherwise we have no information to generate a pairwise dissimilarity measure.

The intersection of times of common observation for two series, L_{ij}, could also be made up of the union of disjoint intervals. This could be generated by a gap in recordings in the trend for a year or when trends for the two years are combined into a single analysis. Suppose that $L_{ij} = [a_1, b_1] \cup [a_2, b_2]$, $\delta = (b_1 - a_1) + (b_2 - a_2)$ and the dissimilarity measure is $D_{ij}^2 = [\int_{a_1}^{b_1} (x_i(t) - x_j(t))^2 dt + \int_{a_2}^{b_2} (x_i(t) - x_j(t))^2 dt]/\delta^2$. This modification was used to handle a gap of a few days of missing measurements in one time series in 2007 and to perform a cluster analysis of the trends across the two summer field seasons. For missing recordings of less than a day, we can estimate the missing values using the model and then compare the estimated functions, integrating the differences in the continuous functions over those small gaps. Typically single observations were removed; at most, three hours worth of observations were deleted and dealt with in this manner.

2.3 Functional cluster analysis

The dissimilarity measure defined previously can be used in two ways: (1) internally used in the clustering algorithm, or (2) pairwise dissimilarities can be used to form the clusters. k-means clustering of functions is an example of a clustering method that uses the measure internally to define a mean for each group and variability within the group. For the trend functions, $x_k(t)$, it is difficult to define a functional mean (Ramsay and Silverman, 2005) such as $\hat{\mu}(t) = \sum_k x_k(t)/\mathrm{n}$ when the times of observation are not the same for all the functions. This fact makes it difficult to envision a direct application of k-means clustering in the current situation. Also, clustering based on the mean may produce a center or "representative" that may not be representative of any of the observed functions. Summarizing the cluster with a mean is especially sensitive to an individual in

the group that is different from the others, but not different enough to be moved to a different cluster. Hierarchical clustering methods such as Ward's method can be applied to the dissimilarity matrix generated using the previous metric, but suffers from the same issues as k-means in terms of identifying representatives of each cluster. In many of these methods, selection of the optimal number of clusters is difficult. Numerical measures exist for objective selection of the number of clusters (for a review, see Milligan and Cooper, 1985), but there is no consensus on the choice of these measures and many require information that is not available if the clustering is performed only on the dissimilarity matrix, as is required here.

Kaufman and Rousseeuw (1990) provide an alternative clustering algorithm called partitioning around medoids (PAM) that provides a k-means type of cluster analysis, with a few important modifications. First, it can be performed on dissimilarity matrices as opposed to relying on access to the original observations to calculate the means of potential groups as in k-means. Second, the algorithm is more robust to outliers as it clusters relative to central observations, "medoids", instead of being based on the means of the groups. Third, it provides a method (called the average silhouette width) for choosing the number of clusters. Fourth, by exploring the initial clustering results, it is possible to identify possible outliers.

PAM uses k random initial starting points to build a cluster solution of size k around these starting points. This is very similar to how k-means clustering works with initial random seeds used to generate the preliminary clusters. However, PAM involves a two-stage algorithm that Kauffman and Rousseeuw (1990) call "BUILD" and "SWAP" that optimizes the initial choice of the cluster representatives in a way that leads to minimal variation in the cluster solution between different random starting points. It builds the clusters in an agglomerative fashion, but with the option to revisit assignments and switch the "representative" for each cluster. The net result of this is that, compared with k-means, it is not as necessary to repeat the cluster analysis many times to assess the variability in the solution based on the random starting points.

To generate a PAM solution, the algorithm is provided a dissimilarity matrix, $\mathbf{D}$ and the desired number of clusters, G. Define $\mathbf{M} = (M_1, \cdots, M_G)$ as a size G collection of the n functions and $d(x_k(t), M_g)$ as the dissimilarity between function k and cluster g, the difference between the function and the cluster representative which is easily found in $\mathbf{D}$. PAM selects the G medoids $\mathbf{M}^*$ that minimize the sum of the distances between the functions and the medoids, $\sum_i d(x_k(t), \mathbf{M})$. Each of the selected medoids, M_g^*, identifies a cluster, with the functions assigned to the cluster in which they are closest. The silhouette width of an observation is a measure of how well the individual observation matches the other items in the cluster. To calculate the silhouette width, for each function k, calculate the aver-

age dissimilarity of it with the other elements (k') in its cluster, $a_k = avg(D_{kk'})$. Then find the cluster that the function is closest to, the one that it has the smallest average distance to, $b_k = min_g(avg(D_{kk'}))$. The silhouette width of function k is then $(b_k - a_k)/max(a_k, b_k)$. The silhouette width ranges between -1 and 1, with larger values indicating more similarity between the function and the others in the cluster. The average silhouette width over all n observations provides an overall measure of cluster quality; maximizing the average silhouette width as a function of the number of clusters, G, provides a method to select the optimal number of clusters. Plotting the silhouette widths by cluster provides additional information on the quality of the cluster solution. Additional information is also available in the related "CLUSPLOT" (Pison et al., 1999) which illustrates dissimilarities between observations using two-dimensional multidimensional scaling (MDS) to construct n points that approximate the original dissimilarities. All cluster solutions were examined using silhouette and CLUSPLOTs to settle on a final cluster solution, but only the final cluster solutions are reported below.

The silhouette width depends on other clusters and can be inflated when a few very extreme outliers are included in the cluster analysis, making it misleading to include the outliers in the cluster analysis. For cluster solutions with two or three clusters and extreme outliers in the data, the outliers will cluster by themselves (Kaufman and Rousseeuw, 1990). By running each cluster analysis initially with only two and three clusters, it is possible to identify outlying functions as those that are in clusters containing only one or two observations. In these situations, Kaufman and Rousseeuw (1990) recommend removing those outlying observations and repeating the cluster analysis. In terms of identifying groups, it is important to understand functions that are unlike all the others, but more general patterns of behaviors are to be found in clusters containing multiple observations.

In a k-means analysis, results occasionally are summarized using an ANOVA for each variable (this is done by default in SPSS (2001), for example) or possibly a MANOVA to test for evidence of a difference across the groups that were formed in the partitioning. Permutation MANOVA (perMANOVA, Anderson, 2001) provides an alternative to the MANOVA that can be applied in this situation. Regular MANOVA cannot be applied because only pairwise dissimilarities are available; perMANOVA can be applied directly to the dissimilarity matrices and still furnish analogous results to a typical MANOVA. Using permutations to assess significance can provide reasonable inference when the normality assumption is dubious, as it would be here. Additionally, we could use perMANOVA to assess differences in the functions based on the type of location (lake or pond) or other characteristics of the sites (e.g., soil characteristics). We do not perform these analyses here, as we are first interested in exploring groups of patterns in the

hydrology to understand relative roles of the different locations and whether it is possible to detect different types of patterns in the short and long term aspects of the gradient time series.

Because PAM only requires the dissimilarity matrix, the fact that the dissimilarities are generated from functional data is not important for the algorithm. A clustering algorithm will find groups based on the dissimilarity measure used. As noted in the introduction, our methods are similar to Hitchcock et al. (2006) and Hitchcock et al. (2007) in using PAM for clustering functional data. Our dissimilarity measures differ from theirs in terms of the alignment issues, smoothing decomposition, and standardization of the differences based on the time of the pairwise observations. A different approach to clustering functional data is provided in James and Sugar (2003), where a model-based approach to deal with similar missing data issues as ours is used. Their method is more computationally intensive, involving the expectation-maximization algorithm to estimate the unknown group memberships. The method presented in James and Sugar (2003) provides some additional inference benefits possible from their "functional cluster model"; the opportunity for inference for the mean functions in the groups is realized because the method is model-based. It is unclear how their method would handle outlying functional observations, as normality assumptions are used in deriving the procedure. The high dimensionality of our data set could be problematic for their approach, which also involves estimation of covariance matrices. Other approaches to clustering functions such as Tarpey and Kinateder (2003), Abraham et al. (2003), and Serban and Wasserman (2005) rely on the B-spline coefficients and k-means algorithms to generate the clusters. In each approach, the spline bases are the same for all the functions being clustered. Because our method relies on numerical integration of the continuous functions in D_{ij}^2, our procedure easily could accommodate different bases underlying the individual functions. This occurs to a small degree in our data set, where missing observations change the spacing of the knots slightly. Additionally, as the length of overlap changes between different pairs, the basis functions generated by each comparison vary. There are additional years of observations where vertical hydraulic gradients were measured, but much less frequently; to compare results across field seasons with the different sampling rates, different densities of basis functions could be used and results compared using our techniques. These extensions of the methods proposed here are left for future research.

3. Results and Discussion

3.1 GAM results

GAM models were fit to the time series for each location in each year to pro-

vide the required nonparametric model components for the different functional cluster analyses. A few outliers (23) were removed from the entire data set that were believed to have been caused by equipment errors or instrument maintenance. For the 27 sites observed in 2006, there were 4,608 overlapping times of observation (96 days). In 2007, there were 534 overlapping times of observation (11 days) for the 29 sites due to drier conditions leading to more unobserved gradients in pond sites. In 2007, the missing data modifications are more prominent than in the same analyses for 2006.

For the 2006 time series, the adjusted R^2 values of the 27 generalized additive models ranged from 0.78 to approximately 1.00. The median of the R^2_{adj} values of the 27 generalized additive models fit for the 2006 data was 0.99. The minimum of the *edf* for the long-term trend in the models fit for the 2006 data was 137.20 and for the diurnal trend, the minimum *edf* was 5.57. The maximum *edf* for the 27 models fit for the long-term and diurnal trend in the 2006 data were 144.90 and 19.12, respectively.

For the 2007 time series, the R^2_{adj} values of the 29 generalized additive models ranged from 0.56 to approximately 1.00 with a median of 0.94. The *edf* for the long-term trend had a minimum of 147.80 and a maximum of 184.90. The *edf* for the 29 model fits for the diurnal components in 2007 ranged from 0.10 to 24.74. One location that had effective degrees of freedom less than one (0.10) was deleted from the diurnal clustering; thus, the minimum *edf* for the functions used was 4.10.

The diurnal variation generally had a larger amplitude in 2007 than the same functions in 2006. The changes in the vertical hydraulic gradient trends were larger for many locations in 2006 than in 2007, but the trends in 2007 were more variable over time than in 2006. The general differences in the functions between the years suggests that the system was operating differently in the two years. The most clear difference between the years was that 2007 was wetter than 2006, which may be a major factor in explaining the higher variability in the trends and stronger diurnal components.

3.2 Cluster analysis results

For the 2006 trend data, initially one functional trend was flagged as an outlier in a two group PAM analysis; it was removed from the cluster analysis. Clustering on the remaining 26 locations, a solution with five clusters is identified. Focusing on the medoids for each group in Figure 5(b), three types of shapes of trends are present in the five clusters. Three clusters (2006T-2, 3, and 4) indicate negative gradients early in the season that become positive later in the season. Additionally, one site (2006T-5) in the five group cluster analysis has a somewhat large positive gradient, suggesting that it is a discharge site, with a slow decrease

in vertical hydraulic gradient over time. The remaining cluster corresponds to the locations that have generally small positive gradients (discharge) that do not change much over the season. The representative of this group drops below zero for a few weeks in the later part of the record, but generally the members of this cluster have small positive gradients.

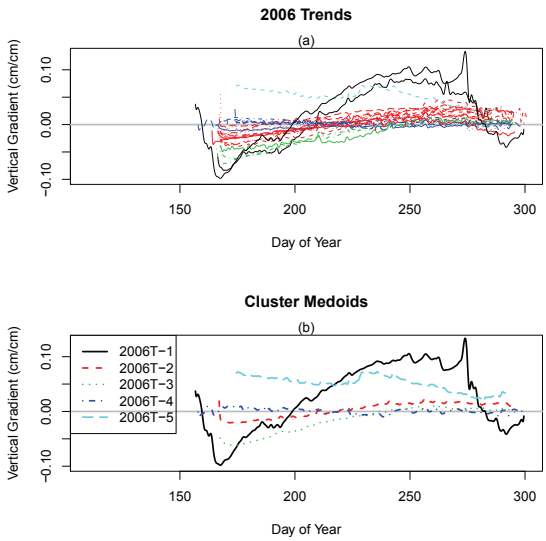

Figure 5: Plot of clustered 2006 trends (a) and cluster medoids (b).

The three clusters that change from recharge (negative gradient) to discharge (positive gradient) may be the most interesting in this group of locations. The first cluster that has the largest range of gradients (2006T-1) contains a lake and a pond site that are quite similar. The largest cluster (2006T-2) has only moderate fluctuation over the season, containing two pond sites and eight lake sites, and transitions between negative and positive gradients sooner than the third cluster (2006T-3). The third cluster has more negative gradients early in the season and barely attains positive gradient values, and contains only lake sites. These trends share common shapes but also have very different implications for vegetation in each location depending on the time during the season that each site is either in discharge or recharge. The common shapes suggest a common mechanism driving the gradients at a set of locations even though the length of time each group of locations is in recharge or discharge differs systematically. Since the differences in timing of transition from recharge to discharge are not easily explained by type of site or location in the study area, further exploration of site characteristics will be required to fully understand this combination of similarities and differences.

For the 2007 functional trends, the higher variability in the trends makes detecting more subtle groupings difficult. Two locations, both pond sites, were

initially identified as being outliers in a three group cluster analysis, both having extremely large positive gradients that were unlike all the other locations and each other. After removing those two observations, only two additional clusters were selected, with twenty-one locations in the first cluster (2007T-1), demonstrating a similar pattern to many of the observations as in the previous year (Figure 6) that start with a negative gradient and becoming positive later in the season. Seven of the sites in this cluster were pond sites. The second cluster (2007T-2) contained five ponds sites which had generally much higher positive gradients than the other group early in the season. Distinct groups did not emerge from this set of locations when the solution was pushed to have more clusters, instead the new clusters introduced more negative silhouette values which made the average silhouette width worsen dramatically. This suggests that further subtle differences in 2007T-1, sites that transition from recharge to discharge, were not detectable, likely because of the higher variability in trends in 2007.

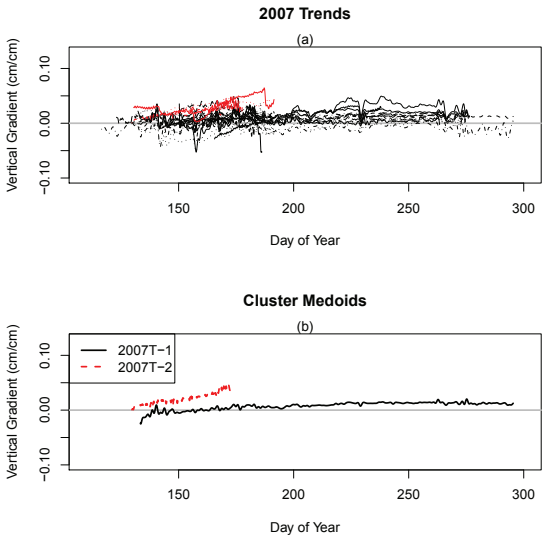

Figure 6: Plot of clustered 2007 trends (a) and cluster medoids (b).

Performing a cluster analysis of the trends across the two years of the study requires the modified dissimilarity measure for the union of disjoint intervals. Observations that are unusual in either 2006 or 2007 must be removed as they continue to be identified as outliers in the analysis spanning the two years. There are 20 sites (10 of each type) that exhibit shared behavior in the cluster analysis with at least one other location in this analysis. The first cluster (0607T-1) in Figure 7 contains sites that transition from recharge to higher discharge levels. The second cluster (0607T-2) contains sites that tend to be closer to zero gradient

in 2006 and are generally discharge sites throughout 2007, with these sites generally changing less over time than those in the first cluster. The links between the transitions made in 2006 and resulting shapes of similar locations in 2007 are some of the most profound results of the study, showing that behaviors of the same locations are different in two years, but that some locations responded to the differing climatic conditions in a similar fashion across the two years. Said another way, the climatic conditions differed between 2006 and 2007, so the vertical hydraulic gradients differ between the years. Identifying groups of locations that responded with different patterns to 2006 and 2007, but were similar within the group across the two years, is an interesting finding.

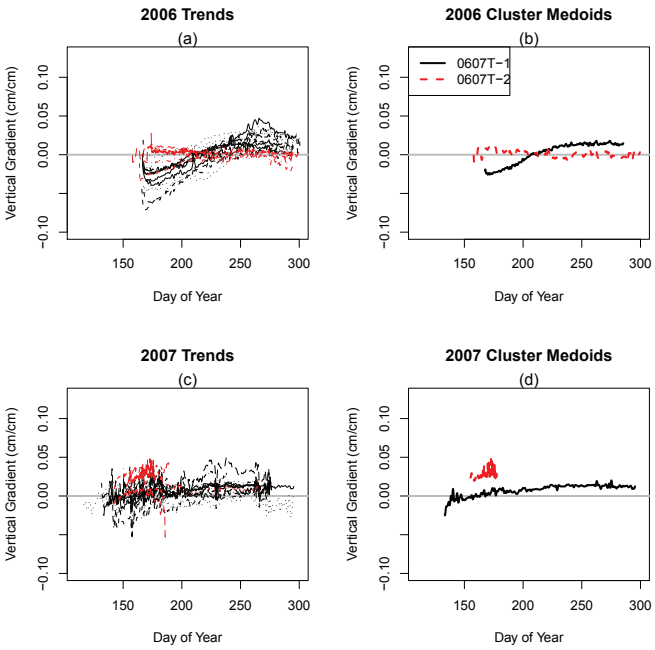

Figure 7: Plot of clustered 2006 (a) and 2007 (c) trends with medoids for the combined 2006 and 2007 trend clusters in (b) and (d).

For the diurnal functions, the amplitude of the components is larger in 2007 than 2006 but the cluster analyses of both year's functions are interesting. For the 2006 diurnal components, one location had a larger amplitude than the other locations, was identified as an outlier, and was removed. The final cluster analysis has eight clusters, as shown in Figure 8. While the amplitude varies between locations, the main differences in the functional cluster medoids are in the phase and shape of the functions. The first cluster (06C-1) contains two lake sites and attains a peak gradient in the early evening, and has both a distinct negative and positive amount of variation. The next cluster (06C-2) also contains two

locations, but one is a lake site and one is a pond site. The shape of these functions is similar to a step function; the functions are either constant values that are positive or negative, with positive values occurring at night and negative values occurring during the day. The third cluster (06C-3) contains four locations, two of each type, and increases slightly but never attains large positive gradients. The largest cluster (06C-4) has seven members, all lake sites, and has a similar pattern to the first cluster with a much smaller amplitude, peaking around early afternoon and containing a distinct negative gradient component. The fifth cluster (06C-5) has five members but only one lake site, and a unique phase with a peak positive gradient in the morning and a peak negative gradient in the early evening. The sixth cluster (O6C-6) contains one of each type of site and also has a step function shape, with a different phase than the second cluster, but it is otherwise very similar. The seventh cluster (O6C-7) resembles the first cluster in shape, but attains its maximum around noon. It contained three sites, with one being a lake site. The last cluster (06C-8) contains only a single location, and is displayed in gray in Figure 8(b). The rich set of different diurnal functions that this cluster analysis detected indicates the great variety of fluctuations that were possible to observe in these locations during this year.

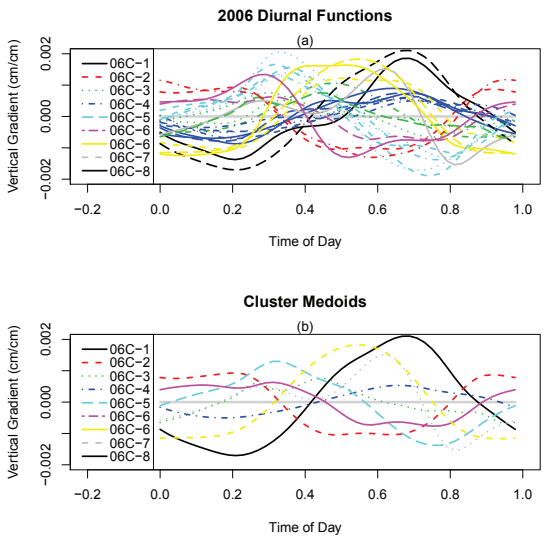

Figure 8: Plot of clustered 2006 diurnal functions (a) and cluster medoids (b). Eight clusters identified, with final cluster only containing a single observation.

For the 2007 diurnal functions, where the amplitude of the functions was generally larger than for the 2006 data, results from the cluster analysis in Figure 9 are very simple. Fundamentally, the clusters reflect whether the locations are pond or lake sites. A difference in the timing of the peaks drives the differences

between the locations that were detected in the cluster analysis. The lake sites make up the first cluster (07C-1) and have a large negative gradient early in the day and a peak positive gradient late in the day. The pattern is reversed for the pond sites (07C-2), with a peak positive gradient in the early morning and a peak negative gradient late in the day. Note that one pond location did not have any variation in the diurnal component and was not used in this cluster analysis. Removing this location corresponds to recommendations made in Serban and Wasserman (2005), where functions with no variation are removed prior to further cluster analysis. Across all the clustering performed, these are the only results that seem to have any direct relationship to the type of site (i.e., lake or pond). There are systematic differences in the diurnal functions between the two years that include a greater magnitude of diurnal fluctuations in 2007 and their coherency by type of site.

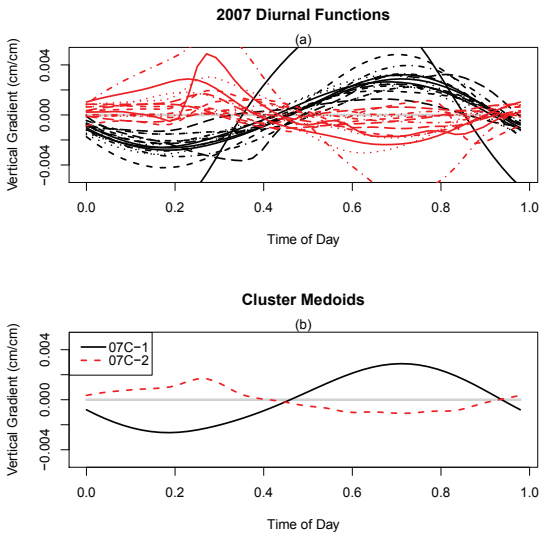

Figure 9: Plot of clustered 2007 diurnal functions (a) and cluster medoids (b).

The diurnal patterns for the lake sites potentially can be explained by this variation being driven primarily by evaporation and not transpiration. Daily evaporation would draw the lake water levels down, enhancing the apparent discharge in the latter parts of the day. However, this explanation is complicated by surface water coming into the lake. In this snowmelt-dominated system, one would expect greatest volume of inflow to occur in the late afternoon or early evening. For the smaller pond sites, the ratio of perimeter to surface area is larger so transpiration plays a larger role. Vegetation pumps water out of the sites earlier in the day and then may be shutting down later in the day. These re-

sults suggest further research on the sources of this variation would be of interest to isolate the local influences on this variability.

4. Conclusions

Cluster analysis techniques prove useful in separating locations based on hydrological characteristics. Cluster analyses and regression trees were previously employed for similar data across the 2003, 2004, and 2005 sample years for investigating wetland hydrology and vegetation in the Red Rock Lakes National Wildlife Refuge (Sojda et al., in preparation). The results of these analyses were based on yearly averages of hydrologic variables and vegetation, thus excluding any daily, weekly, or seasonal fluctuations in the series from consideration. The research presented here is significant in advancing the field of functional clustering when missing observations are encountered and in comparing multiple time series on different scales.

4.1 Practical implications

Determination of causal factors for the differences in trends and daily cycles of the observed hydraulic gradients between sites and years is beyond the scope of this paper, but some preliminary conclusions can be made. In 2007, the wetter of the two years studied, the diurnal fluctuations had a larger amplitude and the patterns observed were related to the type of site, lake or pond. In contrast for the similar analysis in 2006, the amplitudes were smaller and many different shapes (amplitude and phase) were observed with no obvious link between the clusters and the type of site. The differences in shapes and organization in the diurnal patterns between the two years could be explained by a combination of the wetness providing more surface water to be evapotranspirated, the wetter year having more or healthier vegetation to enhance the evapotranspiration, and a stronger snowmelt input in 2007. The fact that the patterns organize by type of site only in 2007 and have different timing between the two clusters suggests that surficial features are critical in this scale of results.

The groups detected in clustering the trends from each year or across years were not related to the results for the diurnal components, suggesting that variation on the two scales is driven by different processes. If diurnal processes, which are shown to differ between the two years, are tied to evapotranspiration, then these hydrologic processes are related to the surface water and vegetation. The long-term patterns, which seem to either make a transition from recharge to discharge during the summer or be discharge throughout, are more directly related to interplay between the local groundwater and surface water, but less impacted by the neighboring surface characteristics. This conclusion is based on the lack

of organization of the trends relative to the type of site in any of the analyses but the diurnal results that can be related to the types of sites.

The explanations for the specific patterns in longer-term trends are more elusive. The groundwater delivery system is very complicated and may be changing over time. The study area encompasses alluvial fan deposits, deltaic sands, lacustrine sediments, etc., which makes it complicated to know how groundwater moves, especially at the fine time scales considered in this study. The differences in groups across years and scales suggests that the system is not constant over time as it responds to differences in climate that affects the delivery of surface water and groundwater, evapotranspiration rates, and vegetation patterns. This study provides a glimpse into the complexity of the hydraulic system by studying the vertical hydraulic gradients on different scales.

4.2 Statistical implications

The first derivatives of the trends also could be considered to be able to link sites with similar shapes but different average responses, similar to methods in Tarpey and Kinateder (2003). However, using the first derivatives would remove information regarding the mean of the vertical hydraulic gradient which is directly related to a location's status as a recharge or discharge site, a key component of this research. We believe that the magnitude of the vertical hydraulic gradient is as important as its variation, and our clustering methods reflect that idea.

In the definition of the functional dissimilarity metric, two advances are provided. First, a simple method is provided for performing functional clustering when functions are observed over different times. This approach maximizes the use of available information for pairs of functions in making comparisons. Second, the dissimilarities are calculated using numerical integration without specific reference to common basis functions as in some other functional clustering research. Different bases may be required when encountering missing observations as in our application or when different locations may have different sampling rates. This could occur by design or due to equipment malfunction. Especially as technology evolves, sampling rates could increase in a long term record. This metric could adapt to the differences in resolution and still provide meaningful results.

Other researchers may find applications for our cluster analysis of functions from GAMs in situations where the cyclic component is of greater interest than the trend or where the derivative of the trend is of greater interest but some higher frequency variation is not of interest. These methods can easily be adapted to highlighting different features of a process. In our application, the trends dwarfed the diurnal variation. If we had wanted to re-combine the two measures to create a cluster analysis across the trends and diurnal components, we would have had to increase the weight on the differences in the diurnal comparisons. A

weighting coefficient could easily be included in a combined dissimilarity and used to tune the cluster analysis across these comparisons. We avoided this arbitrary choice in the methods presented, keeping all of our comparisons on a scale of a difference in the vertical hydraulic gradient per day. This would, however, provide a framework for clustering multivariate functional data since more than one function is observed per individual.

5. Future Work

Since many of the observations contain trends that change from negative vertical hydraulic gradients to positive vertical hydraulic gradients, the estimated times of transition from negative to positive gradient could also be extracted from the trends and compared. It would be possible to explore this aspect using our methods by first defining basis functions that allow for discontinuities, then estimating functions that are defined over the study period but are discontinuous based on the crossing of 0. Then the same dissimilarity measures discussed above could be employed to cluster based solely on this aspect of the time series. Detections of statistical evidence of zero-crossings might also be of interest. Methods to identify crossing-points have been explored in Significant Zero Crossings (SiZer, Chaudhuri and Marron, 1999). These methods could also be applied to the cluster medoids to obtain a better understanding of the clusters by providing detailed evidence for when the representative has crossed from recharge to discharge.

The next step in our research is to attempt to relate these results to information on vegetation and soils at each site, both in terms of the long term trends and the differences in diurnal variation. That will require a further set of tools to use these groups of functions and understand how they can predict multivariate vegetation species compositions at different locations or are related to soils information.

Acknowledgements

R. H. Buxton and J. Warren were instrumental in safely installing wells and piezometers in harsh, wilderness conditions, as well as downloading and processing data. K. Pierce and S. Custer provided key understanding of the geohydrology of the Centennial Valley. L. Fredrickson's encouragement to study the role of groundwater in this wetland complex was pivotal. The logistic support and technical help of Red Rock Lakes staff was outstanding. T. Preston assisted with data organization. D. Paullin's original random sampling scheme in Lower Red Rock Lake also is recognized. Review by A.L. Gallant and B.M. Troutman of the USGS and an anonymous referee substantially improved the manuscript. This project was funded by the USDI-Geological Survey.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Abraham, C., Cornillon, P.A., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *The Scandinavian Journal of Statistics* **30**, 581-595.

Amend, J., Dolan, D., Haglund, B.G., Hansen-Bristow, K., Locke, W. and Montagne, J. (1986). A model for information integration and management for the Centennial Ecosystem. *Greater Yellowstone Coalition*, Bozeman, MT, 183.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46.

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 484-499.

Dingman, S.L. (1994). *Physical Hydrology.* Macmillan Publishing Company, New York.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* Springer, New York.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 623-637.

Greenwood, M. (2004). *Functional Data Analysis for Glaciated Valley Profile Analysis.* Dissertation, University of Wyoming, Laramie.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall / CRC Press, Boca Raton.

Hitchcock, D. B., Booth, J.G. and Casella, G. (2007). The effect of presmoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation* **77**, 1043-1055.

Hitchcock, D., Casella, G. and Booth, J. (2006). Improved Estimation of Dissimilarities by Presmoothing Functional Data. *Journal of the American Statistical Association* **101**, 211-222.

James, G. and Sugar, C. (2003). Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association* **98**, 397-408.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

Kim, Y. J. and C. Gu (2004). Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society B* **66**, 337-356.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159-179.

Montana Natural Resources Information System (2005). Available at URL http://nrix.mt.gov/wis/data/precip.htm.

National Oceanic and Atmospheric Administration Western Regional Climate Center. (2005). Available at URL http://www.wrcc.dri.edu/summary/climsmmt.html.

Nielson, E.C. and Farnsworth, P.H. (1966). *Soil Survey Handbook for Red Rock Lakes Migratory Waterfowl Refuge in Survey Area 061, Dillon, Montana.* U.S. Department of Agriculture Soil Conservation Service.

O'Neill, J.M. and Christiansen, R.L. (2002). *Geologic map of the Hebgen Lake quadrangle, Montana Bureau of Mines and Geology Open File Report 464. 21 page(s), scale 1:100,000.* Available at http://www.mbmg.mtech.edu/sbcitation.asp?cit1=MBMG cit2=464.

Paullin, D. G. (1973). The ecology of submerged aquatic macrophytes of Red Rock Lakes National Wildlife Refuge, Montana. Thesis, Montana State University, Bozeman, Montana.

Pierce, K. L. and Morgan, L. A. (1992). The track of the Yellowstone hot spot: Volcanism, faulting, and uplift, in *Regional Geology of Eastern Idaho and Western Wyoming: Geological Society of America Memoir*, edited by Link P. K., Kuntz, M. A., and Platt, L. B., 179, p.1-53.

Pierce, K. (2006). Personal communication.

Pison, G., Struyf, A., and Rousseeuw, P. J. (1999). Displaying a clustering with CLUSPLOT. *Computational Statistics and Data Analysis* **30**, 381-392.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews*: *Genetics* **2**, 418-427.

R Development Core Team (2009). *R*: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York.

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2009). fda: Functional Data Analysis. *R package version 2.1.2.* http://CRAN.R-project.org/package=fda.

Serban, N. and Wasserman, L. (2005). CATS: clustering after transformation and smoothing. *Journal of the American Statistical Association* **100**, 990-999.

Sojda, R. S., Greenwood, M.G., Sharp, J. L., Rosenberry, D. O. and Warren, J. M. (2010). Statistical classification of water levels and hydraulic gradients in relation to plant communities in a northern Rocky Mountain wetland complex. In preparation.

Sonderegger, J.L., Schofield, J.D., Berg, R.B. and Mannick, M.L. (1982). The upper Centennial Valley, Beaverhead and Madison Counties, Montana. *Montana Bureau of Mines and Geology Memoir.* 50.

SPSS for Windows, Rel. 11.0.1. 2001. Chicago: SPSS Inc.

Tarpey, T. and Kinateder, K. (2003). Clustering functional data. *The Journal of Classification* **20**, 93-114.

Tippy D.L., Hahn, J.P., Tribehoun, R.M. and Nimlos, T.J. (1978). Soil survey, Dillon Resource Area Resources Inventory. Montana Forest and Conservation Experimental Station, University of Montana, Missoula, Montana.

U. S. Fish and Wildlife Service (2004). National Wetlands Inventory website. *U.S. Department of the Interior, Fish and Wildlife Service.* St. Petersburg, Florida. URL http://www.nwi.fws.gov.

United States Soil Conservation Service (1977). Average annual precipitation for Montana for the 1941-1971 base period. *United States Natural Resource Conservation Service Snow Survey Unit.* Bozeman, Montana.

Wood, S. (2006). *Generalized Additive Models*: *An Introduction with R.* Chapman and Hall / CRC Press, Boca Raton.

Wood, S. (2009). mgcv *R Package Version 1.5-5.* http://CRAN.R-project.org/package=mgcv.

Received February 20, 2010; accepted June 22, 2010.

Mark C. Greenwood
Department of Mathematical Sciences
Montana State University
Bozeman, MT, USA 59717-2400
greenwood@math.montana.edu

Richard S. Sojda
Northern Rocky Mountain Science Center
U.S. Geological Survey
Bozeman, MT,USA 59715
sojda@montana.edu

Julia L. Sharp
Department of Applied Economics and Statistics
Clemson University
Clemson, SC, USA 29634
jsharp@clemson.edu

Rory G. Peck
Department of Mathematical Sciences
Montana State University
Bozeman, MT, USA 59717-2400
rorypeck@gmail.com

Donald O. Rosenberry
U.S. Geological Survey
MS 413, Bldg. 53, Box 25046
Denver Federal Center
Lakewood, CO, USA 80225
rosenber@usgs.gov