Greedy Copula Segmentation of Multivariate Non-Stationary Time Series for Climate Change Adaptation

Taemin Heo[1*], Lance Manuel[1]
[1]*Department of Civil, Architectural and Environmental Engineering*
The University of Texas at Austin, 301 E Dean Keeton St C1700, Austin, TX, USA, 78712
[*]*Corresponding Author: <taemin@utexas.edu>*

Abstract

An assumption of stationarity in climate-related processes is often made in the risk assessment of civil infrastructure systems. Such an assumption is difficult to justify in a changing climate. In this study, to optimally adapt to a changing climate, given time series data, we propose a computationally efficient algorithm called Greedy Copula Segmentation, GCS, that could potentially be used in a climate change adaptation (CCA) strategy. The GCS algorithm partitions a multivariate time series into disjoint segments such that each of the segments is described by a stationary copula process, but independence is assumed across segments. An optimal strategy for climate change adaptation, which we will refer to as GCS-CCA, considers the last or most recent segment as containing the most informative data for near future climate pattern prediction. By only using such informative data to build a probabilistic model for the near future, our method effectively accounts for climate change. We provide an algorithmic formulation for greedy segmentation and validate the performance of our GCS-based strategy by applying it to an illustrative benchmark problem and a realistic drought example.

Keywords*: time series segmentation, greedy algorithm, climate change adaptation, non-stationary stochastic process*

1. Introduction

Climate data such as precipitation, wind speed, etc. make up a fundamental source of information for the risk assessment of any civil infrastructure system. Often, the climate parameter is represented as a stationary stochastic process, which then implies that the risk assessment makes use of all the available historical data in prediction. Temporal patterns that include extreme climate events such as droughts, floods, storms, etc. can be quite variable due to inherent non-stationary characteristics and as an outcome of human-induced climate change (Lee and Ouarda, 2010; Sheffield et al., 2012; Dai, 2013; Garcia Galiano et al., 2015; Li et al., 2015; Cid et al., 2016; Van Loon et al., 2016; Ouarda and Charron, 2018; Liu, S. et al., 2019; Slater et al., 2020). In such cases, there are reasons to perhaps consider climate as a piecewise stationary process. In the outlined approach, we show that it is possible to consider the non-stationary characteristics of the underlying climate process by means of sub-segments that are each stationary but mutually independent. Then, near-future patterns can be realistically assumed to be closest to the most recent sub-segments. Consequently, a model derived from such recent data might be expected to lead to better predictions than what we get with the traditional approach that uses the entire historical sample.

This work demonstrates how a time series segmentation technique can be used to identify the optimal multivariate climate data subset near-future risk assessment. The method involves breaking up the input time series into segments where the data in each segment are treated as independent samples from a copula. We propose the use of what we call the greedy copula segmentation (GCS) algorithm, that systematically searches for optimal-length recent segments using a greedy algorithm. Among the identified segments, the most recent sub-segment is then selected as the optimal data for any future planning such as in CCA.

Our method builds from and extends the greedy Gaussian segmentation (GGS) developed by Hallac et al., 2019. The assumptions and formulation of GGS are well-suited to our problem. GGS assumes non-repeatability of segments; this means that model parameters in each segment are unrelated to parameters in other segments. Considering that we are dealing with climate conditions that are widely acknowledged to be changing, the non-repeatability assumption is justified. GGS formulates the time series partitioning problem based on the maximum log-likelihood of the data. Since we are assuming piecewise stationarity and wish to project near-future patterns based on recent observations, a maximum log-likelihood based approach is most appropriate for our problem.

We extend the applicability and generality of GGS by replacing the multivariate Gaussian distribution assumption with a multivariate copula choice. This extension is especially appropriate for our problem since climate variables often follow non-Gaussian distributions (Zelenhasic and Salvai, 1987; Mathier et al., 1992; Yue et al., 1999; Shiau and Shen 2001; Yue 2001; De Michele and Salvadori 2003; Hao and Singh 2013; Mazdiyasni et al. 2019). The use of multivariate copulas can help to represent many complex multivariate dependence structures both by employing various options for marginal distributions and by selecting different copula families (Saklar, 1959; Salvadori, 2004; Salvadori and De Michele, 2004; Nelsen, 2006; Genest and Favre, 2007). Moreover, most common marginal distribution parameters can be empirically obtained from data using maximum likelihood estimation (MLE). The copula family parameter can also be non-parametrically estimated using the empirical Kendall's rank correlation coefficient, tau (Genest et al. 2011, Manuel et al. 2018). For all of these reasons, we propose the use of a more versatile GCS approach, while not losing advantages of the mathematical tractability of GGS.

For civil infrastructure systems, we demonstrate how we can adopt an adaptation policy based on the proposed GCS approach. A 5- to 10-year cycle of climate data that possibly involves policy amendment (CCA) usually starts by updating the site-specific hazard data. Then, derivative policies are updated accordingly. The projected risk assessment will be best for only a near-future period because, after this period, the policy will need to be amended with any newly discovered information/data. We demonstrate how to employ such new data along with all the available historical data to update temporal hazard patterns and derivative policies. Again, in light of the most recent climate change trends, dated data are unlikely to contain meaningful information for near-future projections. In fact, the use of old data can cause a model to exhibit greater bias and uncertainty due to heterogeneity in the data due to non-stationary character. By using the proposed GCS-identified optimal data, we only use informative recent data to update policies. In other words, GCS-CCA discards outdated data to improve prediction performance. It works more discriminately to detect and account for short-term climate abnormalities.

In this study, we make the following contributions: 1) we derive an extension of Greedy Gaussian Segmentation (Hallac et al., 2019) for use with non-Gaussian climate data and any generalized copula model; 2) we demonstrate our GCS method's possible use in plans for optimal climate change adaptation; and 3) we present realistic experiments that illustrate how a near-future pattern of extreme climate events can be optimally predicted using the proposed approach.

2. Related Work

Seeking optimally useful segments from input time series is a key step in climate change adaptation. Many variations of such time series segmentation, also known as change point detection, have been proposed and studied in different contexts. Comprehensive surveys have been presented by researchers from various fields (Reeves et al., 2007; Esling and

Agon, 2012; Polunchenko and Tartakovsky, 2012; Aminikhanghahi and Cook, 2017; Truong et al., 2020).

Truong et al., 2020 organized state-of-the-art offline change point detection algorithms using a structuring methodological strategy to give a systematic understanding of the strengths and weaknesses. They characterized the algorithms by three elements: a cost function, a search method, and a constraint on the number of changes. In their structure, GCS used a likelihood-based parametric cost function as an optimal search method for the unknown number of change points, a variance-based penalty, and a stopping criterion. Aminikhanghahi and Cook, 2017 surveyed the topic of change point detection in the fields of data mining, statistics, and computer science. They focused on machine learning algorithms that are not included in Truong et al., 2020's survey. They categorized the algorithms studied into supervised and unsupervised methods. Polunchenko and Tartakovsky, 2012 provided a survey of online algorithms for discrete time series spans over all major formulations of the underlying optimization problem—namely, Bayesian, generalized Bayesian, and minimax. In the present work, we use GCS because our climate change adaptation problem is better suited for offline algorithms. Online algorithms are good options for problems that need faster detection of instant state changes. Esling and Agon, 2012 summarized a broad range of theories that included general time series data mining techniques, but considered not only time series segmentation but also clustering, classification, etc. A basic theoretical understanding on such methods can be easily gained from this illuminating article. Reeves et al., 2007 specifically reviewed change point detection techniques for climate data. They discussed algorithms based on hypothesis testing, multiple regression, and hierarchical regression models. All of these studies offered useful insights in the development of the proposed GCS-CCA approach.

To demonstrate steps in the algorithms for GCS and GCS-CCA, an example analysis on a benchmark data set is first presented in Section 3. A real-world application for drought risk assessment follows in Section 4.

3. Methodology

3.1 Greedy Copula Segmentation

Assume we have bivariate climate data, available as time series data, as shown in Figure 1. Without loss of generality, assume that the time series are given at discrete data index values as shown.

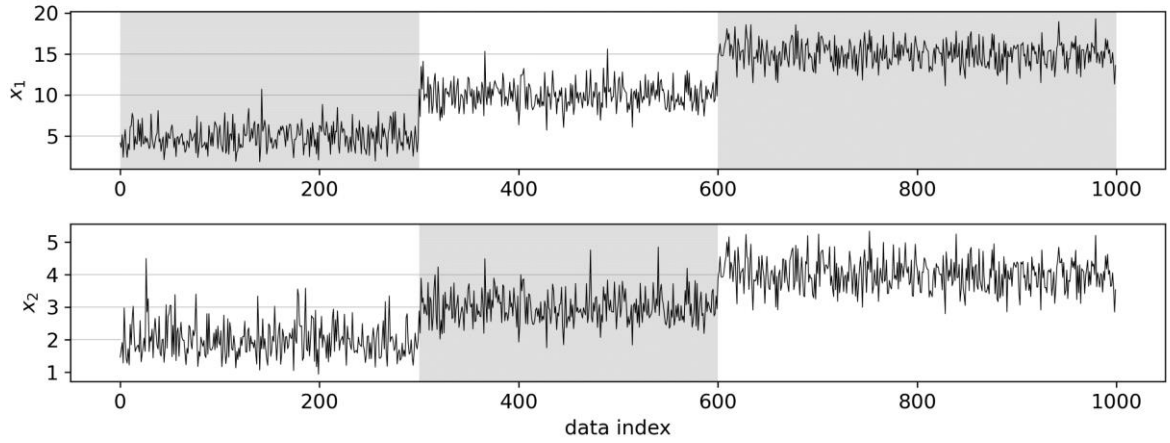

Figure 1. A realization of synthetic bivariate benchmark data time series: 3 separate data segments generated using 3 different parameter settings are highlighted.

In the synthetic data selected for this example, we have two climate-related variables that follow gamma and lognormal distributions, respectively. Their dependence structure is assumed to be represented by a Clayton copula. A total of 1,000 samples were generated with 3 different parameter settings to embed non-stationary character in the data. We have 5 parameters to define the two variables in each of the 3 subsets—they include a copula parameter, α; parameters describing the shape, a, and scale, b, for the gamma variable; and the mean, μ, and standard deviation, σ, for the lognormal variable. Note that the mean and variance of the gamma variable are ab and ab^2, respectively.

For the data, the first 300 samples are synthetically generated using $\Theta_1 = (\alpha, a, b, \mu, \sigma) = (1{,}10{,}0.5{,}2{,}0.5)$, the next 300 samples use $\Theta_2 = (10{,}40{,}0.25{,}3{,}0.5)$, and the final 400 samples are from $\Theta_3 = (50{,}100{,}0.15{,}4{,}0.5)$. For the gamma-distributed variables, the different parameter settings are equivalent to setting different mean values of 5, 10, and 15, and variances of 2.5, 2.5, and 2.25. Figure 2 shows copulas according to the different parameter setting selections. As is clear from Figure 1, the generated time series are non-stationary; the values of both variables are seen to get higher with time (increasing data index value). As such, this synthetic bivariate climate benchmark data set could represent changing extreme climate events – such as storms, floods, droughts, etc. – that get more frequent and severe with time.

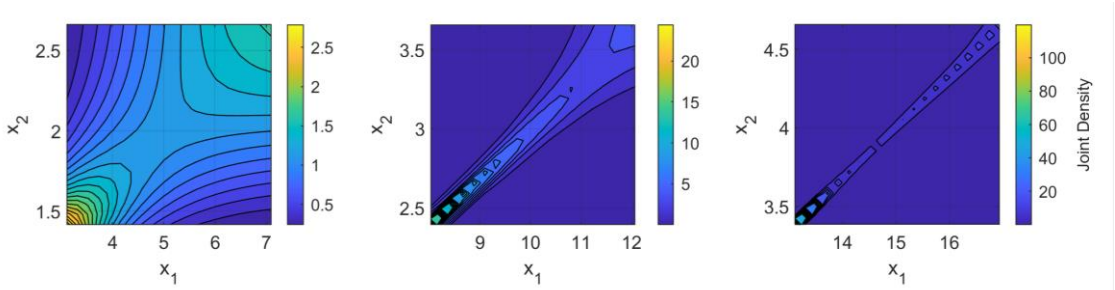

Figure 2. Copulas for the synthetic benchmark data generation using Θ_1, Θ_2, and Θ_3 (left to right).

From the above, one might expect that near-future patterns are most likely to be similar to the last 400 samples. The earlier 600 samples are likely to be deemed outdated and would increase uncertainty in any near-future prediction. Our goal is to find and uncover the last stationary sub-segment from the data. To achieve this goal, we iterate the greedy segmentation approach until no further segmentation on the last segment offers any advantage.

3.1.1 Iteration 1

The GCS algorithm starts with the benchmark data that can be denoted as $X = [\mathbf{x}_1, \dots, \mathbf{x}_{1{,}000}]^T$, where $\mathbf{x}_i = (x_1(i), x_2(i))$. Also, $x_1(i)$ and $x_2(i)$ represent the ith index values of the first and the second variable, respectively. Note that $\mathbf{x}_i$ represents a 2-dimensional vector containing these ith index values of both variables and X represents the entire bivariate data set.

We consider the data as a segment and, thus, the number of current segments $K = 1$; by splitting the data into more segments, the value of K will be changed. In every GCS iteration, we will consider a new breakpoint that then divides one of the current segments into two sub-segments. In the first iteration, we have 999 possible new breakpoints denoted as $b_{1\backslash 2}, b_{2\backslash 3}, \dots, b_{999\backslash 1{,}000}$, where the location of a breakpoint is indicated by the subscript. For

instance, $b_{k\backslash k+1}$ is a breakpoint that divides the data into two sub-segments $X_1 = [\mathbf{x}_1, \dots, \mathbf{x}_k]^T$ and $X_2 = \left[\mathbf{x}_{k+1}, \dots, \mathbf{x}_{1,000}\right]^T$. Figure 3 shows an example with $\mathrm{b}_{\mathrm{k}\backslash \mathrm{k}+1}$, where $k = 500$.

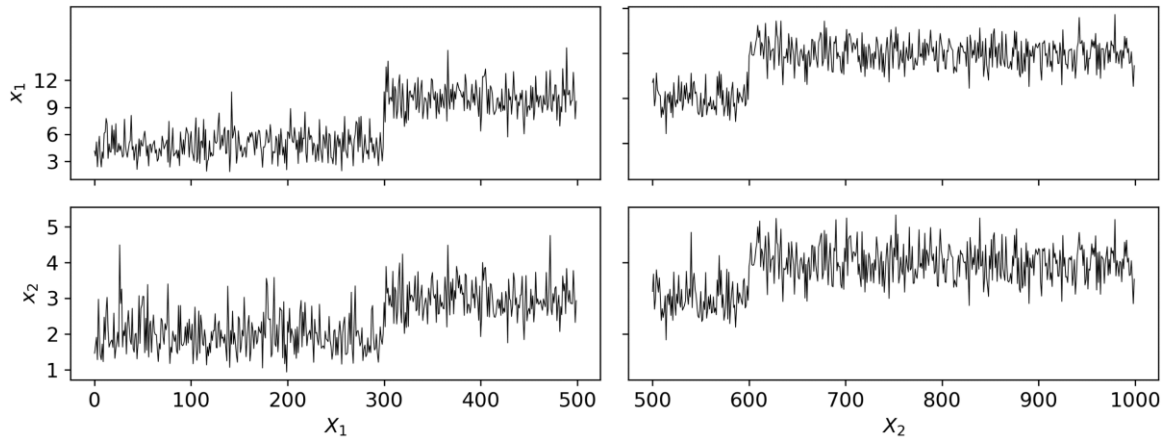

Figure 3. Example sub-segments generated by breakpoint, $b_{500\backslash 501}$.

Next, we compare two scenarios: 1) where X represents independent bivariate samples from a multivariate copula C_Θ based on all the data; and 2) where X_1 and X_2 represent separate bivariate samples from two different copulas, $C_{\Theta_{(1)}}$ and $C_{\Theta_{(2)}}$, respectively. For both scenarios, we assume that the same Clayton copula family and Gamma and lognormal marginal distributions, although different distribution and copula parameters apply in the two scenarios. Scenario 1 leads to fixed model parameters, while Scenario 2 considers that the model parameters change when one considers data before and after the breakpoint, $b_{k\backslash k+1}$. Using maximum likelihood, we will evaluate and maximize the following objective function:

$$\Psi_{k\backslash k+1} = \psi(X_1) + \psi(X_2) - \psi(X), \tag{1}$$

where $\psi(\cdot)$ is a function computed based on the regularized maximum log-likelihood of the available data with regard to the predefined copula family and marginal distributions.

Note that $\psi(X)$, first, employs MLE model parameters, Θ, based on the assigned data, X. The MLE method allows estimation of the marginal distribution parameters and the copula family parameters; MATLAB provides functions named `fitdist` and `copulafit` that accomplish this task. The regularized maximum log-likelihood function is obtained as follows:

$$\psi(X) = \sum_{i=1}^{n} \left(\log c_\alpha\left(F_1(x_1(i)|a,b), F_2(x_2(i)|\mu,\sigma)\right) + \log f_1(x_1(i)|a,b) + \log f_2(x_2(i)|\mu,\sigma)\right) - \frac{\lambda}{s_1^2 + s_2^2}, \tag{2}$$

where n is the length of the input bivariate time series, X; $c_\alpha = \frac{\partial^2 C_\alpha(u_1 = F_1(x_1|a,b), u_2 = F_2(x_2|\mu,\sigma))}{\partial u_1 \partial u_2}$ is the copula probability density function; $u_1 = F_1(x_1|a,b)$ and $u_2 = F_2(x_2|\mu,\sigma)$ are marginal cumulative distribution functions; $f_1(x_1|a,b)$ and $f_2(x_2|\mu,\sigma)$ are marginal probability density functions; s_1 and s_2 are marginal sample standard deviations. To avoid overfitting, marginal variance regularization is applied and $\lambda \geq 0$ is the regularization parameter.

Note that $\Psi_{k\backslash k+1}$, as defined, is the regularized maximum log-likelihood difference between the likelihood function based on data sub-segments divided at the breakpoint, $b_{k\backslash k+1}$, and the

likelihood function based on the entire unsegmented data set. We calculate $\Psi_{k\backslash k+1}$ for every possible breakpoint and then select an optimal breakpoint $b_{k^*\backslash k^*+1}$ as follows:

$$k_1^* = \underset{k}{\operatorname{argmax}}\, \Psi_{k\backslash k+1}, \tag{3}$$

and we also ensure that $\Psi_{k_1^*\backslash k_1^*+1} > 0$. If every Ψ returns a negative value, it means that further segmentation has no advantage. In this case, the greedy algorithm stops the segmentation search and we go to the *Return* stage.

Figure 4 shows 999 Ψ values computed with $\lambda = 100$. The maximum Ψ value occurs for $k = 600$. Based on this result, we divide the data set into sub-segments at the breakpoint, $b_{600\backslash 601}$. These resulting sub-segments are shown in Figure 5.

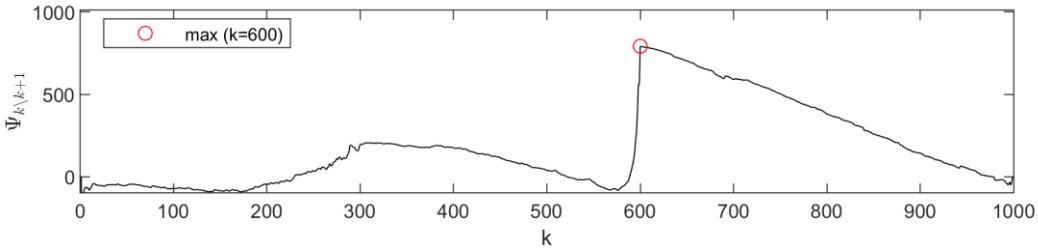

Figure 4. Calculated objective function Ψ for the benchmark data at the first iteration.

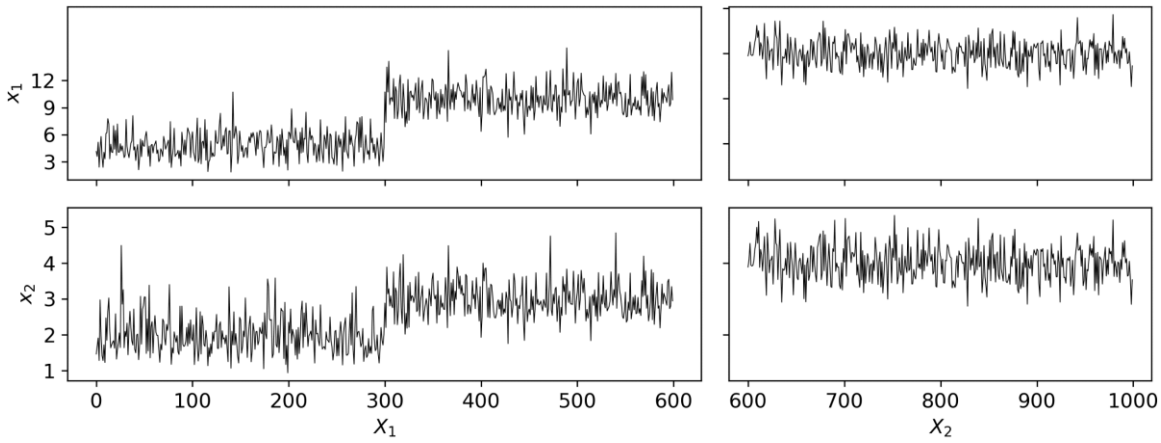

Figure 5. Sub-segments generated by the first identified breakpoint, $b_{k_1^*\backslash k_1^*+1} = b_{600\backslash 601}$.

3.1.2 Iteration 2

After the previous (first) iteration, what we have are new segmented data sets, $X_1 = [\mathbf{x}_1, \ldots, \mathbf{x}_{600}]^T$ and $X_2 = [\mathbf{x}_{723}, \ldots, \mathbf{x}_{1,000}]^T$. Thus, the number of current segments, $K = 2$, and the number of new breakpoints possible is now 998. Again, we compute Ψ for every possible breakpoint and ultimately select a new optimal breakpoint, $b_{k_2^*\backslash k_2^*+1}$. We reject the new breakpoint and terminate the greedy algorithm if all Ψ values have a negative value. An additional termination condition is invoked in Iteration 2 and beyond, if the identified optimal breakpoint is not from the current last sub-segment. This is because our goal with the greedy search algorithm is to find and use only the last stationary sub-segment to be representative of the most likely series for the near future. Therefore, if further segmentation cannot be continued on the current last sub-segment, we terminate the search. On the other hand, if there is a breakpoint, $b_{k_2^*\backslash k_2^*+1}$, within the last sub-segment (in Iteration 2, the last segment =

X_2) and $\Psi_{k_2^*\backslash k_2^*+1} > 0$, we accept this new breakpoint and continue the iteration with the new segmented data sets, $X_1 = [\mathbf{x}_1, \dots, \mathbf{x}_{600}]^T$, $X_2 = \left[\mathbf{x}_{601}, \dots, \mathbf{x}_{k_2^*}\right]^T$, and $X_3 = \left[\mathbf{x}_{k_2^*+1}, \dots, \mathbf{x}_{1,000}\right]^T$. Otherwise, the algorithm moves to what we refer to as the *Return* stage.

3.1.3 Iteration 3+

We repeat the procedure above until any one of the termination conditions: 1) all $\Psi < 0$; 2) k^* does not match an index number in the last sub-segment. After we terminate this iterative greedy search, the algorithm moves to the final *Return* stage.

3.1.4 Return

As final output, the algorithm returns the current last segment as the identified optimal data sub-segment. We denote this data set as X_{opt}. Note that $X_{opt} \subseteq X$.

Figure 6 shows calculated 998 Ψ values for the benchmark data set at Iteration 2. The maximum value occurs at $k = 310$ on the first segment. This means that we have reached the second termination condition. We stop the iterations and send the current last sub-segment $X_2 = \left[\mathbf{x}_{601}, \dots, \mathbf{x}_{1,000}\right]^T$ to the *Return* stage. As a result, the identified optimal data set, $X_{opt} = X_2 = \left[\mathbf{x}_{601}, \dots, \mathbf{x}_{1,000}\right]^T$.

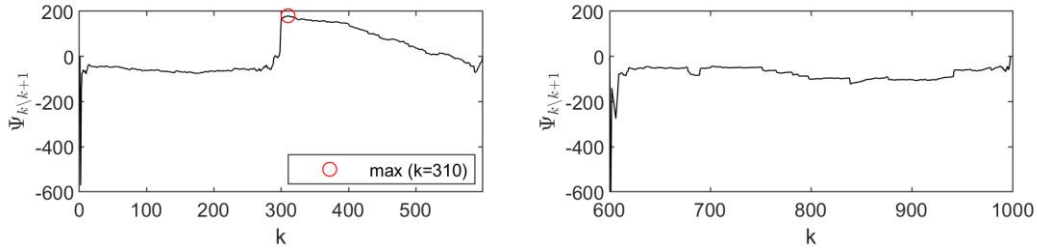

Figure 6. Calculated objective function Ψ for the benchmark data at the second iteration.

The GCS algorithm can be generalized to any d-dimensional multivariate data set, $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times d}, \mathbf{x}_i = \left(x_1(i), \dots, x_d(i)\right)$. Let $f_i(x_i|\theta_i)$ be the probability density function and $u_i = F_i(x_i|\theta_i)$ be the cumulative distribution function for variable, x_i. Multivariate copulas can be denoted as $C_\Theta = C_\alpha(u_1, \dots, u_d)$, where $\Theta = (\alpha, \theta_1, \dots, \theta_d)$. The regularized maximum log-likelihood function for multivariate data, X, is given as:

$$\psi(X) = \sum_{i=1}^{N} \left(\log c_\alpha(u_1, \dots, u_2) + \sum_{j=1}^{d} \log f_j\left(x_j|\theta_j\right) \right) - \frac{\lambda}{\sum_{j=1}^{d} s_j^2}. \quad (4)$$

Figure 7 shows the general GCS algorithm flowchart based on the preceding discussion.

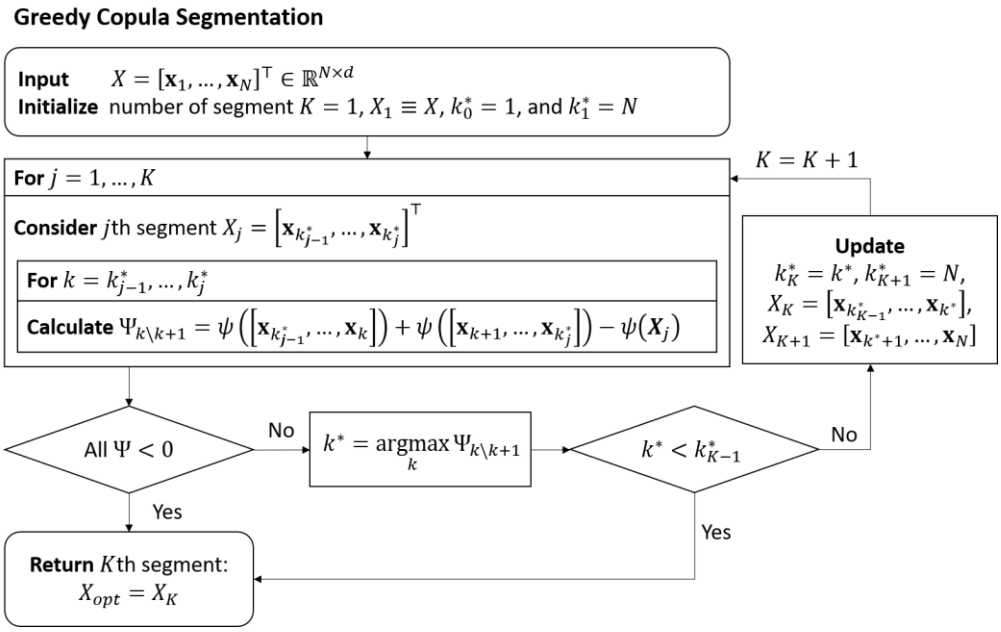

Figure 7. Greedy Copula Segmentation (GCS) algorithm flowchart.

3.2 Climate Change Adaptation with the Benchmark Data

We are interested in attempting a climate change adaptation strategy using GCS assuming that the bivariate data in Figure 1 describe climate parameters of interest. Suppose the benchmark data set, X, represents a 100 year-long set of observations with 10 records per year. Let us first consider a situation where only the first 40 year-long set (400 samples) represent the base data. The traditional approach would develop the base joint copula, $C_{\Theta_{(0)}}$ using all the base data, but our optimal approach will use the GCS-identified optimal data only for near-future projections. Then, such a derived joint distribution will be used for any risk assessment until the new data are obtained, or the existing data set from 40 years is updated. Suppose this distribution is updated in increments corresponding to 10-year cycles. Again, the traditional approach would use all of the now 50 year-long set (500 samples) to obtain a new updated version of the joint copula, $C_{\Theta_{(1)}}$, but our optimal approach will again use the GCS-identified optimal data only. The procedure can be repeated every 10 years and two different joint copulas can be developed based on the two different approaches (traditional vs. GCS).

To highlight the comparative prediction performance of the two approaches, we compute log-likelihoods for m update cycles, each of 10-year length as follows:

$$LL_{trad}(m) = \log \prod_{i=1}^{n_m} C_{\Theta_{trad\_m}}(\mathbf{x}_i), \qquad LL_{opt}(m) = \log \prod_{i=1}^{n_m} C_{\Theta_{opt\_m}}(\mathbf{x}_i). \tag{5}$$

Two different joint copulas, $C_{\Theta_{trad\_m}}$ and $C_{\Theta_{opt\_m}}$, are derived using the base data and the same number of new 10-year data updates, $\mathbf{x}_i, i = 1, \dots, n_m$, is applied to calculate the log-likelihood in Equation 5. As such, the calculated log-likelihoods are fair performance measures to allow comparisons between traditional and GCS approaches. The copula and corresponding approach that yields a higher likelihood when the new data are included is more accurate than the alternative. In other words, the traditional and GCS approaches offer models based on the base data that are then used to assess how well they perform against different lengths of update cycle data increments; relative comparison is possible using Equation 5.

A general formulation can be defined using t_{base} (the base period) and t_{cyc} (the period covered in each update cycle). At cycle m, the traditional approach uses all the data collected from the beginning until $t_{base} + m \cdot t_{cyc}$ to update the distribution, whereas GCS-CCA uses X_{opt_m} for the corresponding distribution. Note that each t_{cyc}-long data update can be used to evaluate predictive performance. Figure 8 shows a diagram summarizing the two different approaches with the formulation as presented.

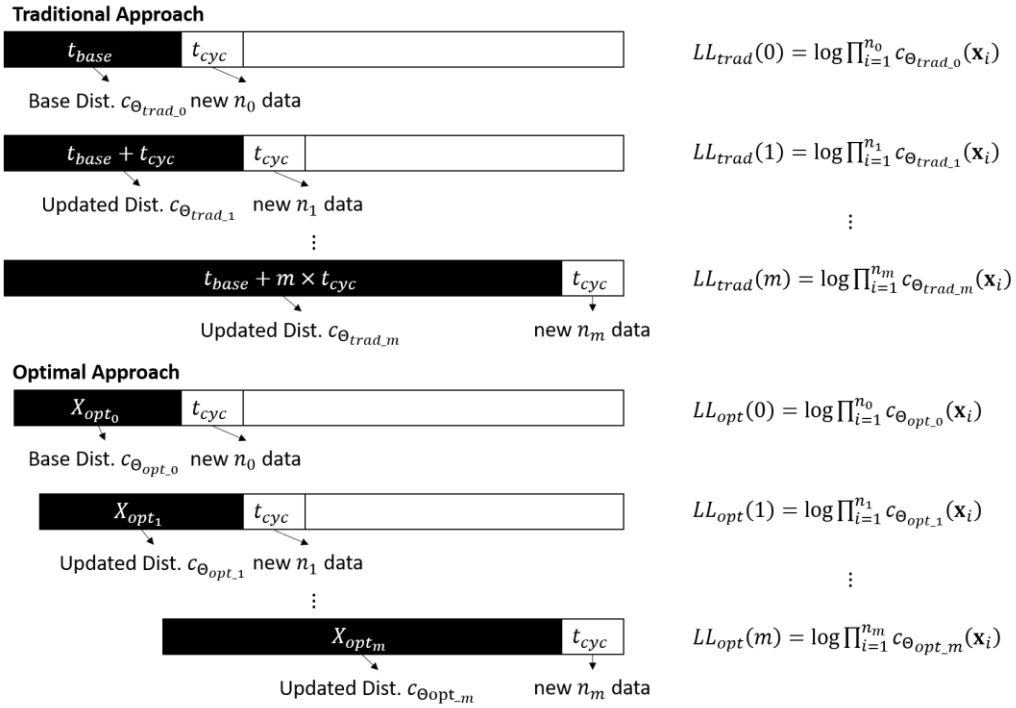

Figure 8. Traditional and optimal GCS approaches for climate change adaptation.

Figure 9 shows results from GCS-CCA with $\lambda = 100$ and $\lambda = 10$, as applied to the benchmark data. The predictive performance is evaluated 6 times since we choose the first 400 samples as the base data and add 100 new samples in each update cycle. We repeat this entire procedure 10 times by synthetically generating (by random sampling) a new benchmark data set each time. Figure 9 shows the (normalized) mean of the predictive log-likelihood difference ratios, $\delta_{LL}(\%) = \frac{LL_{opt} - LL_{trad}}{|LL_{trad}|} \times 100$, calculated with all 10 samplings. The min-max error bars are also shown. We can easily verify that, in the mean, GCS-CCA outperformed traditional CCA in every update for $\lambda = 100$. We also evaluate the influence of the regularization parameter, λ. For the lower value, $\lambda = 10$, GCS-CCA generally performs better than traditional CCA. However, GCS-CCA with $\lambda = 10$ sometimes leads to overfitting, and then its performance is not better than the traditional CCA. Therefore, it is important to use a proper regularization parameter, λ, for GCS-CCA.

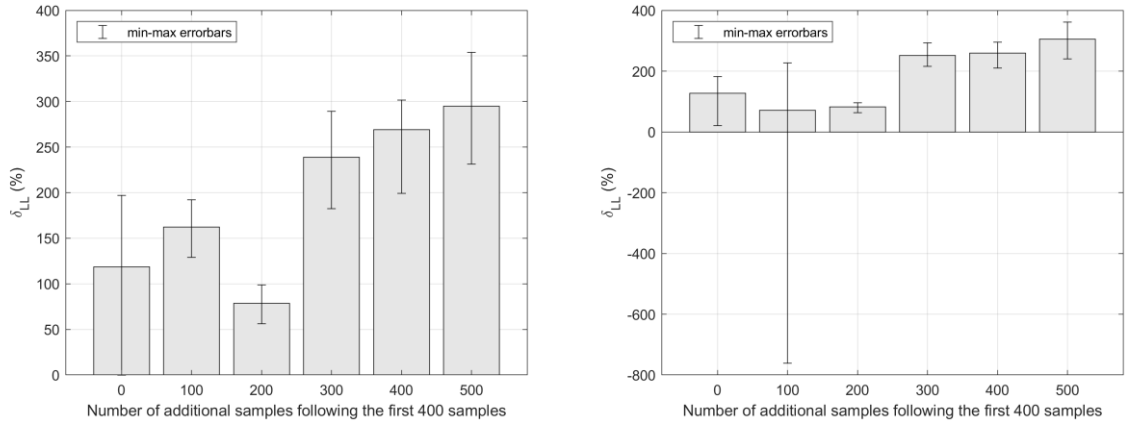

Figure 9. Calculated predictive log-likelihood difference ratios $\delta_{LL}(\%) = \frac{LL_{opt} - LL_{trad}}{|LL_{trad}|} \times 100$ with (Left) $\lambda = 100$ and (right) $\lambda = 10$.

3.3 Regularization Parameter Selection

GCS-CCA leads to more accurate prediction than traditional CCA if we can select the proper regularization parameter, λ. Its value can be chosen by trial and error, using prior knowledge, or using a principled method, such as Bayesian or Akaike information criterion or cross validation (Hallac et al., 2019). In general, one needs a sufficiently high value for λ because this parameter directly influences the extent of segmentation that results. Too high a value for λ results in no segmentation, which is then equivalent to traditional CCA; on the other hand, a low value for λ leads to overfitting, which means that GCS will select a very short recent sub-segment as the optimal data. Then, the joint distribution of the underlying variables is overly fitted to this small amount of data. As we can see from Equation 4, the order of magnitude of the marginal variances affects the regularization along with λ. In practice, one can use several linearly or logarithmically spaced values of λ over a wide range in a search for a sufficiently large regularization parameter.

4. Drought Patterns in CCA

Several hydroclimate variables – e.g., precipitation, air temperature, soil moisture, etc. – simultaneously affect drought scenarios. Indices or scores derived from univariate and multivariate drought indicators that are in turn based on individual or multiple hydroclimate variables have been developed to characterize and quantify drought conditions. Such scores are included in a drought index, and drought index time series can then be used to describe the input data for drought severity-duration-frequency (SDF) analysis.

For a real data analysis and application of GCS-CCA, we collected climate data – representing monthly total precipitation and a monthly average of daily average temperature data – from the Global Historical Climatology Network-Monthly (GHCN-M) Version 3 dataset (Lawrimore et al., 2011). Various types of drought indices were calculated using open-source software originally developed by National Integrated Drought Information System (NIDIS), National Centers for Environmental Information (NCEI), and National Oceanic and Atmospheric Administration (NOAA) (Adams, 2017). The collected climate variables and calculated drought indices cover the geospatial extent: latitude 24.5625 ~ 49.354168 (degrees north), longitude -124.6875 ~ -67.020836 (degrees east), and raster dimensions, (latitude, longitude, time) = (38, 87, 1466). One grid cell near the Austin, Texas area was selected for a regional case study. Figure 10 shows the area covered by the selected grid cell.

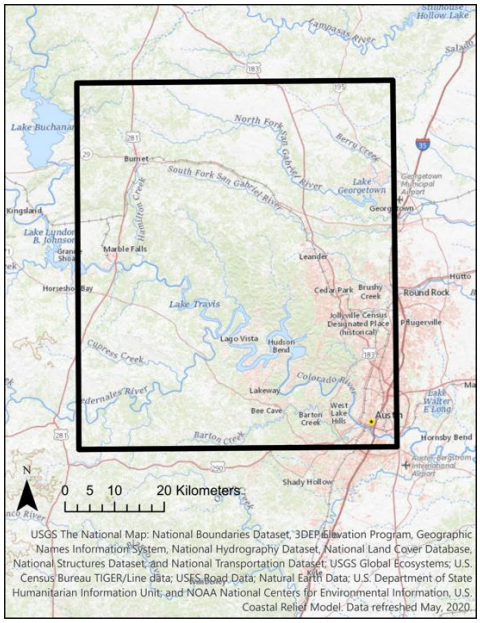

Figure 10. Selected site in the Austin, Texas area.

Among various drought indices, the **S**tandardized **P**recipitation **E**vapotranspiration **I**ndex utilizing a **G**amma distribution with a **3**-month scale (SPEI_G3), developed by Vincente-Serrano et al., 2010 was selected to serve as an indicator of drought events. This selection is justified because studies have shown that SPEI performs better in drought assessments under a global warming trend by combining the multi-scalar character with the capacity of involvement of temperature effects on droughts (Hao and Singh, 2015; Tan et al., 2015; Homdee at al., 2016). Detailed information about SPEI and its calculation can be found in the studies by Vincente-Serrano et al., 2010; Begueria et al., 2014; Hameed et al., 2018.

Figure 11 shows the calculated SPEI_G3 time series, denoted by Z, that is obtained for the period, December 1896 to February 2017. The Thornthwaite equation is used to derive potential evapotranspiration (PET) from air temperature data.

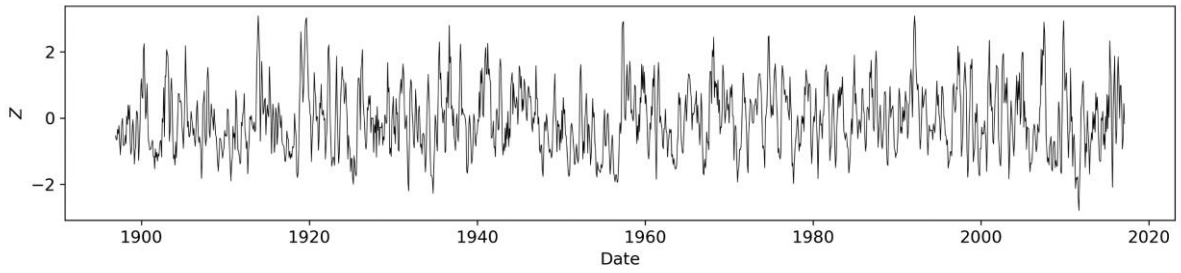

Figure 11. Calculated standardized precipitation evapotranspiration index utilizing Gamma distribution and 3-month scale (SPEI_G3) time series.

This study describes the entire procedure that starts with preparing a bivariate drought pattern time series and proceeds to a final predictive performance evaluation. We provide a step-by-step guide that can be used for not only several types of drought events but also for other extreme climate events and applications that have a similar problem setting and data structure.

4.1 Bivariate Drought Pattern Time Series

To apply GCS, first, we extract drought events from the selected drought index time series using a predefined truncation level. The overall concept of how we define a drought event and its associated duration, d_i, and severity, s_i is illustrated in Figure 12. In this study, drought duration and severity are selected for the analysis since they have been widely used for drought severity-duration-frequency (SDF) analysis. A similar concept can be applied to other climate data time series.

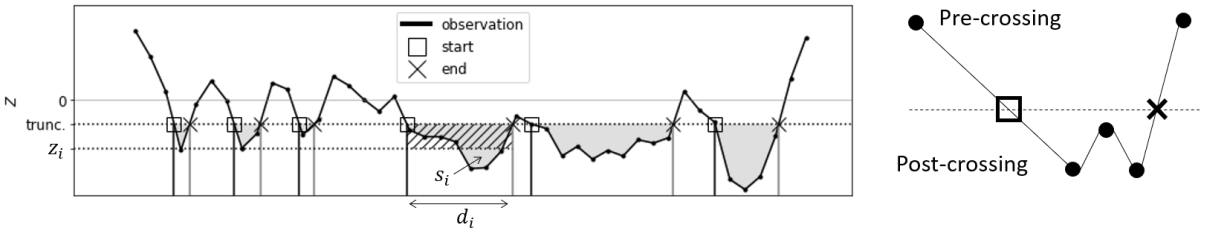

Figure 12. A concept diagram showing definitions of drought event duration, severity, and equivalent intensity, along with indications of a pre-crossing and a post-crossing.

Our definition is a modified version of the Yevjevich (1967) theory of run model. We define the start and end of a drought event by interpolating pre-crossing and post-crossing data points given the data. In this manner, for any drought event, i, the drought duration, d_i – defined as the time difference between the start and end – is real-valued. Then, the absolute value of the integral area between drought index time series and the selected horizontal truncation level from the start to the end of the event is defined as the drought severity, s_i. An equivalent drought index value, z_i, associated with drought event, i, is easily calculated. Mathematically, $z_i = s_i/d_i$, which is sometimes referred to as drought intensity (Cavus and Aksoy, 2020). This drought index when considered at a constant level over the duration of the event leads to an area-based severity that is equivalent to the observed value, s_i, for the same event. This is clear too from Figure 12. To be clear, we refer to z_i as an equivalent intensity.

Suppose we extract N drought events from the given drought index time series. Then, the input data, $X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{N\times 2}$, where $\mathbf{x}_i = (d_i, s_i)$. Each data point can now be considered as data obtained at the start of corresponding drought event. We can now apply GCS-CCA to the input drought data.

Vincente-Serrano et al., 2010 defined various ranges of SPEI values as associated with different intensities of droughts: light drought (-0.5 to -0.99), moderate drought (-1.0 to -1.49), severe drought (-1.5 to -1.99), and extreme drought (-2.0 ≤). In the present study, a truncation SPEI level of -0.5 is selected so as to include even the mildest drought conditions in our assessment. Accordingly, a total of 143 drought events with associated duration and severity (or equivalent intensity) are extracted from the SPEI_G3 time series.

Figure 13 shows the duration, severity, and equivalent intensity values considering all the drought events extracted over the period of measurements (1896-2017) in the selected Austin, Texas region. Average and standard deviation values are shown for the data and are also shown using a 5-year moving window. The moving average and standard deviation variation clearly indicate non-stationary characteristics in the drought pattern. Figure 14 shows scatter plots of the collected data, showing two of the drought-related variables at a time. Based on similar assumptions in past studies, exponential and gamma distributions are selected as marginal probability distributions for duration and severity, respectively. The Gumbel copula family is selected to model the pairwise dependence structure for these two variables (Zelenhasic and Salvai,1987; Hao and Singh, 2013).

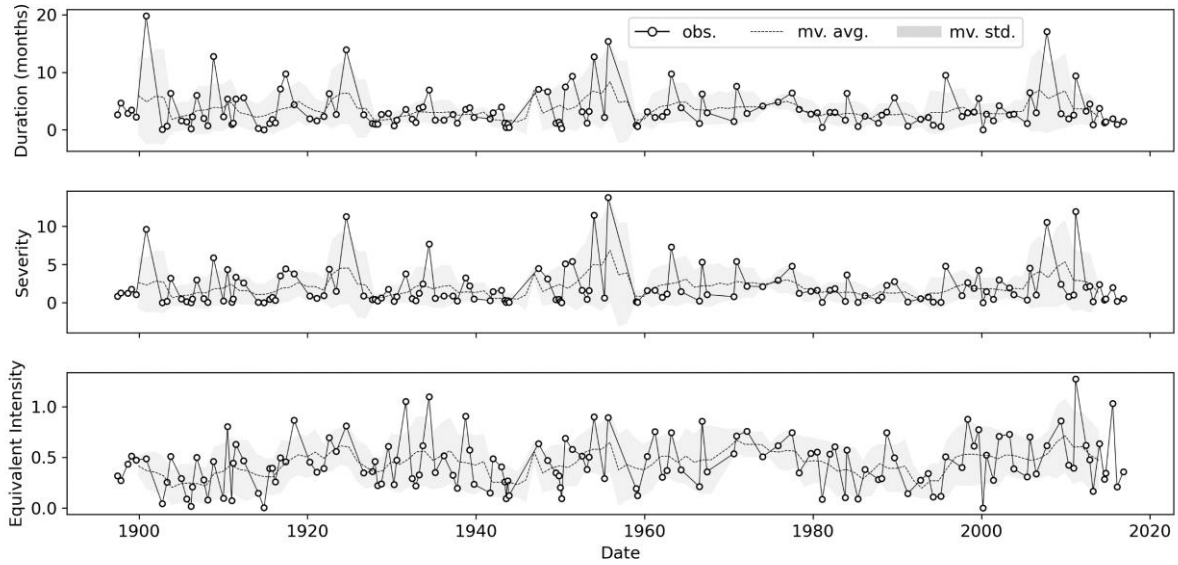

Figure 13. Duration, severity, and equivalent intensity values from 143 extracted drought events, using a -0.5 truncation level with the SPEI_G3 data.

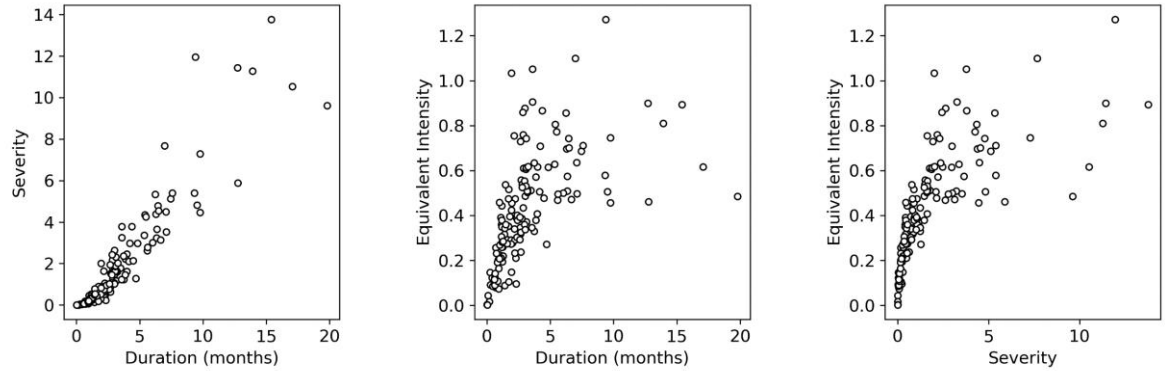

Figure 14. Pairwise scatter plots showing duration, severity, and equivalent intensity for all the drought events in the data set.

We begin by considering only the initial 20-year data as base data and then include 10-year increments as update cycles in projections to be used in possible climate change adaptation, where the GCS-CCA approach seeks to optimize justified use of only the most recent data. The overall input data covers about 120 years (December 1896 to February 2017) and, thus, there are 10 predictive performance evaluations of GCS-CCA versus a traditional that ignores non-stationary trends.

Figure 15 shows results summarized in terms of the predictive log-likelihood difference ratio, $\delta_{LL}(\%) = \frac{LL_{opt} - LL_{trad}}{|LL_{trad}|} \times 100$. We can easily see that GCS-CCA generally has better performance than traditional CCA with $\lambda = 150$. This result implies that the GCS-identified optimal data sub-segments explain near-future drought patterns better than when all of the historical observed data are used. Figure 15 also shows GCS-CCA performance with $\lambda = 100$; the lower λ leads to overfitting. Pre-processing of the data and an appropriate regularization parameter is recommended for such analyses.

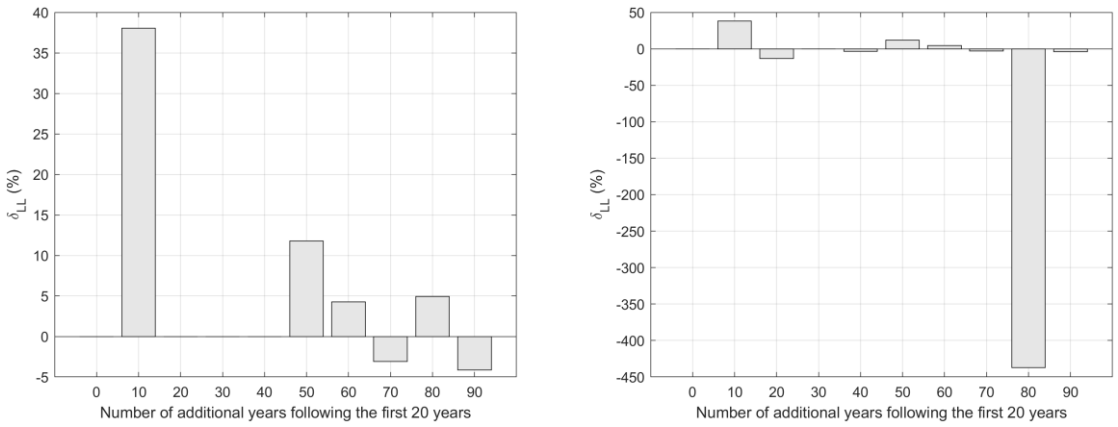

Figure 15. Computed predictive log-likelihood difference ratios, $\delta_{LL}(\%) = \frac{LL_{opt} - LL_{trad}}{|LL_{trad}|} \times 100$, for the drought patterns data using $\lambda = 150$ (left) and $\lambda = 100$ (right).

5. Conclusions

In this work, we extended Greedy Gaussian segmentation (GGS) developed by Hallac (2019) by allowing multivariate Gaussian distributions in the copula definition; we refer to this extended approach as greedy copula segmentation (GCS). Our extension is well-suited for use with climate data since many climate-related variables are non-Gaussian and non-stationary. Based on the wide coverage of different dependence structures possible with the copula family choice, it is expected that GCS could be used in various applications that involve long sequences of multivariate time series data. We have explained GCS, iteration by iteration, so as to offer an accessible description of the greedy algorithm.

Using a synthetic data set as well as an observed drought data set, we have shown that GCS can optimize future projections for possible use in climate change adaptation. Climate change adaptation needs to rationally consider periodic updates of the joint distribution of climate variables by focusing on patterns seen in extreme climate events. We introduce the notion of considering trends in any climate parameter as best understood by defining a piecewise process consisting of several stationary sub-segments to represent the data. In such a piecewise stationary representation, the latest (most recent) stationary sub-segment (whose length must be iteratively established, using maximum likelihood with regularization) can predict most rationally and precisely any near-future patterns in the extreme climate that are to be expected. The proposed GCS approach identifies the most informative data sampled from the latest stationary sub-segment; it iteratively evaluates the benefit of further segmentation on the last segment. By doing so, the algorithm greedily searches for the optimal last segment of input data.

We show that the GCS-identified optimal data produce better predictive performance for possible climate change adaptation by illustrative examples using a benchmark synthetic data set as well as a real 120-year drought-related data set from Austin, Texas. GCS-CCA shows superior predictive performance for the non-stationary benchmark problem. For the real-world application, we collect drought index time series data and extract the bivariate drought event (duration and severity) data. The GCS-CCA results suggest that the proposed approach can rationally uncover changing climate patterns in the time series and can produce accurate near-future projection for adaptation plans compared to more traditional approaches that seek to use long or complete historical data sets. We conclude that GCS-CCA optimizes potential climate change adaptation strategies and have provided a detailed algorithm for its implementation.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CMMI-1663044. The authors are grateful for this support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Adams, J. (2017). climate_indices, an open source Python library providing reference implementations of commonly used climate indices. URL: https://github.com/monocongo/326climate%7B%5C_%7Dindices.

Aminikhanghahi, S. and Cook, D. J. (2017) A Survey of Methods for Time Series Change Point Detection. *Knowledge and Information Systems*. 51 (2), 339–367.

Beguería, S. et al. (2014) Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. International journal of climatology. 34 (10), 3001–3023.

Cavus, Y. and Aksoy, H. (2020) Critical drought severity/intensity-duration-frequency curves based on precipitation deficit. Journal of hydrology. 584, 124312–.

Cid, A. et al. (2016) Long-term changes in the frequency, intensity and duration of extreme storm surge events in southern Europe. *Climate Dynamics*. 46 (5), 1503–1516.

Dai, A. (2013) Erratum: Increasing drought under global warming in observations and models. Nature climate change. 3 (2), 171–171.

De Michele, C. and Salvadori, G. (2003) A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas. *Journal of Geophysical Research: Atmospheres*. 108 (D2), 4067–n/a.

Esling, P. and Agon, C. (2012) Time-series data mining. *ACM Computing Surveys*. 45 (1), 1–34.

Garcia Galiano, S. G. et al. (2015) Assessing Nonstationary Spatial Patterns of Extreme Droughts from Long-Term High-Resolution Observational Dataset on a Semiarid Basin (Spain). *Water (Basel)*. 7 (10), 5458–5473.

Genest, C. and Favre, A.-C. (2007) Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*. 12 (4), 347–368.

Genest, C. et al. (2011) Estimators based on Kendall's Tau in Multivariate Copula Models. *Australian and New Zealand Journal of Statistics*. 53 (2), 157–177.

Hallac, D. et al. (2019) Greedy Gaussian segmentation of multivariate time series. *Advances in data analysis and classification*. 13 (3), 727–751.

Hameed, M. et al. (2018) Apprehensive Drought Characteristics over Iraq: Results of a Multidecadal Spatiotemporal Assessment. Geosciences. 8 (2), 58.

Hao, Z. and Singh, V. P. (2013) Entropy-Based Method for Bivariate Drought Analysis. *Journal of Hydrologic Engineering*. 18 (7), 780–786.

Hao, Z. and Singh, V. P. (2015) Drought characterization from a multivariate perspective: A review. Journal of hydrology. 527, 668–678.

Homdee, T. et al. (2016) A comparative performance analysis of three standardized climatic drought indices in the Chi River Basin, Thailand. *Agriculture and Natural Resources*. 50 (3), 211-219.

Lawrimore, J. et al. (2011) Global Historical Climatology Network – Monthly (GHCN-M), Version 3. DOI: https://doi.org/doi:10.7289/V5X34VDR.

Lee, T. & Ouarda, T. B. M. J. (2010) Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. *Journal of Geophysical Research: Atmospheres*. 115 (D13).

Li, J. et al. (2015) Evaluation of Nonstationarity in Annual Maximum Flood Series and the Associations with Large-scale Climate Patterns and Human Activities. *Water Resources Management*. 29 (5), 1653–1668.
Liu, S. et al. (2019) Identification of the Non-stationarity of Floods: Changing Patterns, Causes, and Implications. Water resources management. 33 (3), 939–953.
Manuel, L. et al. (2018) Alternative Approaches to Develop Environmental Contours from Metocean Data. *Journal of Ocean Engineering and Marine Energy*, 4(4).
Mathier, L. et al. (1992) The Use of Geometric and Gamma-Related Distributions for Frequency Analysis of Water Deficit. *Stochastic Hydrology and Hydraulics: Research Journal*. 6 (4), 239–254.
Mazdiyasni, O. et al. (2019) Heat wave Intensity Duration Frequency Curve: A Multivariate Approach for Hazard and Attribution Analysis. *Scientific reports*. 9 (1), 14117–14118.
Nelsen, R. B. (2006) *An Introduction to Copulas by Roger B. Nelsen*. 2nd ed. 2006. New York, NY: Springer New York.
Ouarda, T. B. M. J. & Charron, C. (2018) Nonstationary Temperature-Duration-Frequency curves. *Scientific Reports*. 8 (1), 15493–15498.
Polunchenko, A. S. and Tartakovsky, A. G. (2012) State-of-the-Art in Sequential Change-Point Detection. *Methodology and Computing in Applied Probability*. 14 (3), 649–684.
Reeves, J. et al. (2007) A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*. 46 (6), 900–915.
Saklar, A (1959) Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris 8*. pp. 229-231.
Salvadori, G. (2004) Bivariate return periods via 2-Copulas. *Statistical Methodology*. 1 (1-2), 129–144.
Salvadori, G. and De Michele, C. (2004) Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*. 40 (12).
Sheffield, J. et al. (2012) Little change in global drought over the past 60 years. Nature. 491 (7424), 435–438.
Shiau, J.-T. & Shen, H. W. (2001) Recurrence Analysis of Hydrologic Droughts of Differing Severity. *Journal of Water Resources Planning and Management*. 127 (1), 30–40.
Slater, L. J. et al. (2020) Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2020-576, in review, 2020.
Tan, C. et al. (2015) Temporal-Spatial Variation of Drought Indicated by SPI and SPEI in Ningxia Hui Autonomous Region, China. *Atmosphere*. 6 (10), 1399–1421.
Truong, C. et al. (2020) Selective review of offline change point detection methods. *Signal Processing*. [Online] 167, 107299.
Van Loon, A. F. et al. (2016) Drought in the Anthropocene. Nature geoscience. 9 (2), 89–91.
Vicente-Serrano, S. M. et al. (2010) A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. Journal of climate. 23 (7), 1696–1718.
Yevjevich (1967) An objective approach to definitions and investigations of continental hydrologic droughts. Hydrology Paper No. 23, Colorado State University, Fort Collins, Colorado.
Yue, S. et al. (1999) The Gumbel mixed model for flood frequency analysis. *Journal of hydrology (Amsterdam)*. 226 (1), 88–100.
Yue, S. (2001) A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrological Processes*. 15 (6), 1033–1045.
Zelenhasic, E. & Salvai, A. (1987) A method of streamflow drought analysis. *Water resources research*. 23 (1), 156–168.