EVOLUTION OF EARTH'S CLIMATE ZONES
VISUALIZING SHIFTS ACROSS TIME AND SPACE
Eurogroup Final Report | CSE 6242 | Fall 2020

G. Taylor Brooks
OMSA Georgia Tech
gbrooks34@gatech.edu

Austin Lipinski
OMSA Georgia Tech
alipinski6@gatech.edu

Grégoire Petit
MSCS Georgia Tech
gpetit3@gatech.edu

Alice Dumay
MSCS Georgia Tech
adumay3@gatech.edu

Each member of the group participated fairly and evenly in all parts of this project and particularly in this report.

1 Background

1.1 A Changing Climate

The warming climate and its impacts become more evident every day. As the climate warms, climate zones are shifting polewards. These shifts in climate zones are seen in the contiguous United States [PA16], the tropics [SLG+18, BPRW18], Australia [PA18], and elsewhere in the world.

As the climate zones shift, so too do the plants and animals species that rely on certain climatic conditions. USDA plant cold-hardiness zones are shifting to allow for more widespread cultivation of cold-intolerant crops such as oranges and almonds [PA16], leading to a new USDA plant hardiness zone map development process [DWH+12].

Animal life is shifting as well to remain within preferential conditions. Monllor-Hurtado, et al. [2012] indicate that tuna catches in subtropical latitudes are increasing while tuna catches in tropical regions are decreasing as the oceans warm [MHPSL17]. Sunday, et al. [2012] indicate that range shifts of cold-blooded marine and terrestrial lifeforms do not follow the same pattern [SBD12].

Weather phenomena also are changing as the climate warms. The spatial distribution of tornadoes in the United States, for example, is shifting to the east [ALCM16] while precipitation amounts are becoming increasingly variable in the Parisian basin [DBMC19].

The importance of detecting shifting climate zones is the basis of this work and led us to our problem. Shifting climate biomes have important implications for agriculture production, species conservation, weather prediction, fisheries sustainability: How can we classify climates relatively easily and visualize the shifts that are occuring?

1.2 Classifying Biomes

Human beings have attempted to classify climates on several reprises in the past. One of the first and most widespread climate classification schemes is the Köppen classification scheme, first published in 1884 by Wladimir Köppen. This scheme divides the climate into five "base" climate groups and then subclassifies these groups by precipitation type and heat amounts [KVB11]. The next major climate classification scheme was defined by GT Trewartha in 1966 and addressed some of the shortcomings of the Köppen scheme. In the Trewartha Scheme, temperate and subtropical climates are treated differently and a new boreal climate type was defined [BHHK14]. A final major classification scheme is the Holdridge Life Zone schema, which uses precipitation, temperature, and potential evapotranspiration to classify [Hol67, LBD+99].

While the previous three schemes used climatic characteristics as the basis of classification, other systems also exist that look at plant life resistance to cold temperatures. The most familiar of these to the home gardener is the USDA plant hardiness zone mappings [DWH+12].

All four of the schemes previously mentioned use human-defined heuristic-based methods to define climate regions. Newer approaches use Machine Learning techniques to classify climate but are either too memory-intensive [NS16], focus on specific areas of

the world [LTYL05], or do not have temporal components [ZMH12].

One issue that the newer models face is how to approach seasonality of weather data. Many researchers propose using time-warping to align multiple temporal datapoints [SKDCG16, IPJ15, DBMC19]. Another shortcoming many of the current analyses is that, most, if not all, provide static visualizations such as those given by the PLACE project [Cen12]. Without dynamic visualizations, it is difficult to understand how the climate classifications evolve over time and space.

1.3 Problem and Motivation

Given the complexities of using unsupervised learning to define climate biomes, no universally accepted model exists. Thus, our goal is to develop an interactive, dynamic tool that will help move the state of the art towards Machine Learning-based climate classifications which will be useful to indicate that the climate zones are shifting across time and space.

2 Methodology and Experimentation

Our project consists of taking weather station data from across the world over a long historical period, preprocessing the dataset, running a clustering algorithm on that data and then visualizing the results. The work is divided into two simultaneous and iterative streams: *Determining the Clusters* and *Visualization.*

2.1 Data and Preprocessing

2.1.1 Getting the Raw Data

In order to observe global climate variations over a broad time period, we turned to the National Centers for Environmental Information (NCEI) within the National Oceanic and Atmospheric Administration. Their website makes available hourly weather data from more than 35,000 stations from 1901 to the present in the Integrated Surface Dataset (1.2 TB) [NOA15].

The data are ordered by year and each year has an `html` page listing the `csv` files corresponding to station data for that year.

To retrieve our dataset, we developed a python script retrieving the `html` code of a page corresponding to a year and, with regular expression rules, we got the list of all the station download links for a given year.

As we were writing the script, we realized that a daily summarized version of the hourly dataset was available on this same website with a size of approximately 32 GB. Using the daily dataset had the additional benefit of providing a precipitation column formed from some combination of columns in the hourly dataset. This combination of columns required prior field knowledge which we did not have. This decision is based on [KVB11], which states that taking precipitation into account is key to classify biomes. We removed the years below 1974, as they have low station availability (see Figure 1).

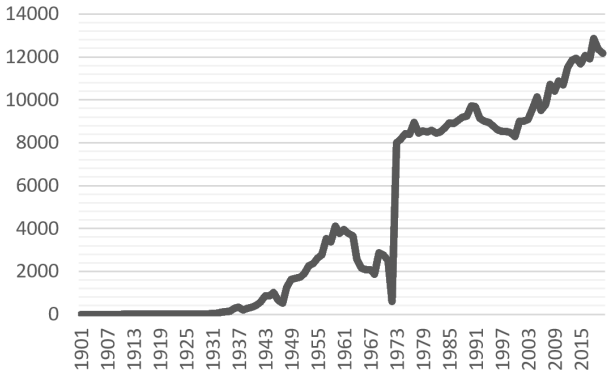

Figure 1: Number of available stations over time

2.1.2 Features

Our dataset had many different - 23 features total - and we wanted to reduce the number of features which our models would be based on. The first criteria for feature selection was avoiding having features with too much missing data. Then, we wanted to select non-correlated features. To do this, we plotted the distribution of the feature values through time, for different stations across the world. We retained five relevant features: dew point, precipitation, temperature, equivalent atmospheric pressure at sea level and wind speed. Among them, three are frequently taken to build biomes in the literature: temperature, precipitation and pressure at sea level.

In our models, we decided to keep the two possibilities, even if *a posteriori* we can see slightly better results in working with 3 features only.

2.1.3 *Inferring missing data*

We observed that a large part of the stations (67%) do not have daily data or that certain attributes (precipitation, temperature, etc.) are not reported for several consecutive days (sometimes for ten consecutive days). To address this we decided to linearly interpolate with existing data, in order to have a consistent data set for each of the station-year tuples.

2.2 Determining the Clusters

2.2.1 *Extracting the features*

To be able to take into account the reversed seasonality in the Northern and Southern Hemispheres, we decided to use a state-of-the-art Dynamic Time Warping method (see Figure 2). To extract the features from our data, we used a Fourier Transform method. With the alliance of Dynamic Time Warping and Fast Fourier Transform [Huo18, SC78], we generated tensors that summarized one weather parameter for a station-year tuple. The tensor was generated with different resolutions (low and high).

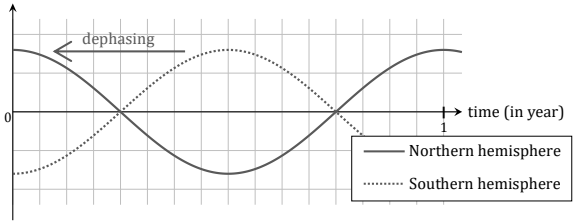

Figure 2: Dephasing during the extraction of features

2.2.2 *Clustering from the features*

Dimensionality reduction

Our goal is to visualize a model showing shifts in climate biomes. To do this we have to determine the type of meteorological characteristics (temperature, atmospheric pressure, precipitation, etc.) characterizing the model. This number of characteristics is then multiplied by the number of Fourier coefficients that the end user can choose, which in some cases leads us to cluster in spaces of `d = (definition` $\times$ `number of characteristics)` dimensions. Sometimes these features are colinear and to reduce the complexity of fitting clusters in a high dimensional space, we attempted dimensionality reductions (PCA & ICA) before clustering the data.

Clustering algorithms

We used two different algorithms to perform our clustering: the broadly used k-means and the more probabilistic Expectation-Maximization. They have different mathematical formulations of what constitutes an optimal cluster. Given that we have no idea of the mathematical distribution underlying our data, we concluded that both could potentially produce interesting climate biomes for our visualization. This conclusion proved true as detailed below.

Measures of Model Quality

After clustering, we have more than 900 raw climate biome models, in total. To select the most promising, we developed two metrics, that we then combined them to have a single score per model. As a baseline for accuracy, we use the Köppen-Geiger model as a ground truth.

For the first error metric, we wanted to quantify the stability of intra-station cluster assignment: we created one list of clusters we found per Köppen-Geiger climate, then computed the entropy of each list and summed the entropies. This metric is a measure of confidence, it gives an insight into the disorder inside a Köppen-Geiger climate.

For the second error metric, we performed the exact same steps, inverting the roles of the Köppen-Geiger model and our model. We wanted to inquire the relevance of the climate zones based on known climate biomes, such as deserts or rainforests (accuracy measure), to measure the disorder of Köppen-Geiger climates inside our model climates. To combine these two metrics, we multiplied the errors found, and scale them so that we could compare the models.

The main challenge of combining the metrics is to be able to compare one model to the others for a given number of clusters. Based on the central-limit theorem, we expect the distribution to be somewhat Gaussian. From that, we wanted to align all the means, and to have a similar area under the curve

to create one unified metric from all the previous singular ones, as presented in Figure 3.

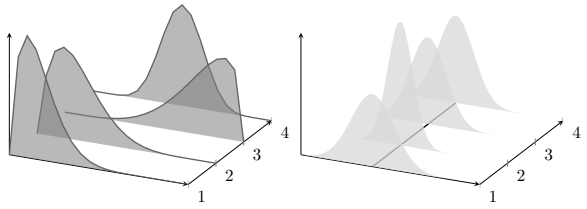

Figure 3: Aligning and scaling the metrics across the clusters

2.3 Visualization

In addition to displaying the clustering results, the visualization serves as a post-processing step for the data, removing temporal and spatial noise while filling in missing areas.

2.3.1 D3 and SVG

After evaluating a number of visualization options, we settled on using Javascript and D3 to develop the interface, due to the universality and flexibility of web apps. We initially developed a prototype of the interface using D3's Canvas elements, but later switched to SVG for easier interactivity (e.g. - mouseover events and tooltips). Significant effort was taken to optimize performance for average computers using Chrome.

2.3.2 Visualization Patterns: Voronoi, Grids, Contours

In deciding how to display the thousands of labeled weather stations on the world map, we considered using the Voronoi pattern, which creates polygonal cells with borders halfway between adjacent points. However, this approach creates minuscule cells in dense areas and massive cells in sparse areas (compare Western Europe and Africa in Figure- 4). Thus, we opted for a gridded approach which divides the map into evenly sized squares. Squares containing multiple stations use the mode for their label; squares with no stations are left blank. This approach solves both the issues of redundancy in dense areas and over-representation in sparse areas. It also looks less visually distracting (see Figure 4). As we moved to determining how to make the visual interactive, we realized that that the gridded approach was too slow as it generated hundreds of cells. Thus, we decided to group the cells into GeoJSON polygon/contour objects, yielding vastly improved performance (see Figure-4).

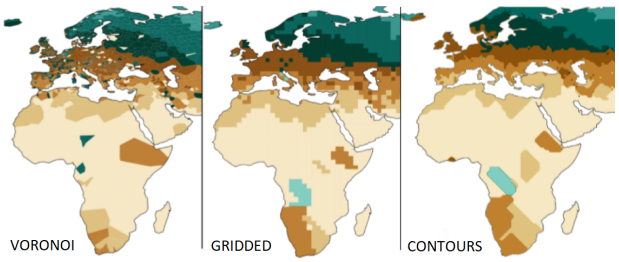

Figure 4: Voronoi, gridded, and contour patterns on the same model

2.3.3 Automatic Color Assignment

Since the labels produced from the clustering algorithm are categorical and arbitrary, we needed a way to consistently color the clusters when switching between models of varying cluster counts (3,5,7,9) for easier comparison. Therefore, we used the clusters' average latitudes to evenly space its labels across our chosen color palette.

2.3.4 Filling Empty Regions

Since weather is typically not localized to the immediate area around the weather stations, we chose to fill empty spaces by "growing" the clusters iteratively. The algorithm takes empty cells that are contiguous to filled cells and fills the empty cell with the mode of its neighbors. By repeating this growing process, empty regions can be filled to varying degrees (see Figure 5).

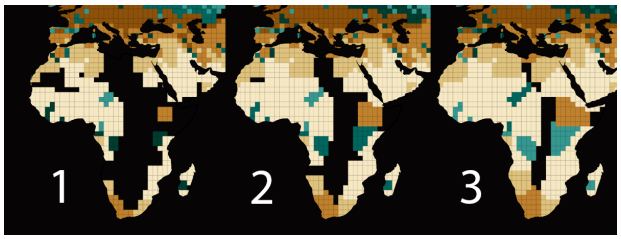

Figure 5: Iterative growth passes to fill empty spaces

2.3.5 *Implementing Smoothing*

Despite the prior data treatment steps, a fair amount of spatial and temporal noise remains, causing the animation to have a chaotic, flickering appearance. We addressed this problem by implementing a custom change detection algorithm similar to *CUSUM* [Mac90], mitigating temporal noise. For spatial noise, we applied the aforementioned contours which use the "Marching Squares" algorithm to smooth out the gridded cells [LC87]. With both smoothing steps applied, the final animation has more easily observable patterns with reduced flickering over time and space.

2.3.6 *Interactive Components*

We created interactive components of two types. The first type is model selection and map animation allowing users to choose how many clusters they want to display on the map and to animate through time to see how clusters change over the years. The model displayed corresponds to the highest scored model for that number of clusters, based on our scoring metrics.

The second type is mouse-based interaction, allowing users to mouseover a climate biome and see basic statistics for the weather variables for that particular biome. As a region is moused-over, the biome cluster contour shades to a contrasting red color and the statistics tooltip appears with the weather variable information (see Figure 6).

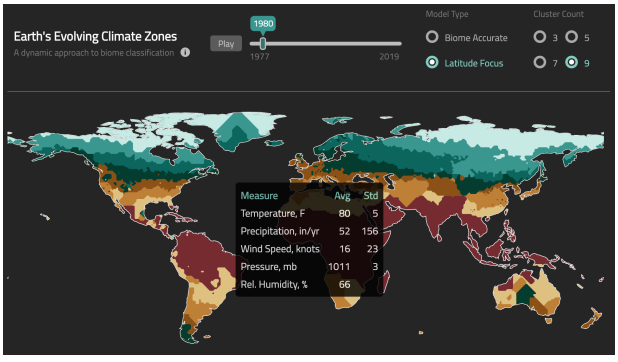

Figure 6: Interactive Components

3 Limitations and Future Directions

Although this work seeks to respond to the problem statement in a robust and comprehensive manner, there are still limitations and future avenues for exploration.

A major limitation in the data lies in the lack of long-term weather station data from certain areas of the worlds. For example, Central and Saharan Africa, parts of the Amazon River basin, Siberia and the Indian subcontinent all lacked weather stations until relatively recently. Although one solution would be to search for other datasets to supplement the one we used (such as satellite data), the only long-term solution is to increase the number of weather stations in these areas and let time do the work required to generate more data. To fill in these areas that don't have many stations, we used the "filling" steps, as outlined above, but a more advanced method may exist, taking into account other factors, such latitude, altitude, or other information.

From an algorithmic angle, although we desired to create a consensus model from the different clustering algorithms we tried, this proved too ambitious for the limited timeframe allotted to the work. Future work can address this shortcoming and try to understand how the strengths and weaknesses of the deterministic k-means clustering (better latitudinal classification) and the probabilistic expectation maximization clustering (better longitudinal classification) might inform a coherent consensus model, as proposed in [DWBA20].

The visualization uses a *CUSUM*-like algorithm to detect changes in the assignment of climate clusters. Although we envisioned trying *CDCStream* [IBPP14], or some other algorithm, due to the time constraints of the project, we were not able to explore this avenue as much as we wanted. We also wanted to have graphs indicating the true distributions of weather variables and be able to compare a certain year's distribution to the overall timeframe distribution. Finally, we would like to add an "advanced" mode to the model selection interactivty, allowing users to pick models based on other attributes besides number of clusters, such as number of parameters, etc.

4 Discussion and Conclusions

The goal of this work was two-fold: 1. to develop a method by which to cluster areas of the Earth into climate biomes using only weather data, avoiding

the use of human heuristics, and, 2. to examine if those clustered biomes exhibit a tendency to shift toward the poles.

4.1 Viewing the Global Climate

The approach outlined in this work provides a novel method to view the rapidily shifitng climate biomes.

Many other climate classifications are based on physical or heuristic models (Köppen and Trewartha) [KVB11, BHHK14] - our model instead uses a cluster analysis that relies on the data it is fed, thus we are not limited to arbitrary, predefined, or human-based definitions. Such a model is also adaptive to a changing climate in that, should a new climate biome arise, the model will automatically classify it.

Further, the number of climate zones to cluster out can be tweaked which allows us to perform more abstract or, inversely, more fine-grained analyses. Additionally, because of the speed by which the models can be updated (simply continue to plug in new observations), new year-based clusters can be developed on the fly. This would allow policy makers in the domains of conservation, forestry, and agriculture to see how climate biomes are changing and formulate policy actions quickly. This is opposed to the current process to create the USDA Hardiness Zone maps, for example, which takes years [DWH+12]. Although the tool detailed in this work would not replace that process, it would provide an informative tool for inter-map periods.

Both algorithms (EM and k-means) performed well, appearing among the top-scoring models, with k-means favoring stability and EM favoring accuracy. Our model scoring algorithm indicated that the best model was the 9-cluster, k-means, low resolution, 3 parameters model without dimensionality reduction. The corresponding EM model also performed well as it is very close to the Trewartha reference. In Figure 7, we compare the known reality of the Trewartha model with both model types.

Figure 7 illustrates how our models match the known reality based on current human models, providing evidence that our models are not fallacious.

Finally, the tool itself provides an animated, interactive visualization that shows climate changes over time. To our knowledge, no other animated visualization of climate biomes exists; all other examples that we have found have been static.

Figure 7: Comparison of Trewartha model (center) with our k-means (left) and EM (right) models. Trewartha image credit: [Pet19]

4.2 A Climate, Changed

The visualization tool that we provide gives evidence that the climate zones are shifting poleward, fulfilling the second objective of this project (see Figure 8)

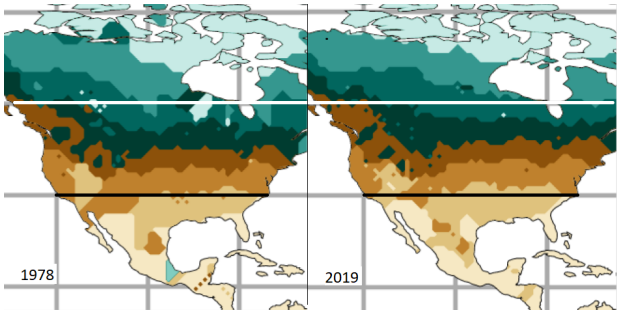

Figure 8: With the white and black lines as references, the northward shift of climate zones in the USA from 1978 (left) to 2019 (right) is evident.

Although the shifts are not very extreme, they are discernible and provide sobering visual evidence for what we already know is true: The climate has changed.

From simple weather station data, indicating precipitation, temperature, dew point, wind speed, and pressure, we have created dynamic models that accurately classify climate biomes and that illustrate how those biomes are shifting towards the poles.

We believe that the ideas and methodology presented here can help inform the public and policy makers by making it easier and quicker to visualize the changing climate.

REFERENCES

[ALCM16] Ernest Agee, Jennifer Larson, Samuel Childs, and Alexandra Marmo. Spatial redistribution of U.S. Tornado activity between 1954 and 2013. *Journal of Applied Meteorology and Climatology*, 55(8):1681–1697, 2016.

[BHHK14] Michal Belda, Eva Holtanová, Tomáš Halenka, and Jaroslava Kalvová. Climate classification revisited: From Köppen to Trewartha. *Climate Research*, 59(1):1–13, 2014.

[BPRW18] Michael P. Byrne, Angeline G. Pendergrass, Anita D. Rapp, and Kyle R. Wodzicki. Response of the Intertropical Convergence Zone to Climate Change: Location, Width, and Strength. *Current Climate Change Reports*, 4(4):355–370, 2018.

[Cen12] Center For Earth Science Information Network-CIESIN-Columbia University. National Aggregates of Geospatial Data Collection: Population, Landscape, And Climate Estimates, Version 3 (PLACE III), 2012.

[DBMC19] Mohamed Djallel Dilmi, Laurent Barthès, Cécile Mallet, and Aymeric Chazottes. Iterative multiscale dynamic time warping (IMs-DTW): a tool for rainfall time series comparison. *International Journal of Data Science and Analytics*, 10(1):65–79, 2019.

[DWBA20] Derek DeSantis, Phillip J. Wolfram, Katrina Bennett, and Boian Alexandrov. Coarse-grain cluster analysis of tensors with application to climate biome identification. *Machine Learning: Science and Technology*, 2020.

[DWH+12] Christopher Daly, Mark P. Widrlechner, Michael D. Halbleib, Joseph I. Smith, and Wayne P. Gibson. Development of a new USDA plant hardiness zone map for the United States. *Journal of Applied Meteorology and Climatology*, 51(2):242–264, 2012.

[Hol67] L. R. Holdridge. Life zone ecology. *Tropical Science Center*, page 206, 1967.

[Huo18] Juan Huo. Dynamic time warping and FFT: A data preprocessing method for electrical load forecasting. *International Journal of Advanced Computer Science and Applications*, 9(2):1–6, 2018.

[IBPP14] Dino Ienco, Albert Bifet, Bernhard Pfahringer, and Pascal Poncelet. Change detection in categorical evolving data streams. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing - SAC '14*. ACM Press, 2014.

[IPJ15] Hesam Izakian, Witold Pedrycz, and Iqbal Jamal. Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39:235–244, 2015.

[KVB11] Wladimir Koppen, Esther Volken, and Stefan Brönnimann. The thermal zones of the Earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world. *Meteorologische Zeitschrift*, 20(3):351–360, 2011.

[LBD+99] Ariel E. Lugo, Sandra B. Brown, R. Dodson, T. S. Smith, and H. H. Shugart. The Holdridge life zones of the conterminous United States in relation to ecosystem mapping. *Journal of Biogeography*, 26(5):1025–1038, 1999.

[LC87] William Lorensen and Harvey Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21:163–, 08 1987.

[LTYL05] Joseph C Lam, CL Tsang, L Yang, and Danny HW Li. Weather data analysis and design implications for different climatic zones in china. *Building and Environment*, 40(2):277–296, 2005.

[Mac90] Gordon MacDonald. *Global climate and ecosystem change*, pages 72–73. Springer Science+Business Media, LLC, New York, 1990.

[MHPSL17] Alberto Monllor-Hurtado, Maria Grazia Pennino, and José Luis Sanchez-Lizaso. Shift in tuna catches due to ocean warming. *PLoS ONE*, 12(6):1–10, 2017.

[NOA15] NOAA National Centers for Environmental Information. Federal Climate Complex Data Documentation for Integrated Surface Data. Technical report, NOAA National Centers for Environmental Information, 2015.

[NS16] Pawel Netzel and Tomasz Stepinski. On Using a Clustering Approach for Global Climate Classification. *Journal of Climate*, 29(9):3387–3401, 04 2016.

[PA16] Lauren E. Parker and John T. Abatzoglou. Projected changes in cold hardiness zones and suitable overwinter ranges of perennial crops over the United States. *Environmental Research Letters*, 11(3), 2016.

[PA18] Julia Piantadosi and Robert S. Anderssen. Maintaining Reliable Agriculture Productivity and Goyder's Line of Reliable Rainfall. In Robert S. Anderssen, Philip Broadbride, Yasuhide Fukumoto, Kenji Kajiwara, Matthew Simpson, and Ian Turner, editors, *Agriculture as a Metaphor for Creativity in All Human Endeavors*, volume 28. Mathematics in Industry, chapter 9, pages 125–138. Springer Singapore, fmfi 2016 edition, 2018.

[Pet19] Adam Peterson. *Trewartha climate types of the Contiguous United States*. Jan 2019.

[SBD12] Jennifer M. Sunday, Amanda E. Bates, and Nicholas K. Dulvy. Thermal tolerance and the global redistribution of animals. *Nature Climate Change*, 2(9):686–690, 2012.

[SC78] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[SKDCG16] Saeid Soheily-Khah, Ahlame Douzal-Chouakria, and Eric Gaussier. Generalized k-means-based clustering for temporal data under weighted and kernel time warp. *Pattern Recognition Letters*, 75:63–69, 2016.

[SLG+18] Paul W. Staten, Jian Lu, Kevin M. Grise, Sean M. Davis, and Thomas Birner. Re-examining tropical expansion. *Nature Climate Change*, 8(9):768–775, 2018.

[ZMH12] Jakob Zscheischler, Miguel D Mahecha, and Stefan Harmeling. Climate classifications: the value of unsupervised clustering. *Procedia Computer Science*, 9:897–906, 2012.

ADDITIONAL UNCITED WORKS

[CSZ+07] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 153–162, 2007.

[EE16] Christien J. Engelbrecht and Francois A. Engelbrecht. Shifts in Köppen-Geiger climate zones over southern Africa in relation to key global temperature goals. *Theoretical and Applied Climatology*, 123(1-2):247–261, 2016.

[Jon18] Nicola Jones. Redrawing the Map: How the World's Climate Zones Are Shifting, 2018.

[LS05] David M. Lawrence and Andrew G. Slater. A projection of severe near-surface permafrost degradation during the 21st century. *Geophysical Research Letters*, 32(24):1–5, 2005.

[Mát10] Csaba Mátyás. Forecasts needed for retreating forests. *Nature*, 464(7293):1271, 2010.

[PAC20] Ivens Portugal, Paulo Alencar, and Donald Cowan. A Framework for Spatial-Temporal Trajectory Cluster Analysis Based on Dynamic Relationships. *IEEE Access*, 8:169775–169793, 2020.

[SKTH17] James H. Stagge, Daniel G. Kingston, Lena M. Tallaksen, and David M. Hannah. Observed drought indices show increasing divergence across Europe. *Nature - Scientific Reports*, 7(1):1–10, 2017.