Analysis of Fort Collins Weather Using Persistent Homology

Casey Martin, Prayash Ghimire, Ethan Greef, Chien Lin, Jayed Alshammari

DSCI 475

May 10, 2024

Abstract

A data analysis technique that is useful for detecting periodicity in time series involves embedding data into a higher dimension and using persistent homology to detect the presence of loops. This technique was applied to a variety of weather measurements from Fort Collins. All persistence diagrams show the presence of a large and prominent loop. This includes the noisy measurements (relative humidity, wind speed, gust speed, wind direction, and gust direction). Wind speed, wind direction, and gust speed show the presence of several smaller subloops. These results show that weather has periodic behavior throughout the seasons and that persistent homology is useful for noisy datasets where periodicity is not obvious.

Introduction

Weather is an extremely complex system that involves many atmospheric variables that each evolve in their complex, but interrelated ways. To even glimpse into what is going on with the weather involves examining many different atmospheric measurements. This study looked at several of those measurements.

The ambient temperature of the air in °C is included and is well known to vary considerably throughout the year. Dew point is the temperature the air must be to become saturated with water vapor and depends on the pressure and water content of the air. Solar energy is the amount of energy in watts from sunlight per square meter. Relative humidity measures the current humidity of the air as a percentage of the maximum humidity the air could be at a given temperature. Wind speed and wind direction measure how quickly air is moving in m/s and what direction it is going relative to north. Similarly, gust speed and gust direction are measured in the same way, the difference being that gusts are strong and abrupt rushes of wind.

Analyzing the periodicity of the time series first involves embedding the data into a high-dimensional point cloud. It uses the variables M for the dimension, and τ for the time delay. It looks at a window of the time series that is the size of M τ and selects M points from that window. The window then slides over by τ and repeat the process until the end of the time series is reached. The M points are put into the sequence of vectors:

$$f_i = [f(t_i), f(t_i + 2\tau), \ldots, f(t_i + M\tau)] \in R^{M+1}$$

Each vector contains the position of the points for its respective axes, which together contain the positions of all the points in the M+1 dimensional space. Principle Component Analysis (PCA) is then used to reduce the point cloud into 3 dimensions (assuming the point cloud is higher than 3 dimensions) to be able to visualize the embedded data and to compute persistent homology on the most salient parts of the data.

Another parameter called the stride, controls the number of points that are embedded by selecting time series points that are some set distance apart. This is important for the computational cost of these methods because embedding and examining the persistence of every single point in the original data can take extremely long.

The PCA embedding is then converted into a Vietoris-Rips complex to measure persistent homology. Spheres with radius r are created around each point and r is gradually increased. When two spheres touch each other, a connection is drawn between points to form a simplicial complex. 1 simplices consist of edges between points, 2 simplices consist of a filled-in triangle, and 3 simplices are filled-in tetrahedra. A group of points connected with edges is a connected component. 2-dimensional spaces that are surrounded by edges but not filled in are loops, and 3-dimensional spaces that are surrounded by surfaces are voids.

To measure persistent homology, r is increased until the entire complex fills in. The values of r for when connected components, loops, and voids appear and disappear are plotted on a diagram where the x-axis represents when they appeared, and the y-axis represents when they disappear. Points that are close to the y = x line are noisy topological features that disappear as quickly as they appear, while points that are far away correspond to actual topological features of the dataset.

Methods

The weather dataset was downloaded from the Colorado State University Department of Atmospheric Science. This dataset contains a variety of hourly weather measurements taken

from Christman Field in Fort Collins. The dataset ranges from April 24, 2020, to April 24, 2024. Time series corresponding temperature (°C), dew point (°C), solar energy (W/m2), relative humidity (%), wind speed (m/s), gust speed (m/s), wind direction (°N) and gust direction (°N) were selected (Fig. 1-8). Pressure (hPa) was not included due to a ton of missing data, and precipitation (mm) was not included due to it being a discontinuous measurement.

These time series were individually embedded into higher dimensions and projected back down into three dimensions using PCA. The parameters were chosen based on trial and error to get embeddings that showed the clearest and least distorted loops. The stride was set to 30 to be able to compute persistence diagrams quickly. The embedding dimension was chosen to be much higher than the time delay to reduce the number of points skipped over (Fig. 9, 11, 13, 15, 17, 19, 21, 23).

The embeddings were put into an algorithm that computes a Vietoris-Rips complex for different radius scales to create a barcode of when certain topological features appeared and disappeared. These features include connected components (H0), loops (H1), and voids (H2). Points corresponding to when these features appeared and disappeared were placed onto persistence diagrams.

Results

As we delved into our data, we observed some interesting patterns across all persistence diagrams (Fig. 10, 12, 14, 16, 18, 20, 22, 24). Firstly, let's discuss the H0 points. These points consistently emerged near 0 length radii and exhibited a relatively short lifespan, disappearing quite quickly after their birth. Moving on to the H1 and H2 points, we noticed a common trend. Despite appearing fairly early in the diagrams, they vanished almost instantly, indicating a fleeting presence. Interestingly, each diagram contained a single HI point, which appeared early and persisted until very late radii. Additionally, there were instances of H2 points emerging very late in the process, only to vanish instantly.

Several H1 points showed some persistence besides the primary H1 points that persisted for a long time. These are found in the relative humidity, as well as wind / gust speed and wind direction. These correspond to several subloops found in their embeddings.

Discussion

These results indicate that many different aspects of weather change periodically, which makes sense considering that weather tends to change seasonally. The time series for temperature, dew point, and solar show very clear periodicity, so it is not surprising that the persistence diagrams for them have single H1 points that are far away from the center line. In the case of solar, the time series is unusual in that it looks like several sin curves stacked on top of each other. The reason for this stacking is unknown. The embedding looks notably different from the much clearer examples of temperature and dew point in that it has 4 “knobs” in it. Regardless, it has a single central loop as expected.

The time series for relative humidity, wind / gust speed, and direction are extremely noisy, so it was unknown whether or not these would have any periodicity to them. However, they all present a central loop. Additionally, all except gust direction have several subloops visible in the embedding and the persistence diagram (although gust direction might have some short-lived but true subloops in it). It was surprising that these time series have even a central loop to them based on their noisiness, but the presence of the subloops is even more surprising. It is unknown what these subloops might represent and is likely to be a topic of future research.

Overall, these results demonstrate that weather has periodicity to it even for measurements where that periodicity is not obvious from just looking at the time series. This underscores the usefulness of persistent homology techniques as they are useful for showing periodicity in complex datasets where noise is high, and no periodicity is obvious. This makes these techniques immensely powerful for studying complex dynamical systems such as the activity of neurons, gravitational waves, and ocean currents.

Figures

Figure 1: Temperature Over Time

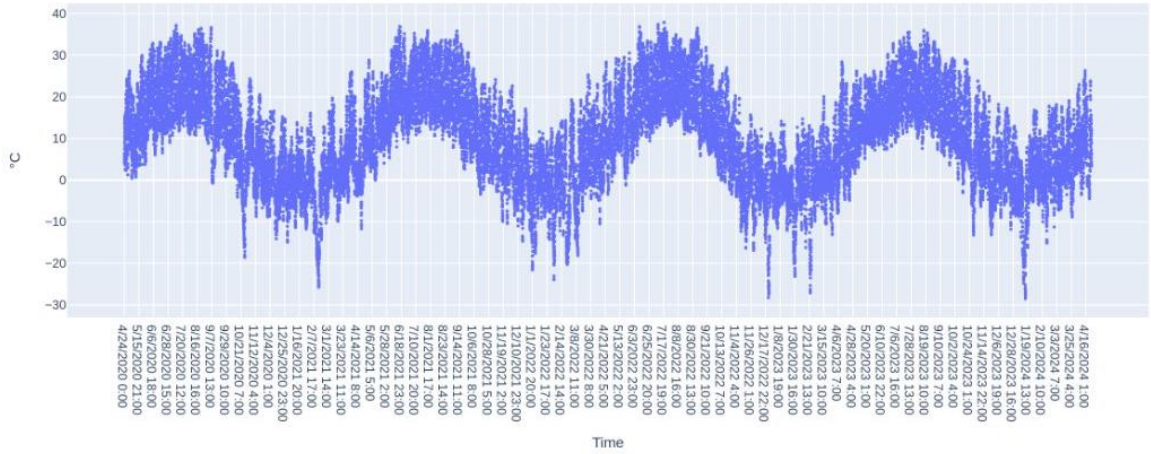

Figure 2: Dewpoint Over Time

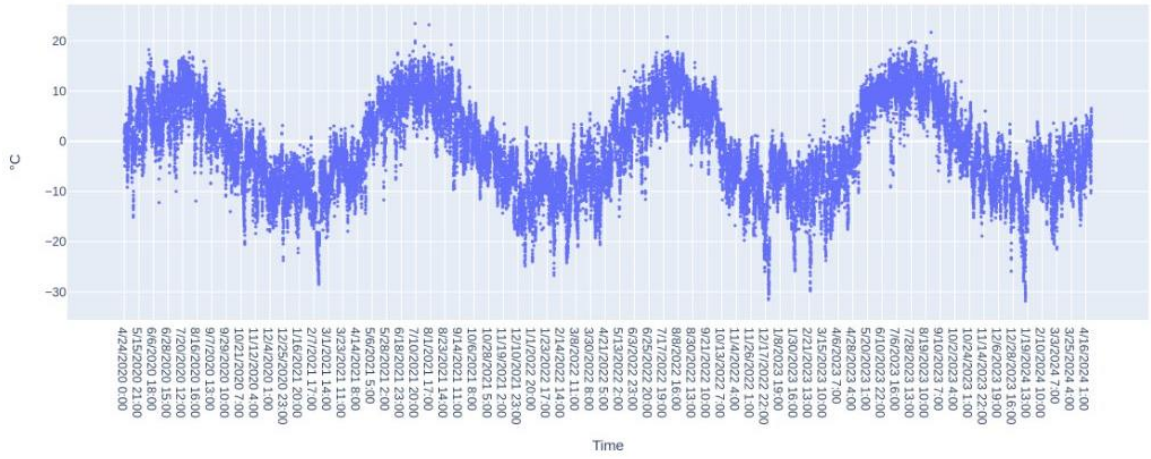

Figure 3: Solar Energy Over Time

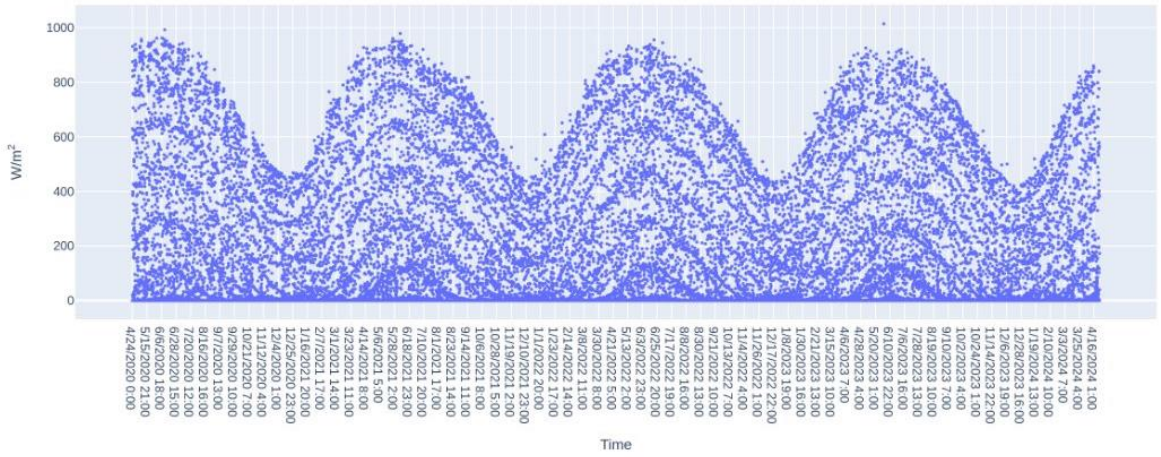

Figure 4: Relative Humidity Over Time

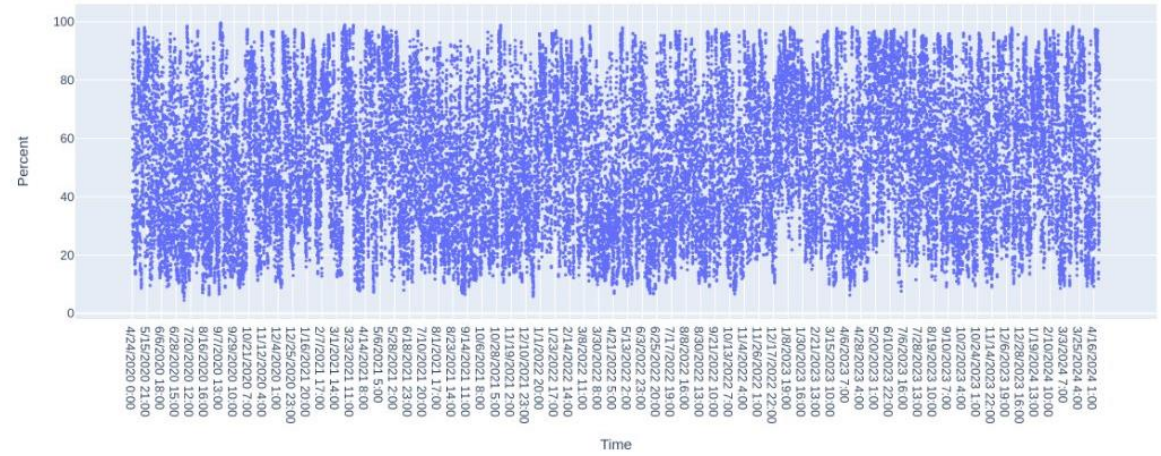

Figure 5: Wind Speed Over Time

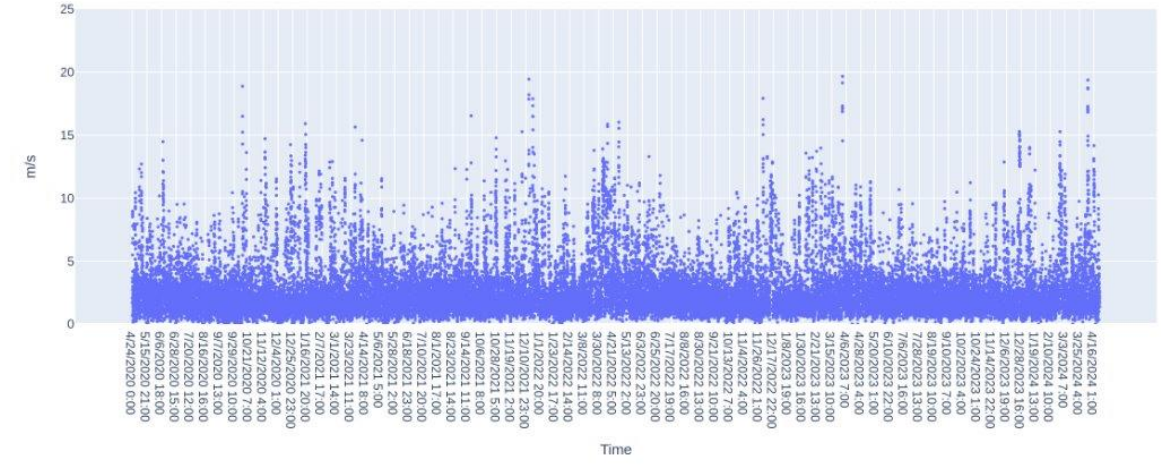

Figure 6: Wind Direction Over Time

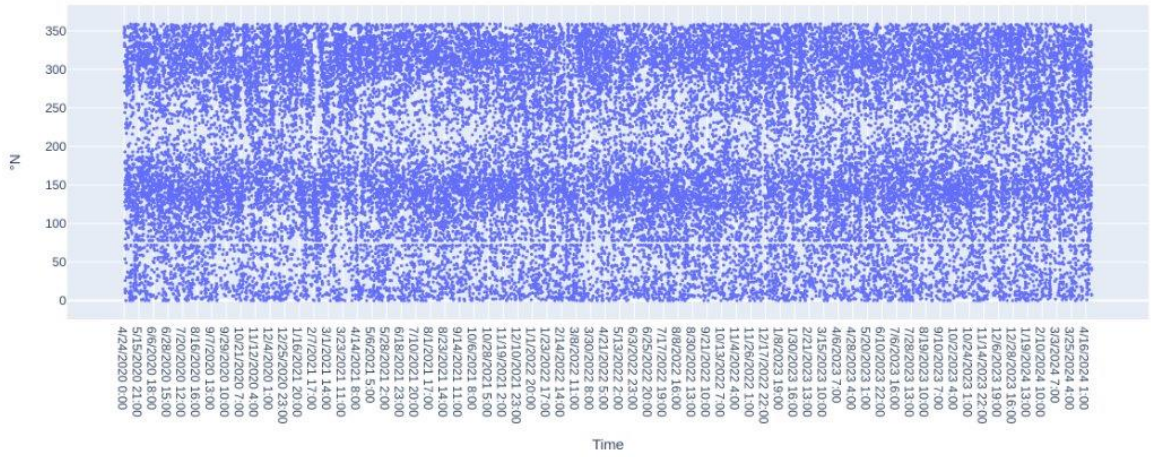

Figure 7: Gust Speed Over Time

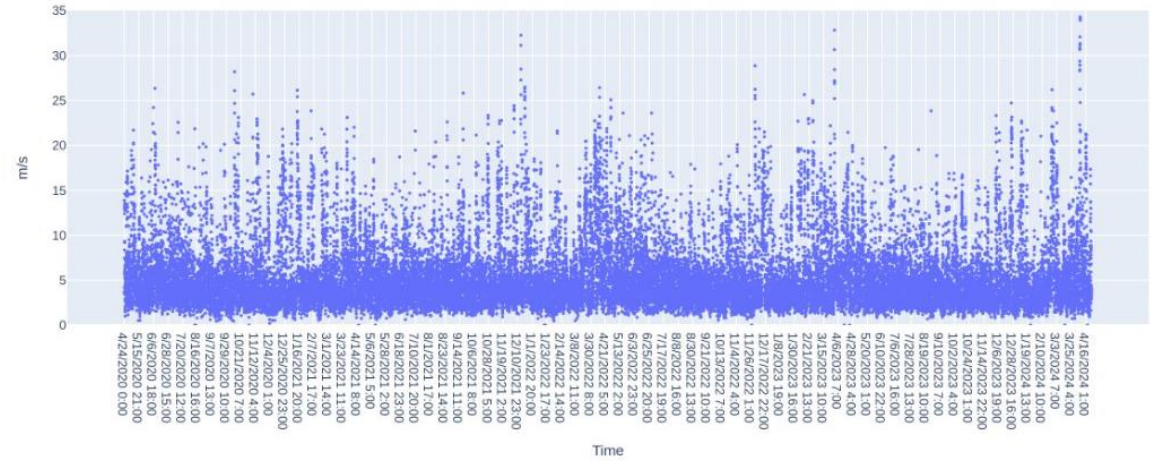

Figure 8: Gust Direction Over Time

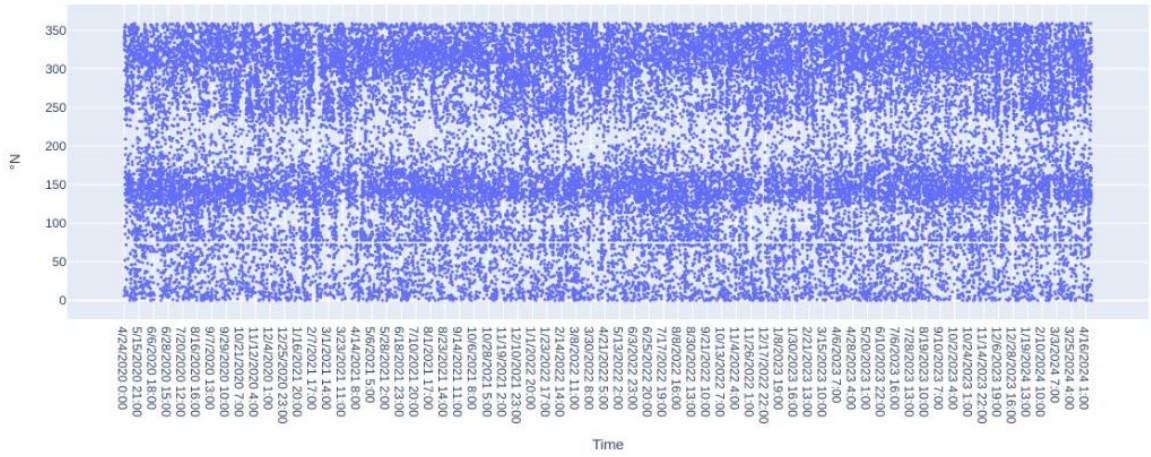

Figure 9: Embedding of Temperature

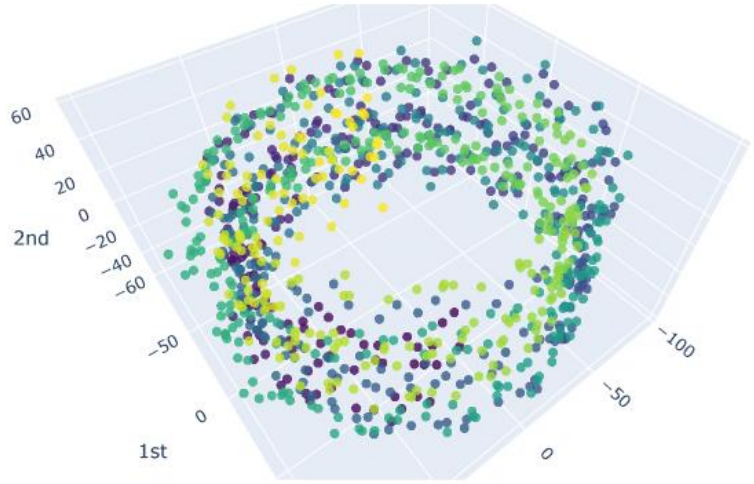

Figure 10: Persistence Diagram of Temperature

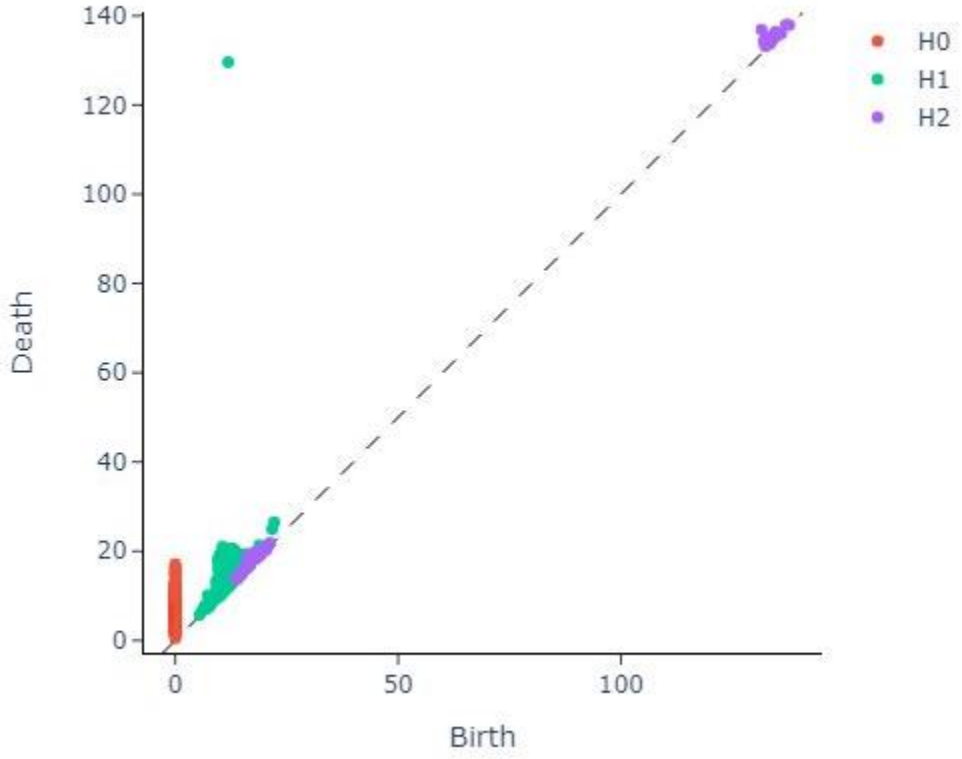

Figure 11: Embedding of Dew Point

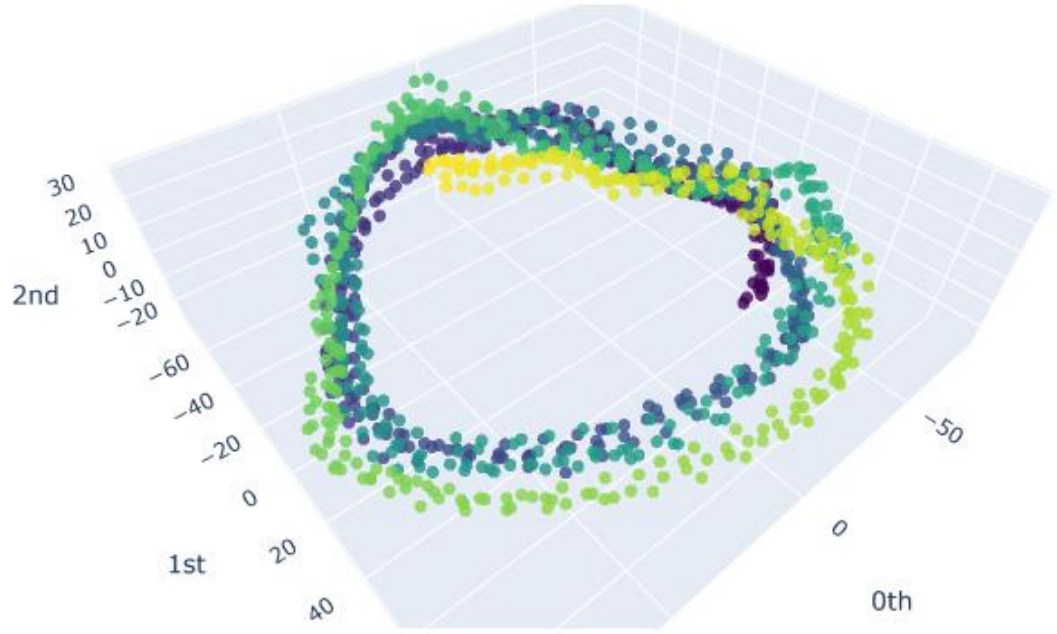

Figure 12: Persistence Diagram of Dew Point

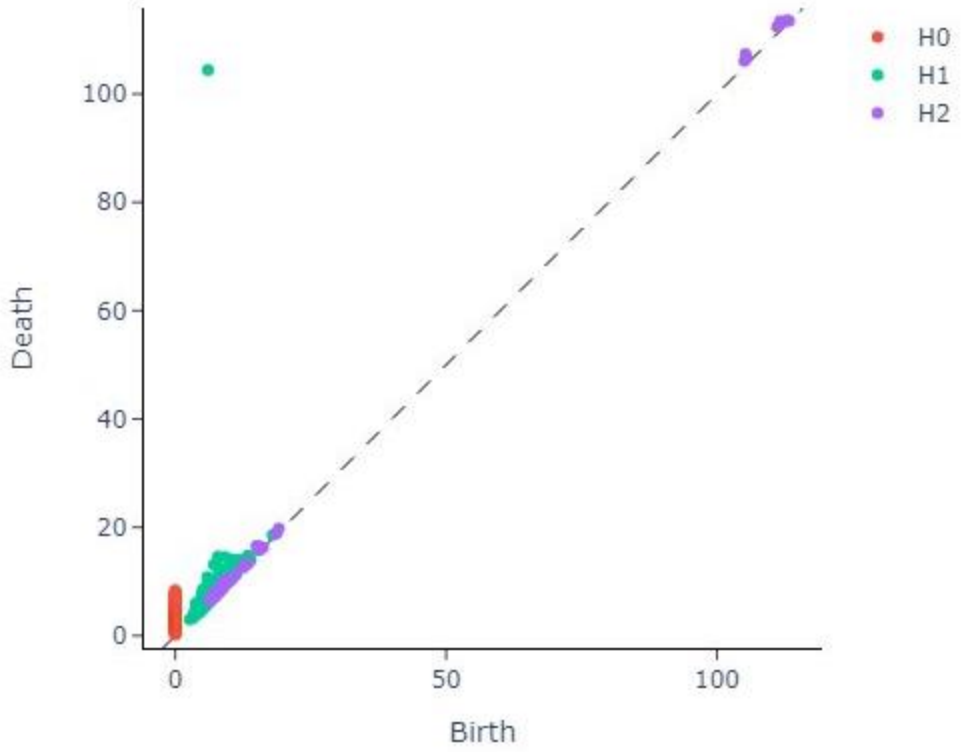

Figure 13: Embedding of Solar Energy

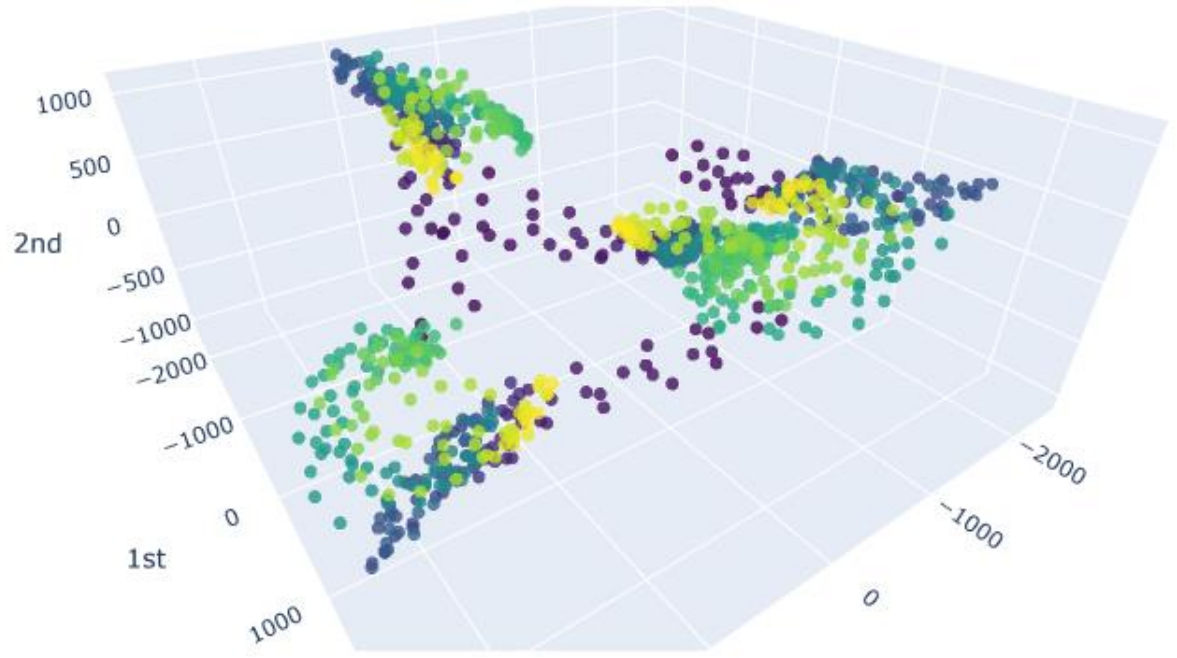

Figure 14: Persistence Diagram of Solar Energy

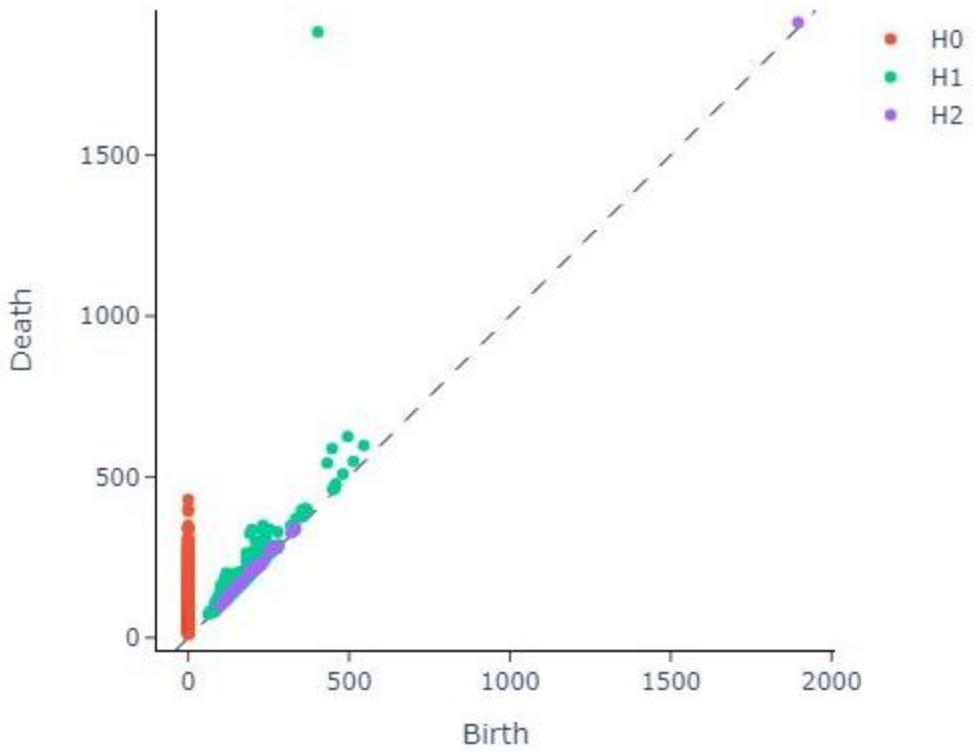

Figure 15: Embedding of Relative Humidity

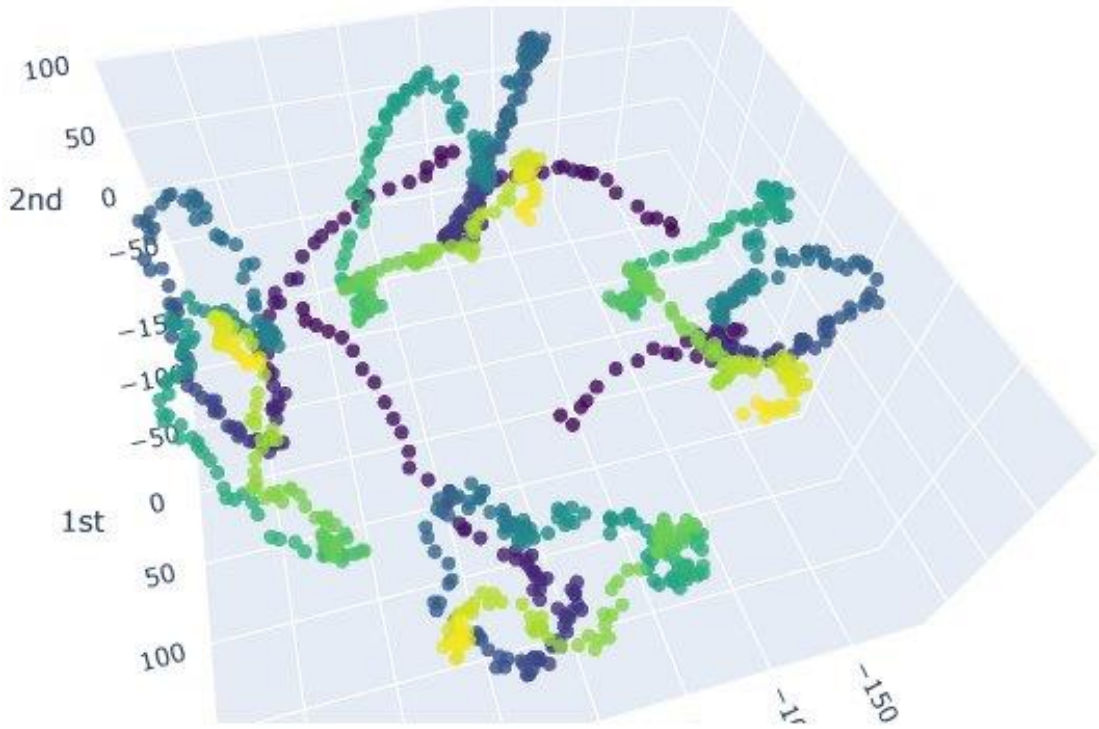

Figure 16: Persistence Diagram of Relative Humidity

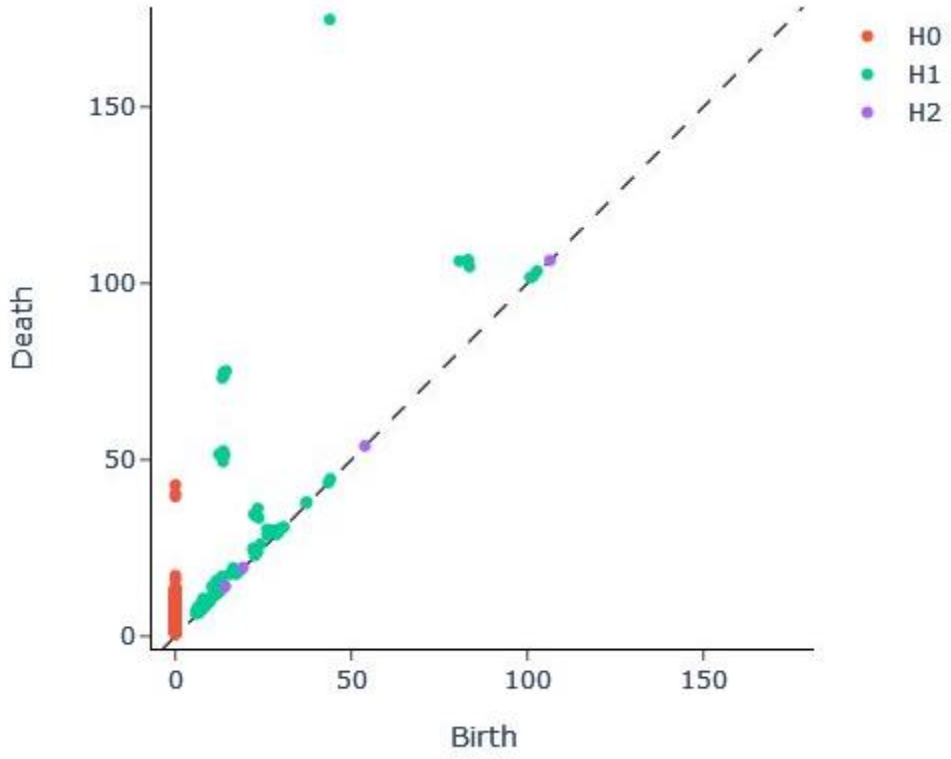

Figure 17: Embedding of Wind Speed

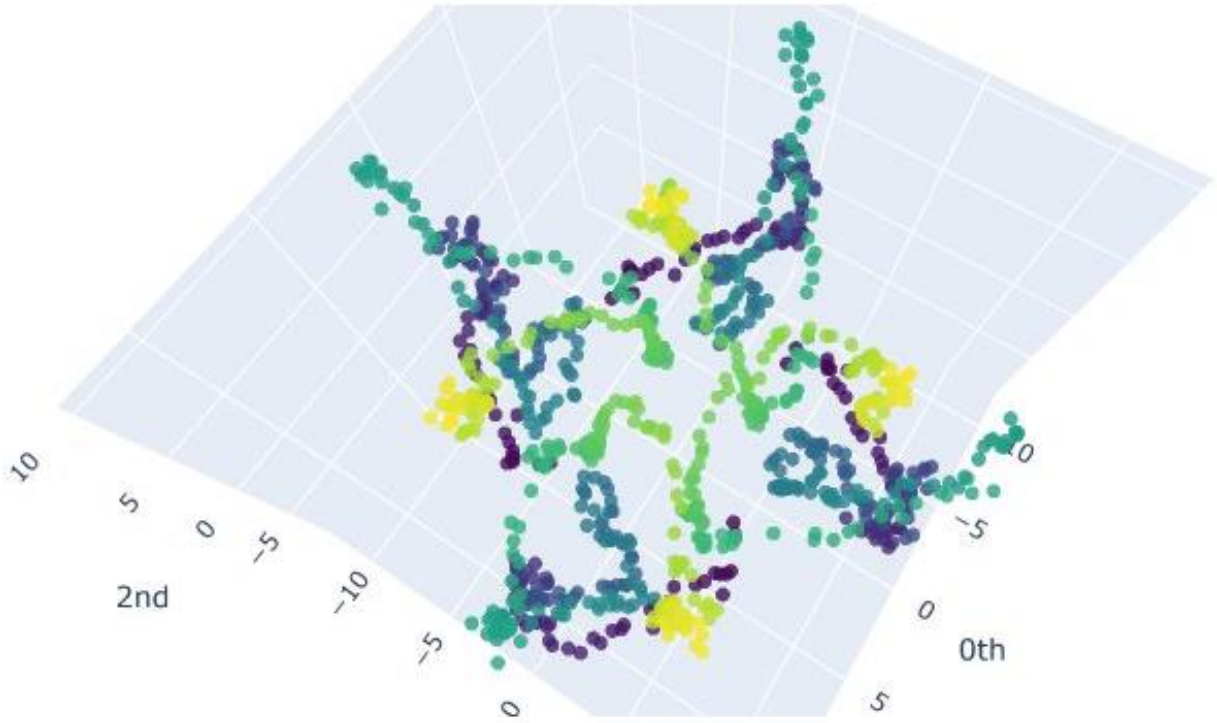

Figure 18: Persistence Diagram of Wind Speed

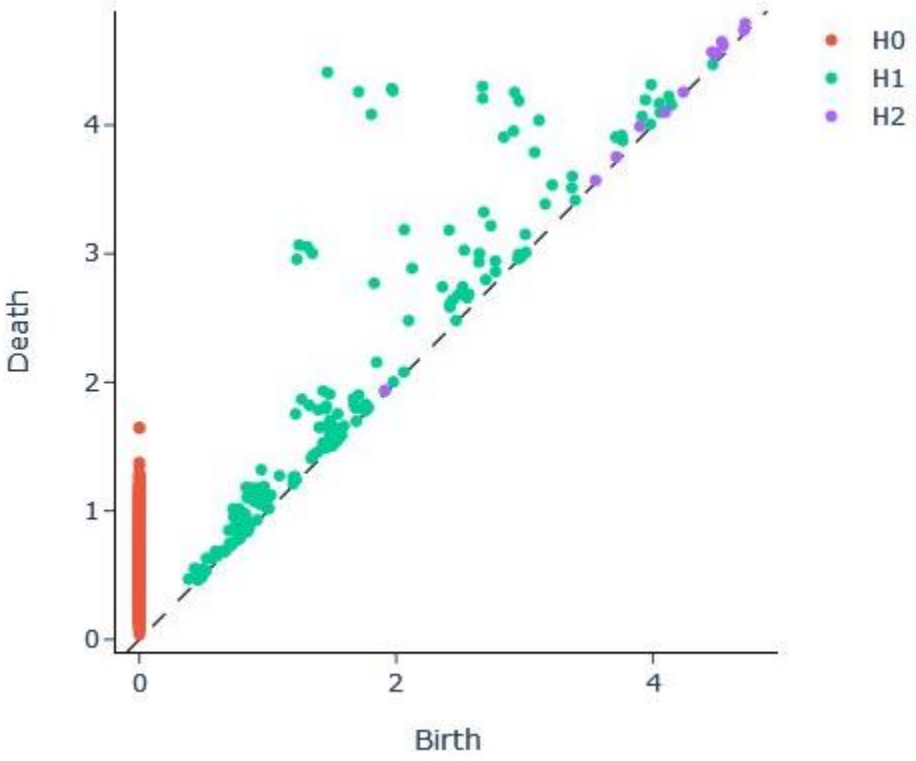

Figure 19: Embedding of Wind Direction

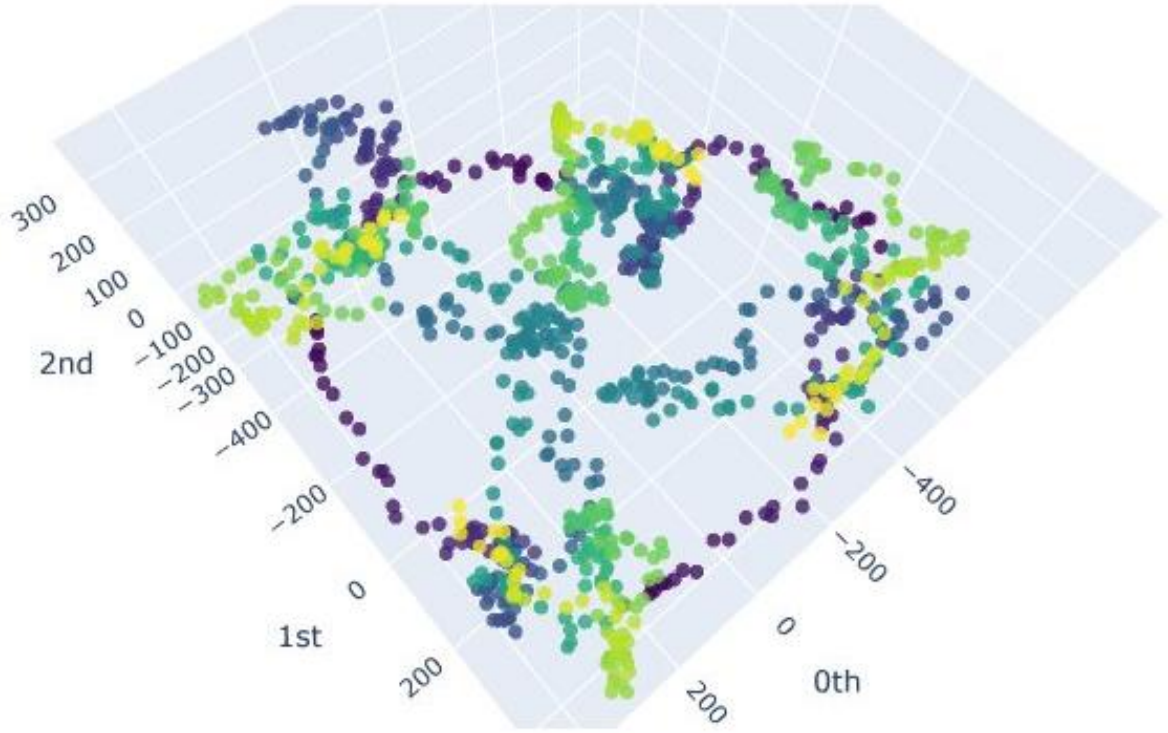

Figure 20: Persistence Diagram of Wind Direction

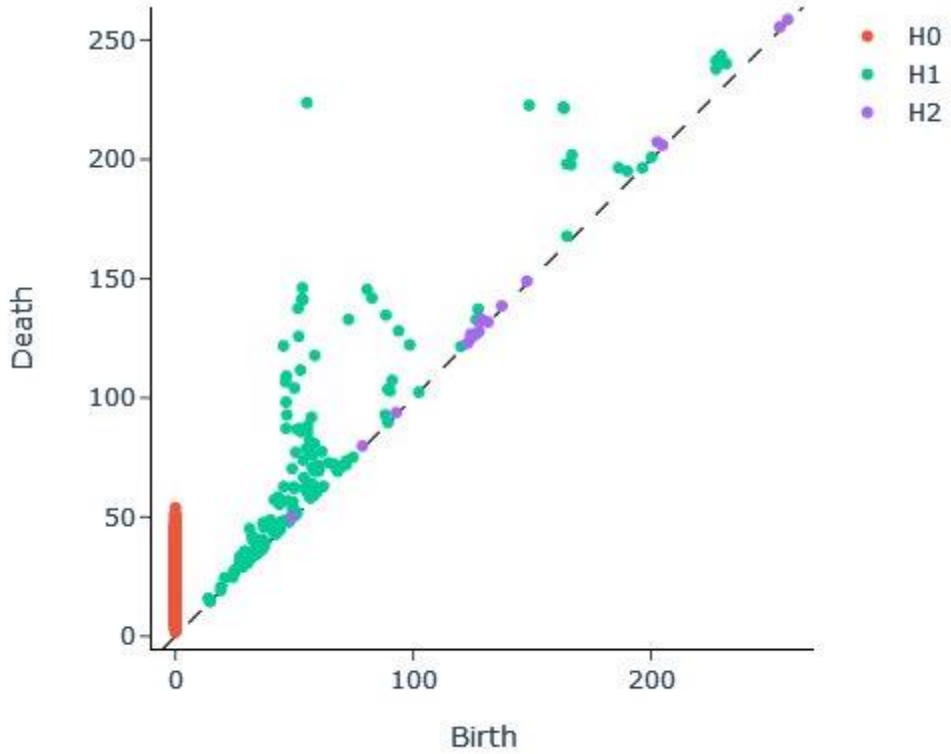

Figure 21: Embedding of Gust Speed

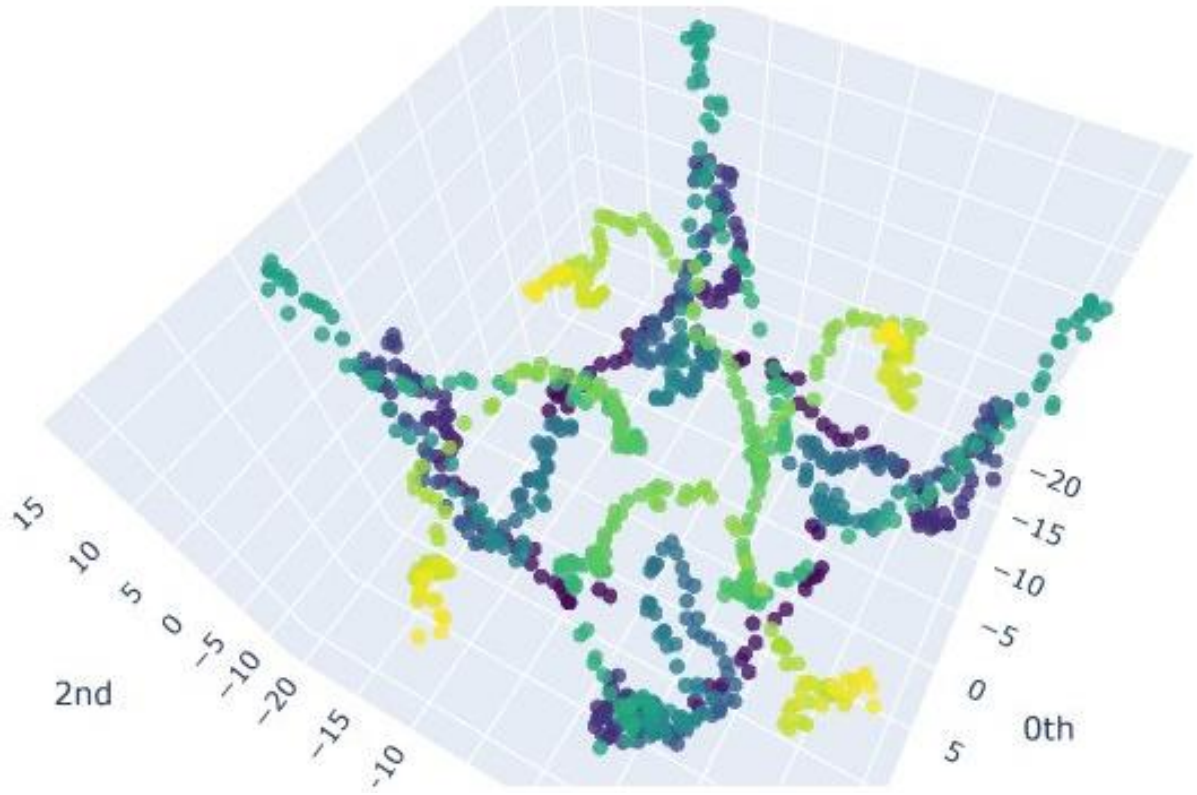

Figure 22: Persistence Diagram of Gust Speed

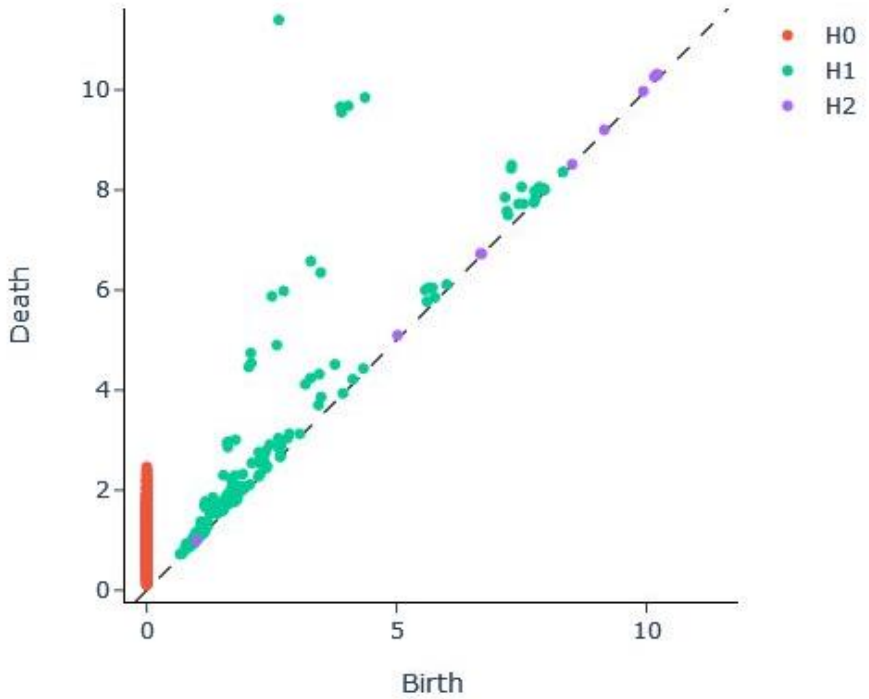

Figure 23: Embedding of Gust Direction

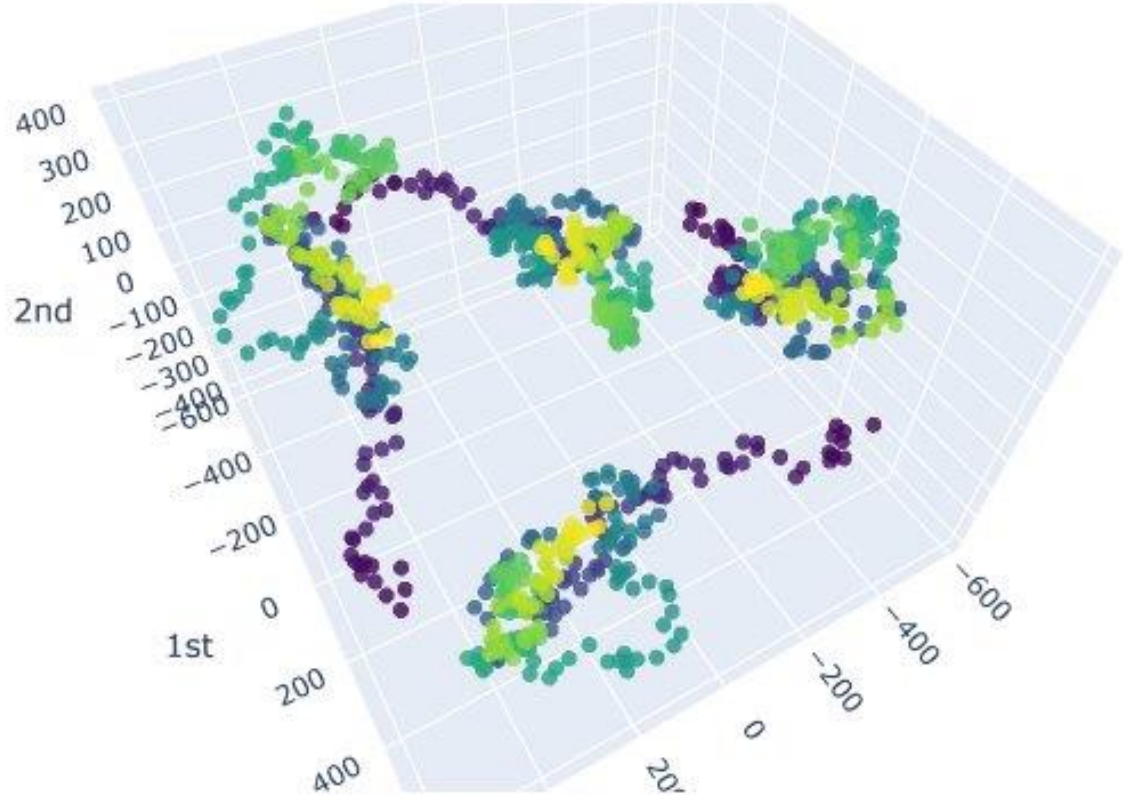

Figure 24: Persistence Diagram of Gust Direction