Working Paper: Detecting Climatic Loop Patterns with Persistent Homology

Anirudh Krovi[1]

[1]PhD, Northwestern University; MBA, NYU Stern; Formerly at McKinsey & Company, anirudh.krovi@stern.nyu.edu

July 11, 2025

Abstract

This paper introduces a topological approach to quantifying and modeling seasonal structure in weather time series. Using persistent homology, we define a scalar measure of cyclicity—*loop strength*—based on the presence of Betti-1 features in 3D point clouds derived from windowed temperature embeddings. Analyzing over a century of global city-level temperature data, we show that loop strength varies widely across space and time, capturing meaningful differences in climatic regularity.

We demonstrate that loop strength is highly predictable from raw monthly temperature profiles: a Random Forest model achieves an R^2 score of 0.9653 with a median absolute error of just 0.0272. December emerges as the most important month, suggesting that the model learns to detect geometric closure in seasonal cycles. Clustering cities based on their loop strength histories reveals latent climate groupings with potential applications in climate modeling, anomaly detection, and urban planning.

Finally, we outline future applications of this framework in financial time series analysis—particularly commodities—where loop structure may serve as a signal of underlying regime shifts. This work highlights the power of topological methods for learning from the *shape* of temporal data, with implications that extend beyond weather to any domain where cycles matter.

1 Introduction

Topological ideas have long played a role in the analysis of geometric and nonlinear systems, but only recently have they gained traction as practical tools for data analysis. This shift has been enabled by developments in *Topological Data Analysis* (TDA), a field that leverages algebraic topology to extract structural features—such as connected components, holes, and voids—from complex datasets. Among the most powerful tools in this area is *persistent homology*, which captures the evolution of topological features across multiple scales and has found applications in domains ranging from image analysis to biology and dynamical systems [4, 3, 16].

One particularly compelling application of TDA is the analysis of natural systems with inherent cyclical or periodic behavior. Weather and climate systems, for instance, often display seasonal cycles that suggest underlying loop-like structure. These cycles, however, may be distorted, broken, or hidden due to anomalies, long-term trends, or measurement noise. As such, uncovering and quantifying these "loopy" effects requires methods that go beyond simple statistical seasonality detection.

This paper investigates whether temperature dynamics at the city-year level exhibit such topological loops, and how persistent they are over time. Rather than assume a predefined structure, we construct short, overlapping windows of smoothed temperature data and embed these into geometric space. The resulting point clouds are then analyzed using persistent homology to assess whether they form topological loops.

Importantly, not all data windows exhibit clear loops, and the degree to which looping is present may itself be a meaningful indicator—reflecting climatic stability, regional seasonality, or even emerging disruption. Persistent homology provides a principled framework to evaluate these features, capturing the birth and death of 1-dimensional cycles (Betti-1 features) across scales [12].

Persistent homology has been applied in several scientific domains, including biological shape analysis [9], neural activity modeling [7], and sensor networks [2], but its use in climate-related time series remains relatively underexplored. This work aims to contribute to that growing area by applying topological methods to the historical temperature records of over 3,400 cities worldwide.

Using the Berkeley Earth GlobalLandTemperaturesByCity dataset, we focus on the period from 1900 to the present, analyzing city-level time series to identify and characterize topological loops. Our goal is to uncover patterns of loop persistence and breakdown over time, which may have implications for climate modeling, long-term risk assessment, and seasonality-aware planning.

The remainder of the paper is structured as follows: Section 2 reviews relevant work on topological data analysis and prior uses in environmental modeling. Section 3 describes the dataset and preprocessing pipeline. Section 4 details the persistent homology workflow and feature extraction. Section 5 presents empirical findings. Section **??** discusses the broader implications of our results, and Section 6 outlines directions for future work.

2 Related Work

Topological Data Analysis (TDA) provides a set of tools to extract qualitative geometric features from high-dimensional data. Among its core techniques is *persistent homology*, which tracks the birth and death of topological features (such as connected components, loops, and voids) across a filtration parameter—typically scale. This approach produces a persistence diagram or barcode that summarizes the dataset's topological structure in a multiscale, noise-robust way [4, 3, 12]. The flexibility of persistent homology makes it especially attractive for settings where traditional statistical features are insufficient to capture global structure.

Persistent homology has been successfully applied across a diverse array of scientific disciplines. In biology, it has been used to characterize protein shapes and folding dynamics [9], while in neuroscience it has helped identify latent structure in neural activity patterns [7]. In robotics and sensor networks, TDA has been used to analyze coverage maps and detect holes in physical space [2]. More recently, TDA methods have begun to appear in finance, materials science, and machine learning, often as a means to detect geometric regularities or transitions in latent representations.

Applications of TDA to time series data constitute a growing subfield. One common strategy involves transforming a time series into a point cloud using delay embeddings, then computing persistent homology on the resulting trajectory in state space [14, 1]. These methods have been used to detect periodicity, regime shifts, and complexity in signals where standard autocorrelation or frequency-domain techniques are limited. In particular, Betti-1 features (loops) can indicate recurring structure or attractor-like behavior, while Betti-0 and Betti-2 features can reflect transitions between stable and unstable modes. In financial data, such methods have been used to detect regime changes; in physiology, to detect arrhythmias.

Environmental and climate applications of TDA are less common but growing. Some work has applied persistent homology to climate model outputs or large-scale geophysical simulations [10], identifying spatial-temporal patterns such as atmospheric rivers. Others have used Mapper or persistence diagrams to analyze the topology of extreme weather clusters or temperature anomalies [11]. These efforts suggest that topology can capture emergent, nonlinear structure in climate systems—but most have focused on large spatial fields or complex simulations rather than localized, real-world sensor data.

To our knowledge, few studies have applied persistent homology to long-term, city-level temperature records. While seasonality is often studied using statistical or Fourier methods, the idea of using topological loops as a structural signature of climatic regularity remains underexplored. Our work builds on recent advances in time series embeddings and loop detection, applying them to over a century of historical temperature data across 3,400+ cities worldwide. In doing so, we seek to assess whether and when temperature patterns exhibit strong loop-like structure, and how these patterns vary across space and time. This represents a step toward using topological signatures as indicators of climate stability, disruption, or transition.

3 Data and Preprocessing

3.1 Dataset Overview

We use the **Berkeley Earth GlobalLandTemperaturesByCity** dataset, which provides monthly average temperatures for over 3,400 cities worldwide, spanning several centuries. Each row contains the city, country, date, and the corresponding average temperature for that month. We restrict our analysis to data from the year 1900 onward to ensure consistency in measurement quality and coverage.

3.2 City-Year Construction

We first aggregate the data by city and year, extracting all years where a complete set of 12 monthly temperature readings is available. For each such valid city-year pair, we store the corresponding 12-dimensional temperature vector (ordered by month). This filtering process ensures that all samples are aligned and seasonally complete. The result is a collection of vectors of the form:

$$\text{Temps}_{city,year} = [T_{Jan}, T_{Feb}, \ldots, T_{Dec}]$$

To ensure robustness and avoid overrepresentation by sparsely sampled locations, we further restrict the dataset to include only those cities for which at least 50 valid year-level records are available. This helps stabilize later analyses involving time trends or regional comparisons.

3.3 Preprocessing for Topological Analysis

Topological features such as loops are not defined directly on one-dimensional temperature vectors. To apply persistent homology, we first convert each 12-dimensional time series into a short sequence of overlapping 3-month sub-windows. Specifically, we extract 10 overlapping 3-month segments per year:

$$[(T_1, T_2, T_3), (T_2, T_3, T_4), \ldots, (T_{10}, T_{11}, T_{12})]$$

Each city-year is thus embedded into a 10-point cloud in $\mathbb{R}^3$. Intuitively, if the underlying temperature dynamics exhibit cyclicity (e.g., warm–cool–warm patterns), these points may form a loop in 3D space.

3.4 Computing Loop Strength

We compute persistent homology on each city-year point cloud using the Vietoris–Rips complex, as implemented in the `ripser` library. From the resulting persistence diagram, we extract the 1-dimensional homology (H_1) features—representing loops—and define a **loop strength** score as the maximum persistence (i.e., lifespan) of any such loop:

$$\text{LoopStrength}_{city,year} = \max_i (d_i - b_i) \quad \text{for } (b_i, d_i) \in H_1$$

This process is parallelized across all filtered city-years using CPU multiprocessing for scalability. The resulting dataframe contains each sample's location, year, loop strength, and temperature profile, and serves as the primary dataset for downstream modeling and analysis.

Why Loop Strength Matters

The loop strength score offers a compact, interpretable summary of how strongly a city's temperature dynamics in a given year resemble a closed seasonal cycle. A high loop strength suggests consistent, well-separated phases in the annual temperature trajectory—akin to distinct seasons—while a low score may reflect disruptions, flattening, or irregular transitions. By quantifying how much topological "loopiness" exists in the embedded data, we gain a new, geometric perspective on seasonality—one that is robust to small fluctuations and agnostic to specific temperature thresholds. This makes it particularly valuable for comparing climate structure across locations and time periods, especially when seeking to identify signals of climate change or anomalous years that deviate from historical patterns.

4 Methodology

4.1 Loop Strength via Persistent Homology

To quantify the presence of cyclic structure in yearly temperature profiles, we compute persistent homology on point clouds derived from monthly temperature vectors. Each city-year sample is represented as a 12-dimensional vector of monthly average temperatures. From this, we extract 10 overlapping 3-month windows (e.g., January–March, February–April, etc.), which are stacked to form a 10-point cloud in $\mathbb{R}^3$.

We apply the Vietoris–Rips complex on this cloud using the `ripser` library, and compute its persistence diagram. Specifically, we focus on 1-dimensional homology (H_1), which captures loop-like features in the data. For each diagram, we define a scalar measure of topological cyclicity—termed **loop strength**—as the maximum persistence (i.e., lifespan) of any H_1 feature:

$$\text{LoopStrength} = \max_i (d_i - b_i), \quad \text{where } (b_i, d_i) \in H_1$$

This value is high[1] when the point cloud exhibits a prominent loop, and low when the data is either scattered or primarily linear. It serves as a topologically grounded proxy for the strength of seasonal cyclicity in the given year.

[1] We chose the maximum persistence rather than, for example, the sum or count of H_1 features, to emphasize the presence of a *dominant*, structurally significant loop. In empirical tests, this yielded more stable and interpretable results, particularly in distinguishing cyclical years from those with noisy or fragmented topology.

4.2 Regression Framework

While loop strength is defined geometrically, we hypothesize that it can be predicted directly from raw temperature values—specifically, the 12-month temperature profile for each city-year. To test this, we formulate a supervised regression problem where:

- The input features $X \in \mathbb{R}^{12}$ are the monthly temperature values: $[T_{Jan}, \ldots, T_{Dec}]$
- The target $y \in \mathbb{R}$ is the computed loop strength

We train a **Random Forest Regressor** to model this mapping. Random Forests are ensemble-based, nonparametric models well-suited to capturing nonlinear dependencies and handling noisy data. The dataset is randomly split into 80% training and 20% testing subsets. Performance is evaluated using standard regression metrics including mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2 score).

4.3 Rationale and Downstream Benefits

Predicting loop strength from temperature vectors serves two purposes. First, it provides an interpretable model of how seasonality manifests in raw temperature data across diverse cities and years. Second, it enables the use of loop strength as a derived feature in downstream tasks without requiring topological computation at inference time. For example, one could use predicted loop strength in time-series segmentation, anomaly detection, or climate change impact analysis.

More broadly, this approach demonstrates that topological structure in time series can be meaningfully learned—and even anticipated—using standard supervised learning techniques. It opens the door to hybrid pipelines that combine geometric insight with scalable statistical modeling.

5 Results

5.1 Distribution of Loop Strengths

We begin by examining the distribution of loop strength values across all valid city-year samples from 1900 to the present. Figure 1 shows a histogram of maximum persistent H_1 lifespans—our scalar measure of topological cyclicity.

The distribution is highly skewed, with a large number of samples showing loop strength near zero. This suggests that many city-year temperature profiles do not exhibit strong cyclicity—either due to climatic irregularity, geographic factors (e.g., equatorial regions), or data noise. However, the long right tail indicates the presence of city-years with pronounced seasonal structure, forming strong and persistent loops in topological space. These high-loop-strength instances likely correspond to temperate regions with consistent seasonal transitions.

This distribution highlights a key advantage of persistent homology: it does not assume that cyclic behavior is universal, but rather quantifies its presence and strength on a per-sample basis.

5.2 Predicting Loop Strength from Temperature Profiles

We next evaluate whether loop strength—a topological feature extracted from 3D embeddings—can be predicted directly from raw temperature values. Using each city-year's 12-month temperature vector as input, we trained a Random Forest Regressor to predict the corresponding loop strength.

The model achieved strong predictive performance on a held-out test set:

- Mean Absolute Error (MAE): 0.1770

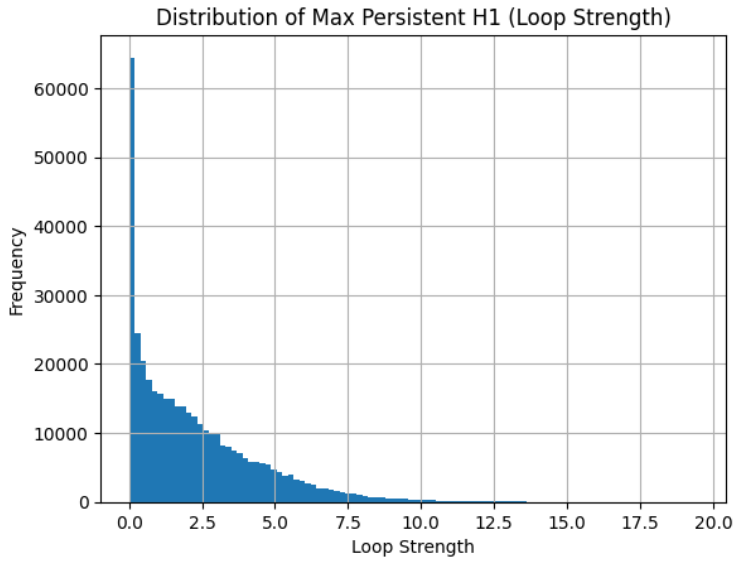

Figure 1: Distribution of loop strength across all city-year samples. The values are computed as the maximum H_1 persistence in a 3D temperature embedding for each year. Most samples exhibit low to moderate loop strength, though a long tail of high values suggests years with strong seasonal structure.

- Root Mean Squared Error (RMSE): 0.4107
- Median Absolute Error: 0.0272
- Coefficient of Determination (R^2): 0.9653
- Explained Variance Score: 0.9653

These results indicate that the geometry of seasonal structure is highly learnable from raw temperatures. The extremely low median error suggests that most samples are predicted with near-exact accuracy, while the R^2 and explained variance values show that over 96% of the variability in loop strength can be explained by monthly temperature alone.

Together, these findings validate our hypothesis: topological cyclicity, though computed geometrically, is strongly correlated with and predictable from temporal structure in temperature data. This also means that loop strength can serve as a tractable, learnable proxy for climatic seasonality in settings where persistent homology computations may be impractical in real time.

5.3 Residual Distribution Analysis

To further analyze the model's performance, we examine the distribution of residuals—defined as the difference between predicted and true loop strength values. Figure 2 shows the histogram of residuals on a logarithmic y-axis.

The residual distribution is sharply centered around zero, with a heavy central spike capturing the vast majority of predictions. The log-scale view reveals a small number of outliers on both ends of the error range, suggesting rare but significant deviations. These likely correspond to city-years with atypical temperature patterns—such as sharp transitions, flat curves, or irregular seasonal behavior—that confound the model's assumptions.

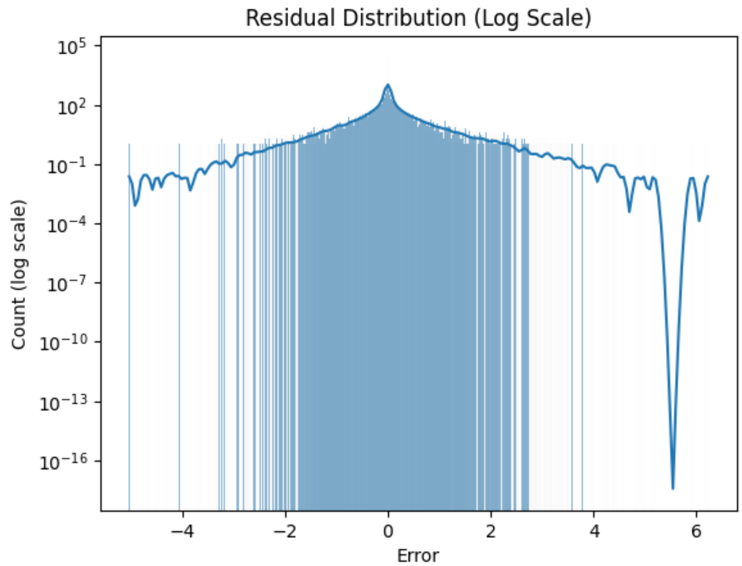

Figure 2: Distribution of residuals (prediction error) for loop strength on a log scale. The sharp peak around zero indicates high accuracy for most predictions, while long but sparse tails reflect a small number of large-magnitude errors.

The overall pattern confirms the model's robustness: while most predictions are highly accurate, the presence of a few long-tailed errors is consistent with real-world climate variability and the limits of learning topological features from raw inputs.

5.4 True vs. Predicted Loop Strength

Figure 3 shows a scatter plot of predicted versus true loop strength values on the test set. Each point corresponds to a single city-year. The red dashed line represents the ideal $y = x$ line, where predictions perfectly match actual values.

The figure demonstrates strong agreement between predicted and actual loop strengths, especially in the mid-range. The scatter is tightly centered around the identity line, indicating the model captures the overall structure of loop strength well. A slight tendency to underpredict at higher values is visible, which may reflect the relative rarity—and potentially greater geometric complexity—of extremely high loop strength instances.

This visual result reinforces the model's ability to approximate a geometric feature using only raw temperature profiles, highlighting both the expressiveness of temperature seasonality and the practical predictability of topological structure.

5.5 Feature Importance Across Months

To better understand what drives loop strength predictions, we examine the feature importances learned by the Random Forest model. Figure 4 shows the relative importance of each month's average temperature in determining loop strength.

The results reveal a striking pattern: December is by far the most important month, followed by November. This suggests that loop strength is highly sensitive to how the year ends, potentially due to the way persistent homology captures closed trajectories in the embedded temperature space.

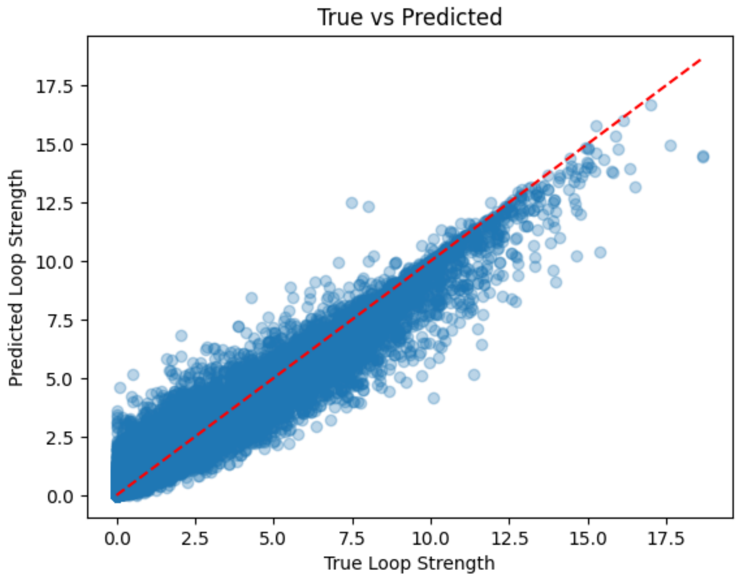

Figure 3: Predicted versus true loop strength for city-year samples in the test set. The red dashed line indicates the ideal $y = x$ relationship. The tight clustering along this line demonstrates high model fidelity and accurate estimation of topological cyclicity.

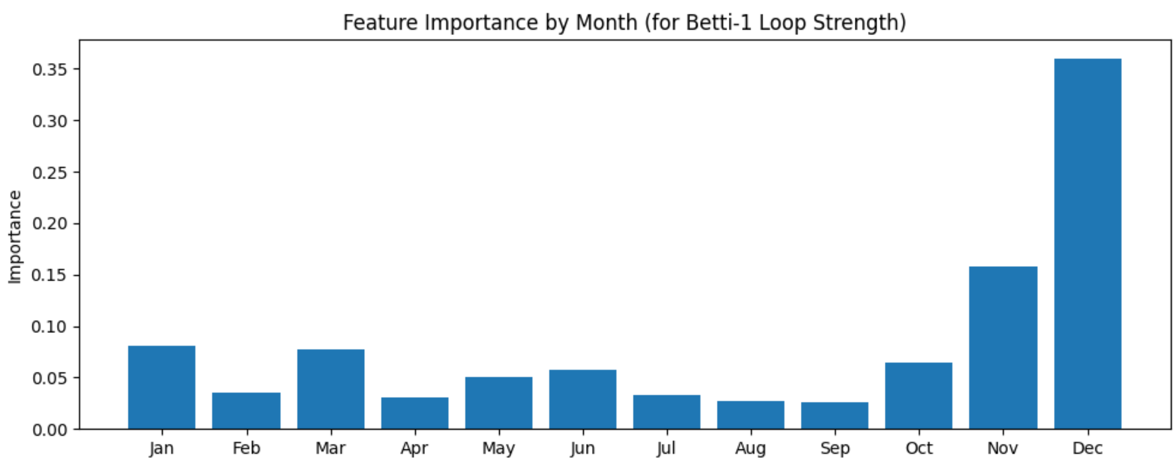

Figure 4: Feature importance by month in predicting Betti-1 loop strength. December, followed by November, emerges as the most influential input, suggesting their role in completing seasonal cycles.

If seasonal patterns begin early in the year and return near their starting point by December, a strong topological loop is more likely to form.

In contrast, mid-year months (e.g., July to September) appear less informative. This aligns with the idea that the beginning and end of the annual cycle anchor the geometric structure, and strong loops arise when there is a full "return" to initial seasonal states. In effect, December helps to "close the loop"—a phenomenon the model has learned as predictive of high Betti-1 persistence.

This also offers a valuable perspective for downstream applications: interventions or disruptions occurring late in the year may have an outsized impact on whether a clean seasonal cycle is preserved or broken.

5.6 Clustering Cities by Loop Strength Patterns

To explore geographic patterns in topological seasonality, we aggregated loop strength values over time for each city. We constructed a pivoted matrix where rows correspond to cities, columns to years, and values to loop strength. To ensure sufficient historical coverage, we retained only cities with data for at least 10 years, filling missing values with zero.

Using this city-by-year loop strength matrix, we applied unsupervised clustering (e.g., KMeans) to identify groups of cities with similar long-term topological patterns. We found that cities naturally cluster into distinct groups—some characterized by strong seasonal recurrence, others by weaker or more irregular patterns.

- **Cluster 0** includes cities like *Liège (Belgium)* and *Van (Turkey)*—locations with moderate, continental seasonal structure.
- **Cluster 1** contains cities such as *Sawangan (Indonesia)* and *Tunis (Tunisia)*—generally tropical or arid climates where seasonality may be present but less geometrically sharp.
- **Cluster 2** includes cities like *Mazatlán (Mexico)* and *Mianyang (China)*, where loop strength may reflect monsoon or subtropical rhythms.
- **Cluster 3**, featuring cities like *Madison (USA)* and *Villeurbanne (France)*, shows pronounced cyclicity, consistent with temperate climates.
- **Cluster 4** includes a more mixed group, such as *Kabankalan (Philippines)* and *Saarbrücken (Germany)*, potentially capturing transition zones or cities with complex seasonal signals.

These clusters reveal how persistent homology can extract high-level climatic fingerprints from raw temperature data. Such typologies could be useful for downstream applications, including regional climate comparison, anomaly detection, or even long-term urban climate planning. Future work could investigate how these loop strength clusters align with traditional climate classifications or with ongoing shifts due to climate change.

6 Conclusion and Future Work

This paper proposed a novel approach to understanding seasonality in time series data through the lens of topological data analysis (TDA). Rather than assuming cycles exist or enforcing parametric seasonality, we used persistent homology to measure the *actual presence and strength of cyclic structure* in empirical temperature time series.

By embedding short overlapping temperature windows into $\mathbb{R}^3$ and extracting Betti-1 features from the resulting point clouds, we defined a scalar summary—**loop strength**—that reflects how strongly a given year's temperature profile forms a closed, recurring trajectory. This measure is purely geometric, robust to noise, and agnostic to calendar conventions, offering a new way to quantify and compare seasonal dynamics.

Our analysis demonstrated several key insights:

- Loop strength is **heavily skewed** across the global dataset: while many city-years exhibit weak or disrupted seasonal loops, a substantial tail of high-cyclicity years persists.
- A **Random Forest model** using only monthly temperature values can predict loop strength with remarkable accuracy ($R^2 = 0.9653$), revealing that the presence of geometric cyclicity is tightly tied to temperature structure, especially at the beginning and end of the year.
- **December** emerged as the most influential feature, hinting at the importance of closure and return in loop formation—a finding that aligns naturally with the geometric intuition behind Betti-1.
- When aggregated over time, cities could be clustered by their long-term loop strength patterns, suggesting a form of **topological climate signature** with potential applications in regional classification and long-term climate trend analysis.

Implications and Broader Potential

Topological features like loops are generally underutilized in applied time series work, particularly in environmental and geophysical domains. This study shows not only that such features are *detectable*, but that they can be *predicted, modeled, clustered*, and potentially monitored at scale.

Recent studies have explored topological approaches to time series classification [15], delay-coordinate embeddings [13], and topological analysis of extreme events in high-dimensional data [6]. These results suggest a growing interest in learning from the geometric and structural aspects of temporal data.

Toward Financial and Economic Applications

While we focused on weather data, the core framework is broadly applicable. Many systems—in finance, economics, health, and biology—exhibit latent cyclicity that may be intermittent, noisy, or nonlinear. Traditional tools (e.g., Fourier transforms, ARIMA) often fail to capture this richness.

In future work, we will apply this topological framework to financial time series, beginning with commodity prices. Commodities such as oil, wheat, or metals frequently exhibit structural cycles tied to supply chains, seasons, macroeconomic shocks, or policy dynamics. Persistent homology has already shown promise in detecting structural transitions in asset dynamics and crash regimes [5], offering complementary insight to classical volatility and trend indicators.

We aim to:

- Use loop strength as a proxy for regime detection or market structure
- Build trading signals based on changes in topological features over time
- Explore "loop-aware" forecasting models that incorporate geometry as a prior

This opens the door to an emerging class of methods we might call **topology-aware forecasting**, where the shape of the data—not just its values—guides learning and decision-making.

Methodological Extensions

Several methodological avenues are also promising:

- Investigating persistent entropy and lifetime distributions rather than just max-persistence
- Incorporating multi-scale homology (e.g., both Betti-0 and Betti-1) to capture richer structure
- Embedding topological penalties or priors into neural models (e.g., VAEs or transformers) [8]
- Using Wasserstein or bottleneck distances between persistence diagrams as inputs to learned pipelines

Closing Thought

This paper illustrates that topological signals exist in the wild—and with the right tools, we can extract, interpret, and even predict them. Whether in climate, commodities, or cognition, learning from the shape of time may help us understand not just what happened, but what kind of world produced it.

References

[1] Julianna Berwald, James Sheehan, Katie Livingston, and Jose A Perea. Revisiting time series with topological data analysis. *Entropy*, 23(2):242, 2021.

[2] Vin De Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. In *Algebraic Topology and Geometric Applications*, pages 65–94. American Mathematical Society, 2011.

[3] Herbert Edelsbrunner and John L Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[4] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[5] Marian Gidea and Bryan Goldsmith. Topological data analysis of high-dimensional time series: The case of foreign exchange markets. *Physica A: Statistical Mechanics and its Applications*, 537:122701, 2020.

[6] Marian Gidea and Yossi Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834, 2018.

[7] Chad Giusti, Robert Ghrist, and Danielle S Bassett. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.

[8] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Learning with persistence homology. *Journal of Machine Learning Research*, 20(55):1–45, 2019.

[9] Xianfeng David Li and Yusu Wang. Persistence-based structural recognition in 3d protein models. *BMC Bioinformatics*, 15(1):1–12, 2014.

[10] George Muszynski, Michael Kirby, Ann Pankajakshan, Amanda Tandarich, and Elizabeth Barnes. Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development*, 12(2):613–628, 2019.

[11] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[12] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):1–38, 2017.

[13] Jose A Perea. Topological time series analysis. In *Proceedings of the 2016 ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 579–583, 2016.

[14] Jose A Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838, 2015.

[15] Lee M Seversky, Steven Davis, and Matthew Berger. Time-series classification using topological data analysis. *Machine Learning*, 107(7):1387–1402, 2016.

[16] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.