Sea ice extent data analysis using statistical and unsupervised learning methods*

Vitalii Alekseiko[1,*,†], Vitaly Levashenko[2,†] , Yurii Voichur[1,†] and Dmytro Medzatyi[1,†]

[1] *Khmelnytskyi National University, Institutska str., 11, Khmelnytskyi, 29016, Ukraine*
[2] *Zilina University, Univerzitná 8215, 010 26 Žilina, Slovakia*

Abstract
The article analyzes the main trends in the change in Sea Ice extent in the Arctic and Antarctic regions. The impact of Sea Ice extent on climate change and ecosystems in both polar regions and more remote regions of the Earth is analyzed. A statistical analysis of historical data is performed and the Time Series are decomposed into the main components: trends, seasonality, and residuals. To determine the stationarity of the numerical series and test the hypotheses, the Dickey-Fuller test is performed. It is used machine learning methods under unsupervised learning to provide clustering for better understandings of sea ice extent patterns and anomalies in data. The direction of further research is outlined based on the results obtained.

Keywords
Sea Ice extent, time series, data analysys, climate change, trends

1. Introduction

Sea Ice extent in the Arctic and Antarctic is one of the seven main climate indicators identified by the World Meteorological Organization (WMO) [1]. This indicator has a significant impact on the ecosystems of the Arctic and Antarctic regions, as well as on other regions of the globe, and plays a significant role in climate change. Determining the main trends and patterns of changes in sea ice extent is extremely relevant in the context of understanding potential risks not only for the polar regions, but also for the entire biosphere. Studying the time series of sea ice extent changes provides extremely valuable information for choosing techniques and models for predicting future indicators.

Arctic and Antarctic ice extent have shown stable trends for decades, but today's observations indicate abnormally low levels [2].

Although nowadays there are many approaches to forecasting, it should be noted that an extremely important aspect is data preparation, which allows to significantly increase the accuracy of the forecast [3]. Thus, there is a need to conduct statistical analysis of data and their pre-processing and preparation for forecasting.

This approach is relevant in the field of machine learning for various forecasting tasks in the fields of economics [4], finance, climate, etc.

2. Literature review

Previous works have examined the impact of surface temperature on climate change, approaches to forecasting temperature time series [5, 6], the impact of such forecasting in the context of developing

IntelITSIS'2025: 6th International Workshop on Intelligent Information Technologies and Systems of Information Security, April 4, 2025, Khmelnytskyi, Ukraine
* Corresponding author.
† These authors contributed equally.
vitalii.alekseiko@gmail.com (V. Alekseiko); vitaly.levashenko@fri.uniza.sk (V. Levashenko); voichury@khmnu.edu.ua (Yu. Voichur); medza@ukr.net (D. Medzatyi)
0000-0003-1562-9154 (V. Alekseiko); 0000-0003-1932-3603 (V. Levashenko); 0000-0003-3085-7315 (Yu. Voichur); 0000-0002-1879-2945 (D. Medzatyi)

sustainable cities and communities [7], and the specifics of developing predictive information systems [8, 9].

This study focuses on the study of time series of sea ice extent. Although the studies share some common patterns, such as seasonality, the specificities of polar regions necessitate a comprehensive analysis that takes into account the climatic characteristics of the regions under study.

The conducted literature review confirms the relevance of the chosen research topic. Nowadays, the main directions of scientific research are the study of the Arctic and Antarctic regions and the impact of changes occurring there on climate change in different regions of the world.

Changes in the ice cover of the Arctic and Antarctica affect the Sea level [2], precipitation [10, 11, 12, 13], coastal wave heights [14] and many other indicators, turning entire regions into areas vulnerable to climate change [15]. Sea Ice also has a significant impact on flora and fauna [16, 17, 18].

The article [19] establishes a correlation between historical trends in the area of sea ice in different seas of the Arctic Circle by analyzing the impact of temperature changes in the Arctic Circle and the average annual concentration of global CO_2 on them. An important aspect of the study of the Arctic region is the analysis of the greening of the area, which demonstrates growth trends [20].

Also, when studying climate data, much attention is paid to the analysis of time series, in particular the study of seasonal patterns.

The work [21] considers seasonal trends of the early twentieth century and conducts a detailed statistical analysis of the data. The study [22] proposes a model that uses the theory of time series decomposition to predict temperature in Chinese cities.

In a number of studies, the Dickey-Fuller test is used to check the stationarity of climate data. In particular, when studying climate trends for the Indian tropical river basin [23], changes in non-stationary extreme precipitation due to climate change in East Malaysia [24], and the incidence of dengue fever due to climate change in Singapore [25].

3. Methodology

3.1. Dataset structure

It was used Daily Sea Ice extent Data dataset from Kaggle [26]. All information provided by Data and image archive of National Snow and Ice Data Center [27]. The calculation of the Sea Ice extent is based on the accepted threshold of the frozen area of 15% [28]. Technologies of the National Aeronautics and Space Administration (NASA) as satellite images analysis are used for data collection [29]. Dataset includes 7 variables:

- Year;
- Month;
- Day;
- Extent (unit is 10^6 sq km);
- Missing (unit is 10^6 sq km);
- Source: Source data product web site [27];
- Hemisphere.

Current dataset is limited by data from October 26, 1978 to May 31, 2019. But all the more recent information is presented by National Snow and Ice Data Center [27].

3.2. Time series data analysis

A number of studies are conducted to analyze Time Series Data. To determine general trends, it is advisable to aggregate the data by calculating the average monthly temperature for each month of each year. Then graphs are plotted and a linear approximation is performed. Next, the angle between the resulting line and the time axis is calculated. The result obtained indicates general trends in the

Sea Ice Extent. A negative result indicates a tendency for it to decrease, a positive result indicates an increase, and a result close to zero indicates no significant changes. Also, for a more comprehensive understanding of the changes and the selection of forecasting techniques [30], it is necessary to decompose the time series into three main components: Trend, Seasonality, and Residuals.

3.2.1. Trend

Trend represents a long-term movement in data. Thus, it may increase, decrease, or remain constant over time. A trend captures a systematic change in data. The important factor is that such changes are not caused by short-term fluctuations or periodic patterns.

In a time series, the trend component is often modeled as a function of time (e.g., linear, quadratic, exponential, etc.).

A linear trend can be represented as:

$$T(t) = \beta_0 + \beta_1 t, \tag{1}$$

where: t – time; T(t) – the value of the trend at time t; β_0 – the intercept (the value of the trend at t=0); β_1 – the slope (indicating the rate of change of the trend over time).

Also there are non-linear trends as polynomial (2) or exponential (3). The exponential form captures the rapid growth or decay.

$$T(t) = \beta_0 + \beta_1 t + \beta_2 t^2, \tag{2}$$

where: β_2 – quadratic coefficient. It captures the curvature or acceleration of the trend over time.

$$T(t) = \alpha e^{\beta t}, \tag{3}$$

where: α and β are constants.

3.2.2. Seasonality

The seasonality component refers to periodic fluctuations in a time series. In most cases, such intervals are fixed. Thus, cycles with regularly repeating intervals can be modeled using periodic patterns, in particular sinusoidal functions.

A simple sinusoidal seasonal model can be represented as:

$$S(t) = \gamma_1 \sin(2\pi f t + \phi_1) + \gamma_2 \cos(2\pi f t + \phi_2), \tag{4}$$

where: S(t) – seasonal component at time t; γ_1 – amplitude of the sine term; γ_2 – amplitude of the cosine term; f – frequency of the seasonality; ϕ_1 – phase shift for the sine component; ϕ_2 – phase shift for the cosine component.

3.2.3. Residuals

Residuals, also known as “noise”, represent the remainder of the time series after trend and seasonality have been subtracted. In effect, they represent the random component of the time series that cannot be explained by trend or seasonal patterns.

$$R(t) = Y(t) - T(t) - S(t), \tag{5}$$

where: R(T) – residual (or noise) component at time t; Y(T) – observed value of the time series at time t;

Residuals are usually considered to be white noise. This means that they should have the following properties:

- Zero mean: the mean of the residuals is zero.
- Constant variance: the variance of the residuals is constant over time (homoscedasticity).

- No autocorrelation: the residuals are independent, meaning that there is no significant correlation between them at different time lags.

Thus, residuals R(t) can be described by the formula:

$$R(t) = N(0, \sigma^2), \tag{6}$$

where: $N(0, \sigma^2)$ represents a normal distribution with zero mean and variance σ^2. If the residuals exhibit autocorrelation, this means that there is some structure missing in the model and it may need to be improved.

3.3. Dickey-fuller test

The Dickey-Fuller test is a statistical test that is used to determine the stationarity of a time series by testing the null hypothesis of the presence of a unit root in an autoregressive time series model.

The Dickey-Fuller test examines the time series y_t using the following autoregressive (AR) model of order 1:

$$y_t = \phi y_{t-1} + \epsilon_t, \tag{7}$$

where: y_t – the value of the time series at time t; y_{t-1} – the value of the time series at time t − 1; ϕ – the coefficient of the lagged variable; ϵ_t – a white noise error term (independently and identically distributed with mean zero).

To perform Dickey-Fuller test it is determined Null (H_0) and Alternative (H_A) Hypotheses:

- Null Hypothesis (H_0) : $\phi = 1$ (The series has a unit root and is non-stationary).
- Alternative Hypothesis (H_A) : $\phi < 1$ (The series does not have a unit root and is stationary).

The Dickey-Fuller test is based on a transformed version of the AR(1) model [31]:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \epsilon_t, \tag{8}$$

where: $\Delta y_t = y_t - y_{t-1}$ – the first difference of the series; α – a constant (optional) to account for a non-zero mean; βt – a trend term (optional) to account for a deterministic trend; $\gamma = \phi - 1$ – the coefficient that measures the presence of a unit root.

Thus, the null and alternative hypotheses are now expressed in terms of γ:

- $H_0 : \gamma = 0$ ($\phi = 1$).
- $H_A : \gamma = 0$ ($\phi < 1$).

The test statistic can be calculated as:

$$\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}, \tag{9}$$

where: $\hat{\gamma}$ – the estimated value of γ; $SE(\hat{\gamma})$ – the standard error of $\hat{\gamma}$.

This statistic is compared to critical values derived from the Dickey-Fuller distribution.

If τ value is less than the critical value, then the series is stationary and H_0 Hypothesis rejected.

If τ value is greater than the critical value, then the series is non-stationary and H0 Hypothesis failed to reject.

The augmented Dickey-Fuller (ADF) test is an extension of this test that includes higher-order lagged differences of y_t to account for autocorrelation in the error terms.

3.4. Data analysis by machine learning techniques

Since statistical analysis does not allow for a holistic understanding of processes, due to its limitations, in particular the inability to capture hidden patterns, it was decided to conduct data analysis using machine learning methods under unsupervised learning [32, 33], in particular K-means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Hierarchical Clustering. Using these methods will provide more information on trends in Sea Ice extent and existing anomalies in the data [34].

3.4.1. K-Means clustering

The K-Means algorithm divides n data points into k clusters, minimizing the variance in each cluster [35].

The first step of this method is initialization, that is initial cluster centroids μ_i are randomly selected. After that, each data point x_i is assigned to the nearest centroid based on the Euclidean distance:

$$C_j = \{x_i : \|x_i - \mu_i\|^2 \leq \|x_i - \mu_m\|^2, \forall m\}, \quad (10)$$

where: C_j – the set of points assigned to cluster j.

Update step is represent calculation of new centroid for each cluster:

$$\mu_i = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i, \quad (11)$$

where $|C_j|$ – the number of points in cluster j.

The convergence criterion is achieved when the centroids no longer change significantly or a fixed number of iterations is reached.

3.4.2. Density-based spatial clustering of applications with noise

DBSCAN performs clustering of points based on density, so the method is robust for data that forms irregular shapes [36].

A point p is a core point if at least minimum number of points required to form a dense region (MinPts) points (including itself) exist within ε-radius.

$$|N_\varepsilon(p)| \geq \text{MinPts}, \quad (12)$$

where: ε – maximum neighborhood radius.

$N_\varepsilon(p)$ is determined as:

$$N_\varepsilon(p) = \{q \in D : \|q - p\| \leq \varepsilon\} \geq \text{MinPts}, \quad (13)$$

where: q and p – points.

If p is a core point and expression (14) is true, A point q is directly density-reachable from p.

$$q \in N_\varepsilon(p) \quad (14)$$

The point r is a density-reachable from p, provided that there exists a chain of directly density-reachable points.

The algorithm of the method consists in choosing an unvisited point and checking whether it is a core point. If the point is a core point, a new cluster is created by expanding its density-reachable point. If the point is not a core point, it is marked as noise. The algorithm is repeated until all points are visited.

3.4.3. Hierarchical clustering

Hierarchical clustering creates a hierarchy of clusters using either an agglomeration (bottom-up) or a divisive (top-down) approach [37].

Agglomerative hierarchical clustering (bottom-up) is based on a distance matrix. The matrix is formed by calculating the pairwise distances between all points using the Euclidean distance:

$$d(x_i, x_j) = \|x_i - x_j\|, \quad (15)$$

where: x_i and x_j – points.

Cluster merging provides by defining each point as a separate cluster. Then, the two closest clusters are iteratively merged based on the connectivity criteria: Single linkage (16), Complete linkage (17), Average linkage (18).

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y), \tag{16}$$

where: x and y – points of different clusters.

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y), \tag{17}$$

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y) \tag{18}$$

The dendrogram is built by merging until one cluster remains. Divisive hierarchical clustering (top-down) is performed starting with a single cluster that contains all points. The clusters are split recursively until each point becomes a separate cluster.

4. Results and discussions

To assess the fluctuations in Sea Ice Extent, the data were aggregated by calculating monthly averages (Fig. 1, 2). It is clear from the chart that for the Northern Hemisphere (Fig. 1) the minimum is observed in September, and the maximum in March, or sometimes in February. At the same time, for the Southern Hemisphere (Fig. 2), the sea ice extent reaches its maximum values in September, and its minimum values in February.

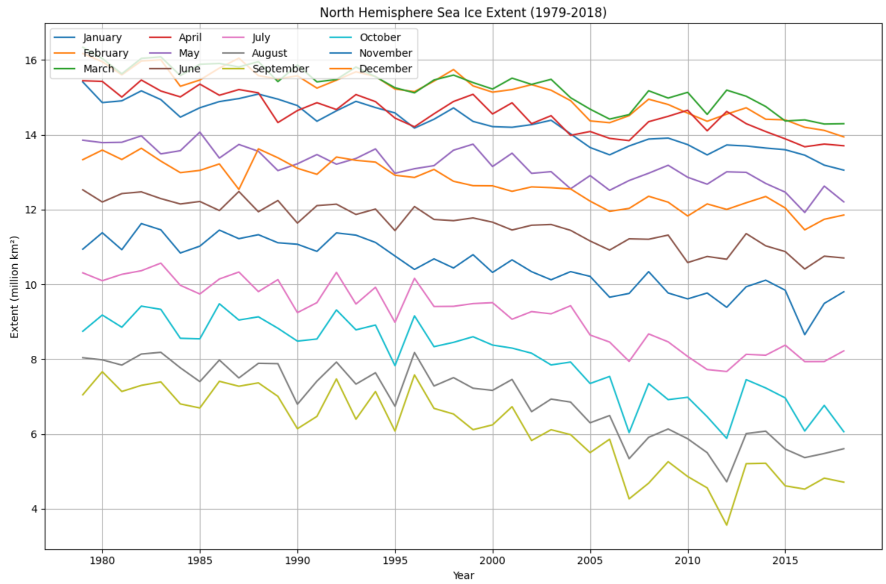

Figure 1: Sea Ice Extent for each month by years in North Hemisphere.

To consider the problem of sea ice extent changes in more detail, we will perform an approximation and construct straight lines that will demonstrate the main trends (Fig. 3, 4). The calculated angles between the resulting straight lines and the time axis are given in Table 2.

Table 1 shows the maximum, minimum and average values of sea ice extent for each month, indicating the year when the highest and lowest values were recorded for the period from October 1978 to May 2019. Statistical data show that the maximums in the Northern Hemisphere were recorded only in the 20th century, while the minimums cover the period from 2016 to 2019. For the Southern Hemisphere, the situation is not so critical.

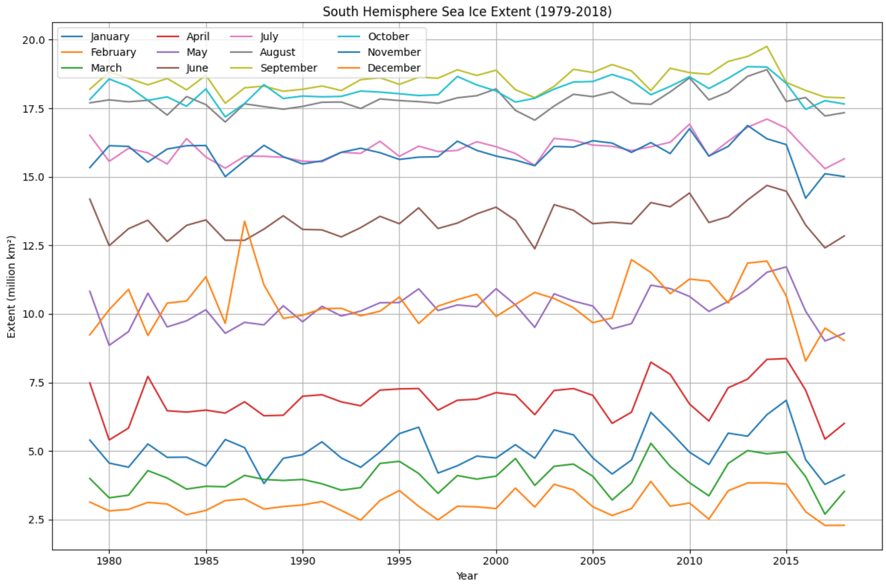

Figure 2: Sea Ice Extent for each month by years in South Hemisphere.

Table 1
Sea Ice Extent Statistics (1978 – 2019)

Month	Sea Ice Extent in North Hemisphere (10^6 sq km)			Sea Ice Extent in South Hemisphere (10^6 sq km)		
	Minimum (Year)	Maximum (Year)	Average	Minimum (Year)	Maximum (Year)	Average
January	12.491 (2018)	15.912 (1979)	14.18	2.514 (2017)	9.393 (2015)	5.00
February	13.741 (2018)	16.579 (1979)	15.04	2.112 (2017)	4.52 (2014)	3.07
March	14.041 (2019)	16.635 (1979)	15.21	2.08 (2017)	6.481 (2014)	4.04
April	12.933 (2016)	15.945 (1982)	14.48	3.653 (2017)	10.164 (2015)	6.88
May	11.132 (2016)	14.786 (1982)	13.07	7.119 (2019)	12.979 (2014)	10.20
June	9.231 (2010)	13.091 (1982)	11.52	10.681 (1980)	16.306 (2014)	13.44
July	6.368 (2012)	11.562 (1979)	9.09	13.883 (2017)	18.16 (2014)	16.04
August	3.648 (2012)	9.224 (1983)	6.78	16.315 (1986)	19.234 (2014)	17.79
September	3.34 (2012)	8.211 (1996)	5.99	17.471 (1986)	20.201 (2014)	18.54
October	4.112 (2012)	10.615 (1982)	7.93	16.475 (1986)	19.67 (2014)	18.16
November	7.219 (2016)	12.695 (1982)	10.45	11.749 (2016)	18.273 (2013)	15.86
December	10.183 (2016)	14.585 (1978)	12.64	5.345 (2018)	14.879 (2010)	10.39

The results show that a significant decrease in sea ice extent is observed in the Northern Hemisphere, especially in September and October. No such trends are recorded in the Southern Hemisphere. The slope is always positive and close to zero. Thus, there has been a gradual, slight increase in sea ice extent.

To examine regional trends by month, the time series data were decomposed into trends, seasonality, and residuals. This decomposition plays a key role, as it helps to break down the underlying patterns and fluctuations in the data, as well as provide a clearer understanding of the process being analyzed.

Trends are used to predict future values based on previous data, which is extremely important when predicting climate indicators based on historical data.

Trends provide insight into future developments, allowing governments and organizations to plan and develop adaptation and mitigation strategies. Thus, if the trend indicates a rise in sea level due to melting ice, cities can prepare for potential flooding or inundation.

Residual analysis allows us to assess how well the model has captured the data. If the residuals show patterns (e.g., autocorrelation), this indicates that the model may need adjustment or additional features. If the residuals are large at a particular time, this may indicate an anomaly or event that was not captured by trend or seasonal patterns. For example, a sudden spike in sea ice extent due to extreme weather would be considered a residual anomaly. If the residuals show consistent patterns, the model may need to be refined to make more accurate predictions.

Linear trends of Sea Ice Extent for each month by years are shown in the Figure 3 for North Hemisphere and in the Figure 4 for South Hemisphere. Figures 3 and 4 show that the area of Sea Ice in the Arctic region has a decreasing trend, while in the Antarctic region, stability of indicators was observed from 1979 to 2019.

To determine the rate of change in sea ice extent, linear interpolation was performed and the angles of inclination to the abscissa axis were determined. The results of the calculations are given in Table 2.

Figure 5 shows the decomposition of the Time Series into the main components: trends, seasonality, and residuals for September in the North Hemisphere, when annual minimum of Sea Ice extent are observed. Random values of residuals, indicating that the decomposition has successfully isolated trend and seasonality.

Results of Dickey-Fuller test are presented in Table 3. The test statistic (−1.0269) is greater than all critical values (−3.4442, −2.8676, −2.5700), and the p-value (0.7433) is much greater than the typical significance level (0.05).

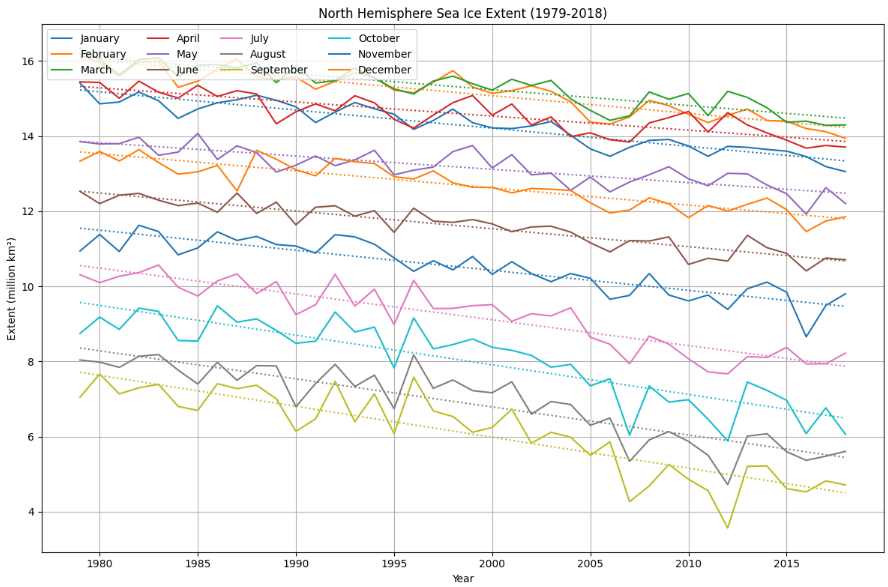

Figure 3: Linear trends of Sea Ice Extent for each month by years in North Hemisphere.

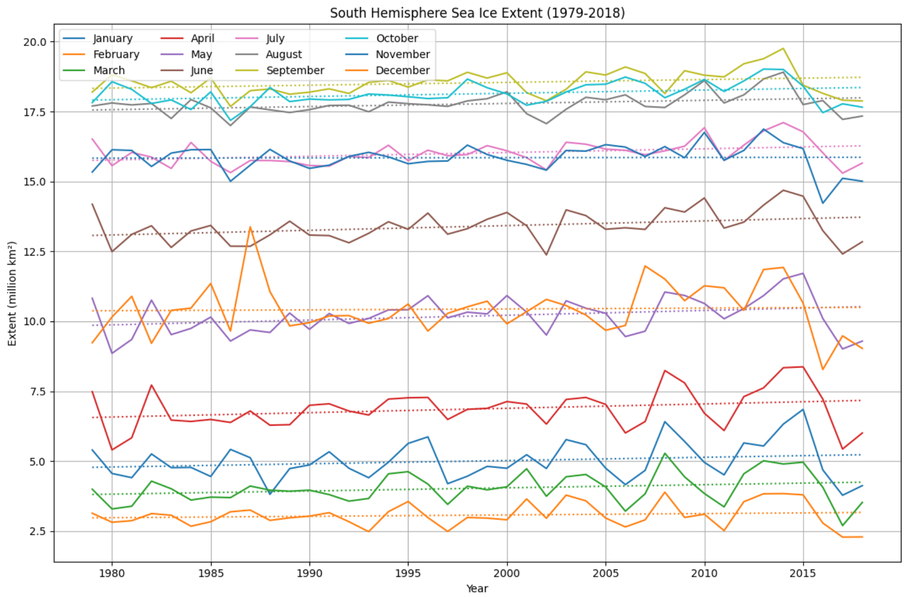

Figure 4: Linear trends of Sea Ice Extent for each month by years in South Hemisphere.

Table 2
Angles of common linear trends to time axis (1979 – 2018)

Month	Angle for North Hemisphere	Angle for South Hemisphere
January	−2.77°	0.66°
February	−2.69°	0.28°
March	−2.41°	0.65°
April	−2.15°	0.89°
May	−2.03°	0.98°
June	−2.72°	0.96°
July	−3.93°	0.76°
August	−4.27°	0.64°
September	−4.71°	0.58°
October	−4.51°	0.66°
November	−3.06°	0.04°
December	−2.61°	0.18°

As a result, it is failed to reject the null hypothesis (H_0) that the Time Series has a unit root. Thus, the Time Series for North Hemisphere Sea Ice extent is non-stationary. It means, that its statistical properties (mean, variance, autocorrelation) may change over time. Therefore, it is might be necessary to provide further transformation (e.g., differencing) to make Time Series stationary before forecasting. The test statistic (−3.4721) is less than the critical value at the 1% significance level (−3.4443), and the p-value (0.0087) is less than 0.05. Thus, the null hypothesis (H_0) is rejected. Therefore Time Series has a unit root. It means, that Time Series for South Hemisphere Sea Ice extent is stationary. Thus, its statistical properties do not change over time.

For better understanding patterns and detect anomalies in data it was performed clustering analysis using unsupervised learning techniques.

The result of K-Means clustering for the South Hemisphere shows three distinct clusters (Figure 6). Cluster 0 highlighted in blue demonstrates high values of Sea Ice extent from 12 to 16 million km². This values were dominant before 2000. Cluster 1 highlighted in green covers values between 6 and 11 million km². This cluster is intermediate and it indicates a gradual transition from high to lower extents. Cluster 2 highlighted in orange indicates low values of Sea Ice extent. They become dominant after the early 2000s. The transition from cluster 0 to 1 and then to 2 shows a steady decline in Sea Ice extent over time.

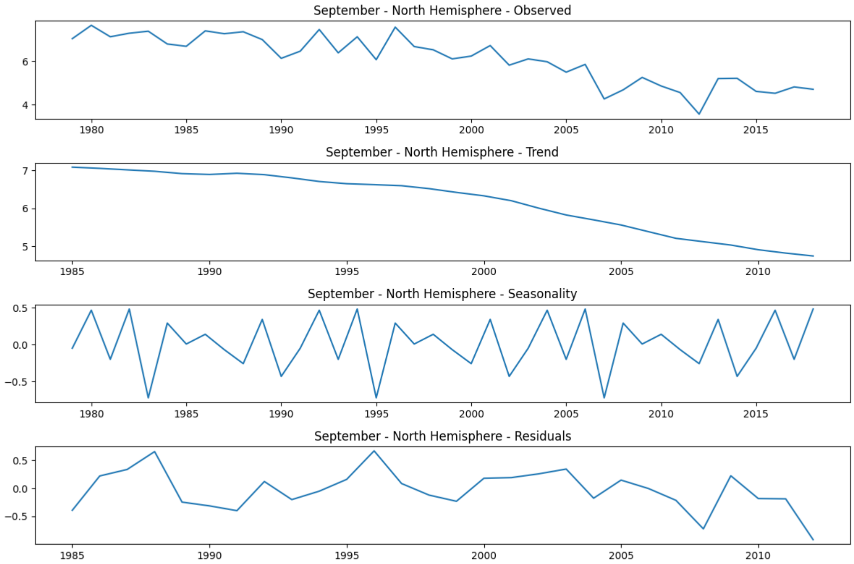

Figure 5: Decomposition of the time series data in North Hemisphere for September.

Table 3
Results of Dickey-Fuller Test

Variable		Value	
		North Hemisphere	South Hemisphere
Test Statistic		−1.026961129357658	−3.4720939348440467
p-value		0.7433042931743226	0.008733379945899898
Lags Used		12	15
Observations Used		475	472
Critical Values	1%	−3.4441920863262863	−3.444280551073031
	5%	−2.8676439813617147	−2.867682902679315
	10%	−2.570021186703601	−2.5700419306592934

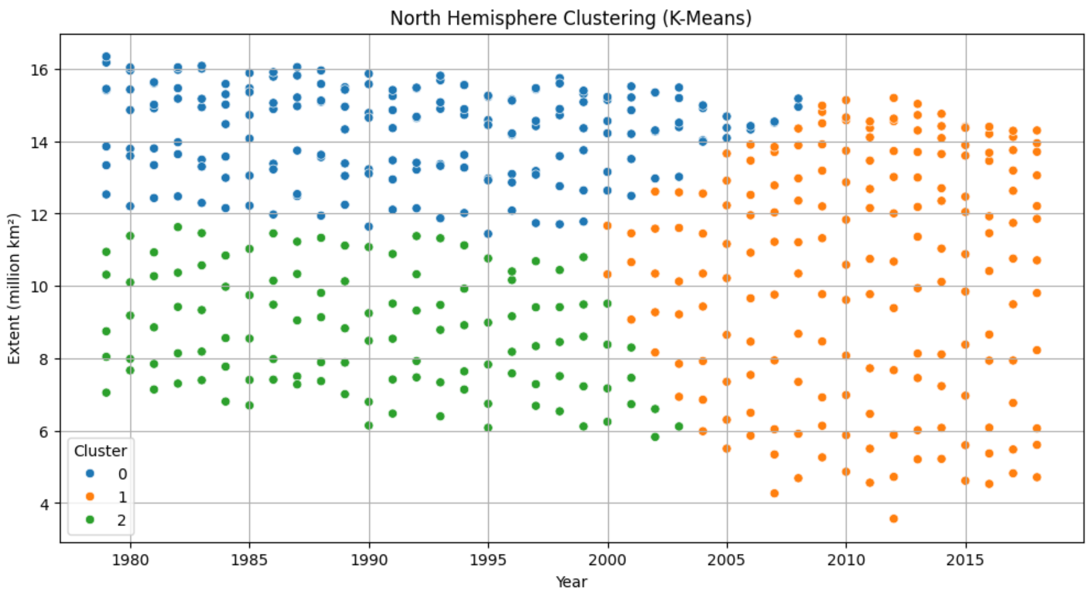

Figure 6: K-Means clustering result for the North Hemisphere.

In the Southern Hemisphere K-Means clustering result shows different patterns (Figure 7) compared to the North Hemisphere.

There are also three distinct clusters. Cluster 0 highlighted in blue demonstrates low Sea Ice extent. Cluster 1 highlighted in orange shows intermediate values, which become dominant after 2000. Cluster 2 highlighted in green represents high values, mostly before the 2000s, then declining over time.

Unlike the Arctic, where observed tendency to losing ice, situation in the Antarctic is more stable. In some periods trends are slightly increasing.

The transition from Cluster 2 to 1 suggests seasonal fluctuations but not a clear long-term decline like in the North.

The clustering suggests that Sea Ice extent fluctuations in the South Hemisphere are more complex and may have more hidden patterns.

The clustering performed using the DBSCAN method did not yield significant results. The method did not recognize differences, assigning all values to one cluster for North (Figure 8) and South (Figure 9) Hemispheres. This result indicates that no anomalous values were detected in the values of both hemispheres.

Hierarchical clustering for North Hemisphere (Figure 9) detects three clusters. There are low values in Cluster 0 highlighted in blue. It represents months with low values of Sea Ice extent among all years. Cluster 1 represents intermidiete values. As for months with the highest extent in each year, it is detected decreasing tendency. Cluster 2 highlighted in green demonstrates high values of Sea Ice extent up to the beginning 1990s. Transition from Cluster 2 to 1 shows main trend to decreasing the highest Sea Ice extent each year.

The results of hierarchical clustering for South Hemisphere are presented in Figure 10. It is detected three clusters. Cluster 0 highlighted in blue shows low values which demonstrates months with yearly minimum Sea Ice extent. Cluster 1 highlighted in orange represents values with high amplitude, which mostly where observed after middle of 1990s. Cluster 2 demonstrates values with lower amplitude in earlier years. Transition from cluster 2 to 1 shows increasing of amplitude and difference between months with high values of Sea Ice extent. Maximum values are increasing and intermediate values are decreasing. Thus, South Hemisphere has more complex patterns compare to North.

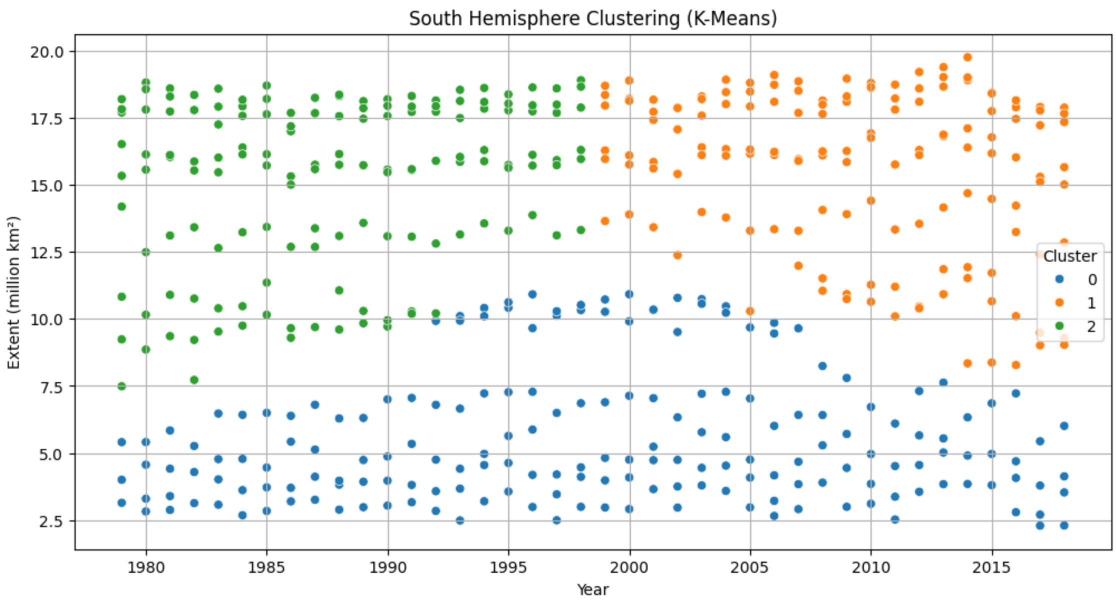

Figure 7: K-Means clustering result for the South Hemisphere.

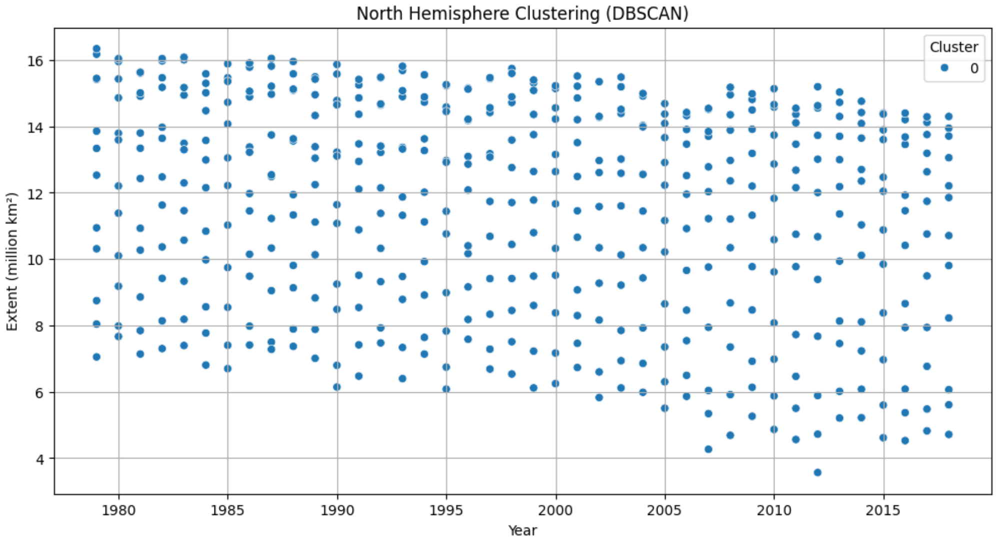

Figure 8: DBSCAN clustering result for the North Hemisphere.

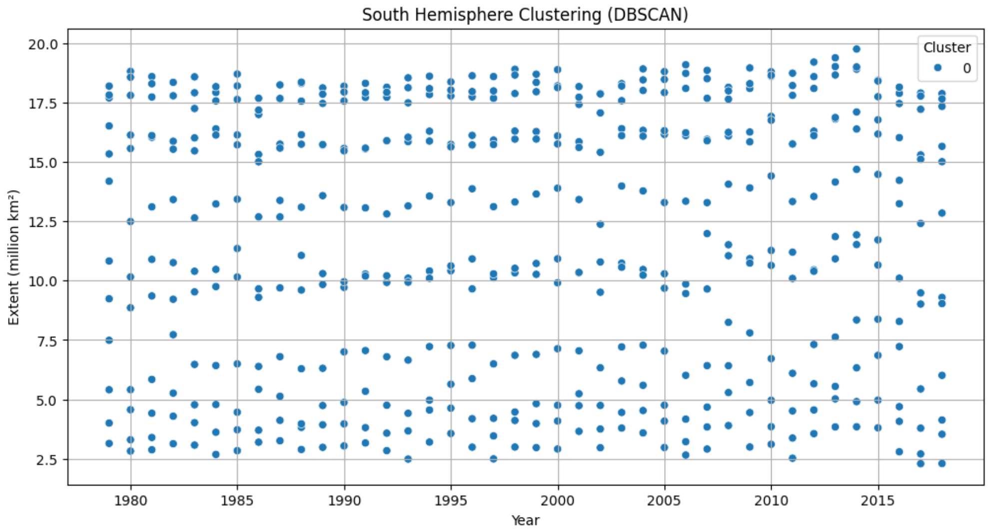

Figure 8: DBSCAN clustering result for the South Hemisphere.

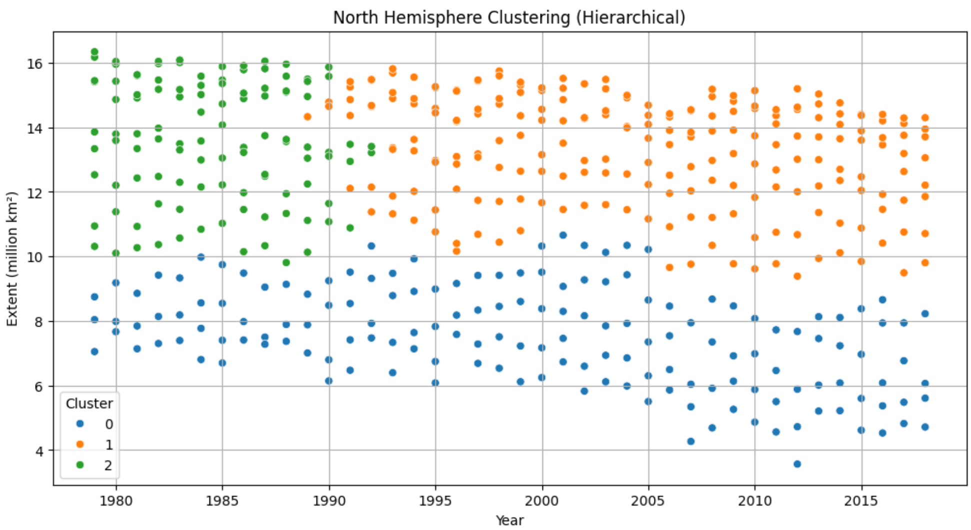

Figure 10: Hierarchical clustering result for the North Hemisphere.

Clustering analysis detects that Antarctic Sea Ice extent has complex patterns and potentially there are some hidden patterns. This region is pretend to be discovered more detailed. Arctic Sea Ice extent are decreasing more quickly each year and it is needed urgent action to avoid negative consequences.

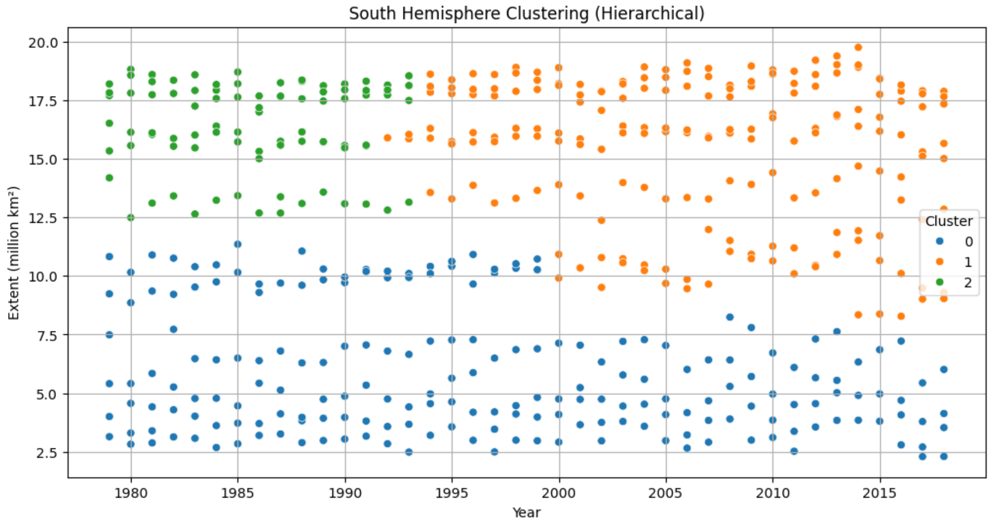

Figure 11: Hierarchical clustering result for the South Hemisphere.

5. Future work

Based on the obtained results, it is needed to provide transformations as differencing, logarithmic or seasonal decomposition for Sea Ice extent in North hemisphere. Then apply re-test to confirm stationarity. For the South Hemisphere series no further transformation is needed, and this data is ready for Time Series modeling and forecasting. Stationarity is a fundamental assumption in most Time Series models (e.g., ARIMA, Vector AutoRegression). Although, machine learning techniques can work with both stationary and non-stationary data, removing trends and seasonality can improve their forecast. Thus, in future work, it is advisable to conduct a comparative analysis of the forecasts of traditional models and machine learning models to determine the accuracy of the forecast.

6. Conclusions

Comprehensive statistical analysis of Time Series allows to identify the main trends of changes in indicators over time. Principal component decomposition (trends, seasonality, residuals) allows to identify the main trends, seasonal fluctuations and potential anomalies in the data. Checking the data for stationarity using the Dickey-Fuller test provides information on the need for further data analysis in preparation for forecasting. The study analyzes fluctuations in Sea Ice extent by identifying monthly trends and identifies the main problems in the Arctic and Antarctic regions. The calculations provide a basis for further research and forecasting of Sea Ice extent. Results of the clustering analysis shows rapidly changes in Sea Ice extent in North Hemisphere. Thus, Arctic region is vulnerable to climate change and needs urgent climate actions. Sea Ice extent in South Hemisphere demonstrate more stable situation. It is needed more long-term observation and more complex analysis to determine hidden patterns in Antarctic Sea Ice extent.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check; DeepL Translate in order to: some phrases translation into English. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] World Meteorological Organization. (2025). https://wmo.int.

[2] J. Turner, C. Holmes, T. C. Harrison, T. Phillips, B. Jena, T. Reeves-Francois, R. Fogt, E. R. Thomas, C. C. Bajish. (2022). Record low Antarctic Sea ice cover in February 2022. Geophysical Research Letters, 49(12). https://doi.org/10.1029/2022gl098904.

[3] N. R. Njeri. (2022). Data preparation for machine learning modelling. International Journal of Computer Applications Technology and Research, 11(06), 231–235. https://doi.org/10.7753/ijcatr1106.1008.

[4] P. Hryhoruk, S. Grygoruk, N. Khrushch, T. Hovorushchenko. (2020). Using non-metric multidimensional scaling for assessment of regions' economy in the context of their sustainable development. CEUR-WS 2713, 315-333.

[5] O. Pavlova, V. Alekseiko. (2024). The concept of an information system for forecasting the temperature regime of the Earth's surface based on machine learning. Computer Systems and Information Technologies, 2, 6–13. https://doi.org/10.31891/csit-2024-2-1.

[6] T. Hovorushchenko, V. Alekseiko, V. Levashenko. (2025). Machine learning methods' comparison for land surface temperatures forecasting due to climate classification. CEUR-WS, Vol. 3899, 55-68.

[7] T. Hovorushchenko, V. Alekseiko. (2024). Land surface temperature forecasting in the context of the development of sustainable cities and communities. Computer Systems and Information Technologies, 3, 6–13. https://doi.org/10.31891/csit-2024-3-1.

[8] V. Alekseiko. (2024). Web-based information system for land surface temperature forecasting using machine learning methods. Science and technology today, 10(38), 17–27. https://doi.org/10.52058/2786-6025-2024-10(38)-17-27.

[9] T. Hovorushchenko, O. Pavlova, V. Alekseiko, A. Kuzmin. (2024). Climate parameters monitoring in the context of natural disasters forecasting. In 2024 IEEE 14th International Conference on Dependable Systems, Services and Technologies, Athens, Greece, October 11-13, 2024.

[10] X. Chen, Z. Wen, Y. Song, Y. Guo. (2022). Causes of extreme 2020 Meiyu-Baiu rainfall: a study of combined effect of Indian Ocean and Arctic. Climate Dynamics, 59(11–12), 3485–3501. https://doi.org/10.1007/s00382-022-06279-0.

[11] M. Müller, T. Kelder, C. Palerme. (2022). Decline of sea-ice in the Greenland Sea intensifies extreme precipitation over Svalbard. Weather and Climate Extremes, 36, 100437. https://doi.org/10.1016/j.wace.2022.100437.

[12] S. Sundaram, D. M. Holland. (2022). A physical mechanism for the Indian summer Monsoon—Arctic Sea-Ice teleconnection. Atmosphere, 13(4), 566. https://doi.org/10.3390/atmos13040566.

[13] B. Zhou, M. Xu, B. Sun, T. Han, Z. Cheng. (2020). Possible role of Southern Hemispheric sea ice in the variability of West China autumn rain. Atmospheric Research, 249, 105329. https://doi.org/10.1016/j.atmosres.2020.105329.

[14] M. Henke, T. Miesse, A. De Souza De Lima, C. M. Ferreira, T. M. Ravens. (2024). Increasing coastal exposure to extreme wave events in the Alaskan Arctic as the open water season expands. Communications Earth & Environment, 5(1). https://doi.org/10.1038/s43247-024-01323-9.

[15] The cryosphere is the frozen water part of the Earth system. (2025). https://oceanservice.noaa.gov/facts/sea-ice-climate.html.

[16] P. T. Fretwell. (2024). A 6 year assessment of low sea-ice impacts on emperor penguins. Antarctic Science, 36(1), 3–5. https://doi.org/10.1017/S0954102024000130.

[17] J. R. Lee, M. J. Waterman, J. D. Shaw, D. M. Bergstrom, H. J. Lynch, D. H. Wall, S. A. Robinson. (2022). Islands in the ice: Potential impacts of habitat transformation on Antarctic biodiversity. Global Change Biology, 28(20), 5865–5880. https://doi.org/10.1111/gcb.16331.

[18] S. A. Robinson, L. E. Revell, R. Mackenzie, R. Ossola. (2024). Extended ozone depletion and reduced snow and ice cover – Consequences for Antarctic biota. Global Change Biology, 30(4).

[19] H. Cao, Y. Zhou, X. Jia, Y. Li. (2024). Analysis of Sea Ice Area Fluctuation in the Arctic Circle Based on Big Data and SARIMA Model. 2024 International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE), Athens, Greece, 376-380.

[20] M. Seo, H.-C. Kim. (2024). Arctic Greening Trends: Change Points in Satellite-Derived Normalized Difference Vegetation Indexes and Their Correlation with Climate Variables over the Last Two Decades. Remote Sensing, 16(7), 1160. https://doi.org/10.3390/rs16071160.

[21] T. Kocsis, R. Pongrácz, I. G. Hatvani, N. Magyar, A. Anda, I. Kovács-Székely. (2024). Seasonal trends in the Early Twentieth Century Warming (ETCW) in a centennial instrumental temperature record from Central Europe. Hungarian Geographical Bulletin, 73(1), 3–16. https://doi.org/10.15201/hungeobull.73.1.1.

[22] X. Huo, N. Sun, L. Ma. (2024). MA-BLTSI model for Land Surface Temperature prediction based on multi-dimensional data. Theoretical and Applied Climatology, 155(7), 6119–6136. https://doi.org/10.1007/s00704-024-05009-2.

[23] S. Dixit, K. K. Pandey. (2024). Spatiotemporal variability identification and analysis for non-stationary climatic trends for a tropical river basin of India. Journal of Environmental Management, 365, 121692. https://doi.org/10.1016/j.jenvman.2024.121692.

[24] J. L. Ng, Y. F. Huang, S. L. S. Yong, J. C. Lee, A. N. Ahmed, M. Mirzaei. (2024). Analysing the variability of non-stationary extreme rainfall events amidst climate change in East Malaysia. AQUA – Water Infrastructure Ecosystems and Society, 73(7), 1494–1509. https://doi.org/10.2166/aqua.2024.132.

[25] M. T. Islam, A. S. M. M. Kamal, M. M. Islam, S. Hossain. (2024). Impact of climate change on dengue incidence in Singapore: time-series seasonal analysis. International Journal of Environmental Health Research, 34(12), 3988–3998. https://doi.org/10.1080/09603123.2024.2337827.

[26] Daily Sea Ice extent Data. (2019, June 10). Kaggle. https://www.kaggle.com/datasets/nsidcorg/daily-sea-ice-extent-data.

[27] Data and image archive. (2025). National Snow and Ice Data Center. https://nsidc.org/data/seaice_index/data-and-image-archive.

[28] Climate change indicators: Arctic Sea ICE | US EPA. (2025, January 17). US EPA. https://www.epa.gov/climate-indicators/climate-change-indicators-arctic-sea-ice.

[29] Current State of Sea ice cover | Earth. (2025, January 23). https://earth.gsfc.nasa.gov/cryo/data/current-state-sea-ice-cover.

[30] J. Korstanje. (2023, July 31). How to select a model for your Time Series Prediction Task [Guide]. neptune.ai. https://neptune.ai/blog/select-model-for-time-series-prediction-task.

[31] The Pennsylvania State University. (2024). Applied Time Series Analysis. Sample ACF and Properties of AR(1) Model. PennState Eberly College of Science.

[32] H. Duan, Q. Li, L. He, J. Zhang, H. An, R. Ali, M. Vazifedoust. (2024). Climate classification for major cities in China using cluster analysis. Atmosphere, 15(7), 741.

[33] K. Sun, T. Lan, Y. M. Goh, S. Safiena, Y. Huang, B. Lytle, Y. He. (2023). An interpretable clustering approach to safety climate analysis: Examining driver group distinctions. Accident Analysis & Prevention, 196, 107420. https://doi.org/10.1016/j.aap.2023.107420.

[34] M. Ali, P. Scandurra, F. Moretti, H. H. R. Sherazi. (2024). Anomaly Detection in Public Street Lighting Data Using Unsupervised Clustering. IEEE Transactions on Consumer Electronics, 70(1), 4524-4535. https://doi.org/10.1109/TCE.2024.3354189.

[35] M. Suyal, S. Sharma. (2024). A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning. Journal of Artificial Intelligence and Systems, 6(1), 85–95. https://doi.org/10.33969/ais.2024060106.

[36] K. Aurangzeb. (2024). DBSCAN-based energy users clustering for performance enhancement of deep learning model. Journal of Intelligent & Fuzzy Systems, 46(3), 5555–5573.

[37] D. Mavaluru, R. S. Malar, S. M. Dharmarajlu, J. P. L. Auguskani, A. Chellathurai. (2024). Deep hierarchical cluster analysis for assessing the water quality indicators for sustainable groundwater. Groundwater for Sustainable Development, 25, 101119.