Advanced Anomaly Detection and Similarity Analysis in Sea Surface Temperatures Across MODIS Orbital Paths Using LSH

Arup Mazumder
Khaleque Md Aashiq Kamal
Sayeb Mohammad Tadvin
Puja Ghosh

December 22, 2023

Abstract

This project introduces an innovative approach to analyzing Sea Surface Temperature (SST) data, utilizing Locality-Sensitive Hashing (LSH) to uncover patterns in a substantial dataset comprising 100,000 samples collected over a year. By adapting LSH, a technique not conventionally applied in oceanographic data analysis, our project aimed to efficiently process high-dimensional SST data for identifying similarities and anomalies across various geographic regions. The methodology involved a comprehensive preprocessing phase, including normalization and feature scaling, followed by meticulous tuning of LSH parameters to achieve a balance between computational efficiency and the precision of results. Our analysis yielded insightful findings, revealing a significant consistency in the overall distribution of SST values, as evidenced by high Histogram Intersection scores. Concurrently, we observed distinct, localized variations in SST patterns, highlighting the diverse nature of ocean temperatures. The project overcame difficulties due to the large size and complex nature of the dataset, and adjusting the LSH parameters. Despite these challenges, the outcomes of our project provide valuable contributions to the understanding of SST dynamics. They hold implications for interpreting global climatic phenomena, enhancing predictive models in climate science, and advancing the field of oceanographic data analysis. This project not only demonstrates the utility of LSH in environmental science but also paves the way for its broader application in climate monitoring and marine studies.

1 Introduction

The study of sea surface temperature (SST) is a crucial component in understanding the myriad dynamics of our planet's climate system. With the advent of satellite technology and the proliferation of big data analytics, the field of oceanography has been revolutionized, offering unparalleled insights into the intricate patterns and phenomena that govern the Earth's oceans. This project introduces a novel approach to the analysis of SST data, employing LSH, a method traditionally underutilized in oceanographic data analysis, to explore new dimensions in this vital area of climate science.

1.1 Objective

The primary objective of this study is to investigate the effectiveness of LSH in analyzing extensive SST datasets, with a specific focus on identifying similiarities and detecting anomalies. LSH, renowned for its computational efficiency in processing large volumes of data, offers a promising alternative to conventional data analysis methods predominantly used in oceanography. This project seeks to leverage the unique capabilities of LSH to identify and categorize complex patterns within SST data and to pinpoint anomalous events among different SST phenomena.

1.2 Motivation

The motivation behind our work originates from the recognition of the immense potential hidden within the vast MODIS SST datasets. Traditional approaches often struggle to cope with the scale

and complexity of these datasets. Basic statistical measures like mean and root mean square (RMS) are commonly used to summarize the central tendency and variability of SST data [1]. While these metrics provide fundamental insights, they may not capture the full complexity and nuances of dynamic oceanic phenomena. On the other hand, setting static ***threshold*** values to identify anomalies or unusual events in SST data is a straightforward but rigid method. It may not adapt well to the variability inherent in large and diverse datasets, potentially leading to false positives or missing important anomalies. Our motivation lies in the desire to overcome these challenges by harnessing the power of LSH—a technique known for its ability to rapidly identify similar patterns in large datasets while maintaining computational efficiency. By developing advanced anomaly detection capabilities, we aim to contribute to a deeper understanding of the complex dynamics inherent in sea surface temperatures across MODIS orbital paths. This project is motivated by the need to uncover hidden patterns, detect anomalies with precision, and pave the way for more sophisticated analyses of oceanic processes. Our work not only addresses the specific challenges posed by SST data but also aligns with broader efforts to enhance anomaly detection methodologies in large-scale remote sensing datasets.

2 Background

SST is a critical parameter in oceanography, climatology, and environmental science. It refers to the temperature of the water's surface layer, typically measured up to a depth of 20 meters. SST data provide invaluable insights into oceanic processes and are integral to understanding and predicting climate change, weather patterns, and marine ecosystems.

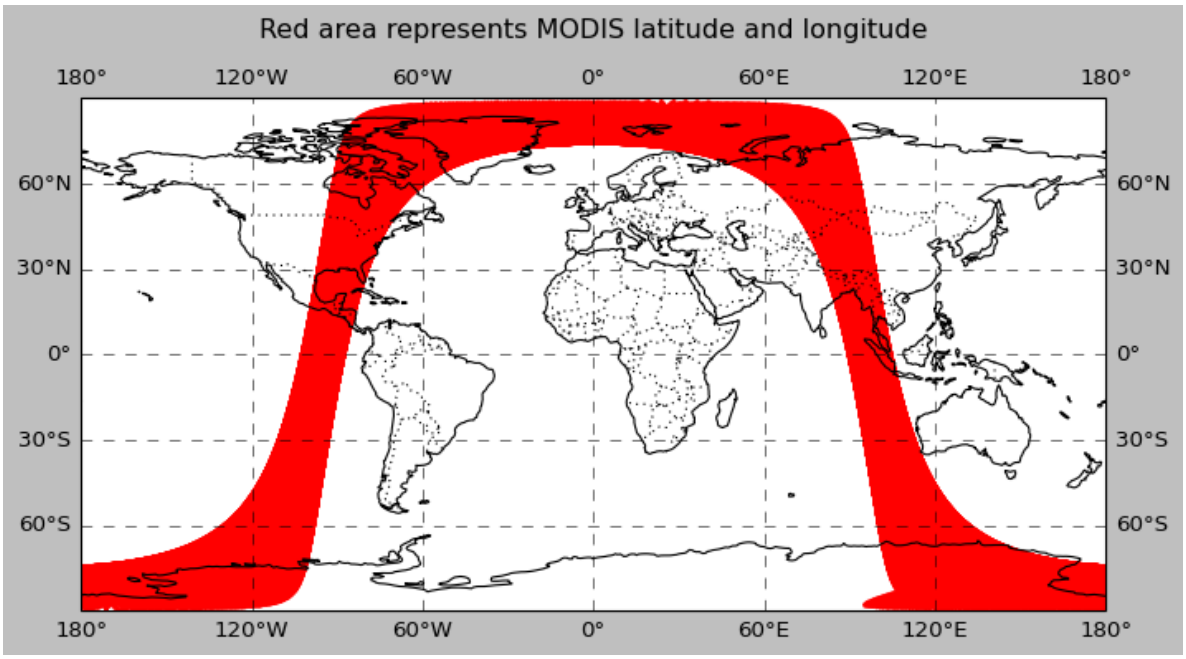
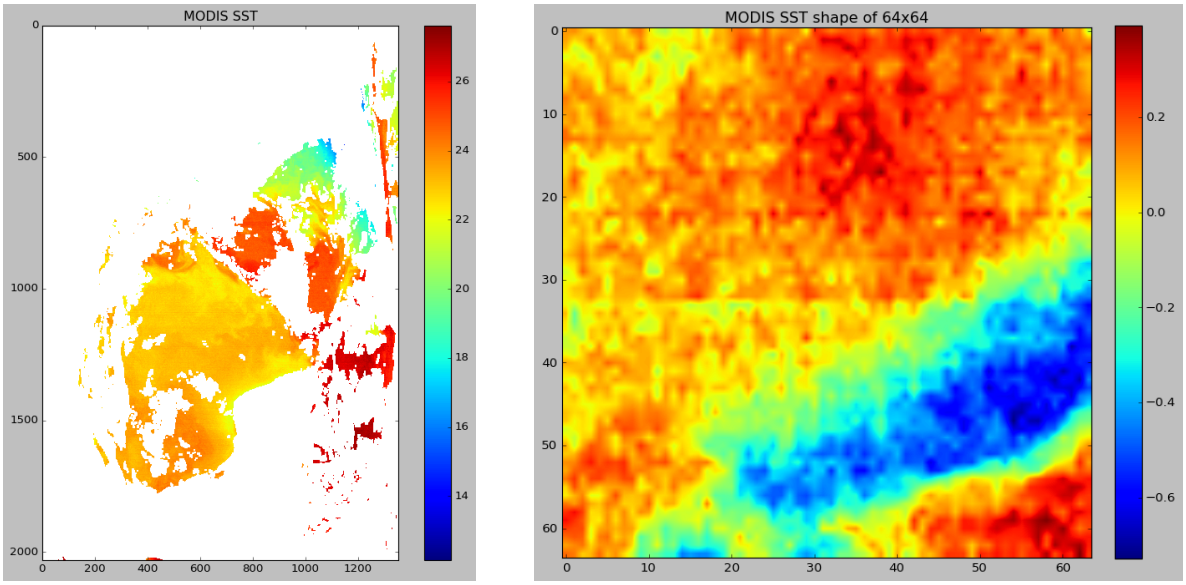


Figure 1: A random orbital path that is followed by a MODIS sensor.

The advent of satellite remote sensing has revolutionized the collection and analysis of SST data. Since the first weather satellites in 1967, there has been a significant advancement in satellite technologies and methodologies for SST measurement. One of the pivotal advancements in this field has been the deployment of NASA's Moderate Resolution Imaging Spectroradiometer (MODIS). MODIS satellites have been instrumental in providing comprehensive, global SST data with a minimal time lag, typically around one day. Figure 1 illustrates the orbital path followed by the MODIS sensor for collecting SST data. The highlighted red regions indicate the specific latitude and longitude points encompassed within the sensor's coverage area. This near-real-time monitoring capability has been essential for tracking dynamic changes in ocean temperatures and has significantly contributed to our understanding of oceanic temperature dynamics.



(a) (b)

Figure 2: (a) An image from an orbit captured by MODIS sensor, (b) A 64x64 clear region from that image.

Satellite-based SST data are characterized by their extensive coverage and high temporal resolution, allowing for the monitoring of ocean temperatures across vast and remote areas. This has enabled researchers to observe and analyze a range of phenomena, from large-scale climatic patterns like El Niño and La Niña to more localized events such as ocean currents and upwelling zones. In Figure 2a, we present the Sea Surface Temperature as captured by the MODIS sensor. This image provides a comprehensive view of SST across a wide area. Complementing this, Figure 2b focuses on a specific 64x64 pixel region extracted from the larger image, offering a detailed look at the SST in a more concentrated area.

In the context of our project, the extensive and detailed nature of SST datasets poses both opportunities and challenges. The vast volume of data, while rich in information, requires advanced analytical techniques for efficient processing and meaningful interpretation. Traditional methods, such as deep learning algorithms, have been used effectively but often require substantial computational resources and expertise.

This is where the application of LSH becomes particularly relevant. LSH, with its ability to handle large datasets efficiently and its suitability for identifying similarities and anomalies within data, presents an innovative approach to SST data analysis. By applying LSH to SST datasets, this project aims to uncover new patterns and anomalies that might not be as easily detected using traditional methods. Furthermore, LSH's potential for scalability and efficiency in processing could make it a valuable tool for ongoing and future SST monitoring and analysis, offering a complementary approach to existing methodologies.

Overall, our study of SST is crucial for a comprehensive understanding of Earth's climate system. By introducing and evaluating the efficacy of LSH in SST data analysis, this research seeks to contribute to the field of oceanography, offering new perspectives and methodologies for studying one of the planet's most vital indicators — its ocean temperatures.

3 Related Study

SSTs have been the subject of numerous research that have looked into anomaly detection and similarity analysis across MODIS orbital trajectories. Detecting anomalies and analyzing similarities are crucial methods for examining trends in SST. A variety of methods, including deep learning models, machine learning algorithms, and statistical techniques, have been developed for anomaly identification. LSH also has become a popular and effective method for finding outliers in high-dimensional

datasets [2].

The authors in [3] present ULMO (Unsupervised Learning of the Multivariate Ocean), a probabilistic autoencoder that combines an autoencoder and a normalizing flow for the analysis of sea surface temperature (SST). The collection included about 12 million 128x128 pixel SST cuts from the NASA Aqua satellite's MODIS sensor between 2003 and 2019. This novel method used deep learning to examine large amounts of remote-sensing data for oceanographic research in an effort to find anomalies and intricate phenomena in the dynamics of the ocean surface. Additionally, in [4], using remote sensing and numerical modeling, the research examines SST anomalies in the Mozambique Channel. Notable changes in SST are found east and west of Madagascar, which are ascribed to variations in wind intensity, depths of mixed layers, heat flow, and the impact of oceanic currents and eddies.

Furthermore, the accuracy and statistical characteristics of sea surface temperature (SST) data in the Arabian Gulf are examined by the author in this research [1]. The study examines trends, seasonal variability, and extreme values by comparing satellite-derived sea surface temperature (SST) with on-site measurements using the Group for High Resolution Sea Surface Temperature (GHRSST) dataset. The study's findings of rising SST extremes highlight the value of this information for regional industrial and environmental applications. Based on MODIS satellite data, a novel deep learning model, NENYA, is used in this research [2] to investigate trends in SST. It leverages a 256-dimensional latent space created from SST data to recognize various SST patterns, particularly in dynamic ocean regions. By highlighting the importance of different submesoscale structures in places like western boundary currents and large-scale SST patterns in equatorial regions, the work provides fresh understanding of oceanic phenomena at large.

Identifying the optimum LSH family for a numerical dataset, such as the sea surface temperature data in the current study, requires several factors. P-stable distributions are often good for Euclidean distances in numerical data, while angular LSH families perform well for cosine similaritiesh [5]. When fine-tuning the perm number in Locality-Sensitive Hashing (LSH), especially for min-hashing algorithms, accuracy and computing efficiency must be balanced. We consider memory and computational limits, and align the perm number with the implementation's specific requirements [2]. Additionally, iterative testing and benchmark comparisons are critical in determining the best perm number. The hashing method implemented is determined by the distinctive analysis or operational needs, such as similarity detection, data distribution, dimensionality reduction, or geographical concerns. Each approach has its own potential and is appropriate for certain areas of data processing and analysis.

These techniques shed light on the temporal evolution and spatial distribution of SST patterns. A complete approach to understanding SST dynamics is provided by combining approaches for anomaly identification and similarity analysis. The work we conducted proposes a sophisticated methodology for the examination of similarity and anomaly identification in SST data derived from MODIS satellite images. Our work effectively finds similarities and anomalies in SST data based on one reference region by LSH.

4 Methodology

4.1 Data Description and Preprocessing

Our project utilizes a comprehensive dataset of SST measurements, encompassing a vast array of geographical locations across the globe. The whole dataset, spanning a year, consists of 700,000 samples, each representing SST readings in a 64x64 grid. In our study, we used 100,000 samples in total. The preprocessing of this substantial dataset involved several critical steps:

- **Data Structuring**: Initially, each 64x64 SST matrix was transformed into a one-dimensional vector, resulting in a new representation where each sample comprised 4096 individual temperature readings. This flattening process was crucial for the subsequent application of machine learning techniques suitable for high-dimensional data.

- **Feature Scaling and Normalization**: Post-flattening, each vector underwent normalization to standardize the range of the temperature values. We employed the StandardScaler from the scikit-learn library, which adjusted each feature to have a mean of zero and a standard deviation of one. This normalization was pivotal in eliminating potential biases arising from disparate

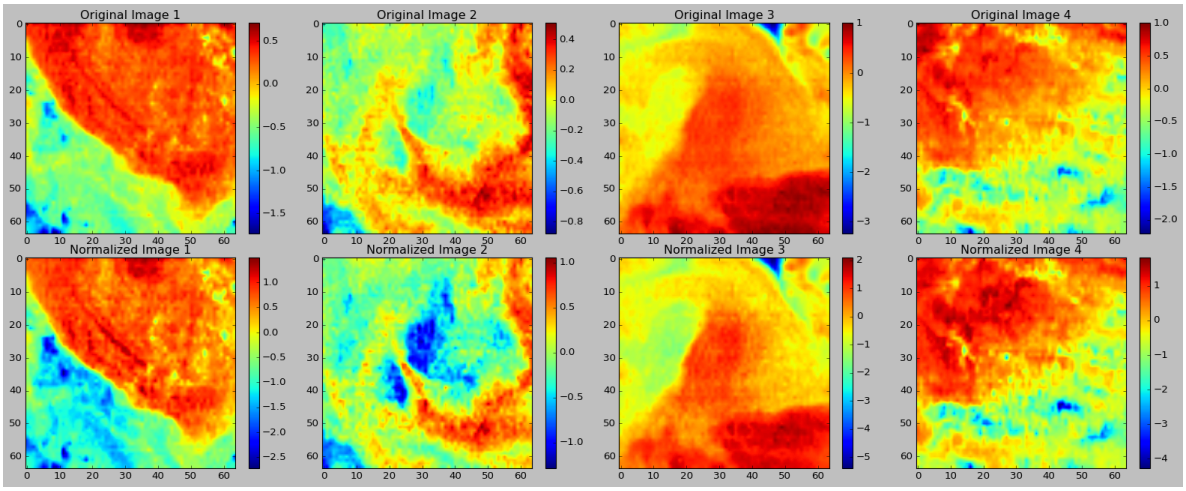


Figure 3: Before and after normalization of the Four random SST samples

ranges of temperature values and ensuring that each feature contributed equivalently to the subsequent analysis.

4.2 Selection and Implementation of LSH

LSH was selected as the primary technique for analyzing the high-dimensional SST data. This choice was guided by LSH's ability to efficiently handle large datasets and its efficacy in identifying similar items in high-dimensional spaces. The LSH implementation involved several nuanced steps:

1. **Choosing the Right LSH Family**: MinHash algorithm was selected because of the high dimensionality of the SST data.

2. **Hashing the Data**: To operationalize LSH, we utilized the datasketch library, a Python-based tool offering efficient LSH implementations. Each normalized SST vector was mapped to a set of hash tables, each representing a unique permutation of the data. These hash tables form the core of our LSH model, enabling the quick retrieval of similar data points while efficiently navigating the high-dimensional space.

3. **Parameter Tuning and Optimization**: The effectiveness of LSH hinges on the fine-tuning of its parameters, particularly the number of permutations and the hash tables. These parameters were initially set to default values and then iteratively adjusted. This tuning process involved a series of trial runs, each time evaluating the precision and recall of the similarity detection, and making incremental adjustments to optimize the balance between computational efficiency and the accuracy of the results.

   In the process of fine-tuning our Locality-Sensitive Hashing (LSH) model for the analysis of Sea Surface Temperature (SST) data, we explored a range of parameter combinations to optimize performance. Specifically, we tested permutations ***num_perms*** at levels 128, 256, and 512. These values were chosen to evaluate the impact of different granularities in the hashing process on our ability to detect similarities and anomalies within the SST data.

   Additionally, we varied the ***thresholds*** parameter across five distinct values: 0.0, 0.25, 0.50, 0.75, and 1.0. This range was selected to cover a broad spectrum, from no similarity ***thresholds*** (0.0) to complete similarity (1.0), allowing us to assess the sensitivity of our model to different degrees of similarity. By systematically iterating over these combinations of ***num_perms*** and ***thresholds***, we aimed to identify the optimal configuration that balances computational efficiency with analytical accuracy in our SST data analysis.

4.3 Detecting Similarity in SST Regions

The detection of similar SST regions through LSH involved a methodical approach, ensuring both the precision of results and computational efficiency:

- **Querying the LSH Index**: The LSH index was systematically queried using specific SST regions as reference points. Each query involved mapping the reference region to its respective hash buckets, followed by retrieving all candidate regions sharing the same buckets.
- **Refining Similarity Detection**: To ensure the relevance of the results, a similarity ***threshold*** was applied to the candidate pairs. This ***threshold*** was static from 0.0 to 1.0; it was adjusted to find the best similarity scores obtained in preliminary tests. This refinement step was crucial in filtering out less similar regions, thereby enhancing the precision of our similarity detection.

4.4 Anomaly Detection in SST Data

In our analysis, a significant focus was placed on detecting anomalies within Sea Surface Temperature (SST) data, specifically aiming to identify regions exhibiting unusual temperature patterns. Unlike the initial approach of using Z-scores based on statistical baselines, the final methodology employed Locality-Sensitive Hashing (LSH) with the MinHash algorithm. This process involved the following steps:

1. **LSH Initialization and MinHash Creation**: We began by initializing the LSH with a specified number of permutations ***num_perm***. For each data point in the normalized SST data, a MinHash object was created. This object generated a compact representation of the high-dimensional data, enabling efficient computation of similarity measures.
2. **Insertion into LSH and Jaccard Distance Calculation**: Each MinHash object was inserted into the LSH index. To detect dissimilarities, we calculated the Jaccard distance, which measures dissimilarity based on the lack of common elements between two sets. For each region in the dataset, we computed its Jaccard distance with every other region.
3. **Identifying Top Dissimilar Regions**: After calculating the Jaccard distances, we sorted these values in descending order to identify the regions most dissimilar to each other. The top 19 regions with the highest dissimilarity scores were then flagged as potential anomalies.

This methodology provided a robust framework for identifying SST regions with atypical temperature distributions. By leveraging the computational efficiency of LSH and the quantification ability of the MinHash algorithm, we were able to process the extensive and complex SST dataset effectively. The use of Jaccard distances offered a precise measure for identifying the most distinct SST patterns, which is crucial for understanding and monitoring anomalies in ocean temperatures.

4.5 Metrics

We used the following metrics to validate the similarity and anomaly between the reference image and output provided by LSH.

1. **Pearson Correlation Coefficient**
   The Pearson correlation coefficient measures the linear correlation between two images. It quantifies how well a linear equation describes the relationship between the pixel intensities of the two images. The coefficient ranges from -1 to +1, where +1 indicates perfect positive linear correlation, 0 indicates no linear correlation, and -1 indicates perfect negative linear correlation. The correlation coefficient is calculated as follows [6]:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

   where m_x is the mean of the vector x and m_y is the mean of the vector y.
   In this project, a higher correlation coefficient would indicate that the temperature patterns in the two SST images are linearly similar. For instance, if one image shows a temperature gradient from warm to cool, a similar gradient in another image would result in a high correlation.

2. **Mean Squared Error (MSE)**
   MSE calculates the average squared difference between the corresponding pixel values of two images. It's a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.
   The Mean Squared Error (MSE) is given by the formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

   where n is the number of observations, y_i is the actual value of the i-th observation, and $\hat{y}_i$ is the predicted value of the i-th observation.

   In this project, A lower MSE indicates that the two images are very similar in terms of absolute pixel values. In terms of SST, it suggests that the temperature values across the two images are closely matched, with little deviation.

3. **Histogram Intersection**
   This metric compares the histograms of the two images. Histograms represent the distribution of pixel intensities (here, temperature values) within an image. A higher histogram intersection value indicates that the two images have similar distributions of temperature values. For example, if both images have a similar range and frequency of temperature values, their histograms will overlap more, resulting in a higher intersection value.

4. **How These Metrics Combine to Measure Similarity:**
   Together, these metrics provide a comprehensive view of similarity. While correlation captures linear relationships in temperature patterns, MSE gives an idea of the absolute differences in temperature values. Histogram Intersection, on the other hand, looks at the overall distribution of temperatures within the images. In the hyperparameter tuning part of the code, a combined score is computed using these metrics (score = correlation - mse + hist_intersection). This scoring function aims to maximize correlation and histogram intersection (indicating similarity) while minimizing the MSE (indicating dissimilarity). The combination of ***num_perm*** and ***threshold*** that results in the highest score is considered the best for capturing the most meaningful similarities in the SST dataset.

4.6 Visualization Techniques

Visualization played a pivotal role in our study, offering intuitive and detailed insights into SST (Sea Surface Temperature) patterns. Our primary visualization technique involved the following:

- **Grid-based Image Display**: We employed a grid-based approach to display images, showcasing both the original SST data and its most dissimilar counterparts. This method was particularly useful for illustrating the differences and anomalies in SST patterns. For each selected image, we plotted a grid of 19 dissimilar images, including the reference image, allowing for a direct visual comparison.
- **Reference Image Display**: The grid prominently featured the original SST image for reference, facilitating a clear baseline comparison with its dissimilar images.
- **Dissimilar Images Comparison**: The dissimilar images, identified based on their Jaccard distances in the LSH analysis, were arranged alongside the reference image. Each image in the grid was accompanied by a colorbar, providing a visual scale for temperature values and enhancing the interpretability of SST variations.
- **Layout and Scalability**: The grid layout was dynamically adjusted based on the number of images to be displayed, ensuring an organized and scalable visualization format. This flexibility was crucial in handling various datasets and comparative analyses.

This visualization technique was integral to our analysis, enabling us to clearly depict and interpret the complex temperature distributions within the SST data. By visually comparing the reference SST image with its most dissimilar counterparts, we could effectively highlight and investigate anomalous patterns, thus enriching our understanding of SST dynamics.

4.7 Addressing Computational Challenges

Given the vast size of our dataset, several computational strategies were employed to ensure efficient processing without compromising the integrity of our analysis:

- **Batch Processing**: To manage the computational demands, we adopted a batch processing approach. This method involved dividing the dataset into smaller, manageable batches, each processed independently. This approach not only optimized memory usage but also improved processing speed.
- **Parallel Computing Techniques**: We leveraged parallel computing techniques to further enhance the efficiency of our computations. This involved distributing the computational tasks across multiple processors, significantly accelerating the data processing and analysis phases.

5 Result and Discussion

5.1 Similarity Analysis

Through extensive hyperparameter tuning of the Locality-Sensitive Hashing (LSH) algorithm, the most effective configuration for analyzing our SST dataset was identified. The optimal parameters were determined to be ***num_perm*** set to 128 and a ***threshold*** of 0.0. This combination was selected based on its performance in three key statistical metrics: Pearson Correlation, Mean Squared Error (MSE), and Histogram Intersection, which were used to evaluate the similarity between SST images.

The best configuration yielded an average Pearson Correlation of approximately -0.0037. This near-zero value suggests a general lack of strong linear correlation between the pixel intensities of the compared SST images, indicating diverse temperature patterns across different samples in the dataset. On the other hand, the average MSE was found to be 0.3639. This metric, indicating the average squared difference in temperature values between images, points to a certain level of dissimilarity in terms of absolute temperature readings in the dataset. Remarkably, the average Histogram Intersection across the dataset was extremely high, at about 0.9994. This result implies that, despite the variations in specific temperature values and patterns, the overall distribution of temperatures is strikingly similar across the majority of images. Such consistency in temperature distribution is a significant finding, particularly in the context of large-scale climatic observations and trends. Figure 4 displays the top 19 regions that exhibit the highest similarity to our reference region, denoted as Image-0. This visual representation effectively highlights the most similar images identified through our analysis, allowing for a comparative study against the baseline provided by the reference image.

Image Index	Min Temp.	Max Temp.	Avg. Temp	Pearson Correlation	MSE	Histogram Intersection
0	-0.729	0.393	-5.439e-07	1.0	0.0	1.0
46	-1.052	0.659	8.27e-07	0.0455	0.065	1.0
99	-1.1	0.656	7.413e-07	0.11	0.10	1.0
200	-0.832	0.526	-7.995e-07	0.018	0.086	1.0
202	-0.971	0.364	-8.158e-07	-0.37	0.0975	1.0
217	-1.056	0.589	-3.539e-08	-0.143	0.108	1.0
230	-1.212	0.485	-1.315e-06	-0.262	0.147	1.0
372	-0.913	0.521	-2.459e-07	-0.153	0.099	1.0
522	-0.739	0.572	5.103e-07	0.249	0.061	1.0
565	-0.816	0.612	-3.39e-07	0.219	0.071	1.0
816	-0.711	0.510	-1.583e-07	-0.212	0.088	1.0
842	-0.804	0.631	8.242e-08	0.173	0.074	1.0
940	-0.507	0.589	-6.724e-07	-0.329	0.103	1.0

Table 1: Statistical analysis of 10 similar regions with reference region (Image 0).

In the table 1 showcasing the analysis of similar Sea Surface Temperature (SST) images, each row corresponds to a specific image identified through Locality-Sensitive Hashing (LSH) as being similar

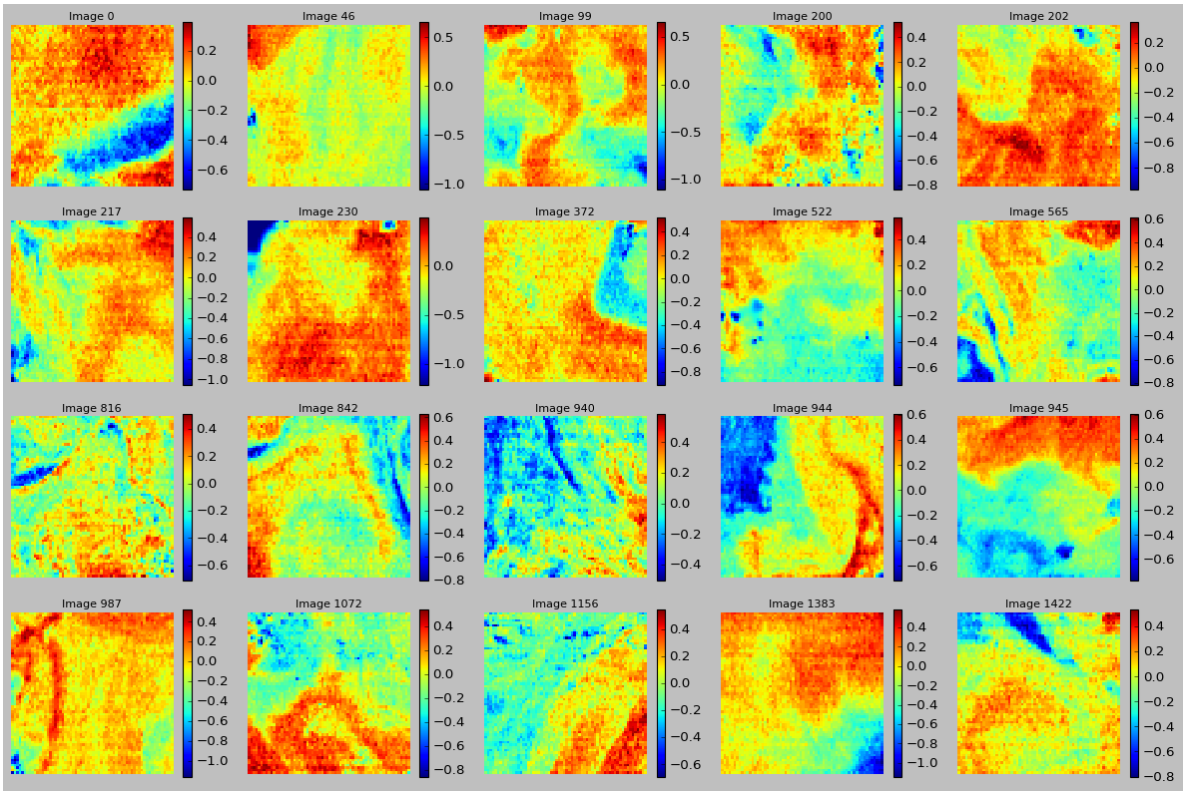


Figure 4: Top 19 similar regions based on reference region (Image 0).

to a reference image. The table columns detail key statistical metrics for these images: minimum temperature, maximum temperature, average temperature, Pearson correlation, Mean Squared Error (MSE), and Histogram Intersection. The minimum and maximum temperatures provide insights into the thermal extremities within each image, reflecting the diverse thermal characteristics of different ocean regions. The average temperature offers an overarching view of the sea surface's thermal state in each image. Pearson correlation values, which range from negative to positive, indicate the degree of linear similarity in temperature patterns compared to the reference image. A high correlation suggests similar patterns, while low or negative values indicate dissimilar or inversely related patterns.

Histogram Intersection consistently scores at 1.0 for all images, indicating a uniform distribution of temperatures across the dataset, despite variations in specific temperature values. This uniformity is critical in understanding broad oceanic and climatic trends. The MSE values, representing the average squared differences between the compared images, add another layer of understanding by quantifying the absolute differences in temperature readings. Together, these metrics paint a comprehensive picture of the similarities and differences in SST patterns across the dataset. They provide a multifaceted understanding of sea surface temperatures, highlighting both the unique characteristics of individual images and the overarching similarities that bind them together, thus contributing significantly to our knowledge of oceanographic and climatic dynamics.

5.2 Anomaly Detection

In our analysis to identify dissimilar regions within the Sea Surface Temperature (SST) data, we determined that the optimal parameters for our Locality-Sensitive Hashing (LSH) model were a number of permutations ***num_perm*** set to 512 and a similarity ***thresold*** of 0.5. This particular combination of parameters was selected after extensive testing and was found to be the most effective in highlighting significant dissimilarities in SST patterns. Figure 5 presents the 19 most dissimilar regions compared to our chosen reference region, Image-0. This figure showcases a selection of images that markedly contrast with the reference image in terms of SST patterns. Each image is labeled with its index

number, identifying its position within our larger dataset of 100,000 images. These visual comparisons underscore the diversity within the SST data and highlight the extent of variation from the reference region.

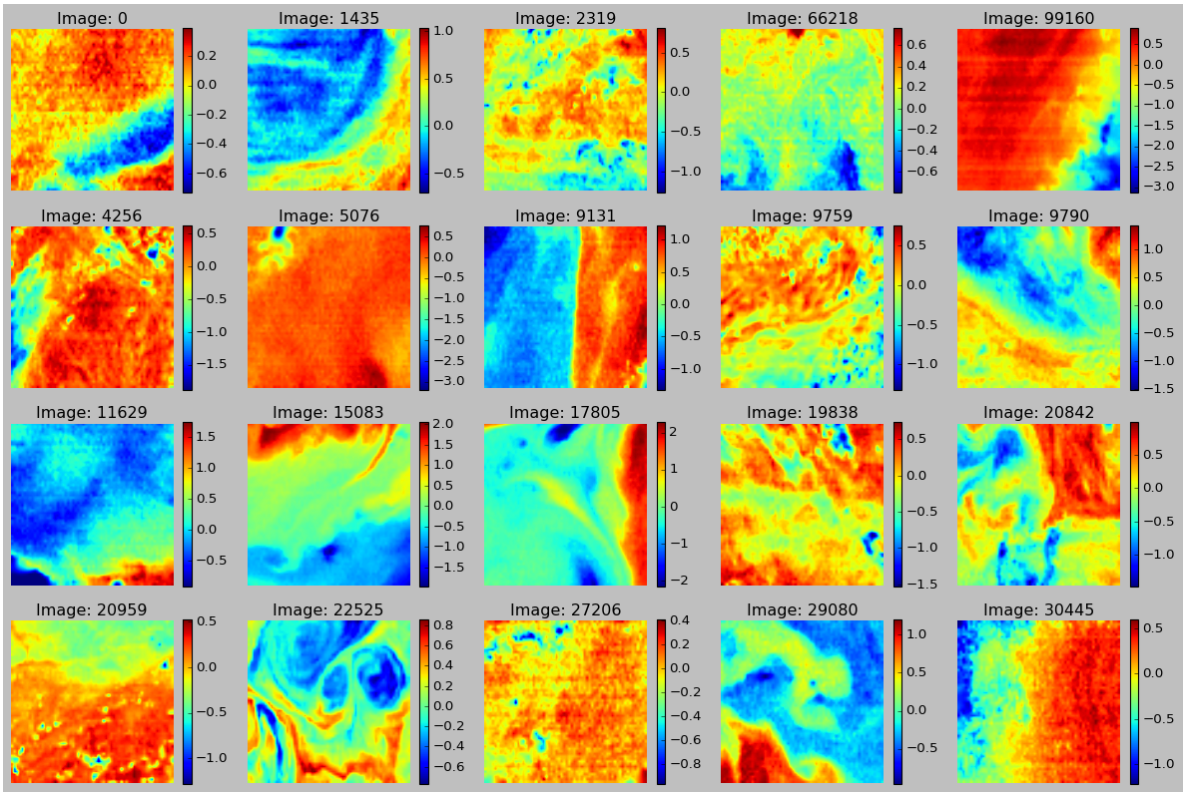


Figure 5: Top 19 dissimilar regions based on reference region (Image 0).

The core of our dissimilarity measurement lies in computing the Jaccard distance between images. For each image, its MinHash is compared against those of all other images, calculating the Jaccard distance, which quantifies dissimilarity (1 - Jaccard similarity). The Jaccard distance is particularly effective here, as it directly measures the lack of commonality between the pixel distributions of two images, with higher values indicating greater dissimilarity. The algorithm then selects the top 19 images exhibiting the highest Jaccard distances for each image in the dataset, identifying them as the most dissimilar. This method yields a comprehensive list of dissimilar images for each entry in the dataset, allowing for an in-depth analysis of anomalies and pattern variations in the SST data. By leveraging the scalability and computational efficiency of LSH and MinHash, this approach provides a robust framework for handling the extensive and complex nature of SST datasets, enabling the exploration of intricate ocean temperature dynamics at a granular level.

6 Discussion and Future Study

The results of our study using Locality-Sensitive Hashing (LSH) to analyze sea surface temperature (SST) datasets present a nuanced view of ocean temperature dynamics. The observed high Histogram Intersection scores across SST images suggest a remarkable uniformity in the distribution of temperature values despite geographic and temporal variations. This uniformity could be indicative of pervasive oceanic patterns influenced by global climatic phenomena such as El Niño or La Niña. However, the variability in Pearson Correlation and Mean Squared Error (MSE) across different regions points to the intricate and diverse nature of ocean temperatures, influenced by localized factors like ocean currents, upwelling zones, and geographical features. This complex interplay between global trends and local conditions underscores the necessity for robust, multidimensional analysis techniques such as LSH, which can effectively handle the vast and intricate nature of SST datasets.

Reflecting on our methodological approach, the application of LSH, particularly with the chosen parameters of ***num_perm*** and ***threshold***, demonstrates its effectiveness in managing large-scale SST data. However, this study also highlights the need for a careful balance between computational efficiency and the accuracy of anomaly detection. Future research could explore the impact of varying LSH parameters or integrating LSH with other data analysis techniques to enhance the detection of subtle SST anomalies. Additionally, given the critical role of SST in understanding climate change and its impacts, extending this analysis to include predictive modeling could provide valuable insights into future climatic trends. The application of LSH in other environmental contexts, such as atmospheric data analysis or integrated Earth system models, could further expand our understanding of the complex interactions within the Earth's climate system. Ultimately, the methodologies and insights gleaned from this study have the potential to significantly contribute to global efforts in climate monitoring, environmental protection, and sustainable management of oceanic resources.

7 Challenges and Limitations

This study, while providing valuable insights into sea surface temperature (SST) patterns using Locality-Sensitive Hashing (LSH), encountered several challenges and limitations that are important to acknowledge.

- **Data Complexity and Volume:** One of the primary challenges was managing the sheer volume and complexity of the SST dataset. While LSH efficiently handles high-dimensional data, the vastness of the dataset posed significant computational challenges. Batch processing and parallel computing techniques were employed to mitigate this, but these approaches have their own limitations in terms of processing power and memory requirements.
- **Parameter Tuning in LSH:** The effectiveness of LSH largely depends on the fine-tuning of its parameters, particularly the number of permutations and the hash tables. Determining the optimal settings was a complex process, requiring multiple iterations and trial runs. There is a possibility that different parameter settings might yield different results, indicating a need for a more robust method of parameter optimization.
- **Generalizability of Findings:** The findings from this study are based on a specific dataset and a particular set of LSH parameters. As such, there may be limitations in generalizing these results to other datasets or different environmental contexts. Future studies should explore the applicability of our methodology across diverse datasets to validate and potentially enhance the generalizability of our findings.
- **Interpretation of Results:** The interpretation of similarity and anomaly detection in SST data requires careful consideration. While statistical metrics provided quantitative assessments, there is a need for a more nuanced understanding of these results in the context of oceanographic phenomena. The complex nature of ocean dynamics means that statistical anomalies might not always correspond to ecologically or climatically significant events.
- **Dependency on Preprocessing:** The preprocessing steps, including data structuring and normalization, were crucial for the analysis. However, these steps might introduce biases or alter the natural variability in the SST data, potentially impacting the results. The choice of normalization technique and the transformation of SST matrices into one-dimensional vectors are particularly critical decisions that could affect the outcome of the analysis.
- **Limitations in Addressing Non-Linear Relationships:** While LSH is effective in identifying linear similarities and anomalies, it may be less capable of capturing non-linear relationships in SST data. Ocean temperature dynamics often involve complex, non-linear interactions, which may not be fully detected through LSH.

8 Conclusion

In this project, we embarked on a sophisticated exploration of SST patterns using LSH. Our objective was to harness the computational efficiency and unique capabilities of LSH in identifying similarities

and detecting anomalies within extensive SST datasets. The findings from our research provide novel insights into the intricate dynamics of the Earth's oceans, revealing both uniformity and diversity in SST patterns across different geographic regions.

Our methodological approach demonstrated the efficacy of LSH in processing large volumes of SST data. The optimal parameter configuration of LSH, determined through rigorous testing, enabled us to effectively identify regions with similar and dissimilar temperature patterns. The high Histogram Intersection scores across the dataset suggest a significant uniformity in the overall distribution of temperatures, while variations in Pearson Correlation and Mean Squared Error (MSE) values underscore the complex and localized nature of SST variations.

However, our study was not without its challenges and limitations. The vast and complex nature of the SST data posed significant computational challenges, and the process of fine-tuning LSH parameters highlighted the delicate balance between accuracy and efficiency in anomaly detection. Moreover, the generalizability of our findings to other datasets and environmental contexts remains an area for further exploration.

Despite these challenges, our study contributes valuable insights to the field of oceanography and climate science. The methodology and findings have practical implications for environmental monitoring, policy-making, and climate research. They also pave the way for future studies, suggesting the potential of LSH in other environmental datasets and its integration into predictive models for climate change.

In conclusion, this research underscores the importance and utility of employing advanced data analysis techniques like LSH in understanding our planet's climate system. By offering a new perspective in SST data analysis, this study contributes to the ongoing efforts in comprehending and responding to the dynamic and complex nature of Earth's oceans, a vital component of the global climate system.

9 Acknowledgment

We extend our heartfelt thanks to Professor Dr. Noah Daniels for his exceptional teaching and invaluable guidance in efficient algorithms for big data, which significantly shaped our project. We are also immensely grateful to Professor Dr. Peter Cornillon for sharing his ideas and providing the polished dataset crucial for our project. Their contributions have been fundamental to our study, and we deeply appreciate their support and mentorship.

10 Code Repository

https://github.com/arupsky/CSC592-Fall23

References

[1] Oleksandr Nesterov, Marouane Temimi, Ricardo Fonseca, Narendra Reddy Nelli, Yacine Addad, Emmanuel Bosc, and Rachid Abida. Validation and statistical analysis of the group for high resolution sea surface temperature data in the arabian gulf. *Oceanologia*, 63(4):497–515, 2021.

[2] Jorge Meira, Carlos Eiras-Franco, Verónica Bolón-Canedo, Goreti Marreiros, and Amparo Alonso-Betanzos. Fast anomaly detection with locality-sensitive hashing and hyperparameter autotuning. *Information Sciences*, 607:1245–1264, 2022.

[3] J Xavier Prochaska, Peter C Cornillon, and David M Reiman. Deep learning of sea surface temperature patterns to identify ocean extremes. *Remote Sensing*, 13(4):744, 2021.

[4] Guoqing Han, Changming Dong, Junde Li, Jingsong Yang, Qingyue Wang, Yu Liu, and Joel Sommeria. Sst anomalies in the mozambique channel using remote sensing and numerical modeling data. *Remote Sensing*, 11(9):1112, 2019.

[5] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262, 2004.

[6] SciPy community. scipy.stats.pearsonr — scipy v1.11.4 manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html. Accessed: 2023-12-21.