Explainability Pipeline for Spatiotemporal Land Surface Forecasting Models

A THESIS

submitted by

Tushar Verma

Roll Number (2022AIY7514)

under the guidance of

Dr. Sudipan Saha

For the award of the degree

Master of Science (by Research)

Yardi School of Artificial Intelligence

Indian Institute of Technology, Delhi

March, 2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Sudipan Saha, for his invaluable guidance and mentorship throughout my Master's research journey. From offering me the opportunity to conduct research at the prestigious Indian Institute of Technology Delhi to providing constant support and insightful discussions that shaped my thinking, his dedication and passion for research have been a continual source of inspiration. I am especially grateful for his encouragement to pursue my extracurricular passion for football alongside my academic endeavors, allowing me the honor of representing this institute at major tournaments.

I am deeply thankful to my parents, whose encouragement and belief in my pursuit of higher education have been my greatest source of strength. Their support and the values they instilled in me have profoundly shaped who I am today.

I also extend my heartfelt gratitude to my big brother Ankit, whose unwavering support and belief in my abilities have been a constant source of motivation. His guidance, both in academics and in life, has shaped my perspective and given me the confidence to take on new challenges. His words of encouragement and wisdom have been a steady anchor, pushing me to strive for excellence even in the face of difficulties.

I am equally grateful to my cousins, Ambuz and Yash, for always standing by my side through every challenge and milestone, and my friends Naimish, Dhruvil, Kshitiz, Gagandeep, Ashish, Hrishabh and Umang for their stimulating discussions, constructive debates, and shared enthusiasm for learning that have enriched my academic experience beyond measure.

A special acknowledgment goes to Nitya for her patience, for always reminding me to take necessary breaks when I needed them the most, and for being my greatest cheerleader through every stressful deadline.

Abstract

KEYWORDS: Climate AI, Explainability, ConvLSTM, spatiotemporal analysis

Satellite images have become an essential tool for studying regional climate change. They provide a clear view of how land surfaces change over time, helping scientists track patterns in vegetation, temperature, and other environmental factors. When combined with meteorological data, they offer valuable insights into how climate evolves and what changes we might expect in the future. One important application of satellite imagery is Earth surface forecasting, which uses remote sensing and weather data to predict shifts in land conditions. This helps researchers monitor things like temperature variations, soil moisture, and land cover changes. However, understanding the complex relationships between weather patterns and surface changes is still a challenge. Many factors, such as temperature, rainfall, and human activities, interact in ways that make climate modeling difficult.

This research presents a novel pipeline that combines principles from perturbation-based techniques like LIME with global explainability methods such as PDP. By addressing the limitations of these approaches in high-dimensional spatiotemporal models, the proposed framework enhances interpretability in complex land surface forecasting tasks. It enables key analyses, including marginal sensitivity, correlation, and lag analysis, providing deeper insights into model behavior and variable influence.

To demonstrate the effectiveness of this approach, we applied ConvLSTM for surface forecasting, examining the impact of key meteorological variables—temperature, pressure, and precipitation—on the predicted Enhanced Vegetation Index (EVI). Our study utilized the EarthNet2021 dataset, which spans a broad region across Central and Western Europe, with most temporal samples covering the period from May to October.

Contents

List of Tables

List of Figures

1. Introduction

1.1 Background and Motivation

The Earth's climate is undergoing a significant transformation, posing a substantial threat to human existence. Its detrimental effects on terrestrial surfaces, which sustain most life on our planet, are becoming increasingly evident. From the depletion of Arctic sea ice [1] to the intensification of fire incidents [2], the repercussions of climate change manifest across diverse and variable geographic regions. Studying the effects of meteorological variables such as temperature, precipitation, and pressure is central to climate change analysis.

Over the past decade, there has been a notable increase in satellite sensors, leading to the availability of Earth observation data on an unprecedented scale. Initiatives like the Copernicus program [3] offer high-resolution data with enhanced temporal coverage, enabling the generation of dense predictions and analyses that were previously unattainable. Land surface forecasting using spatiotemporal forecasting models plays a crucial role in predicting changes in surface conditions over time, such as vegetation growth [4] [5], soil moisture levels [6] [7], and land use patterns [8] [9]. However, interpreting the output of these models and understanding the factors driving their predictions pose significant challenges, particularly in the context of complex spatiotemporal data and high-dimensional feature spaces.

Explainability has emerged as a critical requirement for ensuring the transparency, trustworthiness, and reliability of machine learning models, including those used in the prediction of land surface forecasting. Traditional explainability techniques, such as Local Interpretable Model-agnostic Explanations (LIME) [10] and partial dependence plots (PDP) [11] etc., have limitations when applied to land surface forecasting models. These models often operate in high-dimensional spatiotemporal feature spaces, making it challenging to isolate the effects of individual variables on model prediction.

1.1.1 Significance of Explainability in Land Surface Prediction

- **Enhancing Model Transparency** Explainability helps in making complex land surface prediction models more transparent by providing insights into how predictions are made. It reduces the "black box" nature of machine learning models, allowing researchers and practitioners to understand the underlying factors influencing predictions.

- **Improving Trust and Reliability** Trust in AI-driven land surface predictions is crucial for adoption in scientific and policy-making communities. Explainability enhances confidence by ensuring that model decisions are based on meaningful, scientifically valid relationships rather than spurious correlations or artifacts of data processing.

- **Identifying Key Meteorological Drivers** Land surface changes are influenced by various meteorological factors, including temperature, precipitation, and atmospheric pressure. Explainability techniques help uncover the relative impor-

tance of these variables, enabling better understanding of climate change effects on vegetation, soil moisture, and land use patterns.

- **Enhancing Model Performance and Debugging** Interpretable models facilitate error analysis by identifying biases, inconsistencies, and mispredictions. By understanding the reasons behind incorrect predictions, researchers can refine models, improve feature selection, and adjust training processes to enhance overall accuracy and robustness.
- **Supporting Climate Policy and Decision-Making** Transparent AI models provide interpretable insights that can be used by policymakers for environmental planning and land management. Explainability helps in designing better climate adaptation strategies and disaster preparedness measures by revealing the key factors driving land surface evolution.
- **Ensuring Generalizability and Robustness** An interpretable model allows researchers to assess whether learned relationships hold across different geographic regions and temporal scales. Explainability methods help detect overfitting and ensure that models remain generalizable when applied to new, unseen data.

1.2 Problem Statement

Land surface forecasting models leverage vast amounts of high-dimensional spatiotemporal data to predict changes in environmental conditions. However, the inherent complexity of these models poses a significant challenge in terms of interpretability. Existing explainability techniques struggle to disentangle intricate meteorological interactions, making it difficult to understand the factors driving model predictions. This lack of transparency reduces trust in the model's outputs and limits their adoption in scientific research and policy-making.

Furthermore, current explainability methods do not provide a confidence measure for local analyses. An explanation's reliability should be assessed based on the density of training samples in the vicinity and the model's performance on similar instances. Without a quantifiable metric indicating the trustworthiness of an explanation, model interpretations remain uncertain, hindering their effectiveness in critical applications such as climate monitoring, disaster preparedness, and sustainable land management. Addressing these limitations is crucial for enhancing the usability and reliability of land surface forecasting models.

1.3 Methodology Overview

The Cluster-Segregate-Perturb (CSP) pipeline is an explainability framework designed for spatiotemporal land surface forecasting models. It first downsamples high-dimensional spatiotemporal features into interpretable temporal signals using feature disentanglement techniques like β-VAE. Next, it applies clustering to identify patterns and partitions the data through segregation, ensuring that feature importance can be assessed locally before being aggregated globally. Finally, controlled perturbations are introduced at the latent feature level, allowing for meaningful analysis while preserving

spatial and temporal coherence. This structured approach overcomes traditional XAI limitations, enabling robust, interpretable insights into spatiotemporal model predictions.

In this research, we present the development and evaluation of the CSP pipeline for investigative analyses on land surface forecasting models. We demonstrate its effectiveness through empirical evaluations in uncovering the relationships between meteorological variables and land surface evolution via NDVI.

1.4 Research Objectives

The primary objectives of this research are as follows:

1.4.1 Develop an Explainability Pipeline for Land Surface Forecasting Models

This study aims to design and implement a novel explainability pipeline tailored for spatiotemporal forecasting models. The framework will enhance model interpretability by systematically analyzing meteorological influences on land surface evolution, providing insights into how different environmental factors contribute to predictive outcomes.

1.4.2 Address Limitations of Existing Explainability Techniques

Traditional explainability methods, such as Local Interpretable Model-agnostic Explanations (LIME) and Partial Dependence Plots (PDP), face challenges when applied to high-dimensional spatiotemporal datasets. This research seeks to identify these limitations and develop methodologies that mitigate the complexities associated with interpreting land surface forecasting models.

1.4.3 Analyze the Influence of Meteorological Variables on Land Surface Evolution

A key objective is to investigate the intricate relationships between meteorological variables—such as temperature, precipitation, and pressure—and their impact on land surface changes. This study will focus on **Normalized Difference Vegetation Index (NDVI)** as a primary indicator to assess surface evolution, leveraging the CSP pipeline for comprehensive analysis.

1.4.4 Enhance Transparency and Trust in Machine Learning Models

Ensuring the reliability and transparency of machine learning models in climate science is crucial. By integrating explainability techniques, this research aims to provide a structured approach to understanding model predictions, thereby fostering trust and interpretability in land surface forecasting applications.

1.4.5 Design a Model-Agnostic Explainability Technique

To ensure broader applicability, this research seeks to develop an explainability technique that is not limited to land surface forecasting models but can be extended to other spatiotemporal learning tasks. The method will be designed to function independently of specific model architectures, making it adaptable to a wide range of forecasting applications.

1.5 Organization of Thesis

This thesis is structured into eight chapters, each addressing different aspects of building height detection in underdeveloped regions.

- **Chapter 1: Introduction**

  This chapter introduces the research problem, emphasizing the challenges of explainability in land surface forecasting models. As climate change continues to impact terrestrial surfaces, spatiotemporal forecasting models play a crucial role in predicting changes such as vegetation growth and soil moisture. However, interpreting these models remains difficult due to their high-dimensional feature spaces. Traditional explainability methods often fall short in this context, limiting transparency and trust in model predictions. The chapter also presents the Cluster-Segregate-Perturb (CSP) pipeline, a novel explainability approach designed to analyze model predictions by clustering meteorological variables, segmenting data, and applying perturbations. The chapter also outlines the objectives of this research.

- **Chapter 2: Literature Review**

  This chapter reviews the literature on explainability in spatiotemporal models, focusing on the challenges of interpreting predictions in land surface forecasting. While traditional methods like SHAP and Permutation Importance provide insights into feature contributions, they struggle with the high-dimensional and interconnected nature of spatiotemporal data. Recent approaches, such as the SPLT framework and the Temporally-weighted Spatiotemporal Explainable Neural Network (TSEM), attempt to address these issues but are often constrained to structured time series rather than image-based forecasting. Additionally, techniques like Concept Bottleneck Models (CBM) and TimeSHAP enhance interpretability but face limitations in scalability and generalization. This chapter explores these methods, highlighting their advantages, drawbacks, and relevance to spatiotemporal forecasting.

- **Chapter 3: Cluster-Segregate-Perturb (CSP) pipeline**

  This chapter introduces the Cluster-Segregate-Perturb (CSP) pipeline, an explainability framework designed to assess global feature importance in spatiotemporal land surface forecasting models. Since directly applying explainability techniques to high-dimensional spatiotemporal data is impractical, the CSP pipeline first downsamples the data into interpretable temporal signals, ensuring computational efficiency while preserving fidelity. The chapter discusses the methodology behind CSP and explains how interpretability is maintained across its three

stages: clustering, segregation, and perturbation. Additionally, it highlights key challenges, such as ensuring that the disentangled latent space retains essential spatiotemporal information, maintaining temporal consistency in perturbations, preventing excessive segmentation that may lead to overfitting, managing computational overhead in recursive clustering, and addressing the lack of standard metrics for validating explainability in spatiotemporal forecasting models.

- **Chapter 4: Case Study**

  This chapter demonstrates the application of the Cluster-Segregate-Perturb (CSP) pipeline on a land surface forecasting model. The study employs a ConvLSTM model trained on the EarthNet2021 dataset, a large-scale collection of satellite images and meteorological data. The chapter details the dataset characteristics, preprocessing steps and model architecture, also highlighting on the three steps of the pipeline. The clustering step groups meteorological variables based on their temporal trends using techniques such as K-Means, experimenting with different metric like Dynamic Time Warping (DTW), Soft-DTW, Euclidean distance and Shape-based Distance (SBD). A novel GoodClusterScore (GCS) metric is introduced to optimize cluster selection, ensuring meaningful feature grouping. In the segregation phase, the dataset is recursively partitioned based on the most informative features, using GCS as the partitioning criterion. Additionally several meta data information is also used for meaningful segregation. This step allows for localized feature importance analysis, ensuring robust interpretation across different spatial and temporal conditions. Finally, the perturbation step introduces controlled modifications to the disentangled latent representations, ensuring smooth and semantically meaningful variations in the reconstructed data.

- **Chapter 5: Marginal Sensitivity Analysis**

  This chapter focuses on evaluating the sensitivity of a model's predictions to individual meteorological variables. The study employs the Enhanced Vegetation Index (EVI) instead of NDVI due to its advantages in high-biomass regions and its ability to mitigate atmospheric and soil background effects. Sensitivity analysis is conducted by perturbing one variable while keeping others constant, enabling the isolation of each variable's contribution to the system. Additionally, pairwise perturbation interactions are analyzed to expand the scope of comparisons. The study identifies challenges in sensitivity evaluation for time-series variables, as predictions are influenced by both current and past inputs. To address this, a Partial Autocorrelation Function (PACF) analysis is performed, revealing a dominant influence of lag-1 values. The chapter further proposes subtracting the autoregressive (AR(1)) component from the EVI signals to remove linear dependencies and better isolate the direct effects of meteorological variables.

- **Chapter 6: Conclusion, Limitations, and Future Work**

  This chapter concludes the study by presenting the Cluster-Segregate-Perturb (CSP) pipeline as an effective explainability framework for spatiotemporal land surface forecasting models. The chapter also outlines key limitations, including computational complexity, instability in clustering, potential loss of spatial granularity, and the lack of standardized evaluation metrics for spatiotemporal models. To address these challenges, future work will focus on improving scalability, refining clustering strategies, integrating physics-based constraints, and

expanding the applicability of CSP to broader domains such as biomedical imaging and financial time series forecasting.

- **Appendix A: Cluster Charts**

  The appendix provides a visual representation of temporal cluster patterns formed at various values of cluster size (K) during the clustering process. These clusters were generated using the K-Means clustering algorithm with multiple distance metrics, including Dynamic Time Warping (DTW), Soft-DTW, and Euclidean distance. Additionally, the K-Shape clustering method was explored, and its corresponding cluster formations are also illustrated, offering insights into how different clustering techniques capture temporal dependencies in the data.

2. Literature Review

2.1 Clarifying the Meaning of Spatiotemporal Data

Misinterpretation of Spatiotemporal Data: One common issue in understanding spatiotemporal data is the assumption that it solely refers to **geotagged temporal data**, such as time-series sensor readings from different locations. However, in domains such as *computer vision* and *remote sensing*, spatiotemporal data primarily refers to a **sequence of images over time**, where both spatial and temporal patterns contribute to the data representation.

2.2 Recent Advances in Explainable AI (XAI) for Spatiotemporal Models

While there are no standardized explainability techniques specifically designed for spatiotemporal models, various attempts have been made to bridge this gap with varying degrees of success. Explainable AI (XAI) for spatiotemporal data is an emerging research area that aims to interpret and analyze complex relationships within spatiotemporal predictive models, ensuring transparency and trustworthiness.

2.2.1 Spatiotemporal XAI Technique for SPLT

The work by Huang et al. [12] introduces explainability techniques for Spatiotemporal Predictive Learning Tasks (SPLT), focusing on disentangling complex spatiotemporal dependencies in predictive models. The authors identify key challenges in understanding motion formation and develop a framework for analyzing learned representations.

Methodology

The proposed explainability techniques involve:

- **Component Synthesis:** The model synthesizes multiple independent components to analyze the individual contributions of spatial and temporal features to predictions.
- **State Decomposition and Expansion:** A decomposition technique is introduced to separate intertwined signals within spatiotemporal dynamical systems, enabling a better understanding of motion patterns.
- **Collaboration Mechanism:** The study presents the *Extending the Present and Erasing the Past* (EPEP) principle, which explains motion formation by emphasizing new observations while diminishing past influences.

Mathematical Formulation

Given a spatiotemporal predictive learning model $\mathcal{M}$ that maps an input sequence $X_t = \{x_1, x_2, ..., x_t\}$ to a future state Y_{t+1}, the authors define the decomposition process as:

$$X_t = X_t^{(S)} + X_t^{(T)} \tag{2.1}$$

where $X_t^{(S)}$ represents the spatial components and $X_t^{(T)}$ represents the temporal evolution. The model then applies a learned transformation function $\Phi(\cdot)$:

$$Y_{t+1} = \Phi(X_t^{(S)}, X_t^{(T)}) \tag{2.2}$$

To interpret the contribution of past states, they propose a temporal expansion function:

$$X_t^{(T)} = f(X_{t-k}, ..., X_t) \tag{2.3}$$

where $f(\cdot)$ captures historical dependencies.

Advantages:

- Provides insights into motion formation in spatiotemporal models.
- Separates spatial and temporal components for better interpretability.
- Introduces a novel framework for understanding spatiotemporal dynamics.

Limitations:

- Primarily designed for predictive learning tasks, limiting generalization to other spatiotemporal problems.
- Requires predefined decomposition strategies, which may not generalize across datasets.
- The method assumes clear separability of spatial and temporal components, which might not always be feasible.

2.2.2 Temporally-weighted Spatiotemporal Explainable Neural Network (TSEM)

The TSEM model, proposed by Pham et al. [13], is a novel explainability framework for multivariate time series forecasting that integrates spatiotemporal dependencies. It aims to enhance interpretability while maintaining high predictive accuracy by leveraging recurrent and convolutional architectures.

Methodology

TSEM introduces a unique weighting mechanism that integrates temporal importance into the explainability process. The key components of TSEM are:

- **Recurrent-CNN Hybrid Architecture:** It combines Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to effectively capture spatial and temporal dependencies in multivariate time series.
- **Temporal Weighting Mechanism:** Unlike traditional models, TSEM assigns dynamic importance scores to different time steps, allowing the model to prioritize critical temporal information.
- **Explainability Metrics:** TSEM provides interpretability through feature importance quantification, ensuring alignment with causality, fidelity, and spatiotemporality criteria.

Mathematical Formulation

Given a multivariate time series $\mathbf{X} = \{X_1, X_2, ..., X_T\}$, where $X_t \in \mathbb{R}^d$ represents d features at time t, TSEM learns a mapping function:

$$\hat{Y}_{t+1} = f(\mathbf{X}_t, W_t) \tag{2.4}$$

where W_t is a learned temporal weight matrix that emphasizes important time steps:

$$W_t = \text{softmax}(h_t) \tag{2.5}$$

where h_t is the hidden state derived from the RNN component, determining the temporal importance at each step. The CNN extracts spatial patterns across features, while the temporal weighting function ensures explainability.

Advantages:

- Integrates temporal weighting for improved interpretability.
- Captures both spatial and temporal dependencies in multivariate time series.
- Meets key explainability criteria, including causality and fidelity.

Limitations:

- Requires additional computational overhead due to the weighting mechanism.
- Limited generalization to datasets with highly irregular time steps.
- The interpretability mechanism is tied to the model architecture, reducing flexibility for other neural network designs.

Despite these advancements, a unified and robust XAI framework specifically designed for spatiotemporal datasets remains an open research challenge, necessitating further exploration of novel interpretability techniques.

TSEM and Huang et al.'s SPLT framework, while designed for spatiotemporal data, are primarily suited for structured multivariate time series rather than sequences of satellite images used in land surface forecasting. These methods focus on feature attribution and temporal dependency modeling but lack mechanisms to capture pixel-wise spatial correlations essential for image-based forecasting. TSEM relies on RNNs to weight temporal features, which does not account for the spatial structure inherent in gridded remote sensing data. Similarly, SPLT's decomposition approach is designed for disentangling latent temporal signals rather than explaining localized spatial dynamics.

2.3 Traditional XAI Methods for Evaluating Global Feature Importance

2.3.1 Permutation Importance (PIMP)

Permutation Importance (PIMP) [14] is a method designed to correct biases in feature importance measures, particularly those arising in models like Random Forests. Traditional permutation importance methods can be influenced by feature characteristics, such as the number of categories, leading to biased importance scores. PIMP addresses this by normalizing feature importance through permutation testing, providing significance p-values for each feature.

Methodology

The PIMP algorithm involves the following steps:

1. **Compute Observed Importance:** Train the model on the original dataset and calculate the importance score, I_j^{obs}, for each feature j.

2. **Permute Outcome:** Randomly permute the outcome vector (target variable) to break any associations between features and the outcome.

3. **Compute Null Importances:** For each permuted dataset, retrain the model and compute the importance scores, $I_{j,k}^{\text{perm}}$, for each feature j in permutation k.

4. **Estimate Null Distribution:** Repeat steps 2 and 3 multiple times to build a distribution of importance scores under the null hypothesis (no association between features and outcome).

5. **Calculate *p*-values:** For each feature j, determine the proportion of permuted importance scores that are greater than or equal to the observed importance score:

$$p_j = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(I_{j,k}^{\text{perm}} \geq I_j^{\text{obs}}) \tag{2.6}$$

   where $\mathbb{I}(\cdot)$ is the indicator function, and K is the total number of permutations.

Advantages

- **Bias Correction:** By estimating the null distribution of importance scores, PIMP corrects biases inherent in traditional importance measures, leading to more accurate feature rankings.
- **Significance Testing:** Provides p-values for each feature, facilitating statistical inference regarding feature relevance.
- **Model-Agnostic:** Applicable to any machine learning model that allows for the computation of feature importance scores.

Disadvantages

- **Computational Intensity:** The need to retrain the model on multiple permuted datasets increases computational requirements, which can be demanding for complex models or large datasets.
- **Assumption of Independence:** Permuting the outcome assumes independence between features and the target under the null hypothesis, which may not hold in all scenarios.

2.3.2 Concept Bottleneck Models (CBM)

Concept Bottleneck Models (CBM) [15] are a class of explainable machine learning models that introduce an explicit intermediate layer representing human-understandable concepts. Instead of making direct predictions from raw input features, CBMs first map inputs to a set of predefined concepts and then use these concepts to make the final prediction. This enables transparency and interpretability by allowing users to inspect and manipulate the intermediate concept representations.

Working Principle

CBM follows a two-stage process to ensure interpretability:

1. **Concept Learning:** The model learns to map raw input features X to a predefined set of interpretable concepts C:

$$C = g(X; \theta_c), \tag{2.7}$$

   where g is a neural network parameterized by θ_c that predicts a set of k concepts $C = \{c_1, c_2, ..., c_k\}$.

2. **Prediction from Concepts:** The learned concepts are then used to make the final prediction:

$$Y = f(C; \theta_y), \tag{2.8}$$

   where f is another neural network parameterized by θ_y, responsible for predicting the target variable Y.

Types of CBM

CBMs can be implemented in three different ways depending on how they handle the predicted concepts:

- **Independent CBM:** Assumes concepts are independently predicted and directly used for classification.
- **Sequential CBM:** Introduces dependencies between concepts by modeling their relationships sequentially.
- **Iterative CBM:** Refines the predicted concepts through multiple iterations to improve accuracy and robustness.

Training Objective

The CBM is trained by minimizing a joint loss function that consists of two terms:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_Y, \tag{2.9}$$

where:

- $\mathcal{L}_C$ is the concept prediction loss, typically a binary cross-entropy or mean squared error loss for predicting concepts.
- $\mathcal{L}_Y$ is the final prediction loss, such as cross-entropy loss for classification.

Advantages of CBM

- **Interpretability:** Provides explicit human-understandable concepts for decision-making.
- **Intervention Capability:** Users can manually intervene and correct predicted concepts to influence final predictions.
- **Debugging and Trust:** Allows users to verify whether the model is relying on meaningful features.

Limitations of CBM

- **Concept Annotation Cost:** Requires a predefined set of human-interpretable concepts, which may be expensive to annotate.
- **Concept Bottleneck Constraint:** Forces the model to rely only on the predefined concepts, potentially limiting performance if key features are missing.
- **Concept Correlation Issues:** If concepts are highly correlated, the model may struggle to separate their contributions effectively.

2.3.3 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) [16] is a game-theoretic approach to explain the output of machine learning models by assigning an importance value to each feature. It is based on the concept of Shapley values from cooperative game theory, ensuring a fair distribution of contribution among features. SHAP provides both local (instance-specific) and global (model-wide) interpretability.

Working Principle

Given a machine learning model f and an input instance x, SHAP assigns an importance value ϕ_i to each feature x_i based on the expected change in model output when that feature is included in different subsets of features.

Shapley Value Calculation

The Shapley value for a feature i is computed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{2.10}$$

where:

- F is the set of all features.
- S is a subset of features excluding i.
- $f(S)$ represents the model's output when only features in S are considered.
- $|S|$ and $|F|$ denote the cardinalities of sets S and F, respectively.
- The fraction is a weighting term ensuring fair contribution across all possible subsets.

Types of SHAP Explanations

SHAP provides multiple methods for efficient approximation of Shapley values:

- **Kernel SHAP**: A model-agnostic method that approximates Shapley values using a weighted regression approach.
- **Tree SHAP**: An optimized algorithm for tree-based models like decision trees and gradient boosting models.
- **Deep SHAP**: A method adapted for deep neural networks, leveraging connections to DeepLIFT.
- **Linear SHAP**: An approximation method designed for linear models with additive feature attributions.

Training Objective

SHAP values satisfy several desirable properties:

- **Efficiency**: The total sum of Shapley values matches the difference between the full model output and the expected prediction.
- **Symmetry**: Features with the same contribution receive identical importance scores.
- **Additivity**: Contributions from individual features combine linearly.
- **Null Effects**: If a feature does not affect the model, its SHAP value is zero.

Advantages of SHAP

- **Fair Feature Attribution**: Ensures an unbiased and mathematically justified distribution of feature importance.
- **Local and Global Explanations**: Provides explanations at both instance and overall model levels.
- **Model-Agnostic and Model-Specific Variants**: Can be used with any machine learning model, with optimized versions for trees and deep networks.
- **Handles Feature Interactions**: Accurately captures interactions between input features.

Limitations of SHAP

- **Computational Complexity**: The exact computation of Shapley values is factorial in complexity, making it infeasible for large feature sets.
- **Approximation Errors**: Approximations like Kernel SHAP can introduce inaccuracies, particularly in highly complex models.
- **Dependence on Background Distribution**: The choice of background samples affects the computed SHAP values.
- **Aggregated Global Importance**: It averages local attributions, leading to loss of dynamic relationships. It can be dominated by rare extreme cases, making global SHAP values to be inflated.

2.3.4 TimeSHAP: Explainability for Temporal Models

TimeSHAP [17] is an adaptation of the SHapley Additive exPlanations (SHAP) framework, specifically designed for temporal datasets. It provides local and global feature importance scores while considering the sequential dependencies inherent in time-series data. TimeSHAP applies Shapley values to determine the contribution of individual time steps and features, allowing for transparent and interpretable model decisions over time.

Methodology

TimeSHAP builds on the SHAP framework but adapts it to sequential data by incorporating temporal perturbations. The method involves:

- **Temporal Feature Attribution:** Computes the marginal contribution of each time step to the model's output.
- **Feature Importance per Time Step:** Assesses the importance of each feature across different time steps to understand its dynamic influence.
- **Survival-Based Pruning:** Introduces a pruning mechanism to reduce computational complexity by eliminating features and time steps with low contributions.

The Shapley value for a given feature i at time t is computed as:

$$\phi_{i,t} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}, t) - f(S, t) \right] \tag{2.11}$$

where:

- F is the set of all features.
- S is a subset of features excluding i.
- $f(S, t)$ represents the model's prediction when only the subset S at time t is considered.

Advantages

- Captures both temporal and feature-wise importance.
- Provides local and global explanations for temporal models.
- Reduces computational cost using survival-based pruning.

Limitations

- Computationally expensive for long sequences.
- Requires assumptions about feature independence, which may not always hold in real-world time-series data.
- Works well for models with single-output predictions but struggles with multi-output models (e.g., sequence forecasting models like Transformers, LSTMs with attention).
- Provides local explanations (instance-specific), but aggregating them to obtain global feature importance is non-trivial and computationally expensive.
- Perturbation-based explanations may be sensitive to noise.

2.3.5 Temporal Importance (TIME) Method

Temporal Importance Model Explanation (TIME) [18] is a model-agnostic approach designed to explain temporal black-box models by assessing global feature importance in time-series data. It evaluates the significance of features concerning their temporal ordering and localized windows of influence, providing insights into how features impact model predictions over time.

Methodology

The TIME method involves the following key steps:

- **Temporal Permutation Testing:** Systematically permutes feature values within specific temporal windows to assess their impact on the model's predictions.
- **Hypothesis Testing:** Employs statistical tests to determine the significance of observed changes in model performance due to feature perturbations.
- **Temporal Ordering Assessment:** Evaluates whether the sequence of feature values over time influences the model's predictions, identifying features where temporal order is crucial.

The importance score $I(f, w)$ for a feature f within a temporal window w is calculated as:

$$I(f, w) = \frac{1}{|D|} \sum_{(x,y)\in D} \left[L(y, \hat{y}_x) - L(y, \hat{y}_{x'_{f,w}}) \right] \tag{2.12}$$

where:

- D is the dataset of instances (x, y).
- $\hat{y}_x$ is the model's prediction for instance x.
- $\hat{y}_{x'_{f,w}}$ is the model's prediction for instance x with feature f permuted within window w.
- L denotes the loss function used to measure prediction error.

Advantages

- **Model-Agnostic:** Applicable to any temporal model without requiring access to internal parameters.
- **Temporal Sensitivity:** Captures the importance of features concerning their temporal context and ordering.
- **Statistical Rigor:** Utilizes hypothesis testing to provide statistically significant explanations.

Limitations

- **Computational Complexity:** Permutation-based approach can be computationally intensive, especially for large datasets.
- **Assumption of Feature Independence:** Assumes that permuting a feature does not disrupt the inherent temporal dependencies, which may not always hold true.

2.4 Challenges of Existing XAI Techniques in Assessing Global Feature Importance for Spatiotemporal Models

- **Phantom Samples & Perturbation Artifacts:** Perturbation-based methods like Permutation Importance (PIMP) and Temporal Importance (TIME) often disrupt spatial or temporal coherence, creating unrealistic samples that do not exist in real-world scenarios. This not only distorts model explanations but also increases computational overhead, making them impractical for large-scale spatiotemporal datasets.
- **Lack of True Global Insights:** Many techniques, such as SHAP and TimeSHAP, primarily generate local explanations and aggregate them to infer global feature importance. However, this aggregation often leads to inflated or misleading representations that fail to capture broader patterns or long-term dependencies in spatiotemporal data.
- **Architectural & Domain Specificity:** Methods like Concept Bottleneck Models (CBM) require high-level, human-interpretable concepts, which are often not well-defined in spatiotemporal contexts. Similarly, techniques such as TSEM, designed for recurrent architectures, or the framework proposed by Huang et al. [12] for predictive learning tasks, lack the generalizability required to be directly applied to diverse spatiotemporal forecasting models.
- **Limited Applicability to Models with Temporal Input & Output:** Many temporal XAI techniques, including TimeSHAP, are designed for models with temporal inputs but scalar outputs. However, spatiotemporal forecasting models, such as those used in land surface prediction, involve both temporal inputs and temporally evolving spatial outputs, making adaptation of these techniques challenging and non-trivial.
- **Computational Complexity & Scalability:** Methods like SHAP and TimeSHAP require repeated random perturbations across high-dimensional feature spaces. In large-scale spatiotemporal datasets, such as satellite-based land surface forecasting models, this results in prohibitive computational costs, rendering these techniques impractical.
- **Assumption of Feature Independence:** Many conventional XAI techniques assume feature independence, which does not hold for spatiotemporal models where features are inherently correlated across both space and time. This leads

to misleading attributions and poor interpretability when applied to highly structured data, such as sequential satellite imagery.

- **Inability to Capture Spatial-Temporal Interactions:** Existing XAI techniques often struggle to disentangle spatial and temporal dependencies. While some methods focus on temporal feature importance, they fail to incorporate spatial variations, and vice versa. This results in incomplete explanations for spatiotemporal forecasting models that rely on both dimensions.

To effectively interpret spatiotemporal forecasting models, novel XAI frameworks must be developed that respect spatial and temporal coherence, efficiently handle high-dimensional data, and account for the interdependencies between spatial and temporal features.

3. Cluster-Segregate-Perturb (CSP) pipeline

3.1 Introduction

From the literature review, it is evident that developing a spatiotemporal explainability technique that operates directly on high-dimensional data appears to be the ideal solution for interpreting land surface forecasting models. However, the significant computational challenges associated with such methods make this approach nearly impractical. Instead, a more viable alternative is to transform the data into a lower-dimensional representation while ensuring that the transformed features remain interpretable.

This dimensionality reduction not only addresses computational constraints but also simplifies several key challenges in explainability, such as maintaining coherent perturbations, ensuring scalability, extracting reliable global insights, and achieving model-agnostic applicability. By focusing on interpretable transformations, we can enhance the robustness and efficiency of explainability techniques without sacrificing fidelity in feature attributions.

With these considerations in mind, we propose Cluster-Segregate-Perturb (CSP)—an explainability pipeline designed to assess global feature importance in spatiotemporal land surface forecasting models. This chapter details the methodology behind CSP, its underlying principles, and its advantages over conventional approaches in spatiotemporal explainability.

3.2 Downsampling Spatiotemporal Features into Temporal Signals

Downsampling a spatiotemporal feature into one or more temporal signals is essential for computational efficiency. The choice of downsampling technique depends on the dataset characteristics and the need to retain key spatial variations while reducing dimensionality. Below are several methods for transforming spatiotemporal data into meaningful temporal signals:

- **Global Pooling (Mean, Max, Min, Median, Variance, etc.)** : Aggregates spatial data at each timestamp into a single or a few representative values.
- **Spatial Clustering + Temporal Averaging** : Groups spatial regions based on similarity using clustering techniques (e.g., k-means, DBSCAN). Computes per-cluster temporal signals instead of per-pixel signals, preserving region-specific variations.
- **Feature Weighting via PCA or Autoencoders** : Reduces spatial dimensions while maintaining the most informative variations. Principal components or latent representations act as lower-dimensional temporal signals.

- **Fixed Grid Averaging (Block-wise Aggregation)** : Divides the spatial domain into fixed-size grids (e.g., 2×2, 4×4) and aggregates within each grid cell.Generates multiple temporal signals corresponding to different spatial partitions.

- **Adaptive Grid Aggregation** : Uses adaptive spatial partitioning techniques (e.g., quadtrees, Voronoi partitioning) to dynamically adjust aggregation areas based on feature distribution. Ensures that highly variable regions retain finer details while smoother regions undergo more aggressive aggregation.

- **Spatio-Temporal CNNs or Transformers** : Uses deep learning models to extract compact, meaningful representations of spatial features while maintaining temporal dependencies.

3.2.1 Feature Disentanglement for Robust Downsampling and Explainable Perturbations

The goal of downsampling is not only to reduce dimensionality but also to generate **interpretable** temporal signals that can be **upsampled back** to the original spatiotemporal representation. This constraint is crucial for applying meaningful perturbations before feeding it back to the spatiotemporal model during explainability analysis, as discussed in the *Perturbation Section*. However, ensuring a reversible transformation while maintaining spatiotemporal coherence limits the applicability of many standard downsampling techniques across diverse datasets.

Feature disentanglement using VAE [19] or β-VAE [20] applied to each frame of spatiotemporal data can produce interpretable features for individual frames. These interpretable features enable the reduction of spatiotemporal data into a sequence of temporal features illustrated in 3.1. Furthermore, disentanglement ensures a continuous and structured feature space, allowing for meaningful and interpretable perturbations on the features before upsampling them back to the original dimension. This, in turn, facilitates a range of investigative analyses within the CSP pipeline, making it more robust and versatile for spatiotemporal data analysis.

Training Channel-wise VAE for Spatiotemporal Data

To effectively extract interpretable temporal signals from high-dimensional spatiotemporal data, we propose a **channel-wise VAE** that operates on individual frames, learning a disentangled representation of the spatial features over time. Each frame of the spatiotemporal sequence is encoded into an n**-dimensional latent space**, where each latent dimension represents an independent temporal signal. These latent representations are concatenated across frames to form a sequence of n **temporal features**, effectively reducing the complexity of the spatiotemporal input while preserving meaningful temporal dynamics.

The training process as shown in 3.2 involves **two loss components** to ensure both faithful reconstruction and task-relevant feature extraction:

1. **Reconstruction Loss**: Ensures that the VAE learns an efficient encoding that can accurately reconstruct the original input.

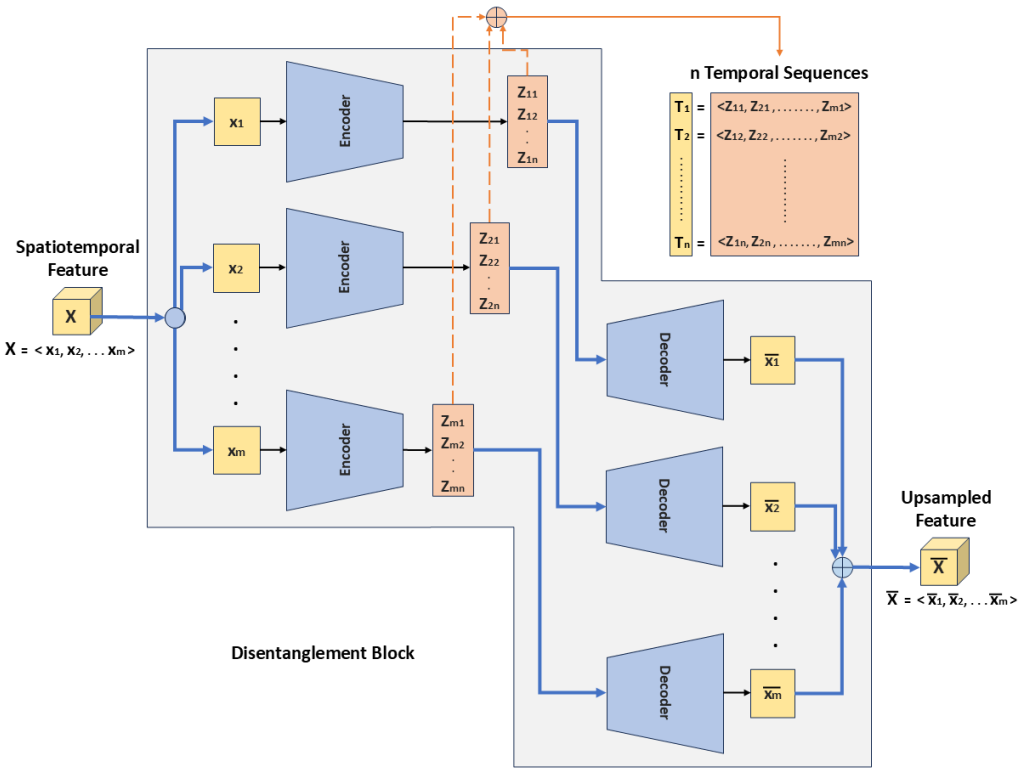


Figure 3.1: Illustrates the disentanglement block for a spatiotemporal feature, all the encoders share weights similarly all the decoders share weights

2. **Downstream Task Loss**: Ensures that the latent features encode meaningful representations essential for the final predictive task. The downstream model, which is the primary subject of the explainability analysis, is typically pre-trained, and only the forward pass is executed to compute the loss. This allows for efficient evaluation of how well the extracted features align with the model's decision-making process.

Given a sequence of spatiotemporal data frames $X = \{X_1, X_2, ..., X_T\}$, where $X_t \in \mathbb{R}^{C \times H \times W}$ (with C channels, H height, and W width), the VAE operates at the **channel level** for each frame. The encoder maps each frame X_t to a latent space Z_t:

$$Z_t = \text{Enc}(X_t), \quad Z_t \in \mathbb{R}^n \tag{3.1}$$

The latent representations across time are concatenated to form n **temporal sequences**:

$$\mathcal{Z} = \{Z_1, Z_2, ..., Z_T\}, \quad \mathcal{Z} \in \mathbb{R}^{T \times n} \tag{3.2}$$

The decoder reconstructs each frame from its latent representation:

$$\hat{X}_t = \text{Dec}(Z_t) \tag{3.3}$$

The **total loss function** consists of two terms:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{task}} \tag{3.4}$$

Reconstruction Loss: Ensuring fidelity to the input.

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{X_t \sim p(X)} \left[\|X_t - \hat{X}_t\|^2 \right] \tag{3.5}$$

Downstream Task Loss: Ensuring task-relevant representations.

Let f_{task} be a downstream model predicting target Y from the decoded frames:

$$\mathcal{L}_{\text{task}} = \mathbb{E} \left[\mathcal{L}_{\text{downstream}}(f_{\text{task}}(\hat{X}), Y) \right] \tag{3.6}$$

The β**-VAE KL divergence term** ensures feature disentanglement by enforcing a prior on the latent distribution:

$$\mathcal{L}_{\text{KL}} = D_{KL}(q(Z|X) \| p(Z)) \tag{3.7}$$

Thus, the final loss function is modified as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{task}} + \beta \mathcal{L}_{\text{KL}} \tag{3.8}$$

where $\beta > 1$ strengthens the disentanglement effect.

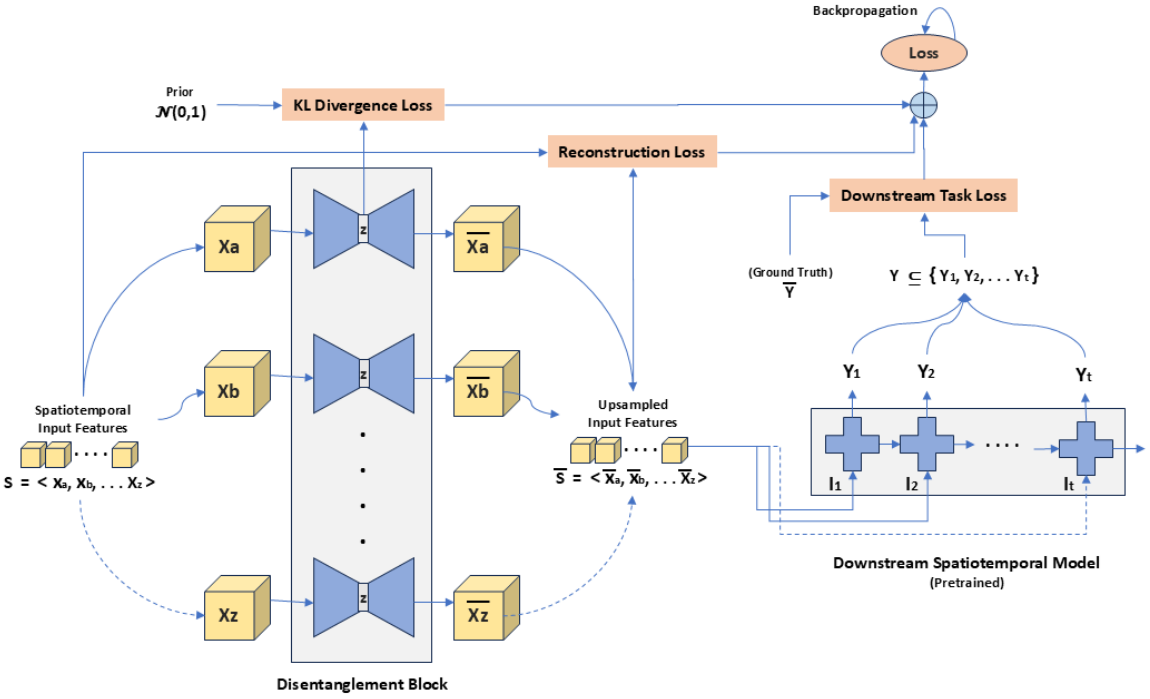


Figure 3.2: Illustrates the training process of VAEs for disentanglement

By enforcing feature disentanglement via β**-VAE**, we transform spatiotemporal data into a **sequence of interpretable temporal signals**, reducing computational complexity while maintaining explainability. The integration of a downstream task loss ensures that the latent representations remain **semantically meaningful**, making them ideal for subsequent modeling and analysis in spatiotemporal forecasting tasks.

After downsampling, each spatiotemporal feature is represented by n temporal sequences. The next step involves human intervention to identify a subset of distinct,

independent, and meaningful latent sequences. Since these sequences retain interpretability, they can be properly labeled by humans, ensuring their semantic relevance even after downsampling. This process is applied to all features, resulting in a refined set of m interpretable temporal features, which can then be fed into the CSP pipeline.

3.2.2 Strengths

- **Interpretable Features**: By disentangling features for each frame, VAE or β-VAE creates a more accessible representation for downstream tasks. This aligns well with the CSP pipeline's requirement for explainable and continuous features.
- **Dimensionality Reduction**: Transforming complex spatiotemporal data into a sequence of temporal features simplifies clustering and analysis without compromising interpretability.
- **Robust Perturbations**: A structured and continuous latent space ensures that perturbations are both meaningful and representative, which is vital for sensitivity studies and exploring correlations with model outputs.

3.2.3 Potential Consideration

- **Domain-Specific Validations**: Ensuring that disentangled features remain meaningful and valid in the context of the domain is crucial. This may require expert input or additional validation metrics.
- **Training the β-VAE**: Training a β-VAE for spatiotemporal data is highly nontrivial due to several factors. First, achieving complete disentanglement in the latent space is inherently challenging, as the model may struggle to separate independent factors of variation in a purely unsupervised manner. Second, the choice of β plays a crucial role—while higher values enforce stronger disentanglement, they may also degrade reconstruction quality, leading to a trade-off between interpretability and fidelity.
- **Scalability and Computational Complexity** : Encoding each frame into a disentangled representation and processing temporal sequences significantly *increases computational cost* compared to simpler dimensionality reduction methods. Training β-VAE requires careful tuning of hyperparameters (e.g., β weight, latent space size) and can be computationally expensive for large datasets.
- **Information Loss from Downsampling** : Although downsampling via disentanglement retains core temporal structures, some *fine-grained spatial details* might be lost. If critical information is not captured in the latent representation, certain perturbations may not accurately reflect their real-world impact.

3.3 Cluster

Clustering involves Identifying the inherent and distinct patterns within each input feature by grouping similar data points together. In our context, where spatiotemporal data has been downsampled into a set of temporal features, clustering specifically

refers to temporal clustering, the process of grouping time-series representations based on their shared characteristics and dynamics over time. This allows for a more structured and interpretable understanding of the temporal variations within the data.

Since our downsampling process results in a set of temporal signals, we can leverage well-established signal processing techniques to effectively cluster these signals. Temporal clustering methods aim to group similar time-series patterns, capturing shared dynamics and trends across different temporal features. Some commonly used techniques for clustering temporal signals include:

- **K-Means for Time-Series** – Applied with dynamic time warping (DTW) or Euclidean distance to group similar temporal patterns.
- **Dynamic Time Warping (DTW) Clustering** – Aligns time-series sequences by computing minimal warping distances.
- **Hierarchical Clustering** – Constructs a dendrogram to capture hierarchical relationships between temporal signals.
- **Spectral Clustering** – Uses graph-based methods to cluster time-series based on similarity in the frequency domain.

3.3.1 Evaluation

To evaluate the quality of the clustering, we devised two complementary metrics: IntraClusterScore (Cohesion) and InterCentroidScore (Separation).

intraClusterScore

The **intraClusterScore** focuses on signals that deviate the most from their assigned cluster centroids. After assigning each signal to a cluster based on the model's predictions, it computes the mean squared Euclidean distance between the signal and its cluster center. these distances are sorted in descending order, and only the top 25% (the farthest signals) are considered. This approach highlights the worst-performing samples in each cluster. Finally these mean distances are then scaled by the proportion of signals belonging to each cluster, accounting for cluster size in the evaluation. This scaled mean loss provides a more balanced metric for clusters of varying sizes. The smaller this value, the tighter the signals fit around their cluster centers.

$$\begin{aligned} intraClusterScore_K = & \sum_{i=1}^{K} \frac{\sum_{j=1}^{j<0.25|d_i|} ||centroid_i^K - d_i[j]||^2 * N_i}{N} \\ & \forall_{(i,j)}\ d_k[i] >= d_k[j]\ when\ i < j \end{aligned} \tag{3.9}$$

interCentroidScore

The **interCentroidScore** quantifies the smallest separation between clusters, offering a worst-case perspective on the distinctiveness of the clustering. A higher value implies better separation and, therefore, better-defined clusters. If the minimum distance is

very small (or zero), it indicates that some clusters are close to or even overlapping, which could signify poor clustering or redundant clusters.

$$interCentroidScore_K = min(\ \forall_i \forall_{j(i \neq j)}\ ||centroid_i^K - centroid_j^K||^2\ ) \tag{3.10}$$

Here K is the cluster size, d_i is the sorted list (descending) of samples assigned to the cluster with centroid $centroid_i^K$ where $i \in \{1, 2....K\}$. In Equation 3.9 $\frac{N_i}{N}$ is the scaling factor. N refers to the total number of data samples, and N_i refers to the count of samples belonging to cluster i.

GoodClusterScore

The **GoodClusterScore** combines the intra-cluster and inter-cluster metrics to provide an overall assessment of the clustering quality. It typically evaluates the balance between cohesion (low intra-cluster scores) and separation (high inter-cluster scores).

$$GoodClusterScore_K = \frac{interCentroidScore_K}{intraClusterScore_K} \tag{3.11}$$

1. A **High** GoodClusterScore indicates well-defined clusters, where the samples within each cluster are tightly grouped around their centroid, demonstrating high cohesion, as reflected by low intraClusterScores. Simultaneously, the clusters are well-separated and distinct, as indicated by high interCentroidScores, signifying minimal overlap. This score implies a high confidence in the quality of the clustering.

2. A **Moderate** GoodClusterScore suggests a trade-off between cohesion and separation. Some clusters may exhibit strong cohesion but overlap with others, or vice versa. This balance may indicate suboptimal clustering or inherent overlap in the dataset, pointing to potential areas for refinement.

3. A **Low** GoodClusterScore, on the other hand, reflects poor clustering performance. High intraClusterScores suggest that the samples within clusters are loosely grouped, while low interCentroidScores highlight significant overlap or poorly separated clusters.

Discriminative Power of GCS(GoodClusterScore)

The **GoodClusterScore (GCS)** can provide insights into the **discriminative power** of different temporal features because it quantifies both **cohesion** (how well signals fit within their clusters) and **separation** (how distinct clusters are from each other).

A **higher** GoodClusterScore indicates well-separated, tightly packed clusters, implying high discriminability. A **lower** value suggests overlapping or loosely structured clusters, meaning the features lack discriminative power.

- If temporal features are **highly discriminative**, the **interCentroidScore** will be large (clusters are well separated), and the **intraClusterScore** will be small (tight clusters).

- Conversely, if the features **lack discriminative power**, clusters will overlap (low $interCentroidScore$) or exhibit high intra-cluster variation (high $intraClusterScore$), leading to a low $GoodClusterScore$.

Trend of GCS Across K-sized clusters also provides insights into the overall discriminative power of a feature.

- If **GoodClusterScore remains high for different K**, it suggests that the variables inherently separate the data well.
- If **GoodClusterScore drops significantly after a certain K**, it implies that increasing the number of clusters does not improve discrimination and may introduce redundancy.

3.3.2 Challenges

- **High Computational Complexity**: As the dataset size or sequence length increases, performing direct clustering on raw temporal sequences becomes increasingly computationally expensive leads to significant memory and processing overhead.
- **Choice of Distance Metrics**: Standard metrics (e.g., Euclidean distance) often fail to capture the nuances of spatiotemporal data. Often the temporal component requires metrics that account for time-series behavior (e.g., DTW, Frechet distance) while the spatial component requires metrics sensitive to spatial arrangements. Combining spatial and temporal similarities into a unified measure is non-trivial.

3.4 Segregate

Segregation involves partitioning the dataset into meaningful segments where the temporal features within each segment exhibit similar trends. This segmentation process enables more refined analyses by ensuring that patterns are preserved within each segment. Additionally, external metadata such as geolocation, climate type, sea level height, and population density in the case of satellite images can provide auxiliary information to further refine segmentation. This localized segmentation is crucial for evaluating robust feature importance at a granular level, which can then be aggregated globally in an informed manner.

3.4.1 Clustering-Based Recursive Segmentation

Segmentation using temporal features inherently aligns with clustering, drawing parallels with decision tree methodologies. The process is structured as follows:

1. **Initial Clustering:** Temporal features are first clustered using an unsupervised method. The optimal number of clusters (K) is determined using the GoodClusterScore (GCS), which balances intra-cluster cohesion and inter-cluster separation.

2. **Recursive Partitioning:** For each feature, we select the K that maximizes the GoodClusterScore:

$$K^* = \arg\max_K \text{GoodClusterScore}_K \tag{3.12}$$

The feature with the highest GCS is then chosen as the partitioning feature.

3. **Partitioning Criterion:** Instead of traditional entropy-based criteria, we use GCS to recursively partition the data:

$$\mathcal{D}_{t+1} = \{X_i | X_i \in \text{Cluster}_j, j \in \{1, 2, ..., K^*\}\} \tag{3.13}$$

where $\mathcal{D}_{t+1}$ represents the dataset at the next iteration of partitioning.

4. **Stopping Condition:** The partitioning process terminates when no remaining feature can attain a maximum GCS for any cluster size K that meets or exceeds the predefined GCS threshold set at the start.

$$\text{GoodClusterScore}_K \geq \tau \tag{3.14}$$

where τ is the threshold set to prevent excessive fragmentation while maintaining interpretability.

This recursive cluster and segmentation algorithm 3.3 ensures that the data is partitioned in a structured and interpretable manner, making the segmentation process analogous to decision tree construction but with clustering-based criteria.

3.4.2 Challenges

- **Cluster Stability:** Since clustering is performed at each recursive step, small variations in the data can lead to different partitions, making the segmentation process sensitive to noise. Ensuring cluster stability across different iterations is crucial for meaningful segmentation.
- **Optimal Cluster Selection:** Determining the appropriate number of clusters (K^*) at each step relies on the GoodClusterScore (GCS). However, if multiple features yield similar GCS values, selecting the best partitioning feature may introduce uncertainty, potentially leading to suboptimal segmentation.
- **Computational Complexity:** Recursive partitioning involves repeatedly performing clustering and computing GCS scores, leading to exponential growth in computational requirements as the dataset size and temporal sequence length increase.
- **Over-Segmentation Risk:** If the stopping criterion is too lenient (i.e., a low threshold τ for GCS), the process may lead to excessive fragmentation, resulting in too many small segments with limited statistical significance. Conversely, a high threshold may prevent meaningful segmentation. Finding the right balance is crucial.

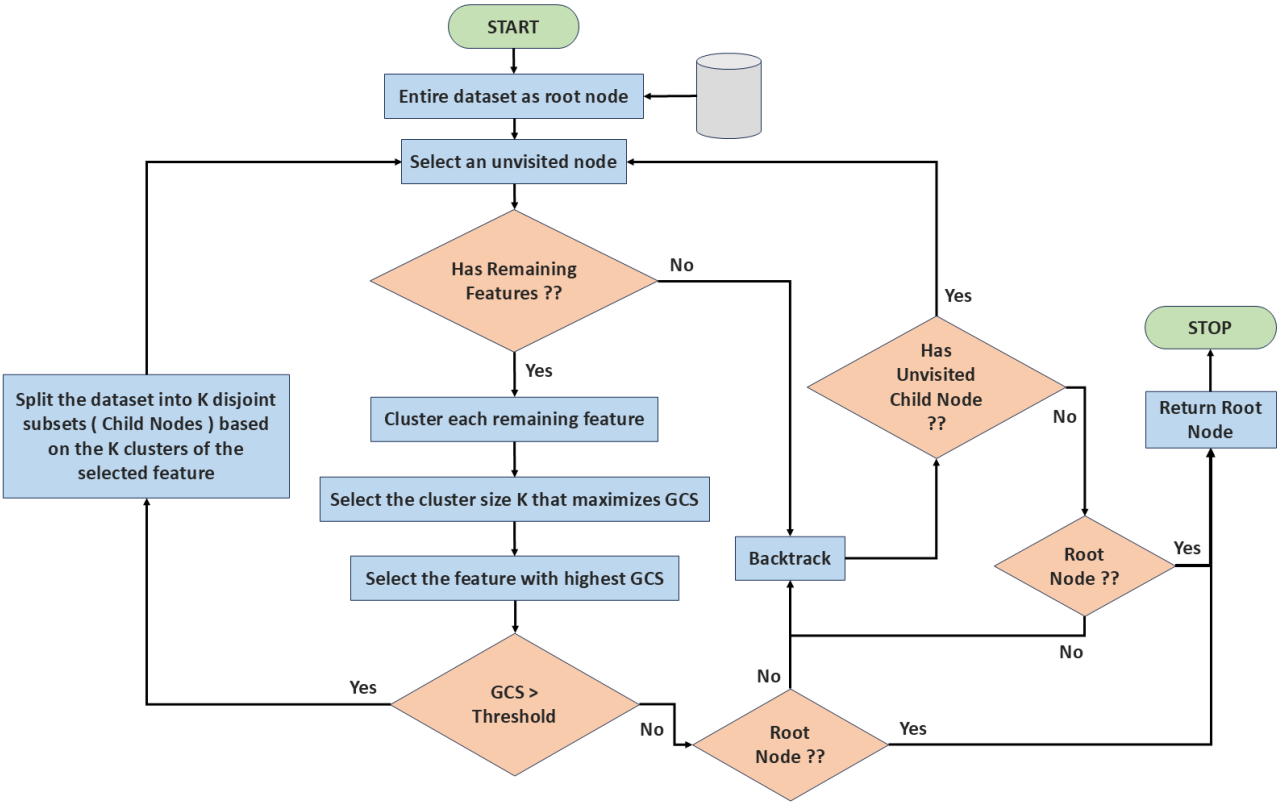


Figure 3.3: The flow chart illustrates the cluster-segregate algorithm

- **Dependency on Metadata:** While auxiliary metadata (e.g., geolocation, climate type) can enhance segmentation, it may also introduce bias if not properly accounted for. Metadata-driven segmentation must be carefully validated to ensure that partitions are meaningful and not artificially constrained.
- **Interpretability vs. Accuracy Trade-off:** While recursive segmentation aims to enhance interpretability, deeper partitioning may lead to highly localized clusters that do not generalize well. A balance must be struck between producing interpretable segments and maintaining global coherence in feature analysis.

3.5 Perturbation

Perturbation plays a crucial role in explainability by systematically modifying input features to assess their impact on model predictions. However, in high-dimensional spatiotemporal data, direct perturbations can disrupt spatial and temporal coherence, leading to unrealistic modifications. Traditional perturbation-based explainability techniques struggle with this complexity, limiting their applicability.

3.5.1 Challenges of Perturbation-Based XAI in High-Dimensional Spatiotemporal Data

- **Limited Perturbation Flexibility:** Raw spatiotemporal data (e.g., sequences of images) only allows random perturbations, which are highly constrained in

their numeric range.

- **Disruption of Spatial and Temporal Coherence:** Large perturbations disrupt either spatial coherence (smoothness and structure in an image) or temporal coherence (continuity over time), making meaningful perturbations nearly impossible.
- **Failure of Perturbation-Based XAI Models:** Due to these limitations, most perturbation-based explainability (XAI) models struggle to provide meaningful insights in high-dimensional settings.
- **Downsampling Trade-offs:** Traditional downsampling techniques reduce dimensionality but often sacrifice interpretability. This makes it challenging to generate meaningful perturbations while ensuring that the resulting modifications remain semantically valid.

To effectively extract interpretable temporal signals from high-dimensional spatiotemporal data, we propose a *channel-wise* β-VAE that operates on individual frames, learning a disentangled representation of spatial features over time. Instead of processing the full spatiotemporal sequence directly—where perturbations often disrupt spatial or temporal coherence—our approach encodes each frame into an n-dimensional latent space, ensuring that each latent dimension represents an independent temporal signal. These latent representations are then concatenated across frames, forming a sequence of n temporal features.

3.5.2 Why Feature Disentanglement Enhances Interpretability

A feature disentanglement model is particularly well-suited for downsampling because its continuous latent space allows for smooth and interpretable perturbations. Unlike raw spatiotemporal data, where random perturbations are both limited in range and likely to break coherence, a disentangled latent space ensures that even small perturbations result in consistent, structured changes in the reconstructed data.

Moreover, representing spatiotemporal features as sequences of temporal features (rather than single scalars) allows for controlled perturbations when upsampling. By modifying the temporal sequence corresponding to a feature, we can directly influence spatiotemporal channels in a structured way, ensuring consistency across different resolutions and reinforcing the validity of feature attribution.

This approach overcomes the limitations of traditional perturbation-based XAI methods, which struggle with high-dimensional data due to the difficulty of performing meaningful, interpretable modifications without disrupting spatial or temporal coherence.

3.5.3 Challenges

- **Latent Space Smoothness and Stability** : Although the latent space is continuous, there is no guarantee that small perturbations always result in *semantically meaningful* changes. If the model learns an entangled or non-smooth latent space, perturbations may lead to *discontinuous or unrealistic reconstructions*, undermining explainability.

- **Temporal Consistency in Perturbations** : While the latent space enforces structure, ensuring that perturbations maintain *long-term temporal consistency* remains a challenge. Modifying a feature at one timestep could have unintended effects on future frames due to the *recurrent nature* of temporal dynamics.
- **Lack of Explicit Supervision** : The model learns *unsupervised latent representations*, meaning there is no direct guarantee that each dimension corresponds to an interpretable real-world factor. Without a way to *align latent dimensions with meaningful physical attributes*, interpretability could still be limited.
- **Difficulty in Evaluating Explainability** : Measuring the *effectiveness of perturbations* in disentangled space is not straightforward. Traditional XAI evaluation metrics may not directly apply, and domain-specific validation methods may be required.

While the proposed approach overcomes fundamental limitations of perturbation-based XAI in high-dimensional spatiotemporal data, addressing the above challenges is essential for ensuring robustness, interpretability, and scalability. Careful *model selection, hyperparameter tuning, and evaluation strategies* will be crucial in deploying this method effectively.

4. Case Study

To demonstrate the application of the CSP pipeline to a land surface forecasting model, we employed a ConvLSTM model trained on the EarthNet21 dataset.

4.1 Dataset and Model Overview

4.1.1 EarthNet2021 - Dataset

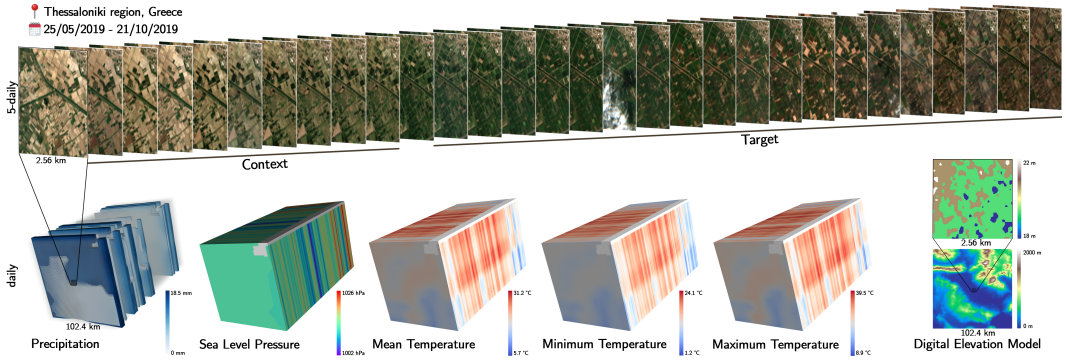


Figure 4.1: Dramatic Visualization of one of the over 32000 samples in EarthNet2021.

EarthNet2021 [21] is a large-scale dataset and challenge for Earth surface forecasting, which involves predicting satellite imagery conditioned on future weather.
The Dataset consists of 32,000 samples within the European region, each comprising a series of 30 Sentinel-2 images, each captured at intervals of 5 days. These images contain four bands (red, green, blue, and near-infrared) with a spatial resolution of 128x128px or 2.56 km^2 and a ground resolution of 20m. Moreover, accompanying weather-related meteorological data is included, such as precipitation, sea level pressure, and temperature (minimum, maximum, and mean), each comprising a series of 150 images for 150 days, at a coarser spatial resolution of 80x80px or 102.4 km^2, sourced from the observational dataset E-OBS [22].
Before the model training, as a preprocessing step, the spatiotemporal resolution of meteorological variables is matched to that of the Sentinel-2 images.

This study used the IID (In-Domain) train set fraction of the EarthNet2021 dataset denoted by D containing 23904 samples denoted by N in our analyses.

$$\begin{aligned} D &= \{x_1,\ x_2,\ x_3,\ \dots\ x_N\} \\ x_i &= (T_{\text{avg}_i},\ T_{\text{min}_i},\ T_{\text{max}_i},\ P_i,\ R_i) \end{aligned} \tag{4.1}$$

x_i is a data sample T_{avg_i}, T_{min_i}, T_{max_i}, P_i, R_i are the 30 timesteps spatiotemporal channels representing average temperature, minimum temperature, maximum temperature, pressure and precipitation respectively each $\in \mathbb{R}^{(30*128*128)}$. Additionally, other channels representing DEM, red, $blue$, $green$, $near-infrared$, $cloud\ mask$, $scene\ classification\ label$, and $data\ quality\ mask$ are also utilized during model training and inference.

4.1.2 ConvLSTM

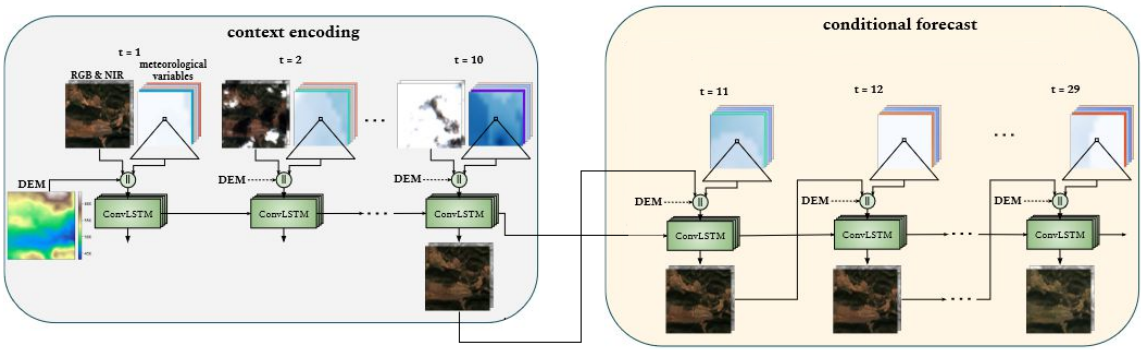


Figure 4.2: The training procedure for ConvLSTM involves encoding 10 four-band context images along with additional inputs. These additional inputs include the five meteorological inputs, which are first cropped and then upscaled, as well as the DEM, which is repeatedly incorporated as input. Using the encoded context, the next 20 images are predicted sequentially, with predictions conditioned on the two provided inputs.

ConvLSTM (Convolutional Long Short-Term Memory) networks have been proposed as effective methods for remote sensing time series analysis. These networks leverage the temporal and spatial contextual information present in time series images to improve classification accuracy. It was first used for precipitation nowcasting in [23] since then it has been used for tasks such as land cover classification, change detection, and time series reconstruction.

In this study, we utilized the ConvLSTM model, as described by Diaconu et al. [24]. We trained the model on the IID (In-Domain) split of the EarthNet2021 dataset for 60 epochs. During this training phase, we attained an EarthNetScore of 0.3257, closely aligning with the score reported in the original paper, 0.3266. Here the EarthNetScore (ENS) is a composite evaluation metric used for assessing the performance of Earth surface prediction models it is the harmonic mean of the four components (MAD, OLS, EMD, SSIM), scaled between 0 (worst) and 1 (best) as described in [21].

4.2 Clustering the Meteorological Variables

4.2.1 Variability Study

Before applying feature disentanglement to the spatiotemporal meteorological variables, it is crucial to conduct a spatial variability analysis for each variable. This analysis ensures that pixel values exhibit minimal spatial variation, allowing some variables to be downsampled into temporal signals using central tendency measures. For the EarthNet21 dataset, each sample represents a small geographical area of just 2.56 km^2, making it an ideal candidate for such an analysis as smaller areas are less likely to exhibit significant spatial heterogeneity, for meteorological variables like temperature, pressure, precipitation. We conducted the study on the entire dataset, and the cumulative probability distributions of the standard deviations are presented in Figure 4.3.

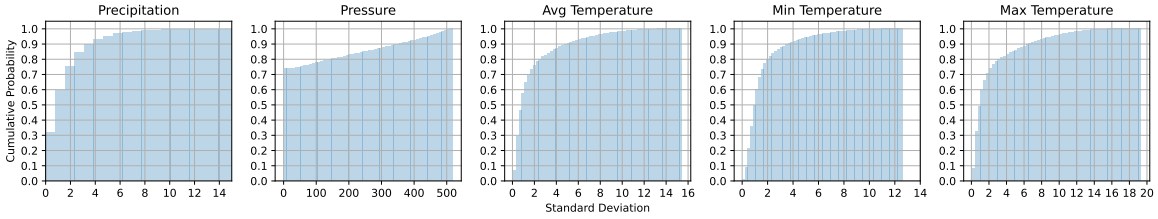


Figure 4.3: The figure shows the distribution of standard deviations across the spatial resolution at each time step of the meteorological variables.

- The **temperature** variations observed are generally consistent with urban heat island (UHI) effects, where localized areas experience temperature increases due to anthropogenic heat sources. A 3-4°C variation in temperature is plausible due to the UHI phenomenon, but such distributions are relatively unusual unless specific localized factors, such as parks or industrial zones, are present.
- Regarding **precipitation**, 90% of the samples show a standard deviation of up to 3 mm, with only 60% within 1 mm and 75% within 2 mm. the distribution indicates that precipitation events exhibit moderate spatial variability, possibly due to convective rainfall, common in Central and Western Europe during summer months.
- Regarding **pressure**, the analysis reveals significant outliers in pressure variability, which are highly unusual for a small spatial area of 2.56 km^2. Typically, pressure remains uniform at such scales, with expected variations of only 1–2 hPa under normal conditions. The extreme values observed (reaching up to 400 hPa) likely point to data quality issues or preprocessing artifacts. To address this, central tendency measures like the median should be preferred over the mean when downsampling.
- *It is important to highlight that the meteorological data in the EarthNet2021 dataset is provided as a time series of 150 images over 150 days. The data is sourced from the E-OBS observational dataset, with a spatial resolution of 80x80 pixels (approximately 102.4 km^2 per pixel), which may also contribute to the aggregation of data at a relatively coarse spatial scale. As such, the observed variability in the dataset may be influenced by both local geographical factors and the resolution at which the data is provided.*

Due to low to Moderate spatial variability, we opted to downsample the meteorological variables to a single average value per time step as shown in Figure 4.4. This approach aligns with our goal of clustering based on temporal trends and underlying patterns over time, emphasizing long-term signals. Additionally, it simplifies the clustering process by focusing exclusively on temporal clustering, rather than first performing feature disentanglement followed by temporal clustering of relevant features.

4.2.2 Preprocessing

Before downsampling the meteorological variables, we performed spatiotemporal alignment of the dataset. Each EarthNet2021 sample minicube includes Sentinel-2 images

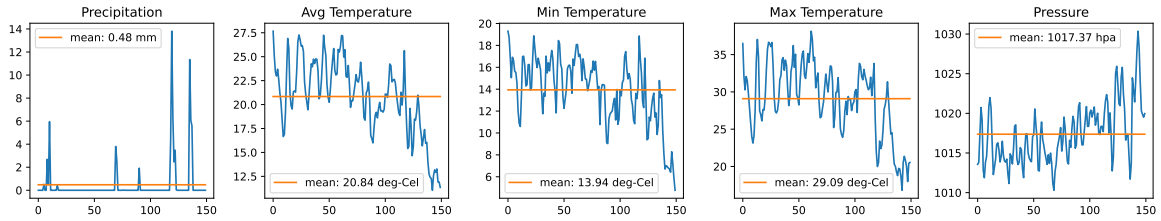


Figure 4.4: The figure shows the downsampled version of the meteorological variables of one of the minicube.

captured over 150 days, with 30 images provided (approximately one every 5 days). In contrast, mesovariables such as precipitation, temperature, and pressure are available daily, resulting in 150 images. To harmonize the temporal resolution of the mesovariables with the Sentinel-2 images, we aggregated the mesovariables into 5-day intervals, as shown in Figure 4.5. Specifically, precipitation values were summed over each 5-day period, while temperature and pressure values were averaged. This process produced 30 mesovariable images, aligning their temporal resolution with that of the 30 Sentinel-2 images.

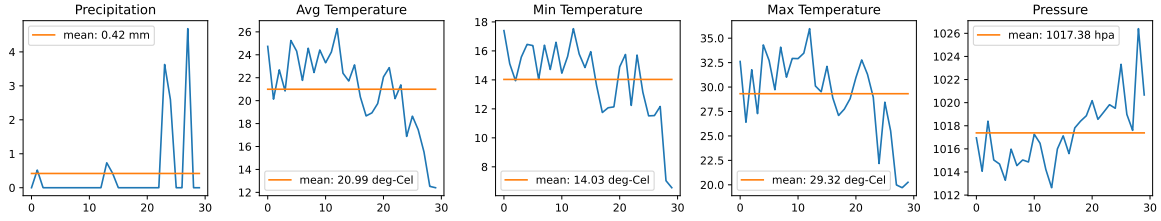


Figure 4.5: Meteorological variables of one of the minicube, downsampled and aggregated from 150 days to 30 days

4.2.3 Clustering Techniques

We standardized the time series using min-max normalization and applied Gaussian smoothing with a gamma value of 1.1 to highlight the underlying trends. To identify the most effective approach, we experimented with four distinct techniques:

1. K-means clustering with Euclidean distance as the metric, focusing on straightforward spatial proximity.

2. K-means clustering with Dynamic Time Warping (DTW), capturing temporal alignments and accounting for time shifts.

3. K-means clustering with Soft-Dynamic Time Warping (Soft-DTW), introducing a smoother variant of DTW to handle minor variations robustly.

4. K-shape clustering with Shape-based Distance (SBD), specifically designed to group time series based on their overall shapes and trends.

Based on the observed scores 4.6, the **k-means with Euclidean distance** as the metric appears to be the most effective technique among those evaluated, whereas **K-shape** is the worst performing one. This observation is further supported by visual inspection at A.1 compared to others.

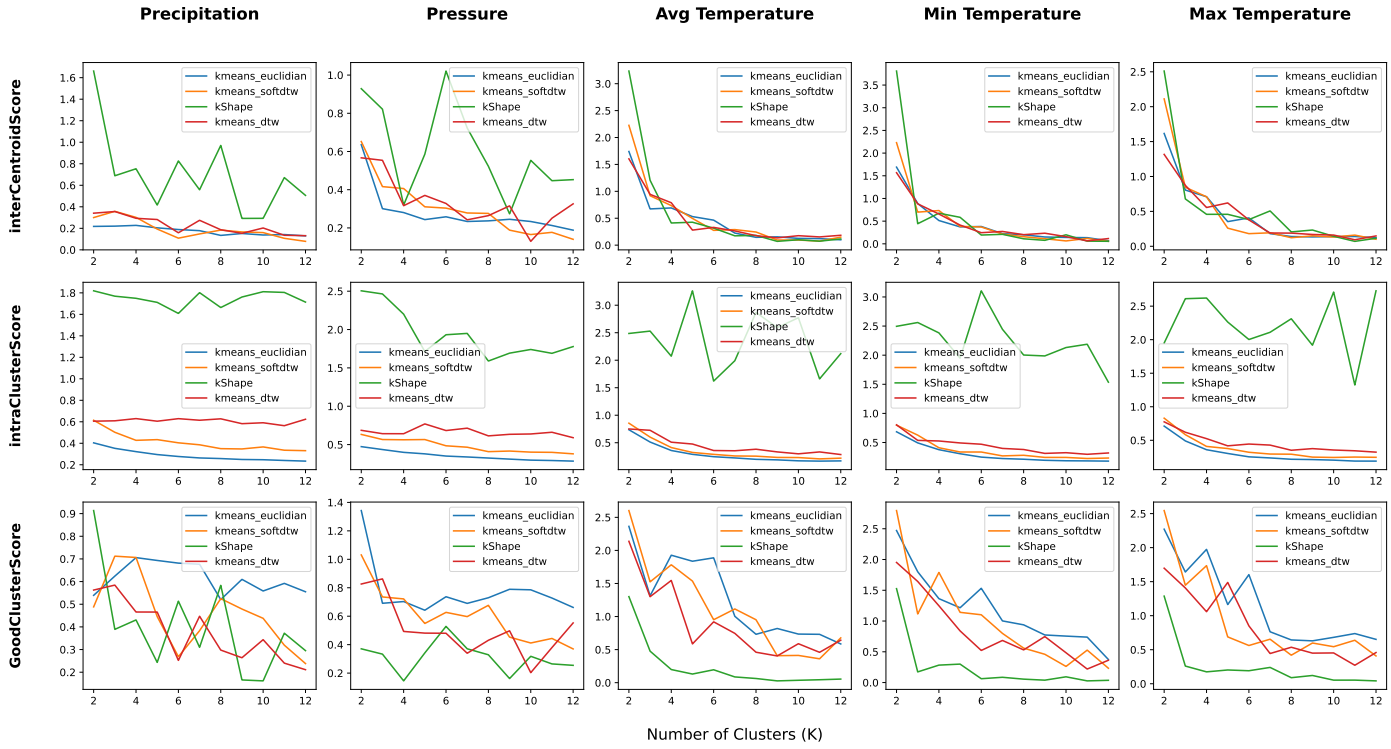


Figure 4.6: The figure shows the performance of different clustering techniques based on interCentroidScore, intraClusterScore and GoodClusterScore for different values of K

4.2.4 Discriminative Power of the Meteorological Variables

GoodClusterScore is not only useful for determining the optimal number of clusters (K), but it's trend also provides insights into the overall discriminative power of different variables in segregating the dataset.

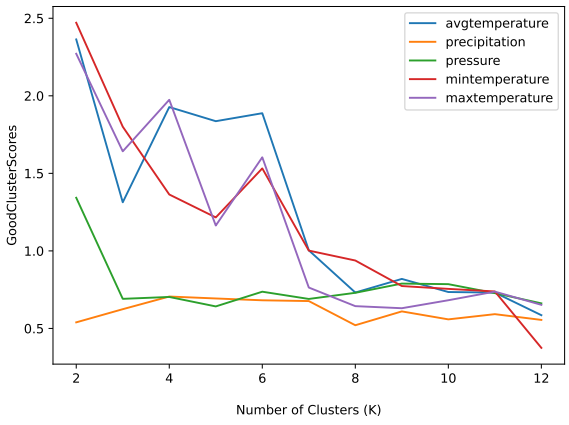


Figure 4.7: GoodClusterScore trends for meteorological variables. on K-means clustering with Euclidean distance

As observed in the graphs 4.7, precipitation's GoodClusterScore primarily ranges from 0.5 to 0.7 with overall constant trend with (k), while for pressure, it falls between 0.6 and 0.9, indicating better discriminative power than precipitation. In contrast, the temperature variable shows a range of 1.0 to 2.0, before dropping dramatically after k = 6, suggesting that it has the highest discriminative power among the three variables. This finding is further supported by the observed seasonal patterns:

- **Temperature** in Western Europe exhibits clear seasonal trends (e.g., warmer summers, colder winters) that are fairly predictable year after year.

- **Atmospheric pressure** is influenced by large-scale weather patterns like the North Atlantic Oscillation (NAO) and the shifting positions of high- and low-pressure systems, which tend to exhibit seasonal patterns (e.g., higher pressures in summer and lower pressures in winter). However, these patterns are not as consistently predictable as temperature.
- **Precipitation** in Western Europe can vary significantly year to year and is influenced by localized weather systems, atmospheric circulation patterns, and ocean currents (e.g., North Atlantic Oscillation).

4.3 Segregation

4.3.1 Cluster(s) based Categorization

The above analysis highlights that precipitation and pressure lack the discriminative power of temperature to effectively Segment the dataset. For temperature, the temporal cluster patterns for average, minimum, and maximum values are nearly identical (see A). Thus, we focus solely on average temperature to select the optimal K. From Figure 4.6, $K = 4, 5, 6$ are strong candidates, while $K = 2$ is excluded despite its high GoodClusterScore due to an excessively high intraClusterScore.

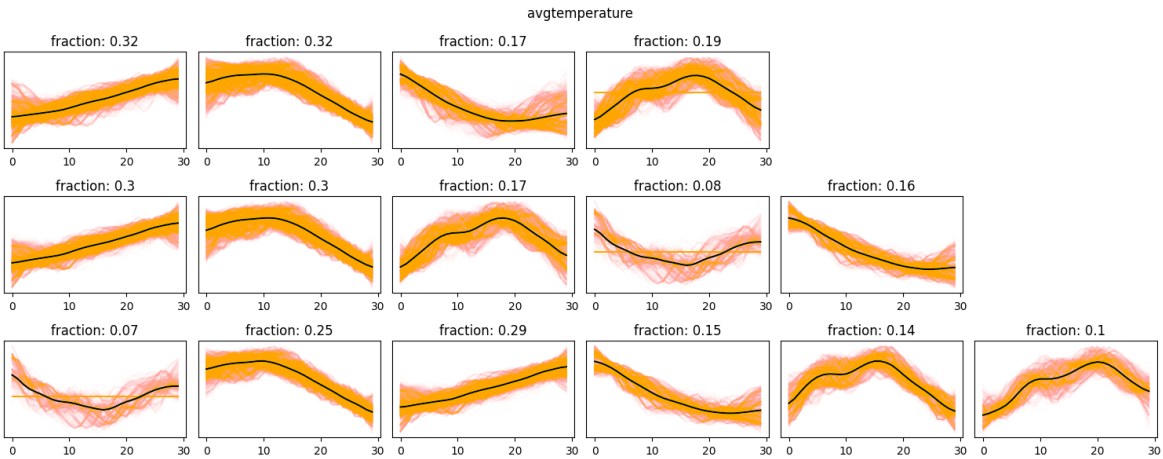


Figure 4.8: The figure illustrates the average temperature patterns identified using the most effective method, K-means clustering with Euclidean distance, for K=4,5,6, which yielded the highest GoodClusterScore.

Although selecting $K = 6$ for its high GoodClusterScore may initially seem appealing, studying the temperature trends in Figure 4.9, reveals a significant overlap between certain clusters. Clusters 1 and 3 show considerable overlap in terms of months, with cluster 3 occurring only 1 to 1.5 months before cluster 1. Similarly, clusters 5 and 6 exhibit a similar pattern, with cluster 6 preceding cluster 5 by about a month. This overlap decreases the distinctiveness of the clusters.

To address this, we selected $K = 4$, which not only minimizes overlap but also provides clearly defined trends across the data. Furthermore, $K = 4$ aligns closely with the four distinct seasons defined by temperature trends in western Europe, as expressed in Table 4.1, improving the interpretability of the results. This seasonal alignment allows for better corroboration with existing studies, many of which analyze

patterns on a seasonal basis, thereby strengthening the reliability and relevance of our findings.

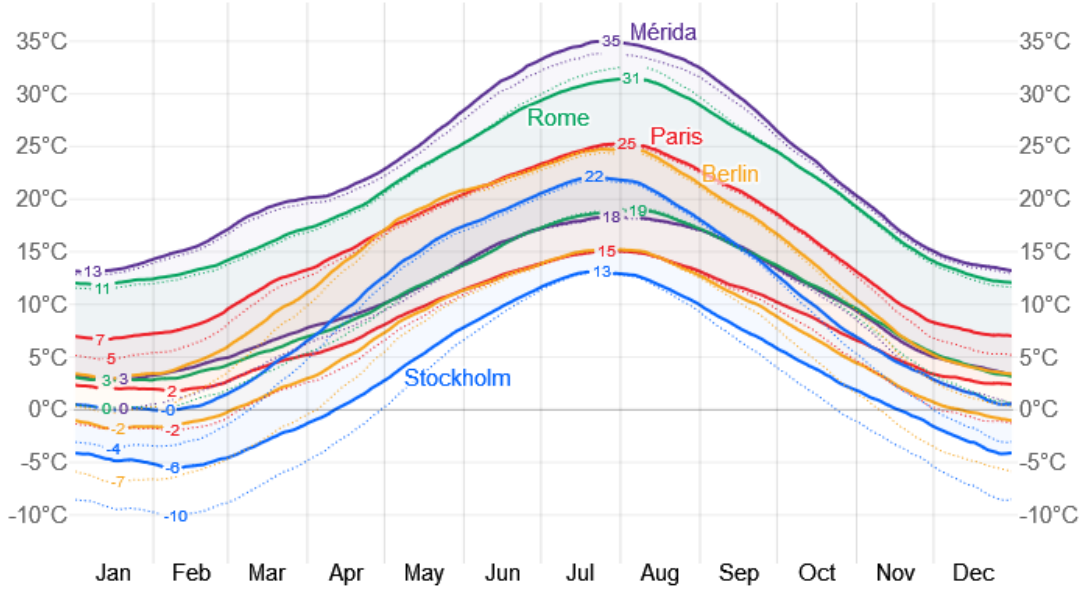


Figure 4.9: Comparison of the Average Weather in Mérida, Rome, Paris, Stockholm, and Berlin. Source: WeatherSpark.com.

Season ($K = 4$) 4.8	**Months**	**Features (Temperature Focus)**
Winter (3)	December–February	Coldest season, temperatures often below freezing in some areas, potential for snow
Spring (1)	March–May	Gradual warming, temperatures rising from cool to mild, marked by variability
Summer (4)	June–August	Warmest season, temperatures can reach highs of 25–35°C, occasional heatwaves
Autumn (2)	September–November	Gradual cooling, temperatures transition from warm to cool

Table 4.1: Seasonal Temperature Trends in Western Europe

4.3.2 Country based Categorization

The minicubes in the EarthNet2021 dataset are organized within folders labeled according to the MGRS (Military Grid Reference System) subgrid. For example, a folder named '29SND' represents a Sentinel-2 tile located at longitude 29° and latitude South, within the ND subquadrant (100 km x 100 km). This systematic arrangement enables precise localization of each MGRS subgrid on the map, along with identifying its country of origin as shown in Figure 4.10.

The distribution of minicubes in the EarthNet dataset 4.10, reflects a highly uneven representation of countries, which could impact the generalizability and fairness of models trained on this dataset.

- **Highly Skewed Representation:** Countries like Spain (6421) and Portugal (4801) dominate the dataset, accounting for a significant portion of minicubes.

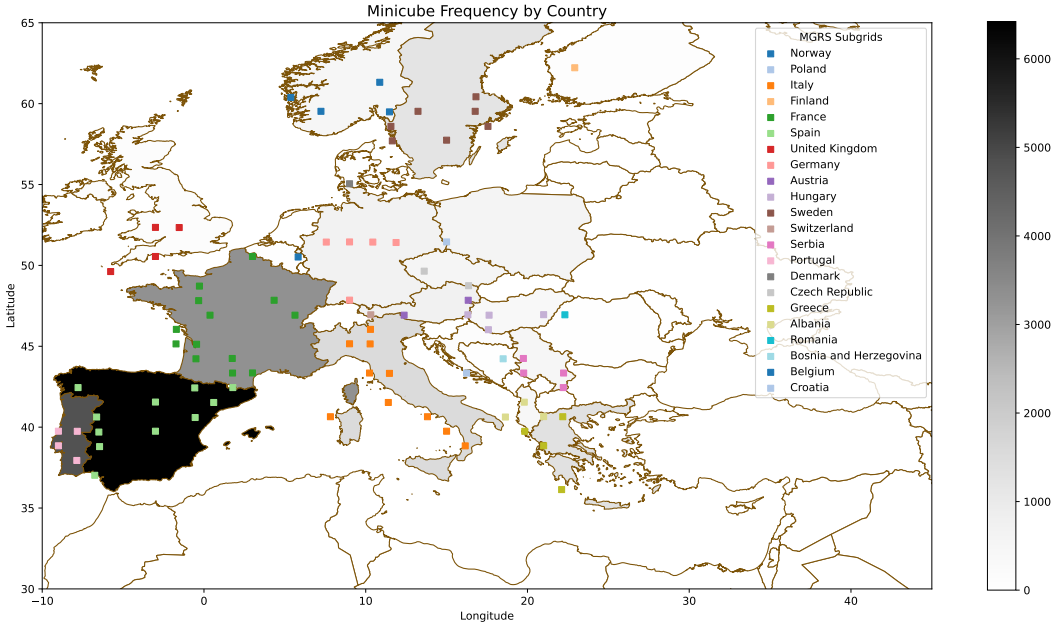


Figure 4.10: MGRS subgrid locations and minicube distribution. *Cube markers indicate the top-left coordinate corner of MGRS subgrids, not their true size, as the original subgrids are significantly larger.*

Underrepresentation is observed in countries such as Switzerland (41), Belgium (57), and Denmark (73).

- **Regional Imbalance:** Southern European countries (e.g., Spain, Portugal, Italy, Greece) have disproportionately high representation.Northern and Central European countries (e.g., Switzerland, Belgium, Denmark) are underrepresented.
- **Possible Bias in Model Training:** Models may perform better in regions with higher minicube representation (e.g., Southern Europe). Bias can affect tasks involving geographically varying physical or ecological properties.

Restricting Segmentation to the country level could lead to biased results due to the significant disparity in the distribution of minicubes across countries. Additionally, country boundaries are artificial constructs and do not necessarily align with natural geographic or climatic divisions. For instance, in a large country with two minicubes located far apart, the weather patterns and scenes they represent could differ significantly. This highlights the need for a more generalized and geography-aware segmentation approach

4.3.3 Köppen Climate based Categorization

The Köppen climate classification is a widely used system for classifying the Earth's climates based on temperature and precipitation patterns. Developed by the German climatologist Wladimir Köppen in 1884, it divides climates into five primary types, each represented by a letter:

A – **Tropical**, characterized by warm temperatures year-round and heavy rainfall.
B – **Arid (Dry)**, with very little precipitation, including deserts and steppes.
C – **Temperate**, featuring mild temperatures and moderate seasonal changes.

D – Continental, marked by large temperature variations and cold winters.
E – Polar, which experiences extremely cold temperatures and minimal precipitation.

The EarthNet2021 dataset exhibits diversity but is significantly imbalanced. Using the Köppen climate classification, we can meaningfully segment the samples into six distinct geographical locales, as visualized in Figure 4.11.

1. **Csa (Mediterranean, Hot Summer)**: Hot, dry summers, mild, wet winters. This category dominates the dataset, especially due to Spain and Portugal, which together account for over 11,000 samples. Countries like Croatia and Albania are moderately represented but pale in comparison to the massive skew caused by Spain and Portugal. This uneven distribution heavily favors Mediterranean climates, potentially biasing any analysis or model training.

2. **Cfb (Temperate Oceanic)**: Mild summers, cool winters, year-round rainfall. France is the most represented, contributing over 3,000 samples, followed by Austria and Germany. Smaller countries like Belgium and Denmark have fewer samples, leading to underrepresentation. While this category has better balance than Csa, the dominance of France still skews the representation within temperate oceanic climates.

3. **Cfa (Humid Subtropical)**: Hot summers, no dry season. This category is moderately represented but largely concentrated in two countries, Czech Republic and Hungary, with combined sample size of 939. The representation is reasonable compared to other categories but lacks broader geographic diversity

4. **Dfb (Humid Continental, Warm Summer)**: Warm summers, cold winters. Countries: Sweden and Poland are the key contributors, with Sweden alone accounting for over 1,200 samples. Romania, however, is significantly underrepresented with only 91 samples, creating an imbalance within this category. The representation is skewed toward Northern Europe.

5. **Dfc (Subarctic)**: Short summers, long cold winters. Countries: This cold climate type is dominated by Norway with 505 samples, with Finland contributing only 118 samples. The representation of subarctic climates is limited.

6. **ET (Tundra)**: Extremely cold, minimal vegetation. This extreme climate category is the most underrepresented, with only 41 samples from a single country, Switzerland. This severe lack of data limits any meaningful insights or generalizations about tundra climates.

4.3.4 Final Decision and Support

The final segregation of the EarthNet2021 dataset based on which further explainability analyses are to be performed are as follows:

- **Season-Köppen:** Segregating the dataset based on the Köppen climate classes identified in 4.3.3, followed by further division of samples within each climate class into four distinct seasons, as revealed by clustering. This hierarchical approach

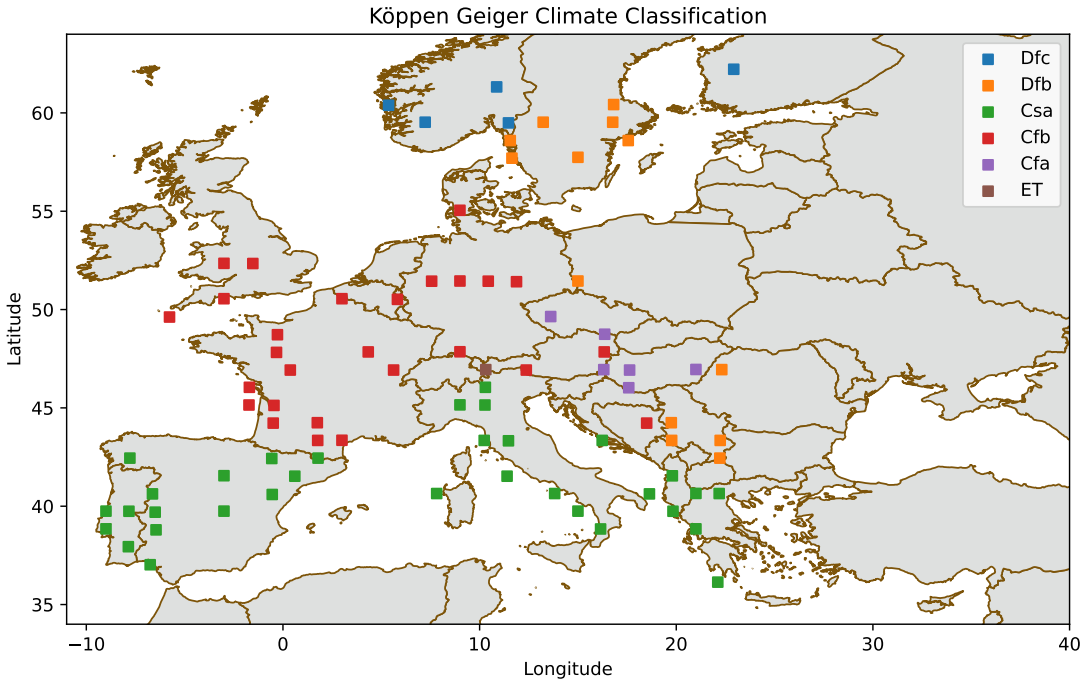


Figure 4.11: Köppen Geiger based Segmentation of MGRS subgrids. *Cube markers indicate the top-left coordinate corner of MGRS subgrids, not their true size, as the original subgrids are significantly larger.*

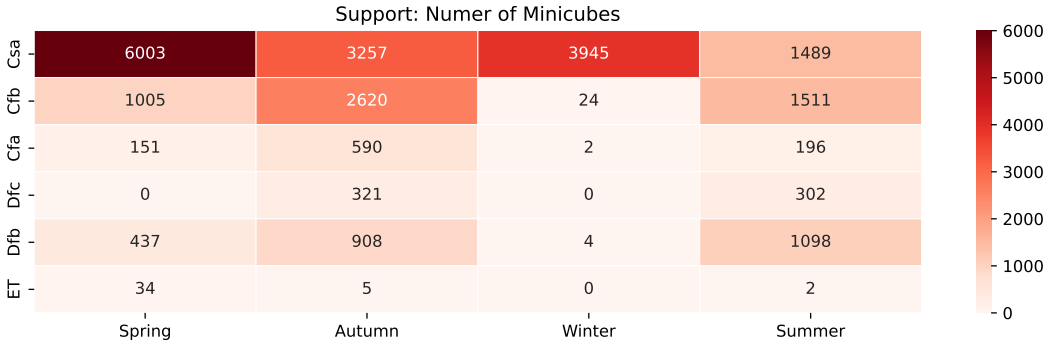


Table 4.2: Season-Köppen Segregation

ensures a meaningful and comprehensive division of the minicubes. Support numbers are given in Figure 4.2.

- **Season-Country:** Dividing the dataset by country boundaries and further splitting the samples within each country into four distinct seasons. Although the support for several countries is low, as shown in **??** and country boundaries are artificial constructs and do not align with the natural forces that drive climatic and ecological variations, this approach still offers insights into localized patterns and seasonal variability, which can be particularly useful for region-specific analyses and comparisons.

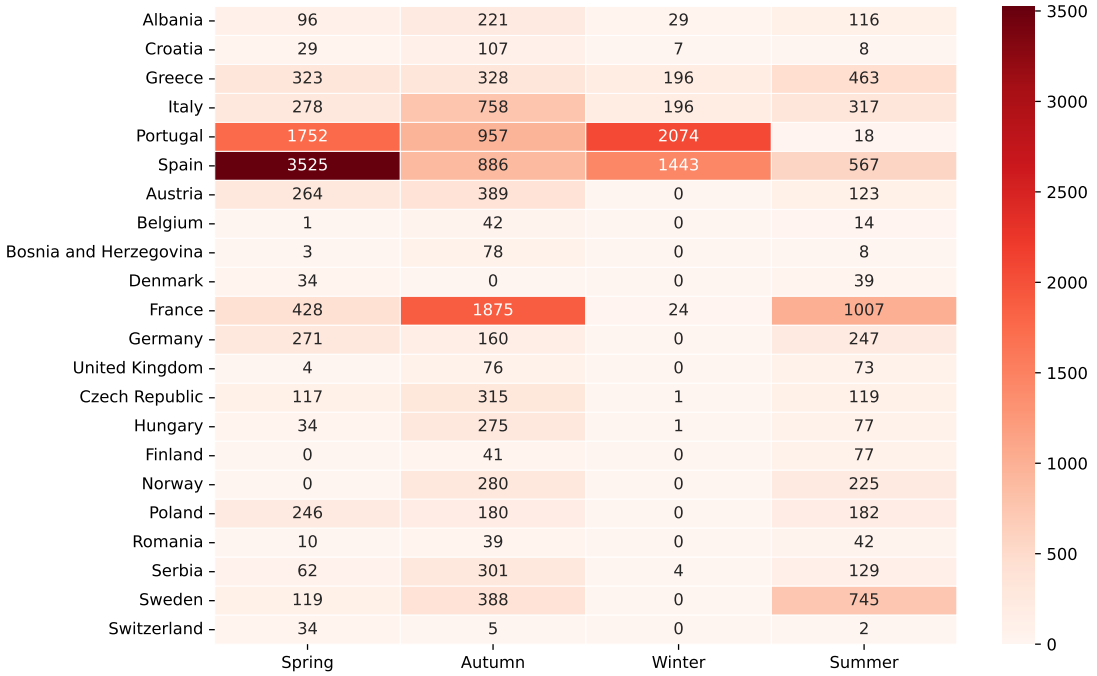


Table 4.3: Season-Country Segregation

4.4 Perturbation

We introduced the following perturbations to the dataset, carefully selecting values to remain within realistic bounds to ensure the integrity and validity of the resulting predictions:

Pressure: [-6, -4, -2, 0, +2, +4, +6]
Temperature: [-6, -4, -2, 0, +2, +4, +6]
Precipitation: [0, +2, +4, +6, +8, +10]

These perturbations were applied uniformly across the entire non-context window of the model. In our ConvLSTM configuration, the context window comprises the first 10 timesteps, while the subsequent 20 timesteps correspond to predictions.

The process involves perturbing one meteorological variable at a time by adding a constant value (+z) uniformly across all non-context timesteps, while keeping all other variables unchanged. This approach isolates the effect of a single variable's perturbation on the model's predictions.

Perturbations were excluded from the context window as it is paired with input Sentinal-2 images, which cannot be adjusted in correspondence with changes to meteorological variables. Once the input is perturbed, predictions for the future timesteps are generated and stored for later evaluation.

5. Marginal Sensitivity Analysis

Marginal Sensitivity Analysis quantifies the influence of individual meteorological variables on model predictions, providing insights into feature importance and model robustness. This is crucial for understanding how climate factors drive land surface changes, improving model trustworthiness, and guiding data-driven decision-making. Sensitivity is assessed by perturbing one variable while keeping others constant, isolating its direct contribution. However, in time-series models, dependencies on past values complicate the analysis, as predictions are influenced by both current and historical inputs.

5.1 Vegetation Index

For this study, we used EVI (Enhanced Vegetation index)5.1 over NDVI because it offers several advantages over traditional NDVI, particularly in high-biomass regions where NDVI saturates, as EVI maintains sensitivity by mitigating saturation effects. It also reduces atmospheric influence, incorporating the blue band to correct for scattering and soil background effects, ensuring greater accuracy in areas with sparse vegetation.

$$EVI = G \times \frac{SR_{\text{green}} - SR_{\text{red}}}{SR_{\text{green}} + C_1 \times SR_{\text{red}} + C_2 \times SR_{\text{blue}} + L} \tag{5.1}$$

Where:
G is the gain factor (typically 2.5), SR_{green}, SR_{red}, and SR_{blue} are the reflectance values for the green, red, and blue bands, C_1 and C_2 are coefficients for the red and blue bands (typically 6 and 7.5, respectively), L is the soil adjustment factor (typically 1).

5.2 Procedure

To quantify the sensitivity of the model's EVI predictions, we analyze the unit change in a single meteorological variable while keeping all other variables constant. This isolates the effect of each variable, providing a clear understanding of its individual contribution to the system.

To enhance the robustness of the analysis, we performed multiple perturbations4.4 and evaluated their interactions (pairwise combinations) on a meteorological variable. By computing the change in EVI values relative to perturbation changes 5.2, we effectively expanded the number of comparisons from n to nC_2, offering a more comprehensive view of the variable sensitivities.

$$Sensitivity_{meso} = \frac{|EVI(\theta(x + permut_i^{meso})) - EVI(\theta(x + permut_j^{meso}))|}{|permut_i^{meso} - permut_j^{meso}|} \tag{5.2}$$

where, θ is the convLSTM model, x is the input minicube, $permut_i^{meso}$ is the i_{th} perturbation for meteorological variable $meso$ 4.4 and function EVI is computing the vegetation index 5.1.

Following the previous step, the next task is to calculate the median of all interactions to derive the unit sensitivity metric. However, there is a critical issue that must be addressed before proceeding.

5.3 Challange

It might be straightforward to evaluate sensitivity for standard numerical variables, but it becomes challenging for time-series variables. In time-series predictions, changes in the output are influenced not only by inputs from the current timestep but also by inputs from previous timesteps. This is evident in our case, as shown in Figure 5.1 'b', which illustrates the trends $\forall$ *perturb* in $EVI\_signal[perturb] - EVI\_signal[0]$ i.e., subtracting the base signal from the perturbed signal. The presence of trends, rather than a consistent value across all timesteps, indicates that older inputs significantly contribute to the output. Such distinct trends are visible for all the minicubes.

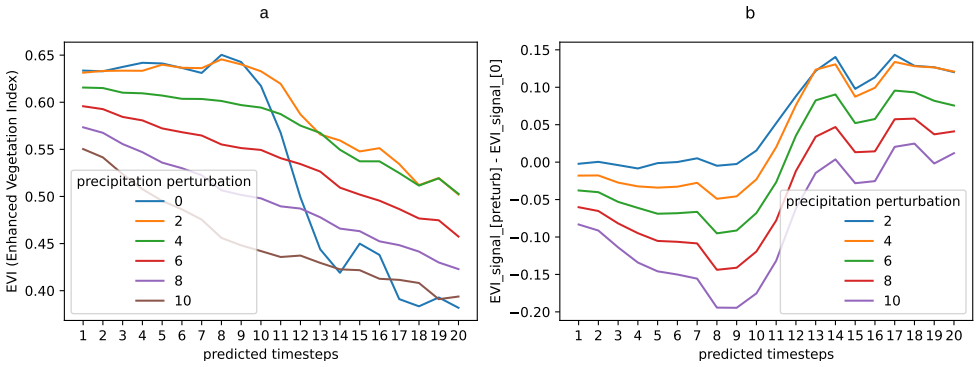


Figure 5.1: **a**: Model's output on precipitation based perturbations, **b**: Difference between the base signal and the perturbed signal

To assess the linear influence of past EVI values on the current EVI predictions, we conducted Partial Autocorrelation Function (PACF) analysis on over 230,000 model output signals (as shown in Figure 5.2). This analysis included signals both with and without input perturbations. The results revealed that lag 1 had a dominant impact, with a median PACF value of 0.8. Additionally, the standard deviation from this median was remarkably low, around 0.06, indicating consistent behavior across the samples. This finding provides strong evidence that the model heavily relies on the immediate past timestep (lag 1) to make predictions for the current timestep. The high median PACF value and low standard deviation highlight a robust and consistent dependency on lag 1 across the dataset. Furthermore, the lack of significant contributions from other lags suggests that the model's temporal dependency is predominantly short-term.

Since the lag-1 timestep contributes the most to the model's predictions, we decided to mitigate its linear influence by approximating an AR(1) model and subtracting it from the EVI signals before computing the differences.

Subtracting a signal's AR(1) model removes the linear effect of its previous value (lag-1 autocorrelation) on the current value. This operation isolates the residual, which represents variations not explained by lag-1 dependencies. The residual highlights immediate or "current" effects, making it particularly useful in sensitivity analysis or time-series modeling.

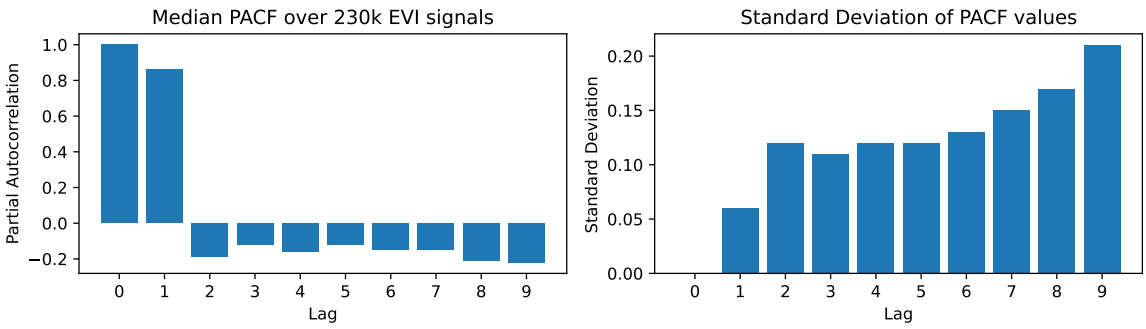


Figure 5.2: PACF analysis

The AR(1) model for a signal y_t is expressed as:

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

where ϕ is the autoregressive coefficient, and ϵ_t is the residual.

By subtracting ϕy_{t-1} from y_t, we compute the residual as:

$$\text{Residual} = y_t - \phi y_{t-1}$$

This residual captures the component of y_t not influenced by y_{t-1}, effectively "detrending" the signal for lag-1 autocorrelation.

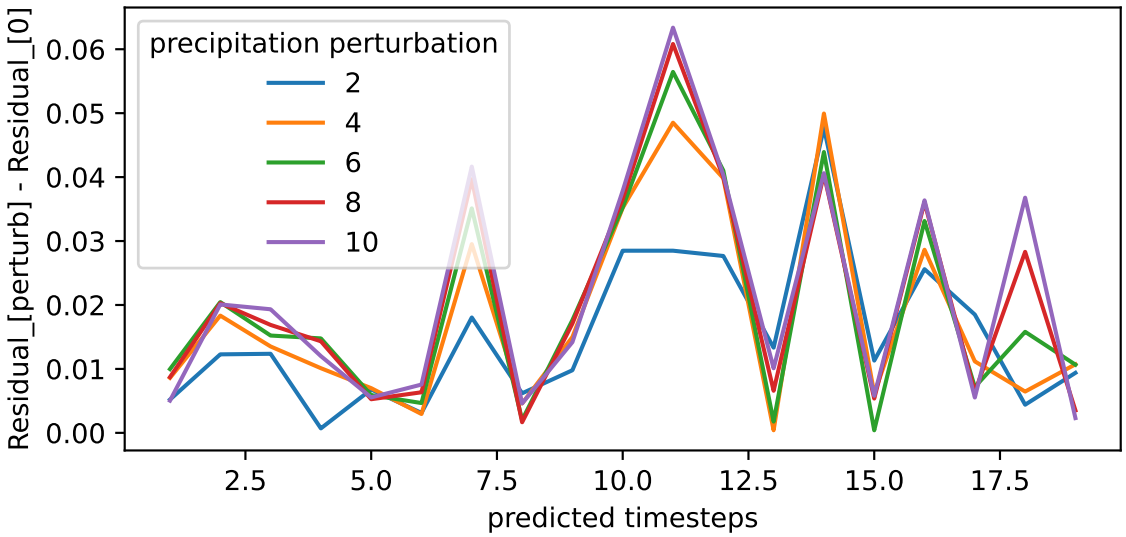


Figure 5.3: Difference of residual of perturbed signal from the residual of base signal i,e., without any perturbation

Implications:

- **Removal of Lag-1 Effects:** The subtraction eliminates the influence of the previous timestep y_{t-1} on the current output y_t, isolating other contributing factors.
- **Focus on Immediate Effects:** This process emphasizes the immediate effects of current input variables or noise, rather than the compounded effects propagated from earlier timesteps.

By isolating these residuals, we aim to better understand the direct sensitivity of the model's predictions to current meteorological inputs.

5.4 Results

5.4.1 Season-Köppen Results

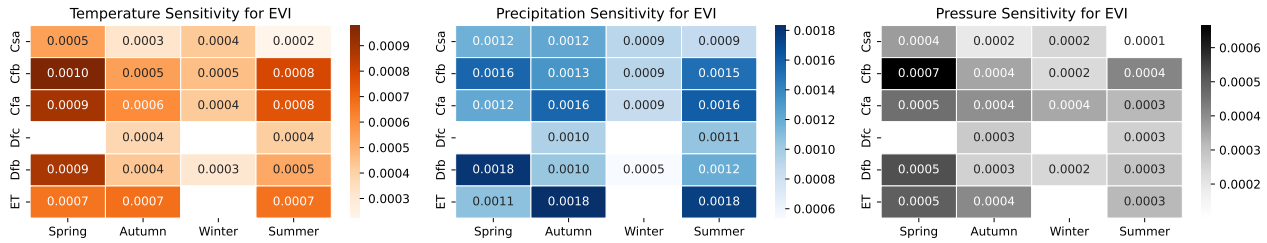


Table 5.1: EVI sensitivity values for unit change in meteorological variables across differnt Köppen regions

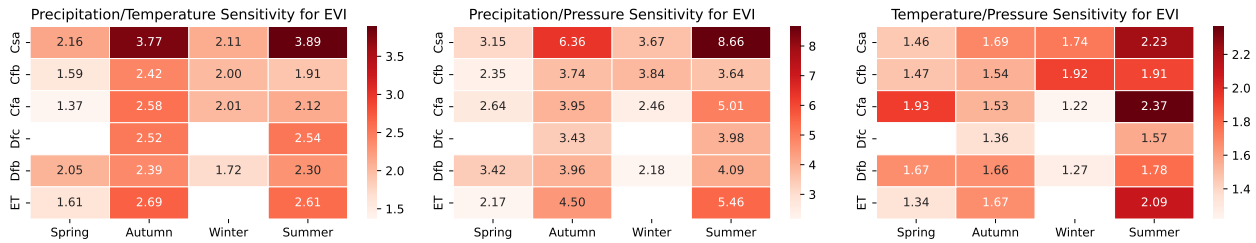


Table 5.2: EVI sensitivity Ratios among different meteorological variables.

Precipitation/Temperature Sensitivity Analysis

- Alignment with Known Research:
  - High Summer sensitivity in `Csa` (3.89) reflects the role of precipitation in mitigating drought impacts in Mediterranean climates.
  - `Dfb` shows relatively high sensitivity in Spring and Autumn, consistent with these seasons driving vegetation dynamics in temperate continental climates.
- Discrepancies:
  - Lower-than-expected Winter sensitivity in `Cfa` (2.01) and `Dfb` (1.71) may indicate underrepresentation in the sample data.

Precipitation/Pressure Sensitivity Analysis

- Alignment with Known Research:
  - `Csa` shows extremely high Summer sensitivity (8.65), reflecting the critical role of precipitation in pressure-driven drought conditions.

- `Dfb` and `Cfa` display moderate sensitivity across all seasons, aligning with the balance of precipitation and pressure in driving vegetation dynamics in these climates.

- Discrepancies:
  - The exceptionally high Autumn sensitivity in `Csa` (6.36) is notable and may warrant further investigation to validate its significance.

Temperature/Pressure Sensitivity Analysis

- Alignment with Known Research:
  - `Csa` and `Cfa` display higher Summer sensitivity (2.22 and 2.36, respectively), which aligns with the dominant role of temperature in driving evapotranspiration and vegetation stress during this season.
  - `ET` and `Dfc` show moderate Spring and Summer sensitivity, reflecting the milder role of temperature in tundra and subarctic climates.

- Discrepancies:
  - The relatively low Autumn sensitivity in `Cfa` (1.53) and `Dfc` (1.36) may suggest underrepresentation of these zones during this season.

Sensitivity Correlation Charts

Quantifies how the response of a dependent variable (EVI, in our case) varies in relation to changes of perturbations in the independent factors such as temperature, precipitation, or pressure. This relationship can be characterized as positive or negative, linear or nonlinear, providing insights into the nature and strength of the dependency between the variables.

Through the correlation study, we discovered that sensitivity exhibits a nonlinear relationship with the magnitude of perturbations across all meteorological variables. This finding highlights the complex and non-proportional dependency of sensitivity on variations in factors such as temperature, precipitation, and pressure. Figure 5.4 illustrates the correlation curves of a specific meteorological variable across different seasons, highlighting seasonal variations in sensitivity patterns. In contrast, Figure 5.5 presents the correlation curves of various meteorological variables within a single season, offering insights into their interdependence and relative influence during that period.

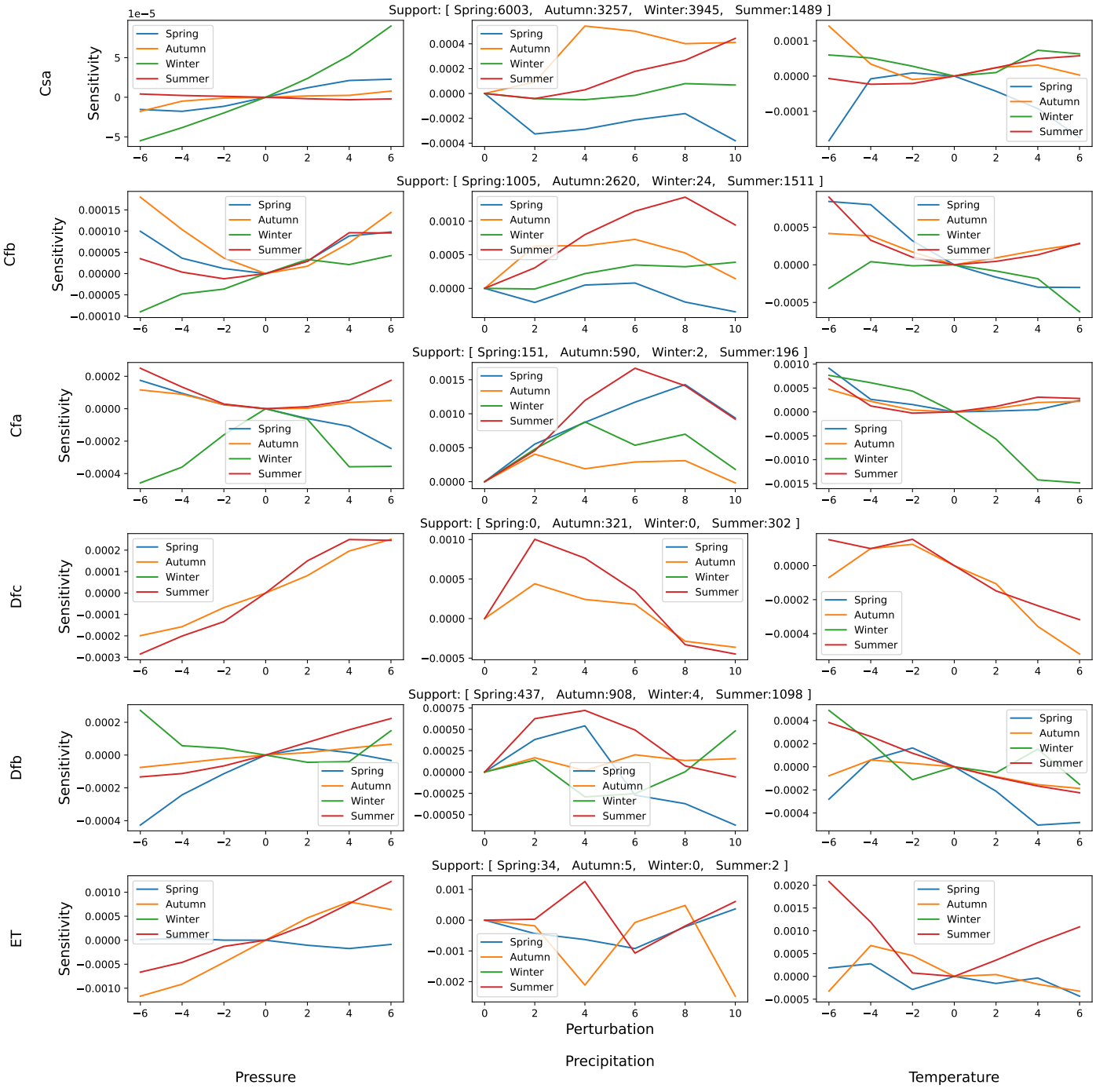


Figure 5.4: Correlation curves of different meteorological variable across different seasons

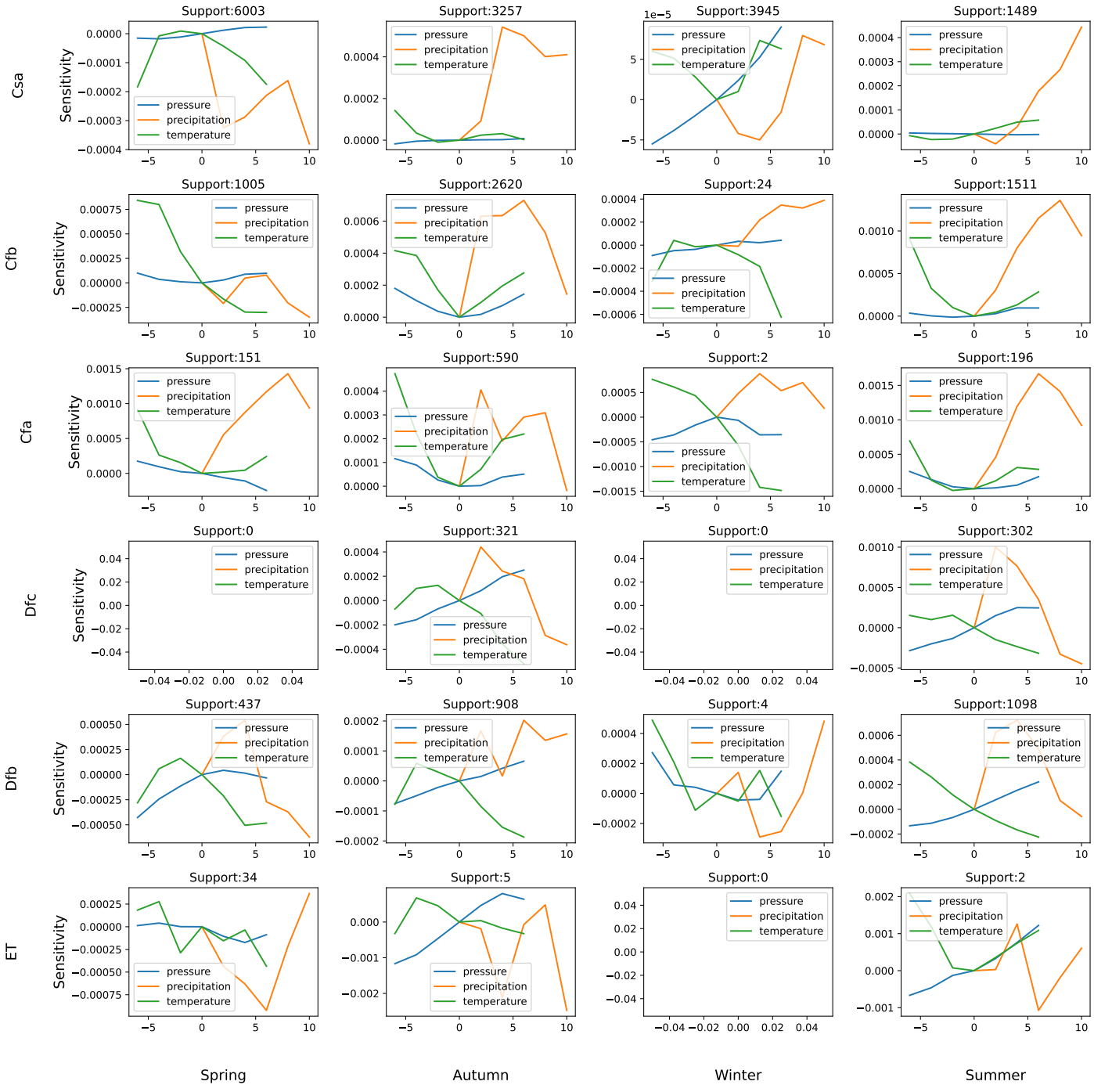
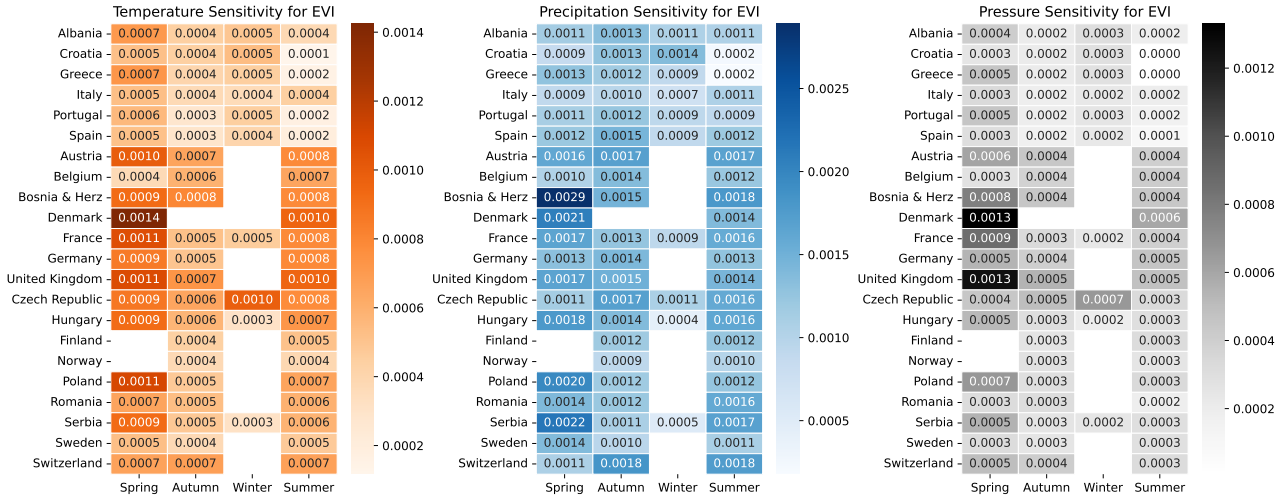


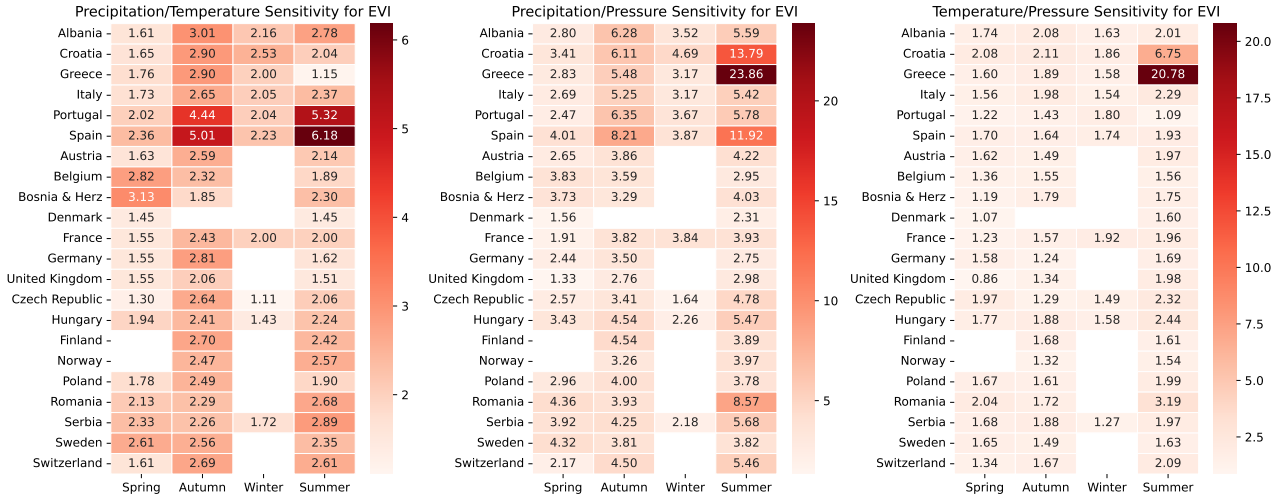
Figure 5.5: Correlation curves of various meteorological variables within a single season

5.4.2 Season-Country Results



	Temperature Sensitivity for EVI				Precipitation Sensitivity for EVI				Pressure Sensitivity for EVI			
	Spring	Autumn	Winter	Summer	Spring	Autumn	Winter	Summer	Spring	Autumn	Winter	Summer
Albania	0.0007	0.0004	0.0005	0.0004	0.0011	0.0013	0.0011	0.0011	0.0004	0.0002	0.0003	0.0002
Croatia	0.0005	0.0004	0.0005	0.0001	0.0009	0.0013	0.0014	0.0002	0.0003	0.0002	0.0003	0.0000
Greece	0.0007	0.0004	0.0005	0.0002	0.0013	0.0012	0.0009	0.0002	0.0005	0.0002	0.0003	0.0000
Italy	0.0005	0.0004	0.0004	0.0004	0.0009	0.0010	0.0007	0.0011	0.0003	0.0002	0.0002	0.0002
Portugal	0.0006	0.0003	0.0005	0.0002	0.0011	0.0012	0.0009	0.0009	0.0005	0.0002	0.0003	0.0002
Spain	0.0005	0.0003	0.0004	0.0002	0.0012	0.0015	0.0009	0.0012	0.0003	0.0002	0.0002	0.0001
Austria	0.0010	0.0007		0.0008	0.0016	0.0017		0.0017	0.0006	0.0004		0.0004
Belgium	0.0004	0.0006		0.0007	0.0010	0.0014		0.0012	0.0003	0.0004		0.0004
Bosnia & Herz	0.0009	0.0008		0.0008	0.0029	0.0015		0.0018	0.0008	0.0004		0.0004
Denmark	0.0014			0.0010	0.0021			0.0014	0.0013			0.0006
France	0.0011	0.0005	0.0005	0.0008	0.0017	0.0013	0.0009	0.0016	0.0009	0.0003	0.0002	0.0004
Germany	0.0009	0.0005		0.0008	0.0013	0.0014		0.0013	0.0005	0.0004		0.0005
United Kingdom	0.0011	0.0007		0.0010	0.0017	0.0015		0.0014	0.0013	0.0005		0.0005
Czech Republic	0.0009	0.0006	0.0010	0.0008	0.0011	0.0017	0.0011	0.0016	0.0004	0.0005	0.0007	0.0003
Hungary	0.0009	0.0006	0.0003	0.0007	0.0018	0.0014	0.0004	0.0016	0.0005	0.0003	0.0002	0.0003
Finland		0.0004		0.0005		0.0012		0.0012		0.0003		0.0003
Norway		0.0004		0.0004		0.0009		0.0010		0.0003		0.0003
Poland	0.0011	0.0005		0.0007	0.0020	0.0012		0.0012	0.0007	0.0003		0.0003
Romania	0.0007	0.0005		0.0006	0.0014	0.0012		0.0016	0.0003	0.0003		0.0002
Serbia	0.0009	0.0005	0.0003	0.0006	0.0022	0.0011	0.0005	0.0017	0.0005	0.0003	0.0002	0.0003
Sweden	0.0005	0.0004		0.0005	0.0014	0.0010		0.0011	0.0003	0.0003		0.0003
Switzerland	0.0007	0.0007		0.0007	0.0011	0.0018		0.0018	0.0005	0.0004		0.0003

Table 5.3: EVI sensitivity values for unit change in meteorological variables across different countries



	Precipitation/Temperature Sensitivity for EVI				Precipitation/Pressure Sensitivity for EVI				Temperature/Pressure Sensitivity for EVI			
	Spring	Autumn	Winter	Summer	Spring	Autumn	Winter	Summer	Spring	Autumn	Winter	Summer
Albania	1.61	3.01	2.16	2.78	2.80	6.28	3.52	5.59	1.74	2.08	1.63	2.01
Croatia	1.65	2.90	2.53	2.04	3.41	6.11	4.69	13.79	2.08	2.11	1.86	6.75
Greece	1.76	2.90	2.00	1.15	2.83	5.48	3.17	23.86	1.60	1.89	1.58	20.78
Italy	1.73	2.65	2.05	2.37	2.69	5.25	3.17	5.42	1.56	1.98	1.54	2.29
Portugal	2.02	4.44	2.04	5.32	2.47	6.35	3.67	5.78	1.22	1.43	1.80	1.09
Spain	2.36	5.01	2.23	6.18	4.01	8.21	3.87	11.92	1.70	1.64	1.74	1.93
Austria	1.63	2.59		2.14	2.65	3.86		4.22	1.62	1.49		1.97
Belgium	2.82	2.32		1.89	3.83	3.59		2.95	1.36	1.55		1.56
Bosnia & Herz	3.13	1.85		2.30	3.73	3.29		4.03	1.19	1.79		1.75
Denmark	1.45			1.45	1.56			2.31	1.07			1.60
France	1.55	2.43	2.00	2.00	1.91	3.82	3.84	3.93	1.23	1.57	1.92	1.96
Germany	1.55	2.81		1.62	2.44	3.50		2.75	1.58	1.24		1.69
United Kingdom	1.55	2.06		1.51	1.33	2.76		2.98	0.86	1.34		1.98
Czech Republic	1.30	2.64	1.11	2.06	2.57	3.41	1.64	4.78	1.97	1.29	1.49	2.32
Hungary	1.94	2.41	1.43	2.24	3.43	4.54	2.26	5.47	1.77	1.88	1.58	2.44
Finland		2.70		2.42		4.54		3.89		1.68		1.61
Norway		2.47		2.57		3.26		3.97		1.32		1.54
Poland	1.78	2.49		1.90	2.96	4.00		3.78	1.67	1.61		1.99
Romania	2.13	2.29		2.68	4.36	3.93		8.57	2.04	1.72		3.19
Serbia	2.33	2.26	1.72	2.89	3.92	4.25	2.18	5.68	1.68	1.88	1.27	1.97
Sweden	2.61	2.56		2.35	4.32	3.81		3.82	1.65	1.49		1.63
Switzerland	1.61	2.69		2.61	2.17	4.50		5.46	1.34	1.67		2.09

Table 5.4: EVI sensitivity Ratios among different meteorological variables.

Precipitation/Temperature Sensitivity Analysis

- Alignment with Known Research:
  - Higher sensitivity values in Spring and Autumn in Mediterranean countries (e.g., Greece, Spain, Italy) align with these seasons being critical for vegetation dynamics.
  - Portugal and Spain exhibit extremely high sensitivities in Summer, reflecting their susceptibility to seasonal drought conditions.
- Discrepancies:

  - Surprisingly high Summer sensitivity in Sweden and Serbia warrants further validation, as these regions typically do not experience extreme summer droughts.

Precipitation/Pressure Sensitivity Analysis

- Alignment with Known Research:
  - High Spring and Autumn sensitivity in Mediterranean and Iberian countries aligns with pressure systems playing a crucial role in their climatic patterns.
  - Moderate Summer sensitivity in central Europe (e.g., Austria, France) reflects stable weather conditions during this season.
- Discrepancies:
  - Extremely high Summer sensitivity values in Greece and Croatia (23.86 and 13.79, respectively) appear anomalous and require investigation.

Temperature/Pressure Sensitivity Analysis

- Alignment with Known Research:
  - Balanced Spring sensitivity in most countries reflects the moderate influence of both temperature and pressure on vegetation indices during this season.
  - Lower Summer sensitivity in Portugal and Spain aligns with temperature playing a dominant role over pressure during heatwaves.
- Discrepancies:
  - Exceptionally high Summer sensitivity in Greece (20.78) is inconsistent with expectations and should be validated.

5.5 Preliminary Discussion

The sensitivity analysis highlights distinct patterns in how spatiotemporal variables influence EVI across different Köppen climate classes and seasons. The relative sensitivities of precipitation, temperature, and pressure varied significantly, reflecting the unique environmental dynamics of each region.

Key findings include the pronounced sensitivity of EVI to precipitation in Mediterranean climates (Csa), particularly during summer and autumn, likely driven by water scarcity. Similarly, the higher temperature sensitivity in continental climates (Dfb) during spring underscores the role of temperature in driving vegetation growth during early growing seasons. Pressure variability exhibited unusual outliers in some regions, warranting caution and indicating potential preprocessing artifacts.

Overall, the results emphasize the importance of regional and seasonal factors in understanding vegetation dynamics.

6. Conclusion, Limitations, and Future Work

6.1 Conclusion

The **Cluster-Segregate-Perturb (CSP) pipeline** presents a novel approach to enhancing explainability in **spatiotemporal land surface forecasting models**, addressing the limitations of existing perturbation-based XAI techniques. By leveraging **β-VAE for feature disentanglement**, CSP transforms high-dimensional spatiotemporal data into **interpretable temporal signals**, allowing structured perturbations while preserving spatial and temporal coherence. **GoodClusterScore (GCS)** optimizes clustering and segmentation, ensuring meaningful partitioning of data to facilitate robust feature attribution. The proposed framework improves **interpretability, scalability, and reliability**, making it a promising solution for understanding complex predictive models in climate science and remote sensing.

6.2 Limitations

- **Assumption of Feature Independence:** Currently, we assume that the features are independent of one another, which may not hold true in practice. To address this limitation, future work will focus on conducting **causal inference tests** among independent variables to identify and quantify highly correlated predictors. This will enable more informed perturbations, isolating the **true sensitivity** of meteorological variables while accounting for potential interdependencies, leading to more accurate and meaningful sensitivity values.
- **Computational Complexity:** Training **β-VAE** for feature disentanglement and performing recursive clustering-based segmentation demand significant **computational power**, limiting feasibility for large-scale datasets without high-performance infrastructure.
- **Stability and Reproducibility Issues:** Clustering-based segmentation is sensitive to **hyperparameters and noise**, leading to **inconsistent partitions** across different runs, potentially affecting the reliability of feature importance estimates.
- **Loss of Spatial Granularity:** While temporal signals retain critical information, **downsampling may discard fine-grained spatial patterns**, reducing the resolution of feature attributions and potentially omitting localized variations.
- **Over-Segmentation and Bias Propagation: Over-segmentation** can lead to **fragmented clusters** with low statistical significance, while suboptimal partitioning may introduce **bias** in feature importance rankings.

- **Perturbation Coherence Across Scales:** Ensuring perturbations remain **realistic and semantically meaningful** across different resolutions is challenging, as small latent modifications may lead to **amplified distortions** in the upsampled spatiotemporal space.
- **Lack of Standardized Explainability Metrics:** Existing XAI metrics fail to **fully capture the spatiotemporal dependencies** in forecasting models, making it difficult to objectively evaluate the quality of feature attributions.

6.3 Future Scope

- **Scalability Enhancements:** Developing **computationally efficient implementations** of CSP using model distillation, tensor decomposition, or distributed computing to handle large-scale satellite and climate datasets.
- **Robust Latent Feature Supervision:** Introducing **self-supervised learning objectives** or **domain-informed constraints** within the β-VAE framework to improve alignment between latent features and real-world phenomena.
- **Adaptive Clustering Strategies:** Implementing **dynamic clustering approaches** that adjust partitioning based on dataset properties, reducing sensitivity to hyperparameters and improving reproducibility.
- **Multi-Resolution Feature Importance Analysis:** Extending CSP to **multi-resolution representations** to retain both **local** and **global** information, enabling a finer-grained understanding of spatiotemporal dependencies.
- **Domain-Specific Validation Frameworks:** Establishing **benchmarking protocols** and **quantitative evaluation metrics** for spatiotemporal explainability, bridging the gap between model interpretability and real-world applications.
- **Integration with Physics-Informed Models:** Combining CSP with **physics-based simulations** to enforce constraints that align with known environmental processes, enhancing the interpretability and reliability of feature attributions.
- **Broader Applicability Across Domains:** Adapting CSP for other high-dimensional, structured datasets, such as **biomedical imaging, financial time series, and autonomous navigation**, to expand its impact beyond land surface forecasting.

By addressing these limitations and advancing the proposed solutions, CSP has the potential to become a **foundational framework for explainability in spatiotemporal deep learning models**.

Bibliography

[1] M. Shokr and Y. Ye, "Why does arctic sea ice respond more evidently than antarctic sea ice to climate change?," *Ocean-Land-Atmosphere Research*, vol. 2, p. 0006, 2023.

[2] S. Lampe, C. Burton, E. Burke, J. Chang, N. Christidis, M. Forrest, L. Gudmundsson, H. Huang, S. Hantson, A. Ito, *et al.*, "The effect of climate change on global wildfire activity," in *EGU General Assembly Conference Abstracts*, pp. EGU–14756, 2023.

[3] D. Geudtner, R. Torres, P. Snoeij, M. Davidson, and B. Rommen, "Sentinel-1 system capabilities and applications," in *2014 IEEE Geoscience and Remote Sensing Symposium*, pp. 1457–1460, IEEE, 2014.

[4] M. Beyer, R. Ahmad, B. Yang, and P. Rodríguez-Bocca, "Deep spatial-temporal graph modeling for efficient ndvi forecasting," *Smart Agricultural Technology*, vol. 4, p. 100172, 2023.

[5] F. van Oorschot, R. van der Ent, M. Hrachowitz, E. di Carlo, F. Catalano, S. Boussetta, G. Balsamo, and A. Alessandri, "Improving the temporal and spatial vegetation variability in land surface models based on satellite observations," in *EGU General Assembly Conference Abstracts*, pp. EGU–6528, 2023.

[6] E. Chrysanthopoulos, C. Pouliaris, I. Tsirogiannis, and A. Kallioras, "Forecasting soil moisture on a spatial and temporal scale using machine learning algorithms," in *EGU General Assembly Conference Abstracts*, pp. EGU–11925, 2023.

[7] Y. Zhang, F. Huang, L. Li, Q. Li, Y. Zhang, and W. Shangguan, "Real-time forecast of smap l3 soil moisture using spatial–temporal deep learning model with data integration," *Remote Sensing*, vol. 15, no. 2, p. 366, 2023.

[8] M. Witjes, L. Parente, C. J. van Diemen, T. Hengl, M. Landa, L. Brodskỳ, L. Halounova, J. Križan, L. Antonić, C. M. Ilie, *et al.*, "A spatiotemporal ensemble machine learning framework for generating land use/land cover time-series maps for europe (2000–2019) based on lucas, corine and glad landsat," *PeerJ*, vol. 10, p. e13573, 2022.

[9] J. Wang, X. Yin, S. Liu, and D. Wang, "Spatiotemporal change and prediction of land use in manasi region based on deep learning," *Environmental Science and Pollution Research*, vol. 30, no. 34, pp. 82780–82794, 2023.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[11] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[12] X. Huang, X. Li, Y. Ye, S. Feng, C. Luo, and B. Zhang, "On understanding of spatiotemporal prediction model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[13] A.-D. Pham, A. Kuestenmacher, and P. G. Ploeger, "Tsem: Temporally-weighted spatiotemporal explainable neural network for multivariate time series," in *Future of Information and Communication Conference*, pp. 183–204, Springer, 2023.

[14] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[15] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International conference on machine learning*, pp. 5338–5348, PMLR, 2020.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[17] J. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, and P. Bizarro, "Timeshap: Explaining recurrent models through sequence perturbations," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2565–2573, 2021.

[18] A. Sood and M. Craven, "Feature importance explanations for temporal black-box models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 8351–8360, 2022.

[19] D. P. Kingma, M. Welling, *et al.*, "Auto-encoding variational bayes," 2013.

[20] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.

[21] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task.," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1132–1142, 2021.

[22] M. Haylock, N. Hofstra, A. Klein Tank, E. Klok, P. Jones, and M. New, "A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006," *Journal of Geophysical Research: Atmospheres*, vol. 113, no. D20, 2008.

[23] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[24] C.-A. Diaconu, S. Saha, S. Günnemann, and X. X. Zhu, "Understanding the role of weather data for earth surface forecasting using a convlstm-based model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1362–1371, 2022.

A. Appendix: Cluster Charts

A.1 K-Means clustering with Euclidean distance

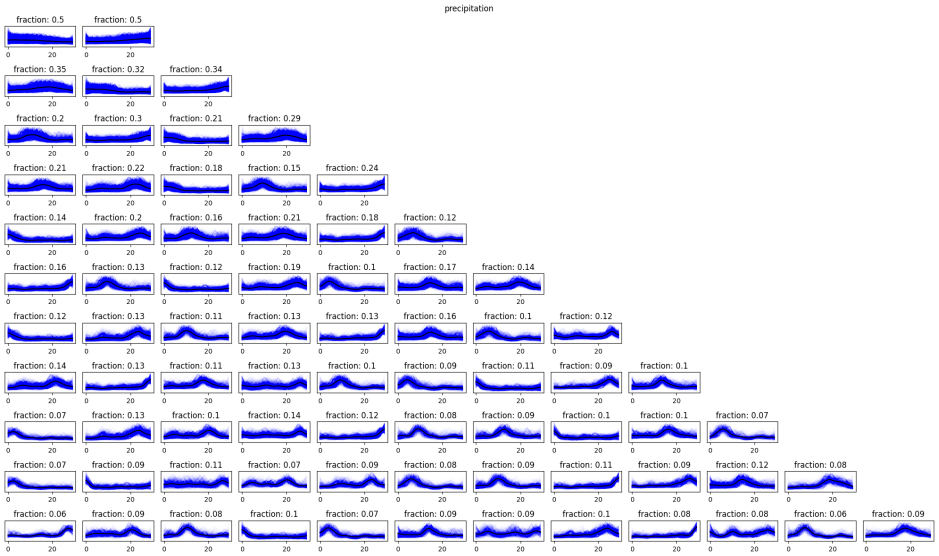


Figure A.1: precipitation clusters for k = [2,12]

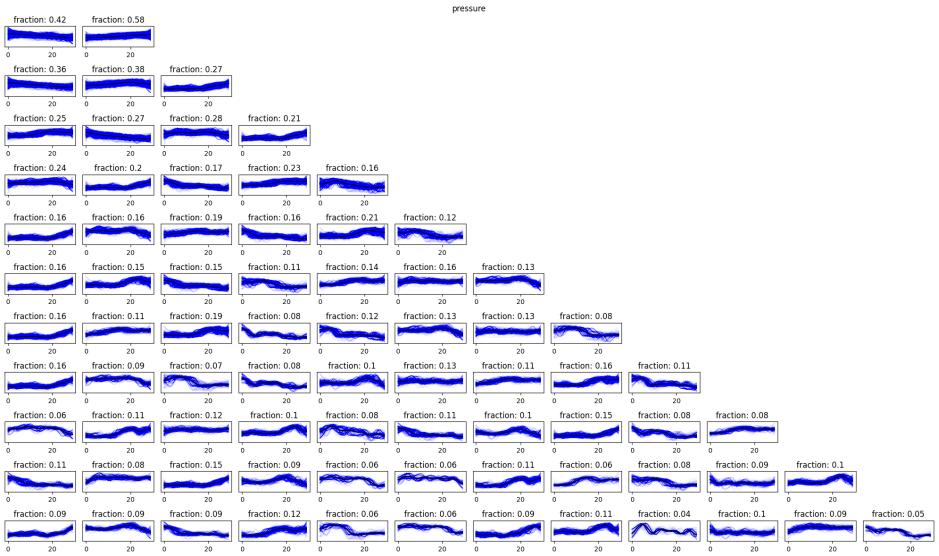


Figure A.2: pressure clusters for k = [2,12]

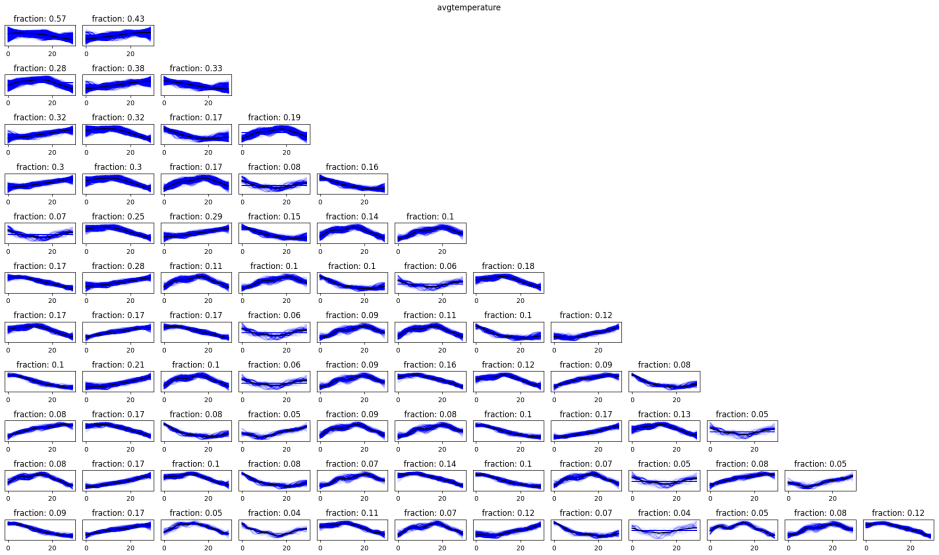


Figure A.3: avg-temperature clusters for k = [2,12]

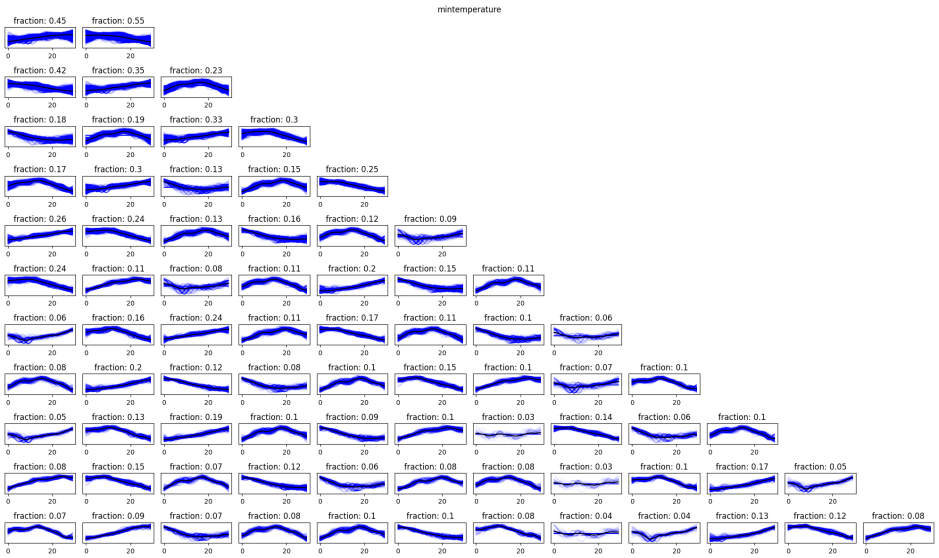


Figure A.4: min-temperature clusters for k = [2,12]

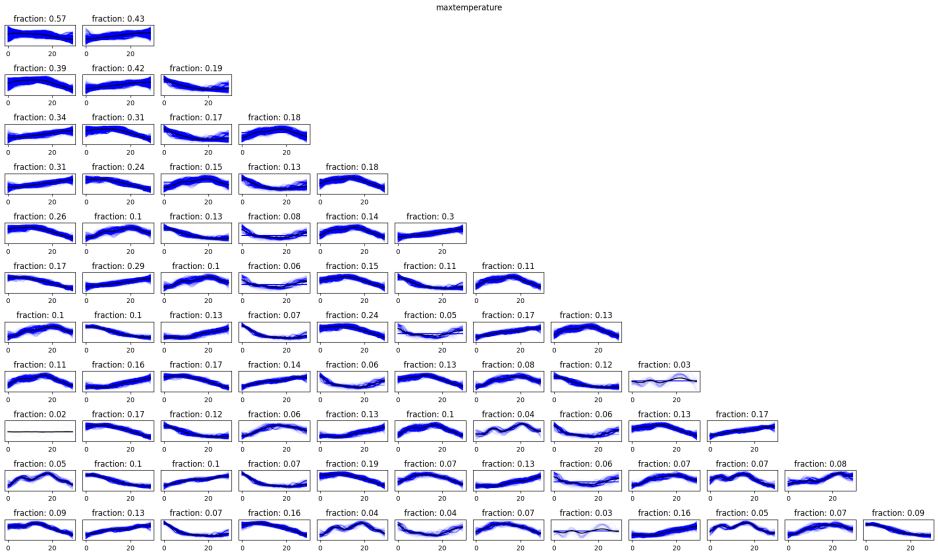


Figure A.5: max-temperature clusters for k = [2,12]

A.2 K-Means clustering with Dynamic Time Warping (DTW)

Figure A.6: precipitation clusters for k = [2,12]

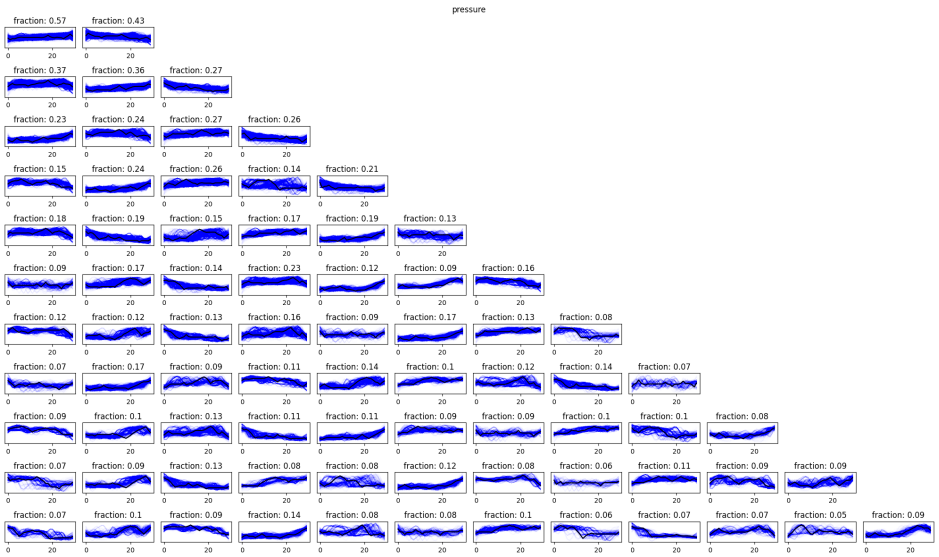


Figure A.7: pressure clusters for k = [2,12]

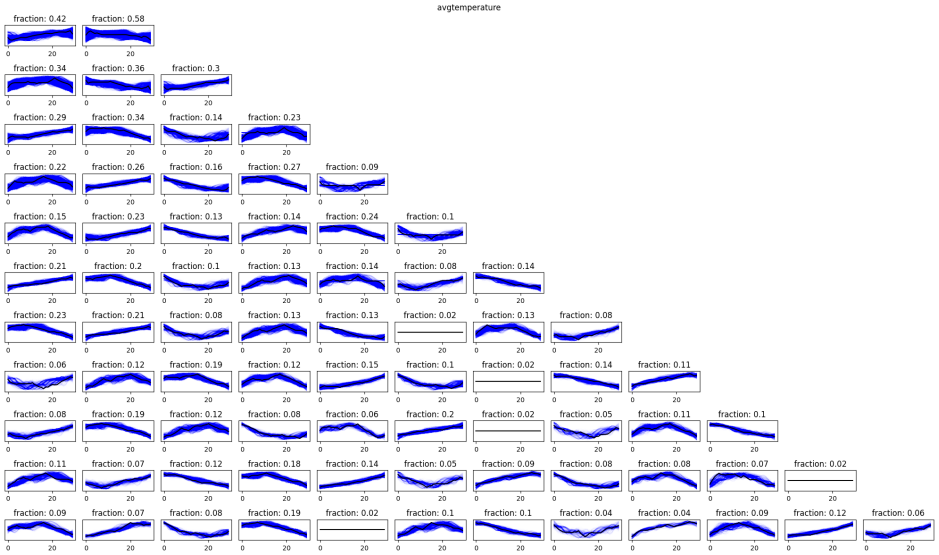


Figure A.8: avg-temperature clusters for k = [2,12]

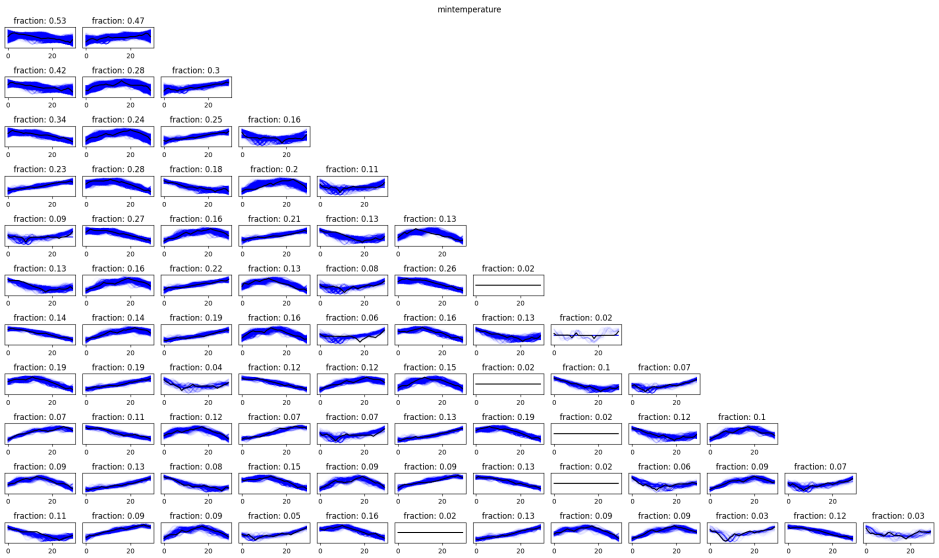


Figure A.9: min-temperature clusters for k = [2,12]

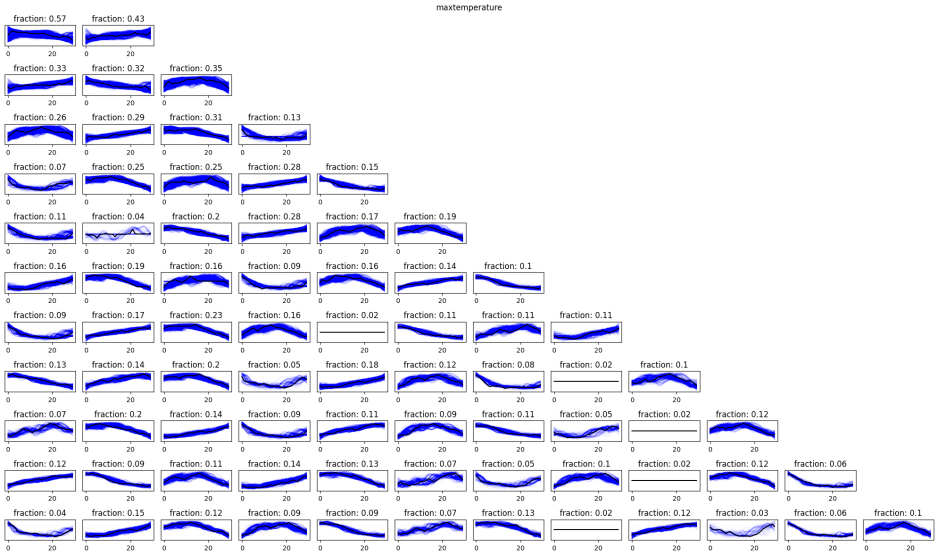


Figure A.10: max-temperature clusters for k = [2,12]

A.3 K-Means clustering with Soft-Dynamic Time Warping (Soft-DTW)

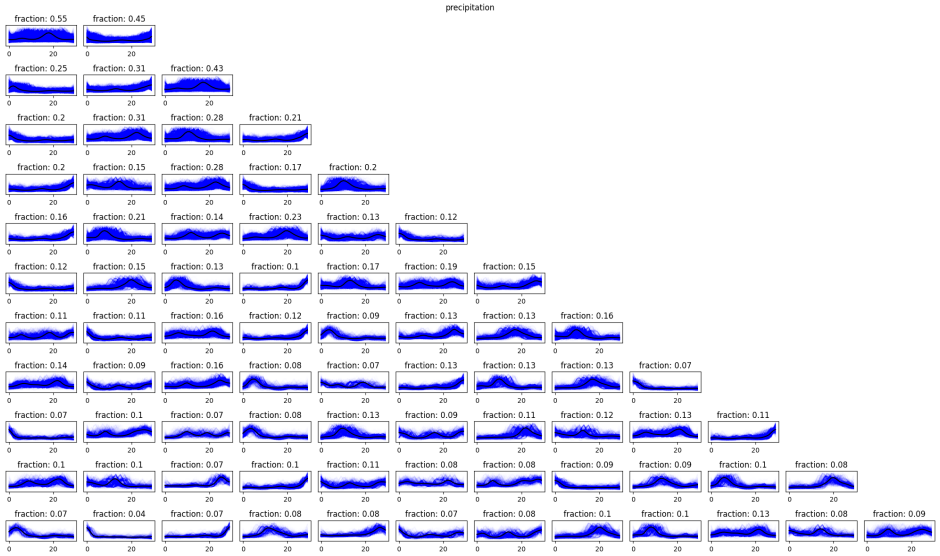


Figure A.11: precipitation clusters for k = [2,12]

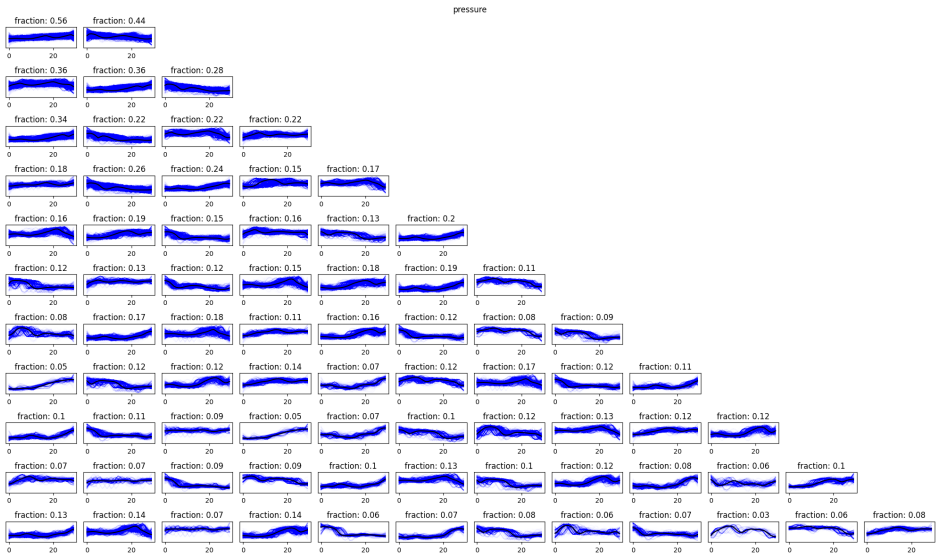


Figure A.12: pressure clusters for k = [2,12]

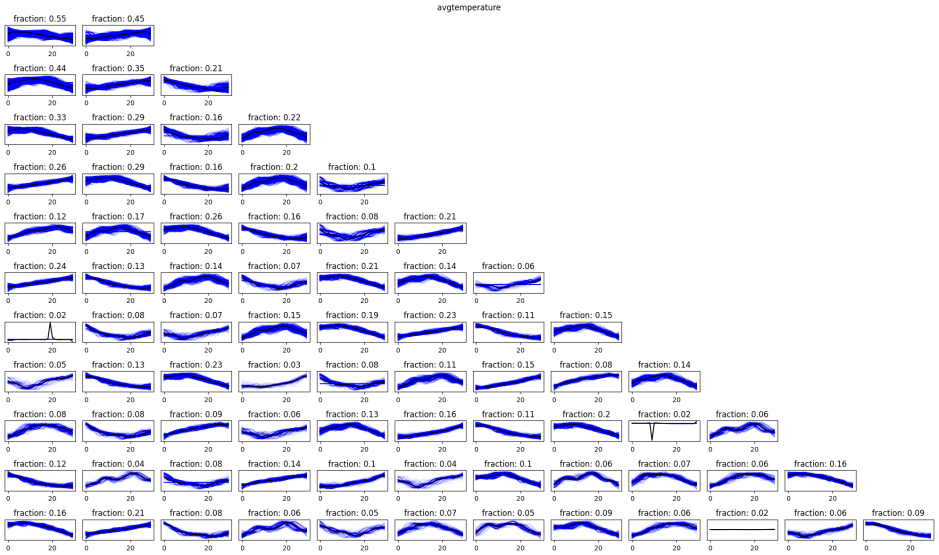


Figure A.13: avg-temperature clusters for k = [2,12]

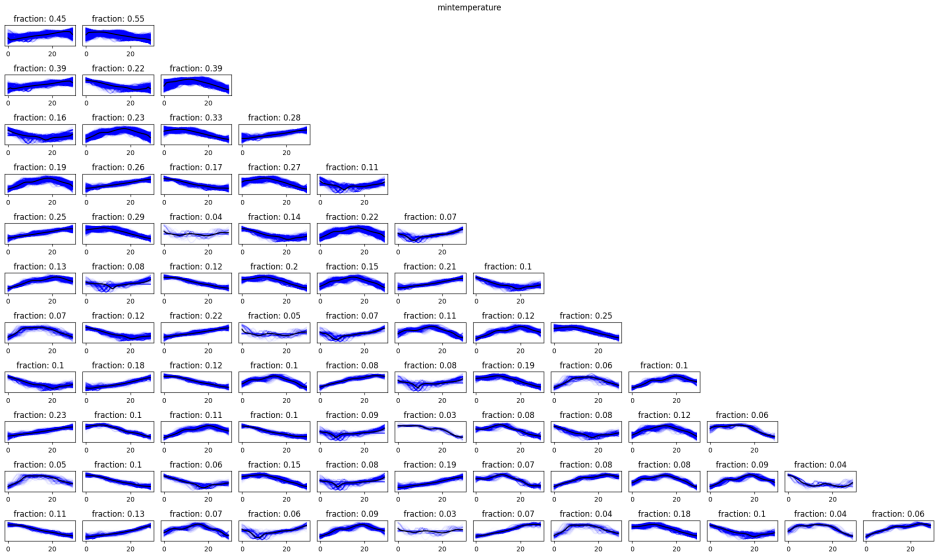


Figure A.14: min-temperature clusters for k = [2,12]

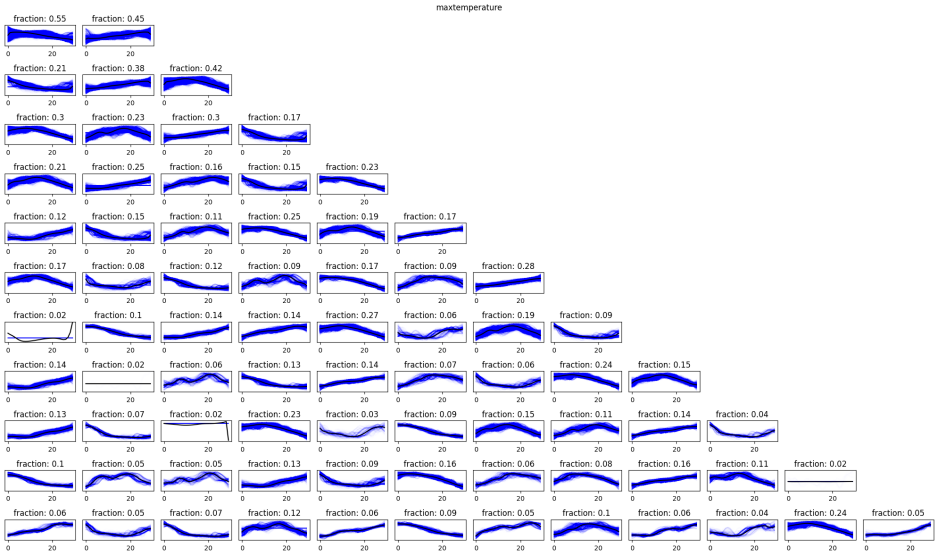


Figure A.15: max-temperature clusters for k = [2,12]

A.4 K-Shape clustering with Shape-based Distance (SBD)

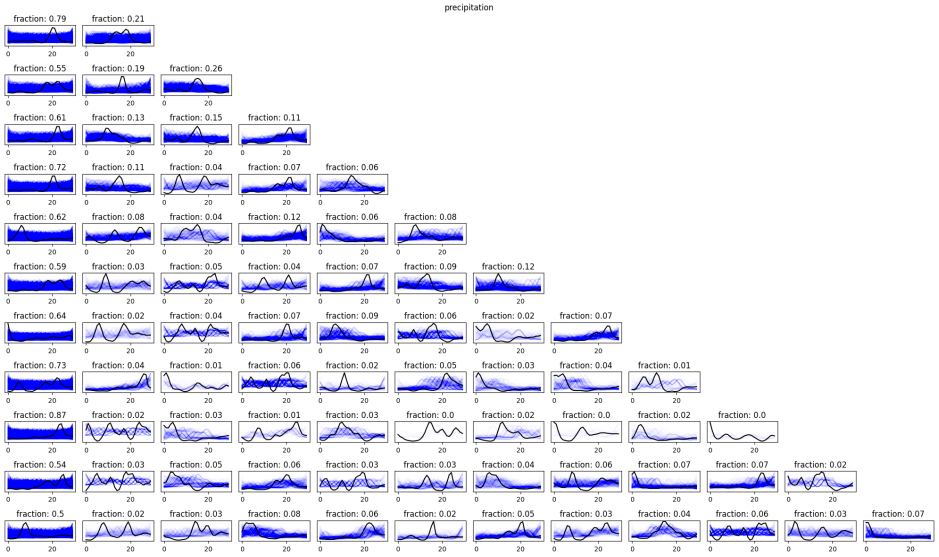


Figure A.16: precipitation clusters for k = [2,12]

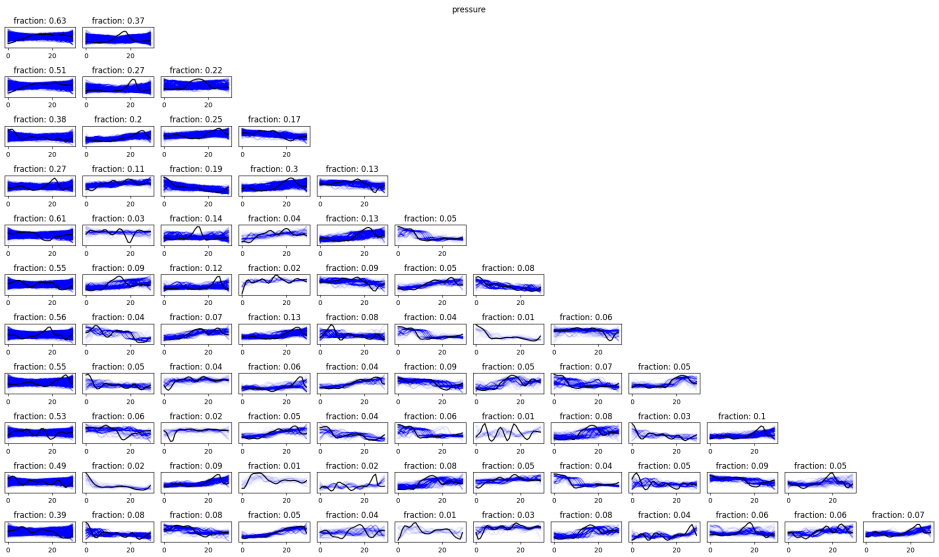


Figure A.17: pressure clusters for k = [2,12]

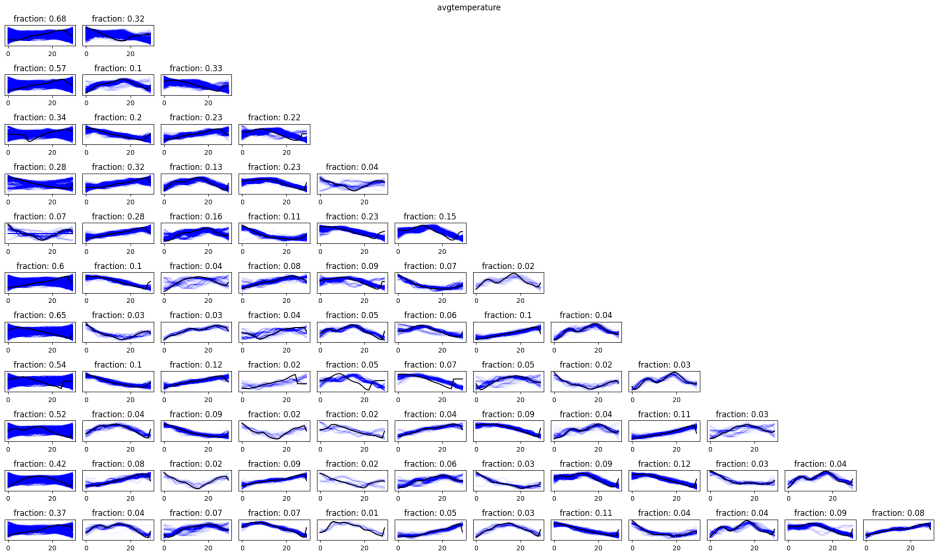


Figure A.18: avg-temperature clusters for k = [2,12]

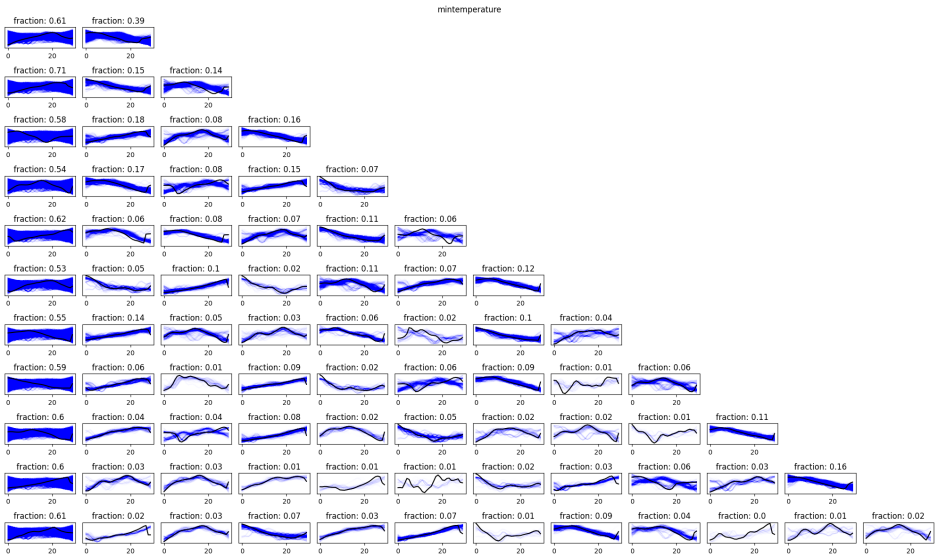


Figure A.19: min-temperature clusters for k = [2,12]

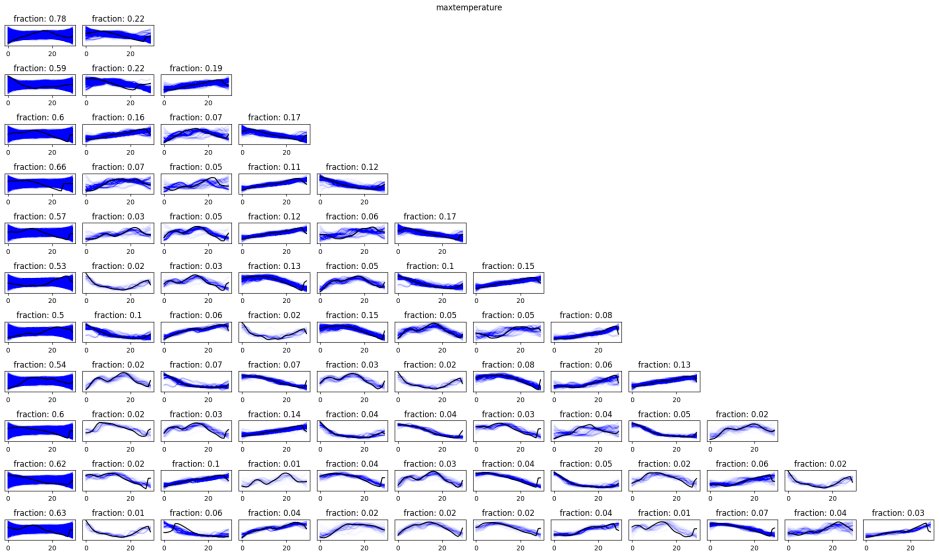


Figure A.20: max-temperature clusters for k = [2,12]