QUANTIFYING VARIABILITY IN LAND-SURFACE HETEROGENEITY AT THE GLOBAL SCALE USING UNSUPERVISED LEARNING

by

BRIJ ROKAD

(Under the Direction of Nandita Gaur & Sheng Li)

ABSTRACT

An understanding of the land surface heterogeneity has been deemed important for modeling earth system processes and developing a mechanistic understanding of soil hydrology. In this study, we have developed a Heterogeneity Index (H-index) that quantifies the soil, topography, and seasonal vegetation based heterogeneity at a global scale. Our study follows the scaling nomograph which was introduced to incorporate the scale and site-specific dependence of soil moisture on geophysical heterogeneity and antecedent wetness conditions at the regional scale. The H-index is based on eigenvalue decomposition of sand, leaf area index (LAI), and flow accumulation, hence quantifies the sub-grid scale variability and co-variability of land-surface heterogeneity for large-scale hydrology. The H-index has been modified for global-scale analysis. The H-index has been found to respond to the changes in vegetation cover and can be used to assess the change in land-surface heterogeneity over time. We introduce a modular and efficient algorithm for index generation that processes partially available data-sets such as region-specific elevation from regional LIDAR surveys and improvements in global or regional soil texture maps. Additionally, the formulation of the H-index has the potential to be incorporated with earth system models. In this work, we present the computed H-index for the period of 2002-2020 and provide a comparison of the H-index with available classification measures such as the Major Land Resource Areas (MLRA) and Common Resource Areas (CRA). To demonstrate *hydrologically similar regions* we cluster Soil Water Retention Parameters (SWRPs) using unsupervised machine learning techniques including K-means clustering, spectral clustering, Gaussian Mixture Model, and Hierarchical clustering. Results show adequate cluster separation that implies a successful classification of near-surface soil moisture dynamics by H-index. Moreover, it demonstrates the potential of replacing representative elementary volume (REV) driven soil hydrology classifiers such as porosity to describe soil hydrology at the remote sensing scale.

INDEX WORDS: Land Surface Heterogeneity, Unsupervised Learning, Geo Science, Heterogeneity Index, Remote Sensing

QUANTIFYING VARIABILITY IN LAND-SURFACE HETEROGENEITY AT THE GLOBAL SCALE USING UNSUPERVISED LEARNING

by

BRIJ ROKAD

B.Tech., Vellore Institute of Technology, INDIA, 2018

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2020

QUANTIFYING VARIABILITY IN LAND-SURFACE HETEROGENEITY AT THE GLOBAL SCALE USING UNSUPERVISED LEARNING

by

BRIJ ROKAD

Co-Major Professor:	Nandita Gaur
Co-Major Professor:	Sheng Li
Committee:	Khaled Rasheed
	Frederick Maier

Electronic Version Approved:
Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2020

ACKNOWLEDGEMENTS

I would like to thank all my committee members for their time and the numerous suggestions they provided throughout my research. In particular, I would like to thank Dr. Gaur and Dr. Li for their guidance, constant encouragement and motivation to keep me going. I owe a great debt to Vinit Sehgal who helped me at the various stages of the project, and whose contributions laid the foundation for my work. I am thankful to my friends and family who kept me sane during these unprecedented times, where the whole world has come to a standstill due to COVID-19. I would like to thank everyone at the *Institute for Artificial Intelligence* who presented me with an opportunity to be a part of this project. I would also like to thank Ms. Tino, whose assistance with the administrative tasks at the university made my life incredibly easy.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Overview

Soil moisture dynamics and it's spatial distribution influences a wide spectrum of factors. The distribution affects hydrologic responses of watersheds, precise agriculture decisions, soil characteristics, temperature and bio-geochemistry. In some cases, temperature becomes the key factor to determine the energy and mass exchange within the atmosphere, which affects the water balance and eco-hydrological processes, such as evapotranspiration ratios and water uptake by plants whereas some properties of soil moisture dynamics helps to determine the suitability of land for growing crops [20]. The health of crops relies upon an adequate supply of moisture and soil nutrients. All of these factors act as a key to determine the flooding potentials, drought patterns, and nutrient transport to the streams and they also affects the development of the atmospheric boundary layer.

Soil comprises of mineral particles, organic matter, water and air. The combination of these determines the soil's properties such as its texture, structure, porosity, chemistry and color. Primarily, soil is made up of different combinations of *sand, silt,* and *clay* particles whose proportions define the soil texture. For example, sandy soils feel gritty whereas silts feel smooth and most clays are sticky and mouldable. Soils that are the mixture of sand, silt and clay are called loams. The texture of the soil helps to identify whether the soil is free draining, or it can hold water and how easy it is for plant roots to grow.

Heterogeneity Index: Background

Various hydrological applications require robust and periodic spatially distributed soil moisture data. However, soil itself has a wide range of properties at a Global Scale and because of that, it is challenging to classify soil moisture present in the soil at a remote sensing scale. The remote sensing satellites such as Soil Moisture Active Passive (SMAP) and Soil Moisture Ocean Salinity (SMOS) have made it easier to monitor spatially distributed soil moisture. The data can be obtained from passive microwave, active microwave, and optical sensors, although the coarse spatial resolution of passive microwave and the inability to obtain vertically resolved information from optical sensors limit their usefulness for watershed-scale applications. The downside of these data sets is that they require some degree of down-scaling before being effectively incorporated into models for operational use. These data sets have the resolution of varying between 30 km and 60 km, but to unlock the the full potential of satellite-based soil moisture estimation we need a much finer resolution. Satellite-based soil moisture watershed models such as Community Land Model (CLM) [16] and Active microwave sensors such as Synthetic Aperture Radar (SAR) [21] can obtain the spatially distributed surface soil moisture at scales of less than 5 km and 10 - 100 m respectively, for watersheds ranging from 1 to 25 km^2. But those models also requires downscaling, which comes with the disadvantage of high uncertainty.

Soil moisture distribution at the watershed scale [14] [11] is a highly nonlinear function of soil, topography, vegetation, land use, and weather. Despite this complexity, there are a multitude of opportunities for satellite-based estimation of soil moisture for critical watershed applications. The combination of soil, vegetation, and topography models could provide distributed and profile soil moisture information with known accuracy at the watershed scale. Therefore describing heterogeneity is very important at the Pixel-level remote sensing. The heterogeneity can be case specific and has already been described in different ways - all of which are qualitative, where the world is divided into hydro-climates [1] that is based on meteorological and climate data, biomes that are mainly divide the world based on vegetation, and land cover land-use. Countries such as the United States have provided a more intensive classification and defined heterogeneity classifications that encompass geology such as Major Land Resource Areas (MLRAs) that are geographically associated land resource units to identify the large areas important in statewide agricultural planning and national planning.

To overcome this challenge, a novel technique that exploits the full potential of the rich satellite-based remote sensing data was developed [12]. The Heterogeneity Index (H-index) developed in [12] follows the scaling nomograph to incorporate the scale and site-specific dependence of soil moisture on geophysical heterogeneity and antecedent wetness conditions at the regional scale.

Modification of Heterogeneity Index

The H-index developed based on nomogrph was limited at most to the regional scale [12] and there is no global scale index thus far to define heterogeneity. Therefore in this work, We have modified the formulation of the Heterogeneity Index to make it applicable for the global scale. The H-index quantifies the soil, topography and seasonal vegetation based heterogeneity at a global scale. The index is based on eigenvalue decomposition of %sand, leaf area index (LAI), and flow accumulation and hence, quantifies the global scale variability and co-variability of land-surface heterogeneity. The major changes made to the original formulation of the H-index include:

1. Normalization of LAI and flow accumulation values at the global scale to ensure equal representation of soil, vegetation and topography in the index.
2. Modification in the resolution of all land surface heterogeneity based factors (%sand, LAI and flow accumulation) to 500m for computing the H-index at 36 km as well as 4 km.
3. Normalization of Heterogeneity index to incorporate high values as well as have them in a well define scale.

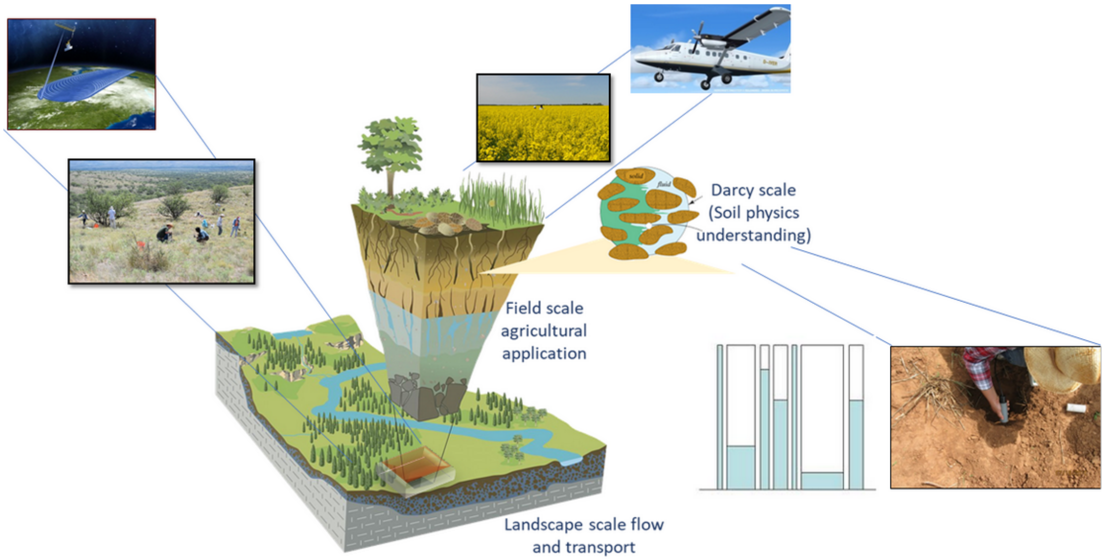


Figure 1: Scaling and Changing Heterogeneity (Derived from [5])

Another addition to the previous studies is *Factor Analysis*. Current work has analyzed the influential factor and their degree during the influence of the various heterogeneity factors, namely,

%sand, LAI and flow accumulation on the H-index land surface heterogeneity formation as well as the *percentage variance* explained during the index formation.

The Heterogeneity Index was further analyzed based on the correlation between different parameters. Additionally, Heterogeneity Index was compared with other heterogeneity definitions that are commonly used such as global aridity index for climate, land cover map and major land resource areas. Global aridity index [36] [37] is used to classify different kinds for arid regions. It ranges from Hyper-Arid to Humid regions. Whereas land cover map [10] classifies regions based on different kinds of land use, for example forest regions, crop land, wet land and etc. The major land resource areas are specifically developed for the United States of America, which characterize a particular pattern that combines soils, water, climate, vegetation and land use.

As shown in [9], this index can especially be valuable for studying soil moisture dynamics at the global scale. Geo-physical heterogeneity in satellite pixels can be defined as the variability and co-variability in sub-pixel topography, vegetation, and soil. These three factors affects the soil moisture spatial distribution at the remote sensing scale. For instance, the exact same percentage of clay fraction will lead to different soil moisture drydown based on whether it is present on a slope or a flat ground. Its effect will also vary with the presence and amount of vegetation on it and these factors can effect on heterogeneity formation. Different topography or elevation can have different kind of soil properties which in turn results in the different soil moisture or water retained (water levels) inside those soil particles (Fig.1). Vegetation or LAI also plays a role in variable heterogeneity. The land-surface heterogeneity changes as the scale of analysis changes. The REV scale only represents soil based heterogeneity while vegetation and topographic heterogeneity also becomes important at the airborne or satellite scale.

Clustering the Soil Water Retention Parameters

To evaluate the utility of the Heterogeneity Index for studying soil moisture dynamic, we have used four unsupervised machine learning algorithms to cluster the Soil Water Retention Parameters (SWRPs). Global surface soil moisture observation from SMAP satellite was used to develop the parameters for the seasonal drydowns [14]. Their new non-parametric approach to identify the canonical shapes followed by a non-linear least-squares for the seasonal drydown was found to be highly effective for each SMAP footprint (36km×36km). The seasonal drydowns were found to respond to the land-surface characteristics, soil, vegetative and atmospheric dynamics.

The Soil Water Retention Parameters (SWRPs) were used in unsupervised learning methods such as K-means clustering, Spectral clustering, Gaussian Mixture Model (GMM) clustering and Hierarchical clustering. Two sets of parameters [SWRPs (*All Parameters)* and $SWRP_{eff}$ (*Selected Parameters*)] were used for clustering, where the optimal number of clusters were determined by Silhouette score, Davies-Bouldin Index and Dunn's Index. Moreover, to find out the statistical significance in cluster separation, Tukey's HSD test was performed.

Related Work

The Soil Water Retention Parameters (SWRPs) were fitted on a non-linear canonical shape, that means that there will be some missing values in the parameters, which can cause problems when clustering these parameters. But these values are type of Missing Not At Random (MNAR) that implies the values that are missing is related to a reason or domain specific, SWRPs are based on the shape of the parameters they were fitted. One of the algorithms often used in clustering is *k-means*, but the k-means requires complete data matrix. *Chi* [4] has come up with a novel technique to address this issue. Their k-POD method represents a simple extension of k-means clustering, which is really effective when the external information is unavailable or there is significant missing value in the data. In order to achieve that, they have formulated the k-means as seeking an optimal rank k-decomposition of the data matrix. The decomposed data matrix was used to formulate the k-means loss for missing data by taking residual sum of square over observed values. This new loss admits a simple majorization that can be inexactly minimized with the k-means algorithm. Additionally, a fuzzy clustering can be effective when it comes to segregate different eco-regions based on soil hydraulic parameters. *Sarkar* [29] has used fuzzy k-means clustering with missing values. Traditionally, data matrix is being completed by doing some kind of imputations or removing the missing entries, instead of following that, they have proposed a technique which exploits the information provided by the patterns with the missing values. Based on the patterns, imputations can be carried out incrementally in each iteration according to the context of the data. Their proposed method was able to out-perform other kinds of imputation with a classification rate of 98.43%.

Bellugi [2] attempted to predicate shallow landslide size using spectral clustering as a search algorithm. Due to the characteristics of spatial data, where the position of the data points matter with the different number of parameters or features, spectral clustering can be useful to segregate those data points. *Stella* [30] has proposed a multi-class spectral clustering on real image segmentation that tries to solve the continuous optimization problem by doing eigen-decomposition on the data. The eigenvector generates optimal solutions through orthogonal transformation. These solution were then solved as a discretization problem for a discrete solution closest to the global-optima. Spectral clustering can be memory intensive when working with a extremely large-scale datasets. To address this issue and to make spectral clustering more scalable and robust, [15] has proposed a novel algorithm on Ultra-Scalable Spectral Clustering (U-SPEC). The proposed algorithm constructs the sparse affinity sub-matrix by combining a hybrid representative selection strategy and a fast approximation method for K-nearest representatives. The proposed algorithm was able to achieve a near linear time and space complexity, additionally they were able to robustly partition ten million levels of non linearly-separable datasets.

Model based clustering were found to be quite effective, when it comes to Geo-spatial data. These types of clustering methods can determine the relation between spatial and non-spatial attributes. In [35], a Spatial Gaussian Mixture Model (SGMM) was introduced in order to cluster the optical remote sensing images. SGMM can be considered as an extension of Gaussian Mixture Models (GMM). A Gaussian Mixture Model is a parametric density function that represents a weighted sum of Gaussian component densities [27]. Statistically, Gaussian Mixture Models are most mature methods for clustering, where each cluster is represented by a Gaussian distribution. The parameters of GMMs are estimated by Expectation-Maximization (EM) algorithm. SGMM was able to incorporate spatial information by generating spatial windows around pixels. Based on SGMM, authors of [35] have proposed two different clustering methods. One of the method was SGMM-MLR, which uses maximum likelihood rule and other one was SGMM-CRF, that combines SGMM and Conditional Random Fields (CRF). Both the methods were found to have high accuracy compared to a K-means and other traditional cluster algorithms. SGMM-MLR was able to achieve the highest AA of 93.63% and OA of 95.37% with the Kappa coefficient of 0.94, where as SGMM-CRF was able to achieve the highest AA of 95.18% and OA of 94.50% with the Kappa coefficient of 0.928.

Hierarchical clustering was also found to be useful for Geo-spatial data, where [22] has implemented a Agglomerative clustering algorithm which is the hierarchical clustering using a bottom up approach. In Agglomerative clustering, each observation starts in its own cluster, and clusters are then successively merged together. The proximity matrix of every data points in a cluster is calculated. At the end of every iteration, the closest points are merged into the clusters and the proximity matrix is updated.

One of the important steps for clustering algorithms is to determine the optimal number of clusters. In general, the validation of clustering algorithms can be classified into external cluster validation internal cluster validation. In [19], authors have presented a detailed study on internal cluster validation for crisp clustering. The study includes eleven widely used clustering validation measures. These eleven clustering validation measures were evaluated on five different aspects: monotonicity, noise, density, sub-clusters and skewed distributions. Additionally, authors of [18] have compared the performance of 16 commonly used Cluster Validity Indices (CVIs) and evaluated their effectiveness on remote-sensing data. To evaluate the performance authors have used Fuzzy C-Means (FCM) algorithm to cluster nine types of remote sensing data-sets.

Research Objectives

This work addresses the two main objectives.

1. Formulation of Heterogeneity Index and its effects on Land Surface Heterogeneity.
2. Evaluate the utility of the Heterogeneity Index as a classifier of soil hydraulic parameters at global scale using machine learning.

Organization of Thesis

The rest of the thesis is organised as follows.

- In Chapter 2, *“Data Description”* presents the detailed description of all the datasets that are used in this study.
- In chapter 3, *“Methods”* presents the scalable heterogeneity index which is developed at a Global scale as well as the necessary information about the newly developed python framework. The python liberary is called ***pyHetro*** and the section describes the workflow for the index formation. The later part of the chapter 3 contains methods on unsupervised machine learning techniques, which are used to cluster the soil water retention parameters.

- In chapter 4, *"Results and Discussion for Heterogeneity Index"* reports the factor analysis for heterogeneity index as well as the comparison of heterogeneity index with different kinds of arid regions and land use regions based on Global Aridity Map and Landuse/Landcover classification respectively. The later section of chapter 4 contains comparison of heterogeneity index with Major Land Resource Ares (MLRAs) and Common Resource Areas (CRAs).

- In chapter 5, *"Results and Discussion for Cluster Models"* reports the experimental results of all the cluster models and discussion on these results.

- In chapter 6, *"Conclusion & Future Directions"* reflects on the implications of the current work by summarizing overall results and identifying limitations and possible future directions to build on this research.

CHAPTER 2

DATA DESCRIPTION

2.1 OVERVIEW

The purpose of this chapter is to introduce data-sets which were used during the course of this research. For the formation of Heterogeneity Index we have used %sand, leaf area index (vegetation) and flow accumulation (topography). Additionally, to validate the formulated index it is compared it with Global Aridity Map, Land cover and Major Land Resource Areas (MLRAs). Lastly, to evaluate the utility of the heterogeneity index we have used Soil Water Retention Parameters (SWRPs). Here, we'll go over the each data-set individually to understand it's properties and characteristics as well as the source and the collection of these data-sets.

2.2 SOIL

Soil data for sand was obtained from SoilGrids where [13] have provided a global predictions for standard numeric soil properties. The grid data is available at different varying depths of 0, 5, 15, 30, 60, 100 and 200 cm, additionally SoilGrids also provides depth to bedrock where they have used 150,000 soil profiles for training and predictions. Apart from that, 158 remote sensing-based soil covariates (soil characteristics) were used to fit an ensemble of machine learning methods—random forest, gradient boosting, and multinomial logistic regression. Their ensemble models were able to explain between 56% (coarse fragments) and 83% (pH) of variation with an overall average of 61%. It can be said that their improvement is due to considerable investments in preparing a finer resolution covariate layers. The results demonstrates that using a series of cross-validation tests, that the new version of SoilGrids represents is a significant improvement upon the previous products at 1 km resolution, especially in terms of spatial detail and attribute accuracy.

For the purpose of this study, soil properties at 5cm depth has been used. The data-set was collected through the *Google Earth Engine*, where spatial resolution of %sand was at 250 meters from SoilGrids. However, %sand at 5cm depth was not available on Google Earth Engine (GEE), to solve that we have averaged out the %sand at 0cm and 10cm depth.

2.3 LEAF AREA INDEX

Leaf Area Index was obtained from 4-day composite MODIS product (NASA Land Processes Distributed Active Archive Center) [24]. The data-set is the product of the best pixel available from all the acquisitions of both MODIS (Moderate Resolution Imaging Spectroradiometer) sensors located on NASA's Terra and Aqua satellites from within the 4-day period.

Leaf Area Index (LAI) describes the canopy structure of the plants, such that LAI quantifies the amount of leaf material in a canopy. By definition, it is the ratio of one-sided leaf area per unit ground area and because it is a ratio of areas, LAI is unit-less. LAI becomes useful for understanding canopy function because many of the biosphere-atmosphere exchanges of mass and energy occur at the leaf surface. For these reasons, LAI is often a key biophysical variable used in biogeochemical, hydrological, and ecological models. Leaf area index is also commonly used as a measure of crop and forest growth at spatial scales ranging from the sub-grid scale to the global scale. Figure 2 gives a simplified explanation of Leaf Area Index, where LAI characterizes plant canopies. The data-set was collected from United States Geological Survey's (USGS) AppEEARS stands for Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) [32]. The collected data ranges form 4th July 2002 to 30th April 2020.

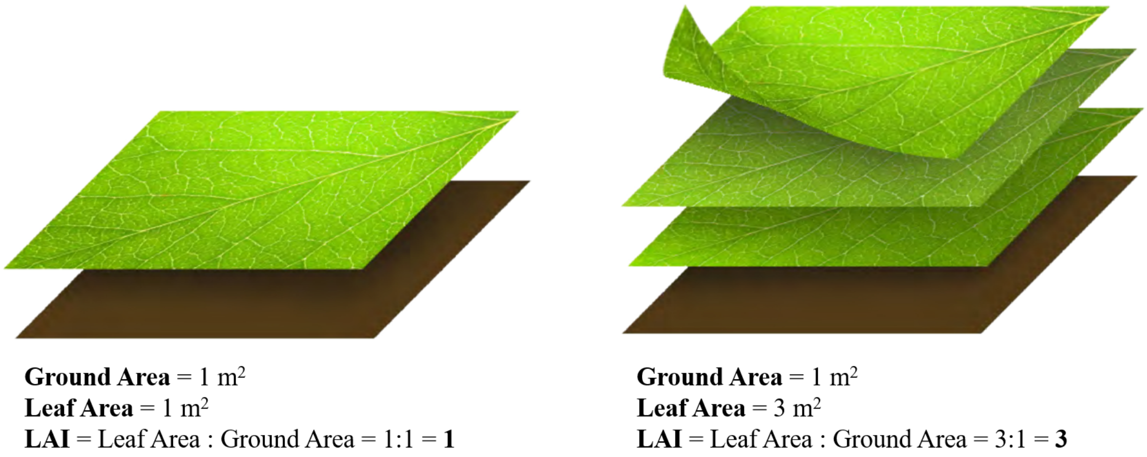


Figure 2: Leaf Area Index (Derived from [3])

2.4 DIGITAL ELEVATION MODEL

Digital Elevation Model (DEM) was produced by the Shuttle Radar Topography Mission (SRTM) [8]. SRTM is a collaborative effort by the National Aeronautics and Space Administration (NASA) and the National Geospatial-Intelligence Agency where they have used dual radar antennas to acquire the interferometric radar-data, processed to digital topographic data at 1 arc (30 meter) resolution. SRTM employed two synthetic aperture radars, a C band system (5.6 cm, C radar) and an X band system(3.1 cm, X radar). C radar was used to generate the contiguous mapping coverage, where as X radar was used to generate data along 50 km wide discrete swaths. These swaths offered nearly contiguous coverage at higher latitudes [9].

DEM was obtain from Google Earth Engine (GEE). The elevation model is necessary to generate topographic map.

Flow Accumulation

Elevation model is used to calculate the flow accumulation (topography) using D-infinity Algorithm [31]. To calculate the flow accumulation, flow directions of the elevation model were calculated using up-slope areas where rectangular grid of the digital elevation models is presented. The entire process is based on representing the flow direction as a single angle taken as the steepest downward slope on the eight triangular facets centered at each grid point. Up-slope area is then calculated by proportioning flow between two down-slope pixels according to how close this flow direction is to the direct angle of the down-slope pixel.

Flow accumulation performs a cumulative count of the number of pixels that naturally drain into outlets such that it can be used to find the drainage pattern of a terrain. For this study, flow accumulation was calculated using ESRI's *ArcGIS* software.

2.5 GLOBAL ARIDITY MAP

Global Aridity Map represents the different hydro-climates present in the globe. The map has assigned aridity index at 1km resolution which provides a high-resolution global raster such that climate data can be related to evapotranspiration processes and rainfall deficit for potential vegetative growth. The aridity model was derived and described in [36] and [37]. Aridity is expressed as a function of precipitation and temperature such that aridity index is used to quantify precipitation deficit over atmospheric water demand. In a generalized term, aridity index is a numerical indicator that represents the degree of dryness at a given location. The data was collected from Global Aridity and PET Database [33] where the aridity is divided into five categories: *Hyper-Arid, Arid, Semi-Arid, Sub-Humid* and *Humid*.

2.6 LAND COVER CLASSIFICATION

Land cover represents regions that are covered by forests, wetlands, impervious surfaces, agriculture, and other land and water types. The data was collected at 36 km from Soil Moisture Active Passive (SMAP) Ancillary Data Report on Landcover Classification. Landcover classification has 16 divisions (excluding water). Divisions' definition were given by IGBP classification scheme [10]. To obtain the global raster of land cover, five consecutive days worth of data was taken from SMAP L3. All five raster were then merge to complete a global raster for land cover.

2.7 MAJOR LAND RESOURCE AREAS

Major land resource areas (MLRAs) are geographically associated land resource units (LRUs) that act as a qualitative classifications of the land-surface which is developed by NRCS national agency [34]. The Major Land Resource Areas (MLRAs) are further subdivided into Common Resource Areas (CRAs). MLRAs are characterized by a particular pattern that combines soils, water, climate, vegetation, land use, and type of farming. There are 204 MLRAs in the United States, ranging in size from less than 500,000 acres to more than 60 million acres. However, these indices are mainly aimed at agricultural planning and lack quantified information that can be very useful in monitoring changes in heterogeneity in a changing land-use land cover scenario.

2.8 SOIL WATER RETENTION PARAMETERS

Soil Water Retention Parameters (SWRPs) are the pathways of seasonal drydowns which were developed using global surface soil moisture (θ_{RS}) observation from SMAP satellite at $36km \times 36km$ [14]. θ_{RS} was represented by six canonical shapes which were identified using data-driven non-parametric approach. These parameters were fitted into soil moisture data or soil moisture loss curve using a piece-wise linear curve (Figure 3). Figure 3 shows a high seasonal variation in soil hydraulic parameter distribution. For each season, the soil hydraulic parameters were very distinct. For instance, a transition curve (m_2) that goes from a point that transition between a wet period (W) to dry period (D) and transition between the transitionary period and the dry period, the slope of this particular curve which shows the ability of the soil to drydown or soil moisture to change.

Table 1: Parameters of Global Surface Soil Moisture Drydown using SMAP [14]

Parameters	Description	Units
l_D	Constant-rate loss during dry phase	$m^3/m^3/day$
θ^{TD}	Transition point between transitional and dry phase	m^3/m^3
m_1	Slope of gravity drainage phase	day^{-1}
θ^{WT}	Transition point between wet and transitional phase	m^3/m^3
l_W	Constant-rate loss during wet phase	$m^3/m^3/day$
θ^{GW}	Transition point between drainage and wet phase	m^3/m^3
m_2	Slope of transitional phase	day^{-1}

Figure 4, 5, 6 and 7 shows the spatial distribution of soil water retention parameters, these distributions shows that there is a wide variation in the slopes across the seasons. The seasonal differences in soil hydraulic parameter are more evident. Additionally, three $SWRP_{eff}$ (θ^{TD}, θ^{WT} and θ^{GW}) indicates the transition point (m_2) between dominant hydrologic regimes of the SM drydown process. This means, that the these four parameters ($\theta^{TD}, \theta^{WT}, \theta^{GW}$ and m_2) should be able to explain majority of the global region.

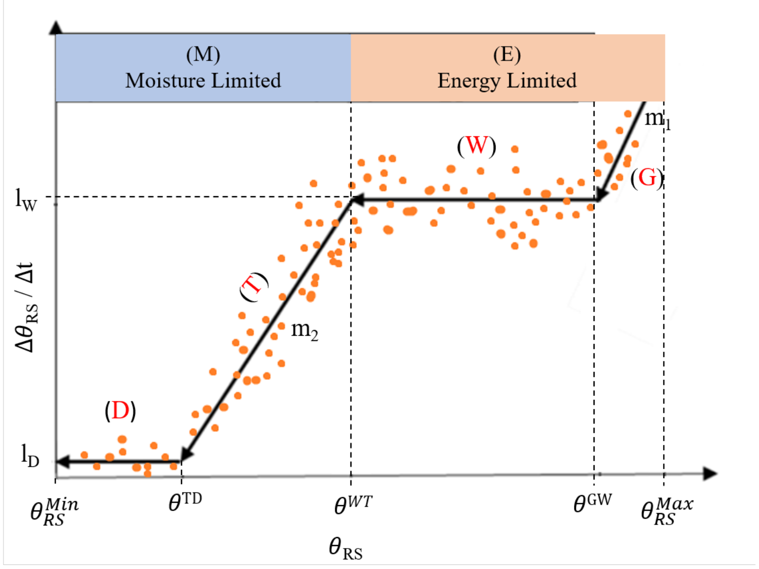


Figure 3: Piece-wise linear fit on SWRPs (Derived from [14]). Labels *D, T, W and G* represents Dry, Transitional, Wet and Gravity drainage respectively.

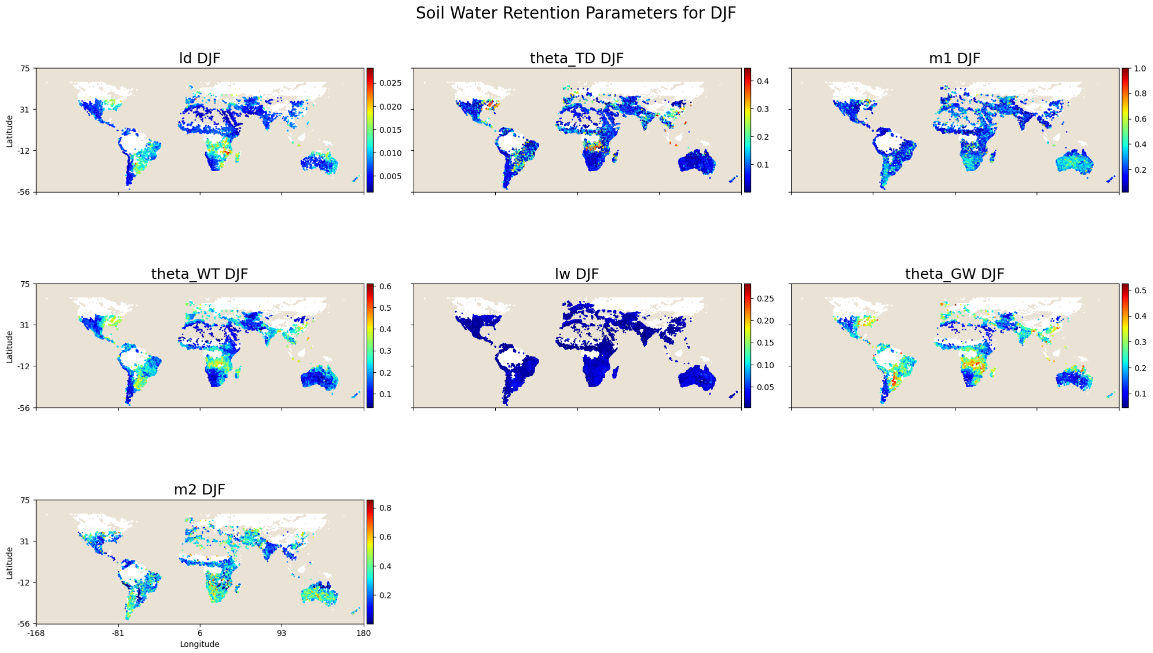


Figure 4: Spatial Distribution of Soil Water Retention Parameters for DJF season

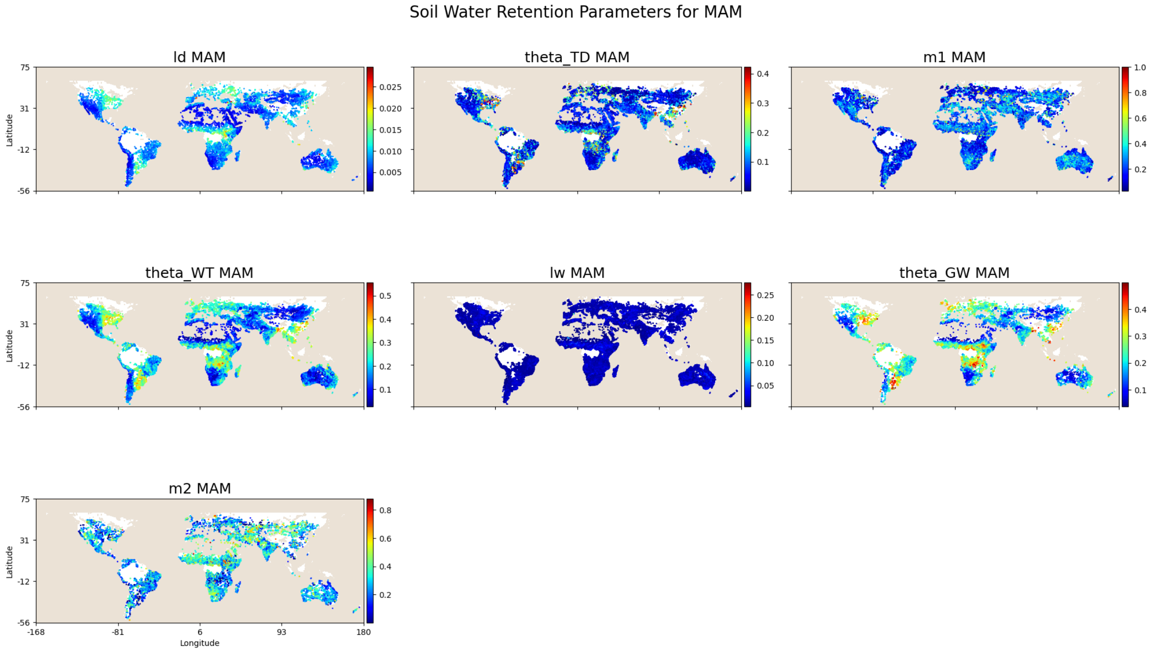


Figure 5: Spatial Distribution of Soil Water Retention Parameters for MAM season

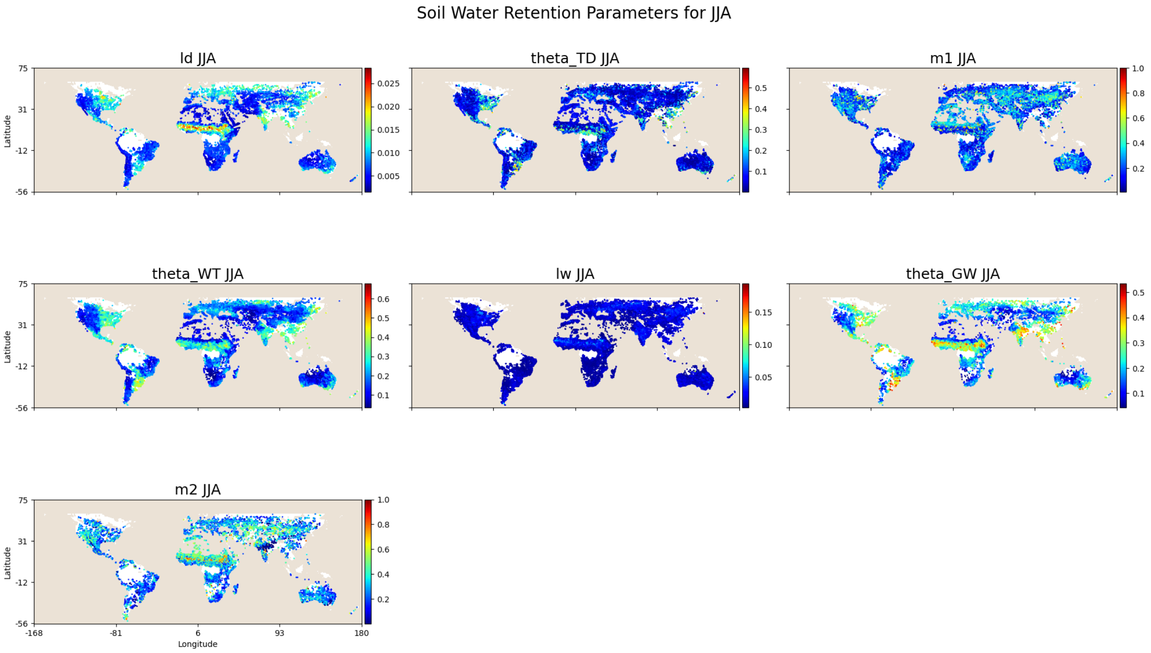


Figure 6: Spatial Distribution of Soil Water Retention Parameters for JJA season

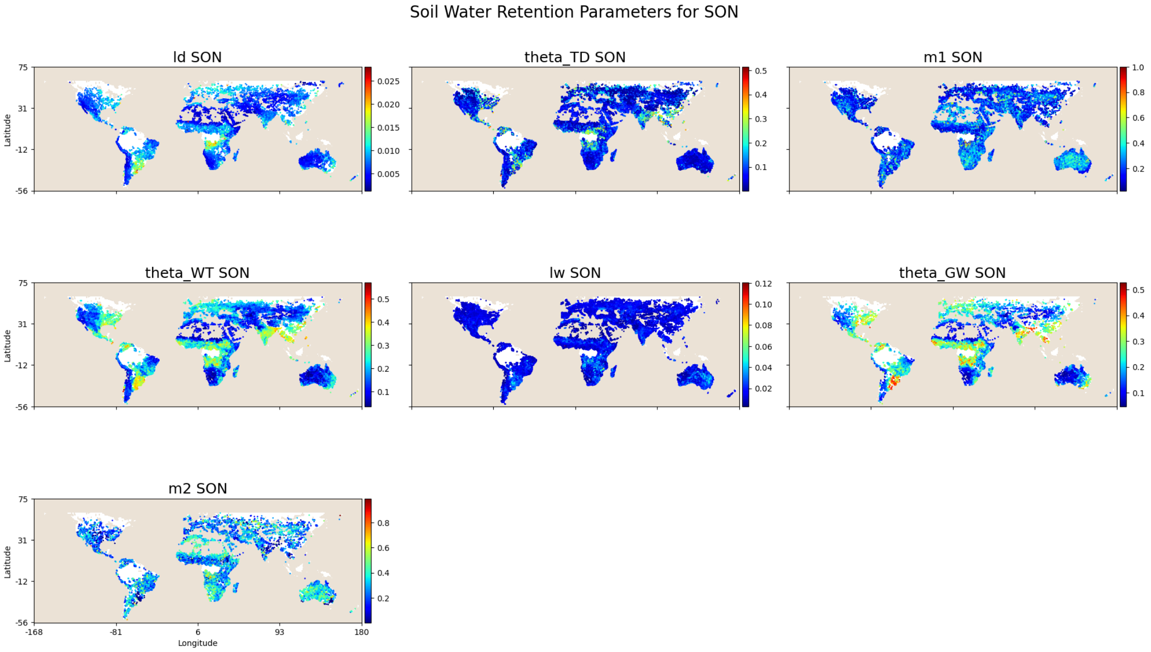


Figure 7: Spatial Distribution of Soil Water Retention Parameters for SON season

CHAPTER 3

METHODS

3.1 OVERVIEW

The Heterogeneity Index is computed using three different data-sets: Soil data where we have taken %sand as our first data-sets, leaf area index (LAI) with 18 years of data from 2002 to 2020, and the elevation model (DEM) to generate the necessary topographic parameters which includes flow accumulation. All of these data-set underwent the pre-processing step to ensure that they have an equal contribution while computing the heterogeneity index. The data-sets were normalized using MinMax normalization such that all three data-sets are in identical range. Afterwards they were resampled using bilinear resampling to make sure that the spatial frame of *%sand, LAI* and *flow accumulation* matches with each other.

The formulation of heterogeneity follows nomograph approach[12]. However, instead of doing normalization based on minimum value, we have transformed the heterogeneity and then normalized it based on the boundary percentiles. The normalized heterogeneity is then transformed again based on the median value over the 18 years of data-set. This kind of transformation will give a S-shaped curve where the data ranges from 0 to 1. Even though normalized heterogeneity is well defined, we still won't know at what degree any of these three variables (%sand, LAI, flow accumulation) are affecting the formation of land surface heterogeneity. To discover that, factor analysis was performed on these three variable, which will essentially give the direct influence over the index formation. The developed heterogeneity was then analyze with different hydroclimate region as well land cover divisions.

The current work also purpose a new framework for heterogeneity formulation called ***pyHetro***. The framework was built in python. The library is capable enough to handle large-scale data, where a size of an image can be $30,000 \times 77,0000$ pixels or greater. ***pyHetro*** successfully address the computation challenges which can arise while working at global scale for remote-sensing images.

Further study includes, evaluation of Heterogeneity and it's utility with machine learning algorithms. Four clustering algorithms were implemented to segregate eco-regions presented in heterogeneity based on SWRPs. The four algorithms that implemented are *K-Means Clustering, Spectral Clustering, Clustering based on Gaussian Mixture Models,* and *Hierarchical Clustering (Agglomerative Clustering).*

3.2 HETEROGENEITY INDEX

Heterogeneity Index is the combination of soil [13], leaf area index [24], and topography [31]. All three data-sets varying in different range and spatial resolution. Current index formation follows scaling nomograph approach from [12]. We have made the necessary changes in order to ensure that the Heterogeneity Index is properly scaled at a Global scale.

3.2.1 PRE-PROCESSING

Pre-processing is done in order to ensure that, there won't be any dominant effect due to very high or very low value from any of the data-sets. For instance, flow accumulation (topography) has very high values compare to %sand and LAI. Our previous analysis at US-Continental scale shows that due to very high flow accumulation heterogeneity index is being dominated by it. This will result in an inaccurate representation of soil moisture. We have divided the Pre-processing into two steps.

1. **Resampling**

   Resampling is done in order to match the spatial resolution of each data-set, so that at the time of the heterogeneity formation spatial frames of each data-set line up exactly. We have used *Bilinear Resampling Method*, which is a distance weighted average of the four nearest pixel values to estimate a new pixel value. The closest four cell from the input raster are taken with respect to cell center of the output processing cell. A weighted average based on the distance from the center is used for output output processing cell.

2. **Normalization**

   Normalization is done in order to equalize the boundary of all data-sets. Here, %sand ranges from 0 to 99 (theoretically 0 to 100), LAI ranges from 0 to 70 and Flow Accumulation ranges from 0 to 1.1×10^7. To address this inconsistency between the data-sets, *MinMax Normalization*

(Eq.1) was applied and all the data-set ranged into 0 to 100. Figure 8 shows data frequency for %sand, LAI and Flow Accumulation, where all three data-set are in a same range which helps during the index formation with the equal representation.

$$norm = \frac{current - min}{max - min} \times 100 \tag{1}$$

norm = Normalized value , current = Current value

min = Minimum value , max = Maximum value

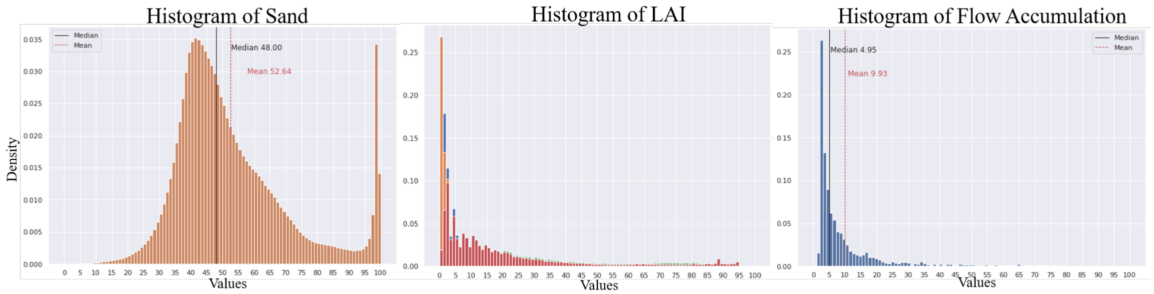


Figure 8: Data frequency of (from left to right) Normalized %sand, Leaf Area Index (LAI) and Flow Accumulation

3.2.2 INDEX FORMATION

This study incorporates the formation of the heterogeneity index. To represents the numerical variability and co-variability in soil, vegetation, and topography at the global scale %sand, leaf area index (LAI), and flow accumulation respectively are used as the chosen factors. To determines the infiltration capacity of the soil, %sand was chosen. To capture the different processes that causes soil moisture to change at the satellite pixel scale, physical factors such as indication of available leaf area for transpiration and rainfall interception are needed. These physical factors can be determined by LAI. Flow accumulation can determine the spatial patterns in the topography such as overland flow, localized evaporation, and infiltration by calculating the accumulated flow in each pixel. The accumulated flow represents the number of surrounding pixels that drain into a particular pixel. Heterogeneity (H) matrix (Eq.(2)) represents the heterogeneity for a sub-region that incorporates the variability and co-variability of the geophysical heterogeneity. The index was counted at an area of 36km(diameter) as well as at a finer resolution of 4km(diameter).

$$H = \begin{bmatrix} \sigma_{s,s} & \sigma_{v,s} & \sigma_{t,s} \\ \sigma_{s,v} & \sigma_{v,v} & \sigma_{t,v} \\ \sigma_{s,t} & \sigma_{v,t} & \sigma_{t,t} \end{bmatrix} \tag{2}$$

H = Heterogeneity Matrix

σ = Statistical Co-Variance

s,v,t represents soil(%sand), vegetation(LAI) and topography (Flow Accumulation) respectively

To scale the heterogeneity matrix, total area of the fine support pixels within the domain has been used (Eq.(3)). The scaling is done in order to account the rectangular extent of the data-set and square shape of the sub-region this way variance in the data-set can be considered with increasing or decreasing number of pixels.

$$H_a = \frac{H}{n \times A} \tag{3}$$

n = Number of fine support scale pixels within a sub-region

A = Area of fine support pixels (km^2)

The eigenvalue decomposition was done on the Heterogeneity matrix in order to have a single-valued index that represents the 3 dimensional heterogeneity. The eigenvalue decomposition done on scaled Heterogeneity matrix (H_a) projects the land-surface heterogeneity on a decorrelated vector space (equation (4)).

$$(H_a - \delta I)U = 0 \tag{4}$$

I = Identity Matrix

U = Eigenvector Matrix

δ = Eigenvalue (scalar)

To represent the maximum variability of the data where it is oriented, dominant eigenvalues (δ_{dom}) and the corresponding eigenvector (u_{dom}) were taken into consideration such that the orientation of this variability in a decorrelated vector space is described by an angle of the dominant eigenvector with a standard reference vector. Thus, making it distinct to the subpixel heterogeneity of each region. The heterogeneity index, H' Eq.(5), is defined as the product of δ_{dom} and the unique angle that u_{dom} makes with the vector, i.

$$H' = \delta_{dom} \times \cos^{-1}(\frac{u_{dom} \cdot i}{||u_{dom}|| \times ||i||}) \tag{5}$$

where, $\cos^{-1}(\frac{u_{dom} \cdot i}{||u_{dom}|| \times ||i||})$ = angle (in radians) between the dominant eigenvector and a reference vector $i([1\ 1\ 1])$. Maximum variability-covariability in heterogeneity cannot be oriented by a small δ_{dom} value along a single direction in the vector space. This infers either heterogeneity in the subregion is more divergent or the data has low variance. In a decorrelated vector space it is not possible to represent the variance-covariance structure in the data set by a single axis which can be termed as divergent heterogeneity. Thus, regions with low variance or less correlated heterogeneity(such as in agricultural domains where natural correlations between soil, vegetation, and topography have been disturbed) should have a relatively lower heterogeneity index.

Normalization and Transformation of Heterogeneity Index

Now that we have the *Raw Heterogeneity Index*, it is still not bounded by a theoretical limit. The Raw Heterogeneity has no upper bound. In order to make it well defined, we consider an approach what we call *Transform-Normalize-Transform* in short Trans-Norm-Trans. The raw heterogeneity index is transformed using the *Logarithmic Transformation* (Eq.6), afterwards the *log* transform index is normalized using boundary based percentile *MinMax Normalization* (Eq.8) and finally, the normalized index is again transformed using *Sigmoid Transformation* (Eq.9) that represents the Heterogeneity Index.

1. ***Logarithmic Transformation***

   Raw Heterogeneity Index has a skewed distribution due to no upper bound. Log transformation vs Raw Heterogeneity in Figure 9 shows the skewed values present in the Raw Heterogeneity. To address that, log transformation has been applied to make it less skewed. Moreover, the log transformation also makes patterns in the data more interpretable.

$$log\_H' = log(H') \tag{6}$$

log_H' = Log Transformed Heterogeneity Index

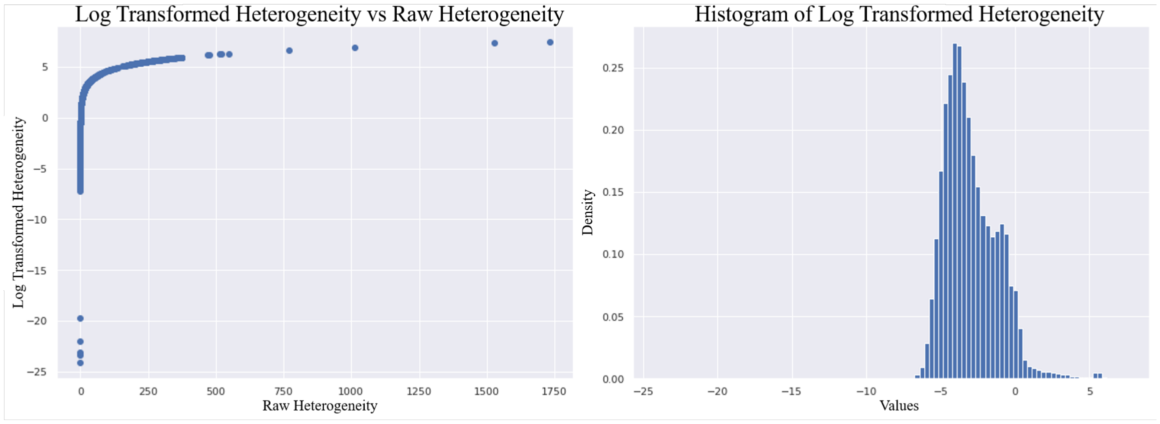


Figure 9: Log Transformed Heterogeneity

2. ***MinMax Normalization***

Even though log transformation gives better interpretation to data patterns, there are still extreme values remaining that affects the land surface heterogeneity at a global scale. Also, log transformation ranges from negative to positive which will create problems when sigmoid transformation is applied. Therefore, to convert the transformation on the positive side of spectrum and to consider those extreme values, percentile boundary based MinMax normalization has been applied, where $globalMin$ is taken as $0.5^{th} percentile$ and $globalMax$ is taken $99.5^{th} percentile$ from 18 years of data. Before the MinMax normalization, extreme values beyond the 0.5^{th} percentile and 99.5^{th} percentile are floored to 0.5^{th} percentile and 99.5^{th} percentile(Eq. 7) respectively. This will make the data linear and as a result, Figure 10 which shows the clear distribution with those extreme values.

$$log\_H' = \begin{cases} globalMin & \text{if } log\_H' < globalMin \\ globalMax & \text{if } log\_H' > globalMax \\ log\_H' & \text{otherwise, for all } log\_H' \end{cases} \tag{7}$$

$$norm\_H' = \frac{log\_H' - (1.01 \times globalMin)}{(1.01 \times globalMax) - (1.01 \times globalMin)} \tag{8}$$

log_H' = MinMax Normalized Heterogeneity Index

globalMin = 0.5^{th} percentile of log_H', globalMax = 99.5^{th} percentile of log_H'

Here, globalMin and globalMax are min and max value of log_H' for entire timeline(18 years of data, 2002 to 2020).

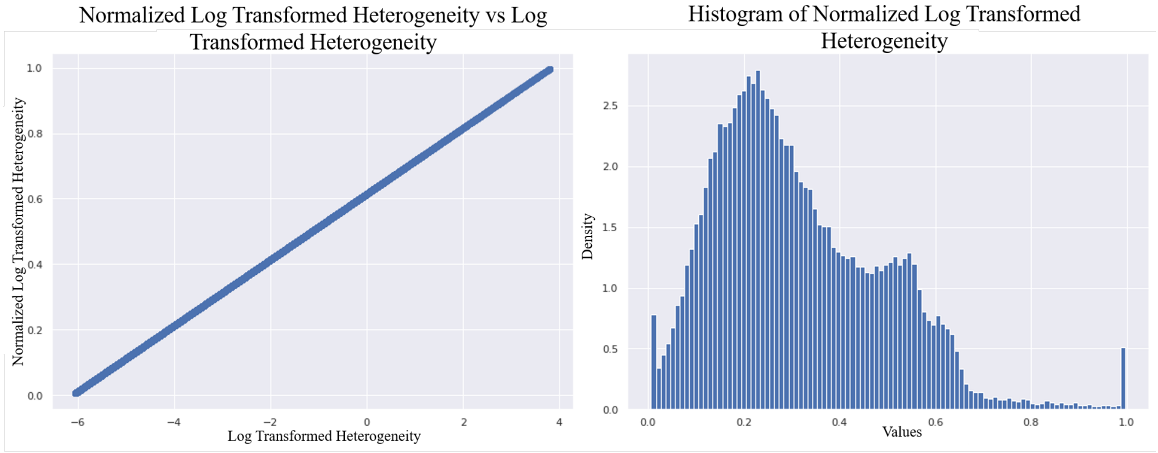


Figure 10: Normalized Log Transformed Heterogeneity

3. ***Sigmoid Transformation***

Finally, to make the distribution of data less skewed and normal, sigmoid transformation is applied (Eq.9). Sigmoid transformation gives a well defined *Normalized Heterogeneity Index* (Fig 11), in which the data ranges from 0 to 1 with the *S-shaped* curve of the data points resulting in a non-linear transformation. The sigmoid transformation is achieved based on the *median* value of 18 years of norm_H' (2002 to 2020).

$$sigmoid\_H' = \frac{1}{1 + (\frac{median}{norm\_H'})^{Alpha}} \tag{9}$$

sigmoid_H' = Sigmoid Transformed Normalized Heterogeneity Index

median = Median value of the entire timeline(2002 to 2020) of norm_H'

Alpha = Constant (Default 3)

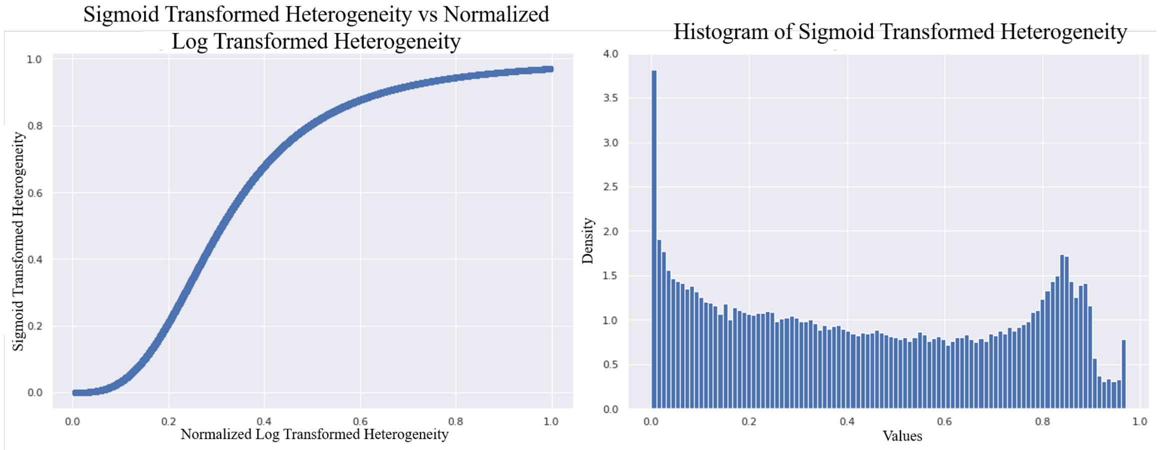


Figure 11: Sigmoid Transformed Heterogeneity based on the Central value of Heterogeneity Index

Figure 12 and 13 represent the H-Index at 4km and 36km respectively, for Land Surface Heterogeneity, which was scaled and then transformed at a Global scale. H-index at 4 km is a much finer resolution which is able to show the changes in Land Surface Heterogeneity more accurately.

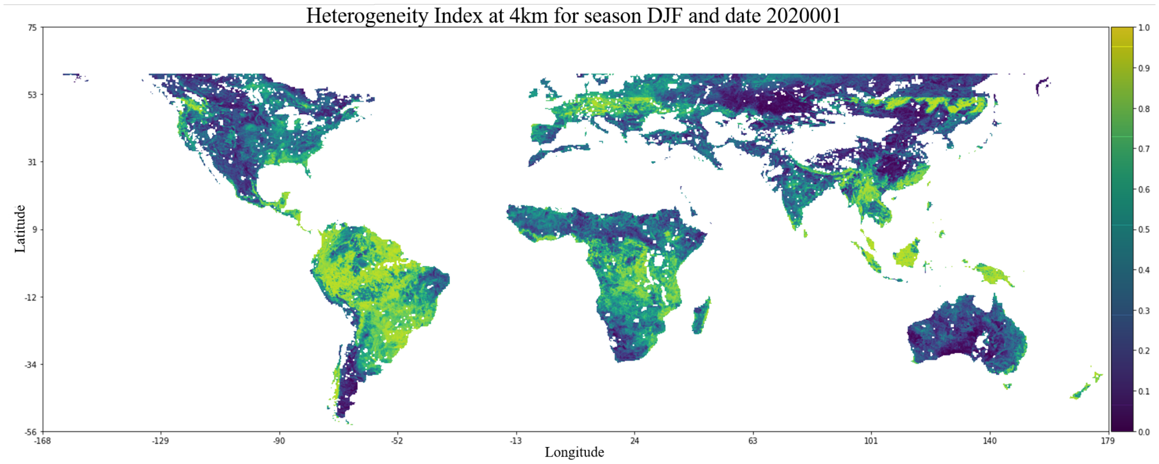


Figure 12: Global Heterogeneity Index at 4 km (Date: 1^{st} January 2020, DJF season)

3.3 PYHETRO

To develop the heterogeneity at a Global scale, this study uses %sand at 250m resolution, Leaf Area Index (LAI) at 500m resolution and Digital Elevation Model (DEM) at 30m resolution. Due to the very high resolution, it becomes challenging to process the data from computational perspective. The size of the images can go upto $30,000 \times 77,000$, that becomes computationally expensive to just

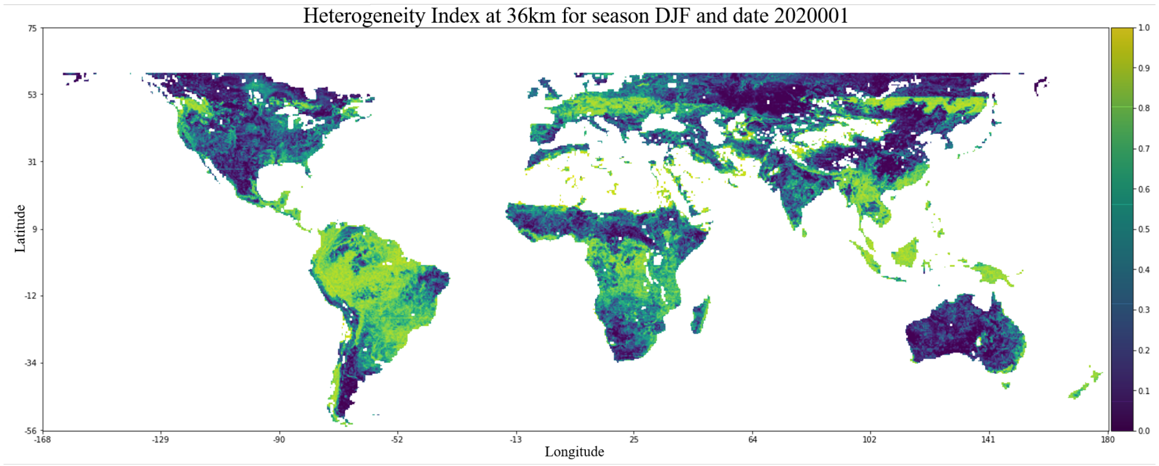


Figure 13: Global Heterogeneity Index at 36 km (Date: 1st January 2020, DJF season)

perform a normal operation on these images.

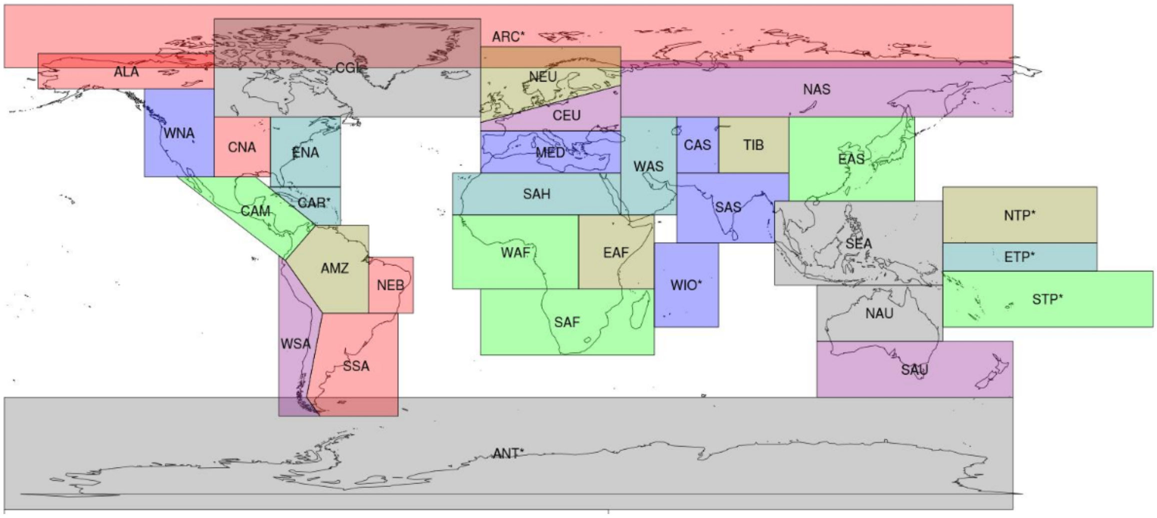


Figure 14: AR5 Regions [23]

To address this issue and standardize the process, data-sets are split based on *IPCC's AR5 Regions*. The International Panel on Climate Change (IPCC) has formed these sub-regions based on different climate models that uses a range of methods to deal with the land and coastal boundaries. Figure 14 shows the regions used to calculate regional summary statistics in AR5 that were split into 26 SREX regions defined by the IPCC's Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (IPCC, 2012: also known as "SREX") [23].

Despite the data being split based on the standardize sub-regions, image size of sub-regions

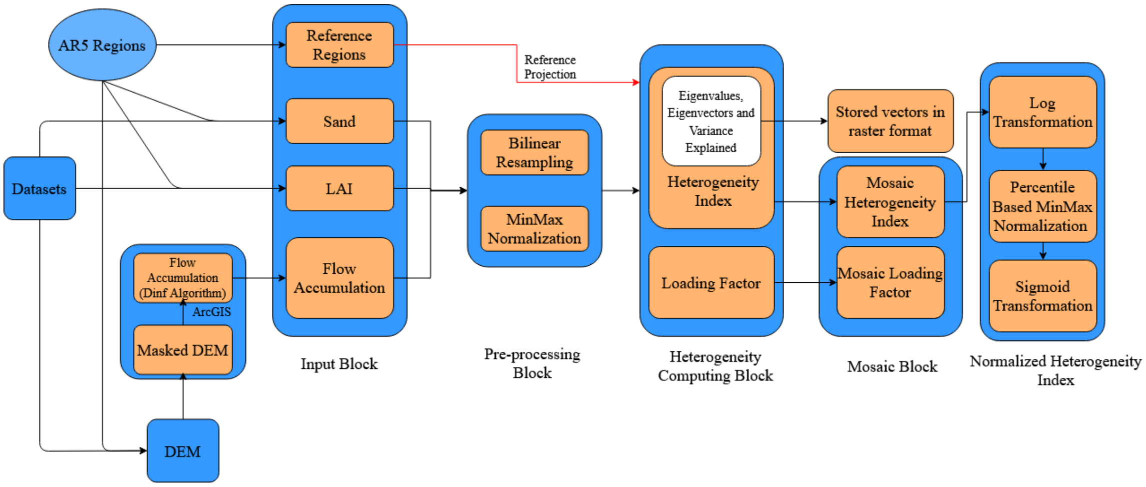


Figure 15: ***pyHetro:*** High-level schema

might remain too big to get the robust results. This can become a huge challenge when it comes to perform a batch processing or to formulate the heterogeneity at a much coarse resolution (e.g. 4km). To address the issue of scalability, ***pyHetro*** incorporates multiple modules that can be used according to the amount of resources available or the computation needed based on spatial resolution of the images.

Figure 15 shows a high-level schema for the ***pyHetro*** framework. The framework is fully modular at the same time it can work as an independent service, given the input data for the heterogeneity index formation, the framework takes care of all necessary steps that are required. To address the scalability and robustness, ***pyHetro*** framework supports parallel processing as well as distributed computing. For the distributed computing, this framework can easily connect to the systems that are on same network. This gives an opportunity to form the local clusters which can be highly scalable. But in the distributed computing, process requires data to be shared among the connected nodes in the clusters, that can reduce the speed of the index formulation due to every large size of the image, though this issue is not a problem for ***pyHetro***, as it doesn't share the entire image across the clusters, instead it only distributes the minimum amount of data required to compute the heterogeneity for any pixel.

3.4 CLUSTERING MODELS

To evaluate the utility of the Heterogeneity Index as a classifier of global scale, we have leveraged the various types of clustering algorithms on soil hydraulic parameters. As discuss in the previous chapters to achieve adequate segregation among the hydraulic parameters, we have performed the following experiments on the following algorithms.

3.4.1 EXPERIMENTAL SETUP

In this chapter, two series of experiments are performed towards the clustering of soil water retention parameters for heterogeneity classification. In the first series, seven hydraulic parameters were used, namely l_D, θ^{TD}, m_1, θ^{WT}, l_W, θ^{GW}, and m_2. For the purpose this study, these 7 parameters are named *All Parameters.* These parameters were able to give the seasonal variability in the soil hydraulic parameters. In the second series, only four hydraulic parameters were used, the parameters were θ^{TD}, θ^{WT}, θ^{GW}, and m_2. For the purpose this study, these 4 parameters are named *Selected Parameters.* The reason being, study in [14] proved that these four parameters (*selected parameters*) were able to indicate dominant hydrologic regimes of soil moisture drydown process.

Several unsupervised machine learning algorithms such as *K-Means Clustering, Spectral Clustering, Clustering based on Gaussian Mixture Models and Hierarchical Clustering* were used for the purpose of evaluating the utility of the heterogeneity index with different Eco-regions. Python based machine learning library *scikit-learn* [26] was used for the implementation of these machine learning algorithms. Internal cluster validity indices were taken into account to find out the optimal number of cluster for all the seasons. To have comparative results, throughout the seasons as well as between different clustering algorithms, a fixed number of optimal clusters were taken into account.

K-means is considered as one of the simplest algorithm for unsupervised clustering, but the requirement of complete data matrix becomes a challenge. The parametric data we have on soil hydraulics are the type of Missing Not At Random (MNAR). Performing imputation on the hydraulic parameter can result in an inaccurate separation. To minimize the affecting factor from imputation, K-Means Clustering was implemented which was inspired from [4] and [29].

For both the types of parameters- all parameters and selected parameters, the missing values were handled by doing a mean imputation for the initial iterations maintaining a map of these missing values. On the next iteration, all the missing values were assigned to their respective cluster centroid. That is, in next iteration, the distance of these missing values will be zero from their respective centroid and the only factor that affects the direction of the centroids will be the original data which were not missing. The algorithm does this iteratively till a convergence is achieved.

For spatial data, k-means might not work properly, even though the weightage of imputations were minimized. Whereas, spectral clustering can prove to have a significant result where the algorithm performs a low-dimension embedding of the affinity matrix between the data. For the current experiment, affinity matrix of the spectral cluster was constructed using a radial basis function (RBF) kernel with the kernel co-efficient of 1. These kernel represents as feature vectors for input data-set. The computation were done using parallel jobs.

Model based clustering were found to have a high degree of segregation. The experiments were done using Gaussian Mixture Model which implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. Due to ellipsoids, shape of GMMs maximizes only the likelihood and it will not bias the means towards zero, or bias the cluster sizes to have specific structures, which acts in favour of spatial data-sets. The co-variance was fully constrained for this experiments that means each component has its own general co-variance matrix.

The last experiment was done using Agglomerative Clustering which is type of Hierarchical clustering using a bottom up approach. Agglomerative Clustering also uses affinity matrix to complete the linkage distance, but compared to spectral clustering affinity matrix that was used here is based on the euclidean distance. The linkage criteria that was used here was the minimizing the variance of the clusters being merged.

3.4.2 ABLATION EXPERIMENTS

Optimal Number of Clusters

One of the vital issues with clustering algorithms is their validation and it becomes essential to have a well defined techniques for cluster validation. In this experiments, internal cluster validity indices have been adapted in order to find out the optimal number of clusters for validating all the cluster algorithms. The experiment incorporates three different cluster validity indices *Silhouettes Score* [28], *Davies-Bouldin Index* [6] and *Dunn's Index* [7]. The combination of these three validity indices were used to find the optimal number of clusters *(N)*. Total 19 different clusters were used ranging from $N = 2$ to $N = 20$. The experiments were performed on all four seasons for SWRPs, using both the sets of parameters separately (*All Parameters* & *Selected Parameters*). The evaluation metrics for validity indices are as follows,

Silhouette Score:

Silhouette Score provides the validation of consistency within the clusters by comparing the similarity of data-points from one cluster to other clusters. The optimal number of cluster is determined by maximizing the silhouette score.

Davies-Bouldin Index (DB):

DB Index is computed by averaging all the highest similarity assigned to the clusters. The highest similarity is taken by computing the similarity between each cluster and all other clusters. The optimal number of cluster is determined by minimizing the index.

Dunn's Index:

Dunn's Index is computed by taking the ratio of inter-cluster separation and intra-cluster compactness. The optimal number of cluster is determined by maximizing the Dunn's Index.

Statistical Significance for Cluster Separation with Heterogeneity Index

Based on the optimal number of clusters, the subsequent experiments were done for statistical significance on cluster separation. For both sets of parameters (*All Parameters* & *Selected Parameters*), experiments were divided into two series. In the first series, experiments were carried forward with a similar number of clusters for all the clustering algorithms. This was done in order to compare the performance of the cluster algorithms and also to determine which algorithm is efficient for Soil Water Retention Parameters. The number of clusters were decided by internal cluster indices for all the algorithms. The second series of experiments were carried forward with a different number of clusters for all the clustering algorithms, but with a similar number of clusters for all the seasons (*DJF, MAM, JJA and SON*) respective to their clustering algorithm. The number of clusters for the second series of experiments were decided based on the internal cluster indices of four seasons respective to their clustering algorithm.

Tukey's HSD (Honestly Significant Difference) test were performed. The clustered points were mapped to the heterogeneity index in order to extract the corresponding values associated to it.

These heterogeneity index values were then used with the cluster groups in order to find the significance between the clusters. Tukey's HSD test was used in order to indicate the cluster significance for all four algorithms and to find out the pairwise difference between clusters groups.

CHAPTER 4

RESULTS AND DISCUSSION: HETEROGENEITY INDEX FORMULATION

4.1 OVERVIEW

There are significant temporal changes in heterogeneity. In this chapter, further analysis was done in order to ensure the validity of the heterogeneity at a global scale. Factor analysis was done in order to find influential factors which were affecting the heterogeneity during the formation. This type of analysis can give a better insight to understand the heterogeneity at a remote-sensing scale. Additionally, correlation between heterogeneity and leaf area index was calculated to determine the degree of similarity between them. The H-index was then compared with different hydro-climate and land cover to determine the it's spatio-temporal characteristics.

4.2 FACTOR ANALYSIS

While the normalized heterogeneity is well formed and scaled properly, we would also like to know the major factors affecting those indices. One way to know that is to perform factor analysis. All though factor analysis is related to Principal Component Analysis (PCA), but they are not identical. Essentially, we would like to know influential factors or variables which are affecting heterogeneity index or how correlated they are with each other.

In a nut-shell, factor analysis describes the variability among observed variables, correlated variables in terms of a potentially lower number of unobserved variables called factors. For instance, it is possible that variations in any one of the three observed variables: %sand, LAI and Flow Accumulation may reflect the variations on the heterogeneity index, which is a unobserved or underlying variable. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus ”error” terms. By doing that factor analysis aims to find independent latent variables.

In order to have a better understanding of temporal changes in the land surface heterogeneity, We have divided our analysis according to four season: Winter - December to February (DJF), Spring - May to March (MAM), Summer - June to August (JJA) and Fall - September to November (SON).

Our analysis follows *Principal Factor Method* also known as *Principal Axes Factor*. Principal Factor Method is a exploratory analysis with a straightforward procedure. Successive eigenvalue decomposition are done on a correlation matrix where the diagonal values are replaced with the sum of diagonal values until they do not change. The *loadings* derived from these factor analysis tells us whether they have strong influence or weak influence on an unobserved variable. The loading factor ranges from -1 to 1, where -1 is for negative influence, 1 is for positive influence, and 0 is for weak influence. In this analysis, we have taken absolute loading factors to clearly understand the changing behaviour between weak and strong influence on heterogeneity through-out the seasons.

The Loading Factors and 1^{st} Variance Explained are taken as seasonal average across all years from 2002 to 2020. Figure 16 represents the loading factor for winter (DJF) season where we have high influence from %sand and low to moderate influence from flow accumulation, where as LAI is ranging form moderate to high influence. When the current analysis compared with Spring (MAM) season (Fig. 17), all though %sand and flow accumulation doesn't change much but the influence of LAI is shifting from low to high for United Continent and major part of the Globe. This is clearly reflected on *percentage variance explained* for winter and spring. Spring season has comparatively less percentage variance when compared to winter season. This indicates that the seasonal changes in heterogeneity is depended on leaf area index.

The similar analysis can be drawn towards summer and fall seasons. In summer (Fig. 18), %sand and flow accumulation has slightly higher influence compared to spring. Leaf area index has higher influence towards Eastern United States as well around middle east region also the percentage variance shifted from high to very high around the Globe. As for the fall season (Fig. 19), leaf area index started to get a high influence where as %sand and flow accumulation only has slight variance towards their influence. On the other hand, percentage variance for fall became lower than summer as well as winter.

From our analysis, its clear that the temporal variability in the heterogeneity index is corresponding to the vegetation.

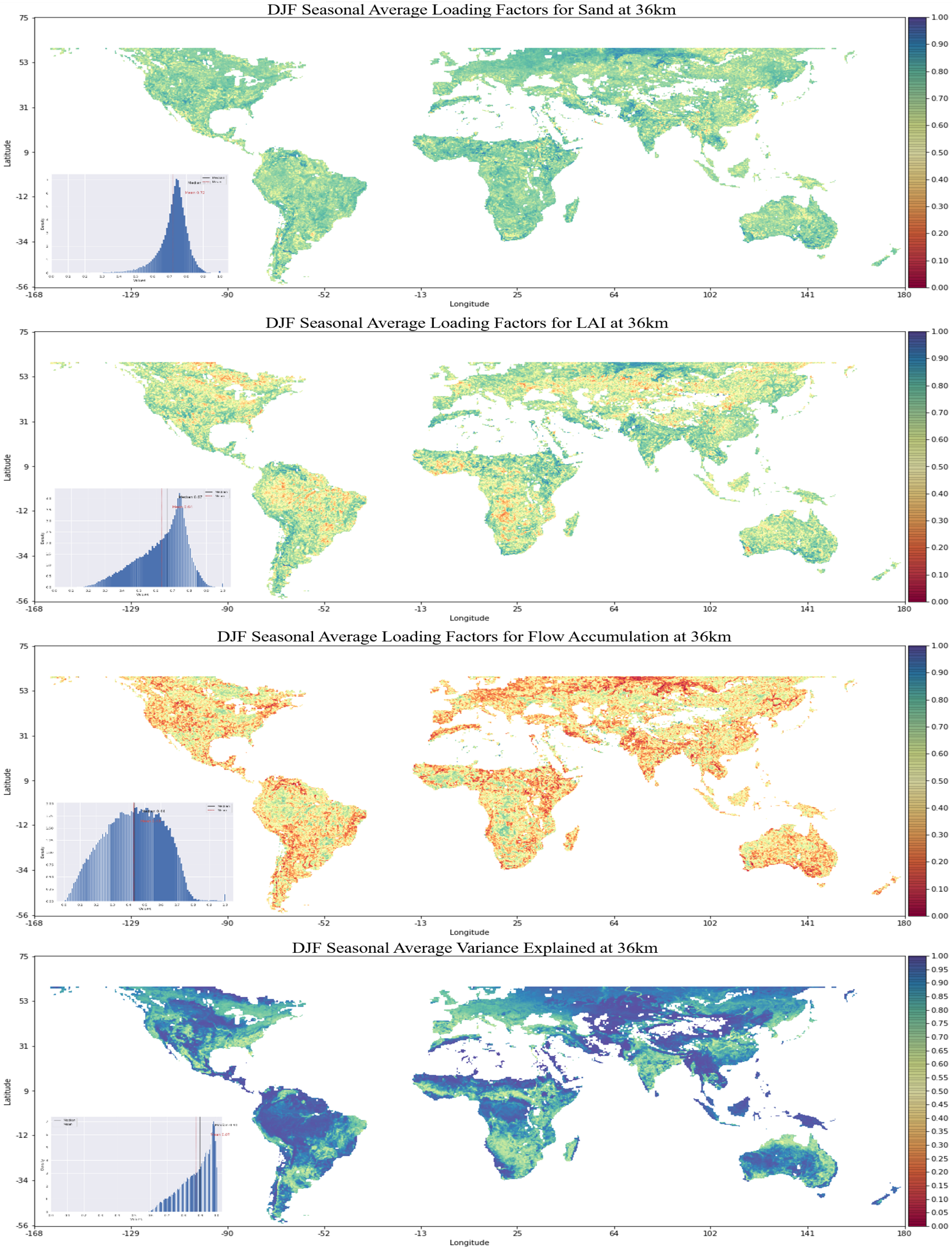
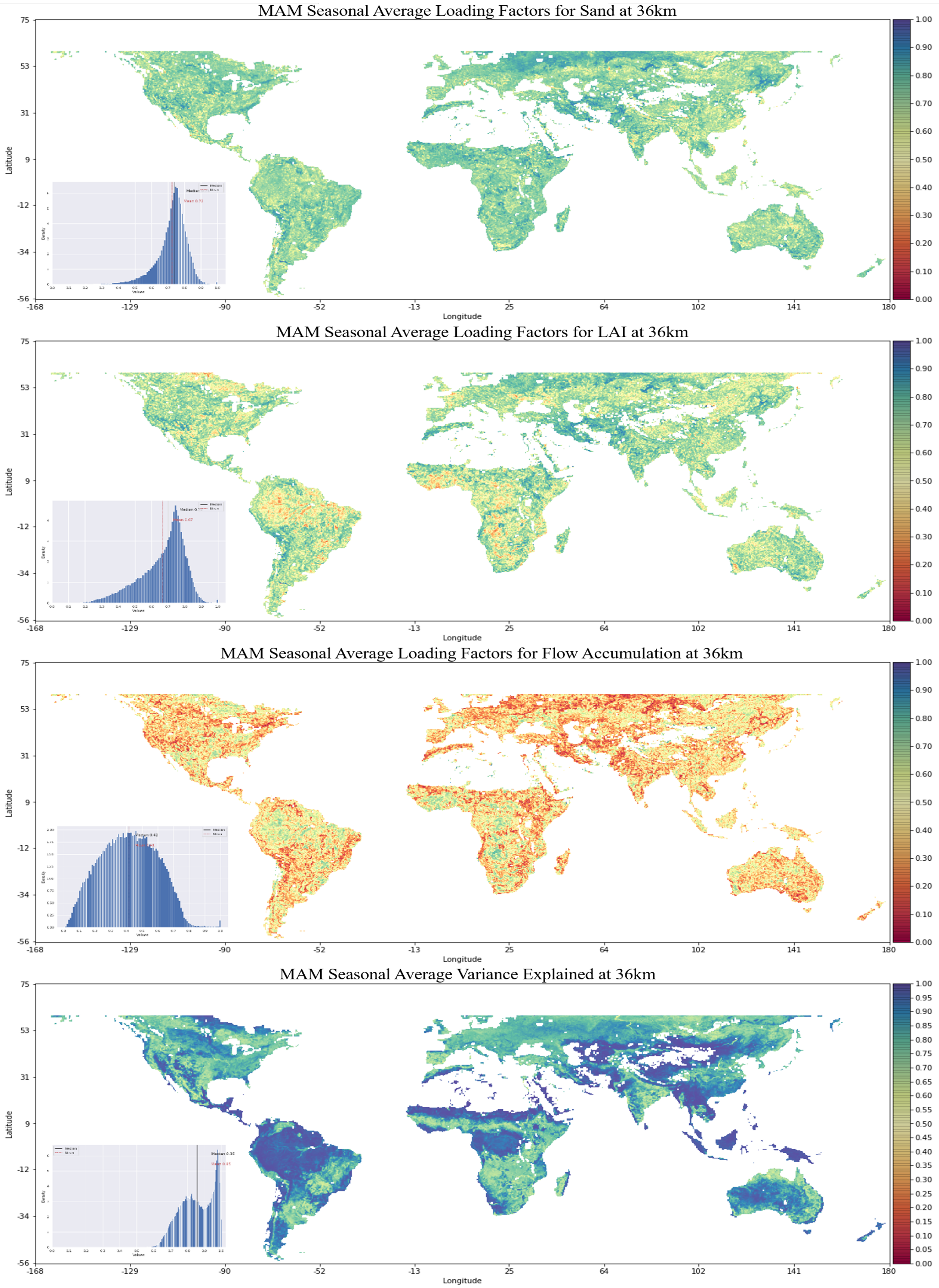

DJF Seasonal Average Loading Factors for Sand at 36km
DJF Seasonal Average Loading Factors for LAI at 36km
DJF Seasonal Average Loading Factors for Flow Accumulation at 36km
DJF Seasonal Average Variance Explained at 36km
Latitude
Longitude
75
53
31
9
-12
-34
-56
-168
-129
-90
-52
-13
25
64
102
141
180
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
Values
Median
Mean
Median 0.67
Median 0.44
Mean 0.87

Figure 16: Loading Factors and 1st Variance Explained are seasonal average across all years from 2002 to 2020 for DJF, (from top to bottom) loading factor for %sand, LAI and Flow Accumulation respectively and the bottom one is 1st Variance Explained



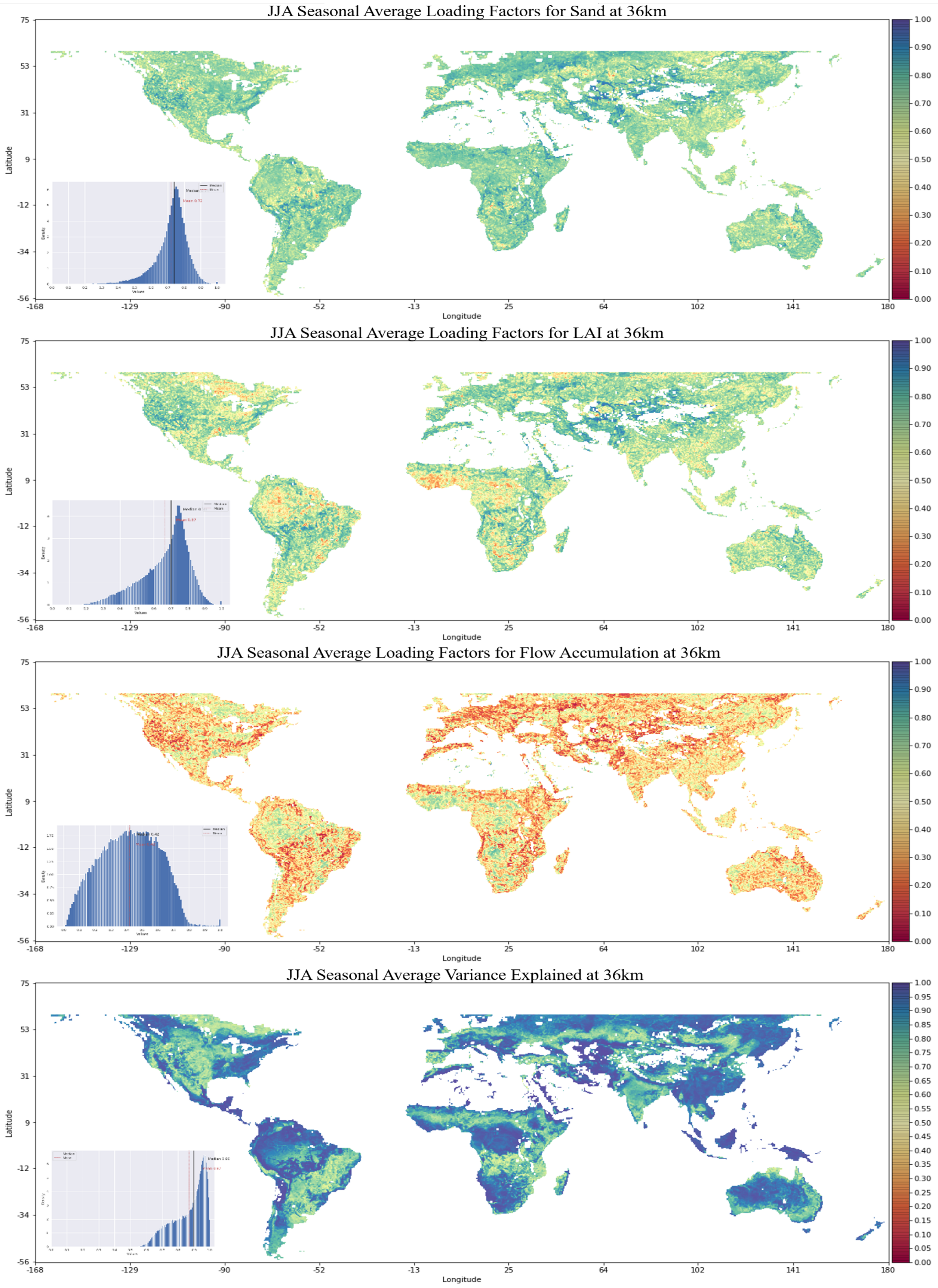
MAM Seasonal Average Loading Factors for Sand at 36km

MAM Seasonal Average Loading Factors for LAI at 36km

MAM Seasonal Average Loading Factors for Flow Accumulation at 36km

MAM Seasonal Average Variance Explained at 36km

Figure 17: Loading Factors and 1st Variance Explained are seasonal average across all years from 2002 to 2020 for MAM (similar to Figure 16)



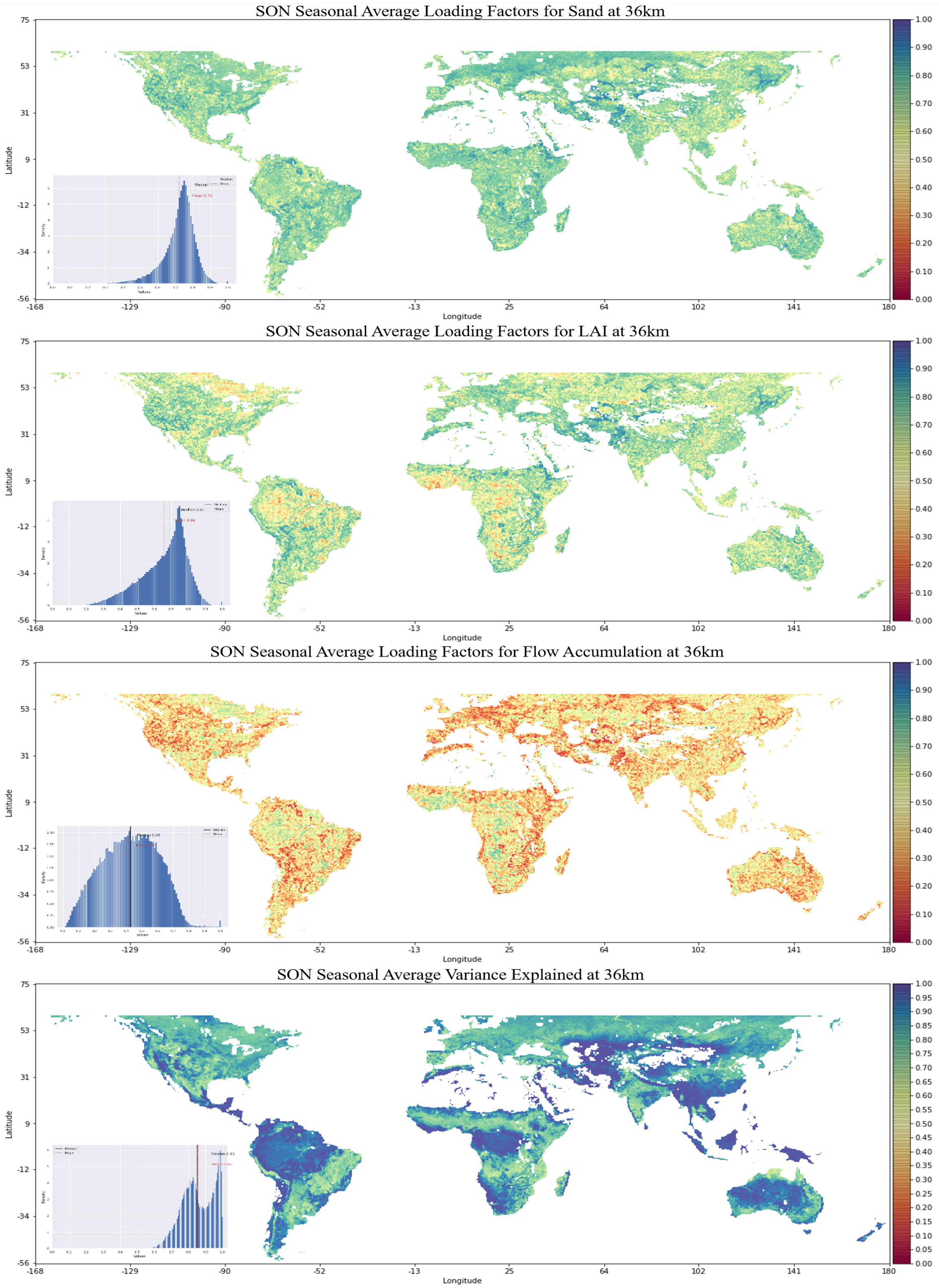
JJA Seasonal Average Loading Factors for Sand at 36km

JJA Seasonal Average Loading Factors for LAI at 36km

JJA Seasonal Average Loading Factors for Flow Accumulation at 36km

JJA Seasonal Average Variance Explained at 36km

Figure 18: Loading Factors and 1st Variance Explained are seasonal average across all years from 2002 to 2020 for JJA (similar to Figure 16)



SON Seasonal Average Loading Factors for Sand at 36km

SON Seasonal Average Loading Factors for LAI at 36km

SON Seasonal Average Loading Factors for Flow Accumulation at 36km

SON Seasonal Average Variance Explained at 36km

Figure 19: Loading Factors and 1st Variance Explained are seasonal average across all years from 2002 to 2020 for SON (similar to Figure 16)

4.2.1 CORRELATION BETWEEN LEAF AREA INDEX AND HETEROGENEITY INDEX

Its also an important analysis to find out, how correlated heterogeneity index is with different parameters: soil, vegetation and topography. But what kind of correlation should be used here? Karl Pearson developed the index that we still use to measure correlation, Pearson's r [25]. On the other hand, [17] have presented different computational and conceptual definition of r which are represented into 13 different formulas. Each formula suggests a different way of thinking about correlation coefficient, it can be from algebraic or geometric, or it can be in trigonometric settings. [17] have shown that Pearson's r (or simple functions of r) may vary depending on the type of approach is taken to the correlation problem. For our analysis, we have used correlation as standardized covariance.

Based on the factor analysis, we found out that spatiotemporal changes happening in heterogeneity are corresponding to vegetation (leaf are index). It is something that we were expecting from that way we have design the index formation. But, we don't want both Heterogeneity and LAI to have high correlation with each other. If they have, then there won't be a point to have heterogeneity index in the first place.

Additionally, to perform an extended analysis on the temporal correlation between LAI and Heterogeneity, it is essential to compare both of them and get a correlation co-efficient for seasonal average values. The seasonal average values are taken based on the season for each year. But before performing the correlation we have to match the resolution of both the rasters. LAI is at 500 meter resolution where as Heterogeneity is at 36km resolution, if we perform a correlation on the current raster images then the results won't be accurate. Therefore, in order to have an equal resolution we have downsampled LAI using *Average Resampling* method. The reason why we did average resampling instead of bilinear resampling is becasuse bilinear interpolation gives a weighted average at the centroid of the upscaled pixel. Therefore, these values are interpolated to a specific location within the large pixel, which is good when we are interpolating values for single latitude and longitude. But in this case, we wanted to find the effective value for the entire pixel and not just the centroid. That's why we have used average resampling instead of bilinear resampling.

Based on the Figure 20, which shows the *r_value* or correlation coefficient for all four season through-out the 18 years of data, it mostly ranges from 0.5 to 0.6 with spring season (MAM) to have the highest correlation while Fall season (SON) to have lowest correlation out of all four seasons. There is not much difference between seasonal correlation, if we look at the seasonal correlation and their distribution, interquartile range of winter season (DJF) and spring season (MAM) lies between 0.54 to 0.57 with the median of 0.55 and 0.56 respectively. Where as, interquartile range for summer season (JJA) and fall season (SON) lies between 0.50 to 0.56 with the median of around 0.54 for both of them. These moderate correlation can be considered appropriate where heterogeneity index is not completely corresponding to leaf area index.

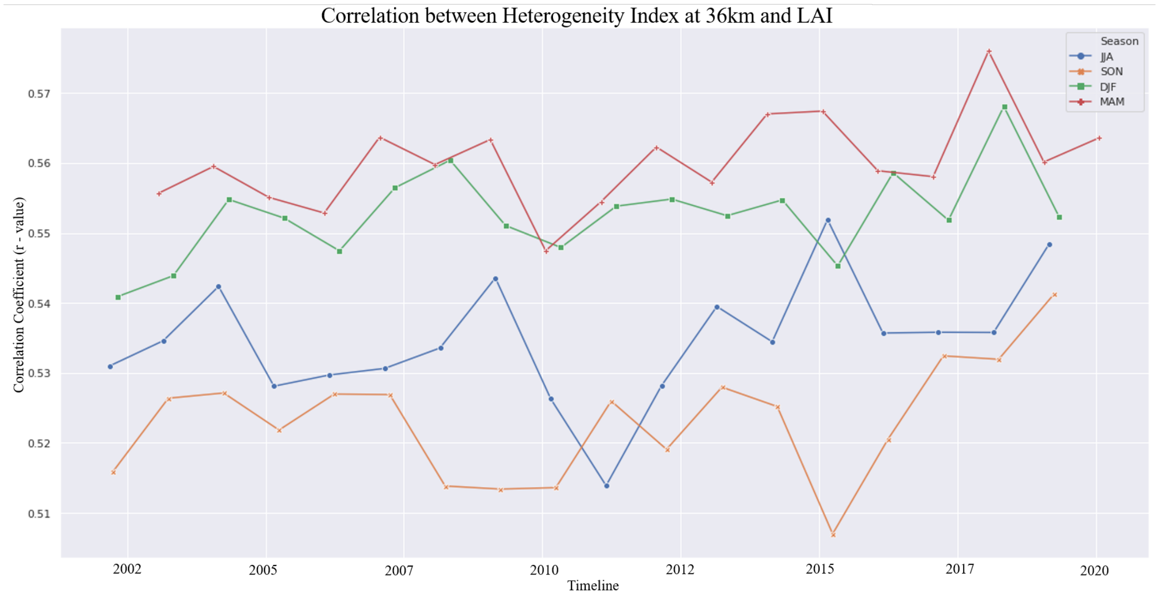


Figure 20: Correlation between Heterogeneity Index and LAI

4.3 COMPARISON WITH HYDRO-CLIMATE

Hydro-climate regions are categorized into different categories. For this analysis we have used Global Aridity Index where the method to derived the dataset is followed from [36] and [37]. Global Aridity hydro-climate has five different categorize and index associated with it. *1 - Hyper Arid, 2 - Arid, 3 - Semi Arid, 4 - Sub Humid* and *5 - Humid*. Hyper Arid regions are mostly complete deserts i.e. Savannah. In this analysis we have excluded the Hyper Arid regions, reason being if we look at the Figure 21 (Aridity Map) and compare it with Figure 13 (Normalized Heterogeneity), there is not much data preset for Hyper Arid region. With the less data points we won't be able to have accurate analysis, thus Hyper Arid region was excluded in the analysis.

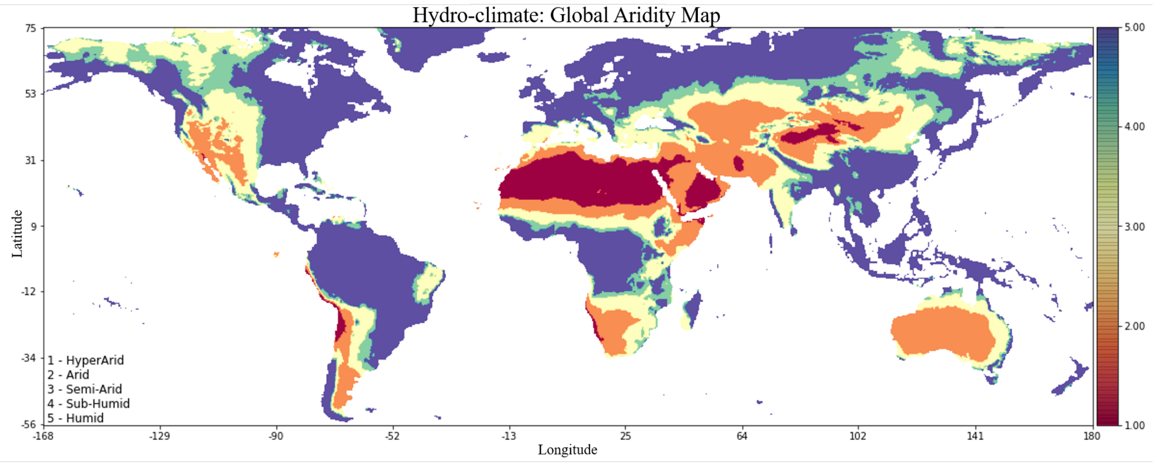


Figure 21: Hydroclimate: Global Aridity Map

So far, our previous analysis using factor loading slightly pointed out that heterogeneity index is corresponding to vegetation (LAI). If that was the case, then it means that the only temporal component in heterogeneity is coming from LAI, so if we classify the temporal variability based on the hydro-climate (Fig. 22), it should have very big differences across the seasons. But that wasn't the case here, if we look at Figure 22 except for a one or two classes (Sub Humid and Humid), there are no significant changes going on in terms of LAI.

It is also possible that we were expecting LAI to have a temporal component due to the only dynamic factor present in the index formation but based on our current analysis Heterogeneity is encompassing not just vegetation (LAI) but also topography (Flow Accumulation) as well as soil (sand). Therefore, vegetation is just one part of it. Now, given soil also dependent on hydro-climate, specially at that temporal-scale soil can develop based on hydro-climates.

When we look at summer season (JJA) in Figure 22 for heterogeneity index, everything remains almost similar except for Sub-Humid and Humid hydro-climates. this indicates that there is a temporal signature present here, the variance of heterogeneity also reduces compare other seasons. whereas loading factors doesn't change much. This proves that the *Heterogeneity* is not only driven by LAI, it also driven by other factors like topography and soil. Though one thing to note is that soil remains dominant in the index formation.

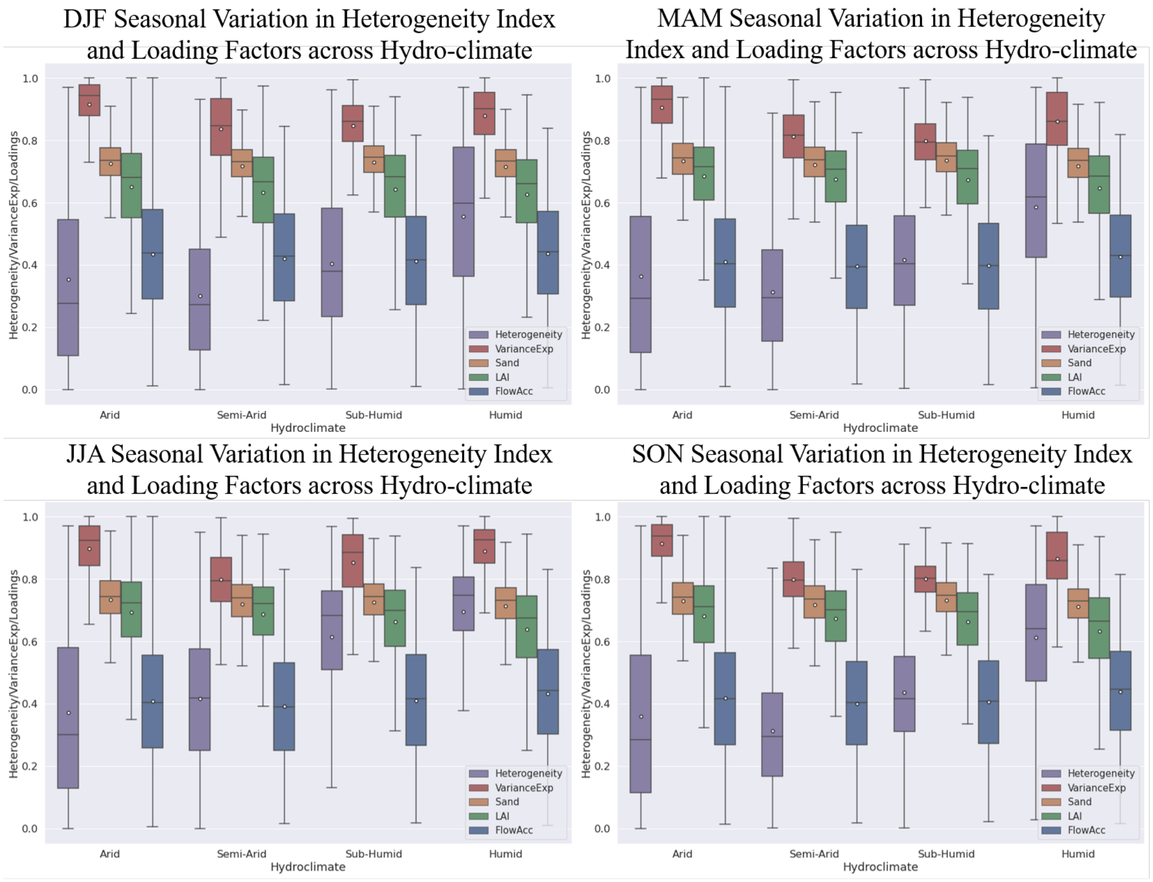


Figure 22: Comparison of Heterogeneity Index and Loading Factors with different Hydroclimates. The comparison is done for all seasons top-left is winter (DJF), top-right is spring (MAM), bottom-left is summer (JJA) and bottom-right is Fall (SON). The Box plots are representing Heterogenity Index(Purple), Percentage Variance Explained (Red), Loading Factor from Sand (Yellow), Loading Factor from LAI (Green) and Loading Factor from Flow Accumulation (Blue).

4.4 COMPARISON WITH LAND COVER

Land cover represents the actual or physical presence of vegetation (or other materials where vegetation is nonexistent) on the land surface. There are 16 types of different land cover (excluding water). These land cover defines different types of vegetation. In this section, we have compared the heterogeneity index and loading factors with the different types of land cover. To summarize the results better we have combined all the different types of forest into a single category called forest. So from Figure 23, all 1 to 5 forest became as *All Forest*. We have also combined Closed Shrublands and Open Shrublands as Shrublands. Current analysis excludes land cover “snow and ice” and “urban” as those land cover are not necessary for the analysis.

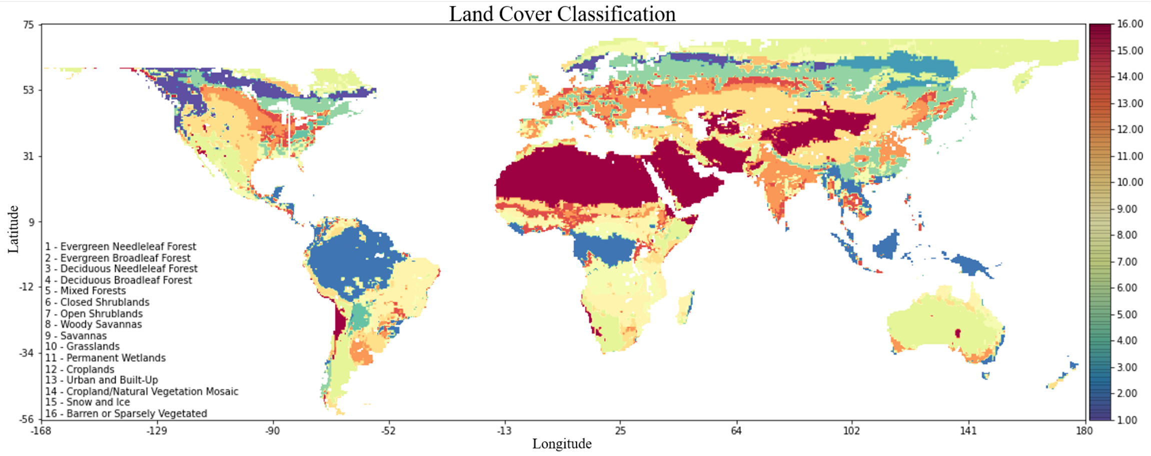
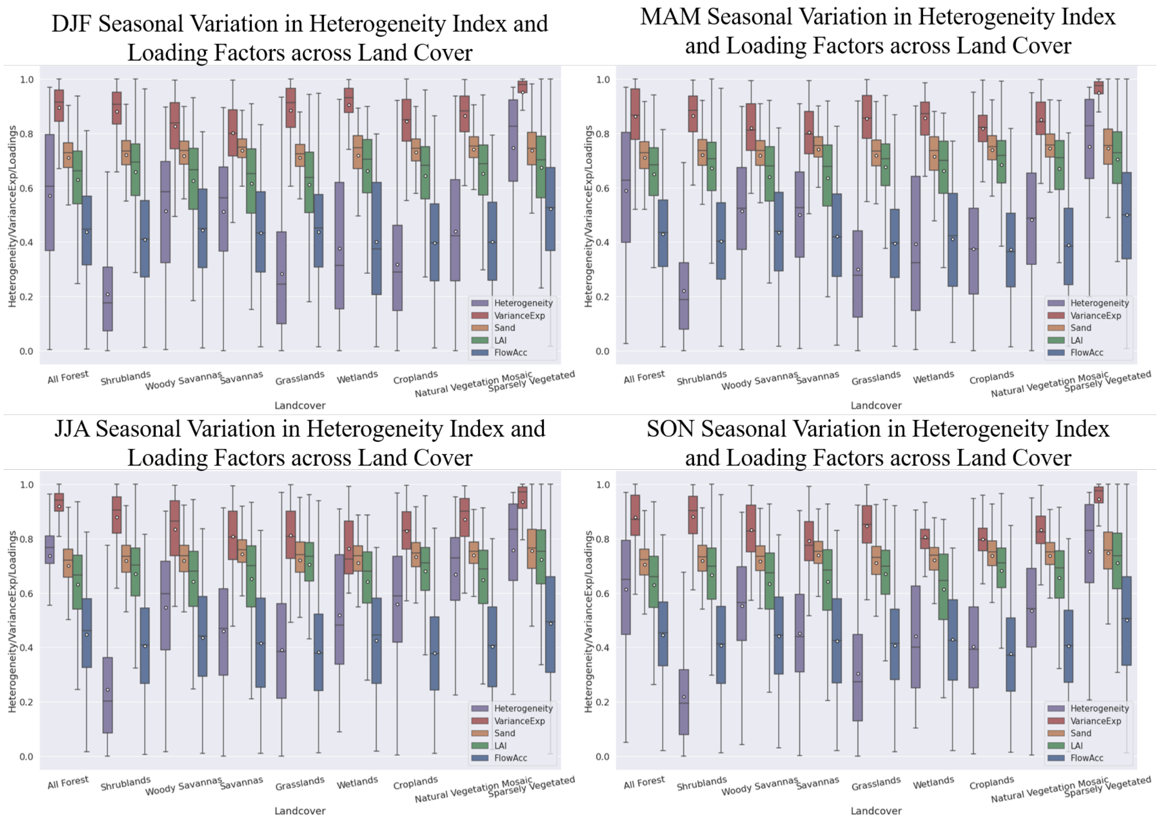


Figure 23: IGBP [34] Land cover Classification

Like our previous analysis, if we try to correlate just vegetation here (due to only temporal factor present), we won't able to specify heterogeneity index. Based on the Figure 24 it's clear that loadings for soil are high everywhere. This means that loading which is the highest on heterogeneity index is from soil properties not from vegetation (LAI).

Its a significant finding which reflects a lot on what heterogeneity index and soil would be as opposed to just land cover. The reason why vegetation is not dominant is because when the land cover is done, the vegetation is already minimized based on the heterogeneity due to all of those divisions. This kind of pattern follows for all land cover division where soil has high influence over heterogeneity compared to vegetation. But one thing to note is that the Shrubland division has the lowest heterogeneity out of all divisions. Though we do see the temporal changes over season, for example in summer crop land division have a high heterogeneity and that's what we actually expect, none the less this temporal changes caused by vegetation.

Thus to summarize the analysis, most of the heterogeneity is coming from the soil, but most temporal changes are coming from vegetation. This kind of characteristics acts in favor of the H-Index. It means that Heterogeneity Index which is developed for Global-scale is able to incorporate different soil properties as well as adapting the temporal changes happening over time.


DJF Seasonal Variation in Heterogeneity Index and Loading Factors across Land Cover

MAM Seasonal Variation in Heterogeneity Index and Loading Factors across Land Cover

JJA Seasonal Variation in Heterogeneity Index and Loading Factors across Land Cover

SON Seasonal Variation in Heterogeneity Index and Loading Factors across Land Cover

Figure 24: Comparison of Heterogeneity Index and Loading Factors with different Land cover. The comparison is done for all seasons top-left is winter (DJF), top-right is spring (MAM), bottom-left is summer (JJA) and bottom-right is Fall (SON). The Box plots are representing Heterogenity Index(Purple), Percentage Variance Explained (Red), Loading Factor from Sand (Yellow), Loading Factor from LAI (Green) and Loading Factor from Flow Accumulation (Blue).

4.5 COMPARISON WITH MAJOR LAND RESOURCE AREAS

MLRA's are geographically associated land resource units delineated by the Natural Resources Conservation Service and characterized by a particular pattern that combines soils, water, climate, vegetation, land use. MLRAs are great for getting qualitative information what they lack is quantified information for soil and vegetation changes through-out the year. In this study, heterogeneity index provides the quantified information. When comparing Figure 25 with Figure 26, heterogeneity index was clearly able to incorporate the Eco-regions presented in Major land resource ares. The advantage with the heterogeneity is that it is able to quantify the temporal changes occurring on those Eco-regions. For an instance, when comparing East and Central Farming and Forest Region (N) and South Atlantic Region(P) with the heterogeneity of the winter (DJF) season, heterogeneity index was able to incorporate these Eco-regions. Similar analysis can be formed for Heterogeneity Index at 36km, Figure 13 shows separation of different Eco-regions for 36km Heterogeneity Index. By taking a closer look at the Georgia State in Figure 25, there are two separation between North Georgia and South Georgia that can be seen in MLRAs and which are divided into (N) and (P), heterogeneity index (Figure 26 and 27) was even able to classify those boundaries. Additionally, heterogeneity introduces a dynamic factor to this Eco-regions by have those temporal changes.

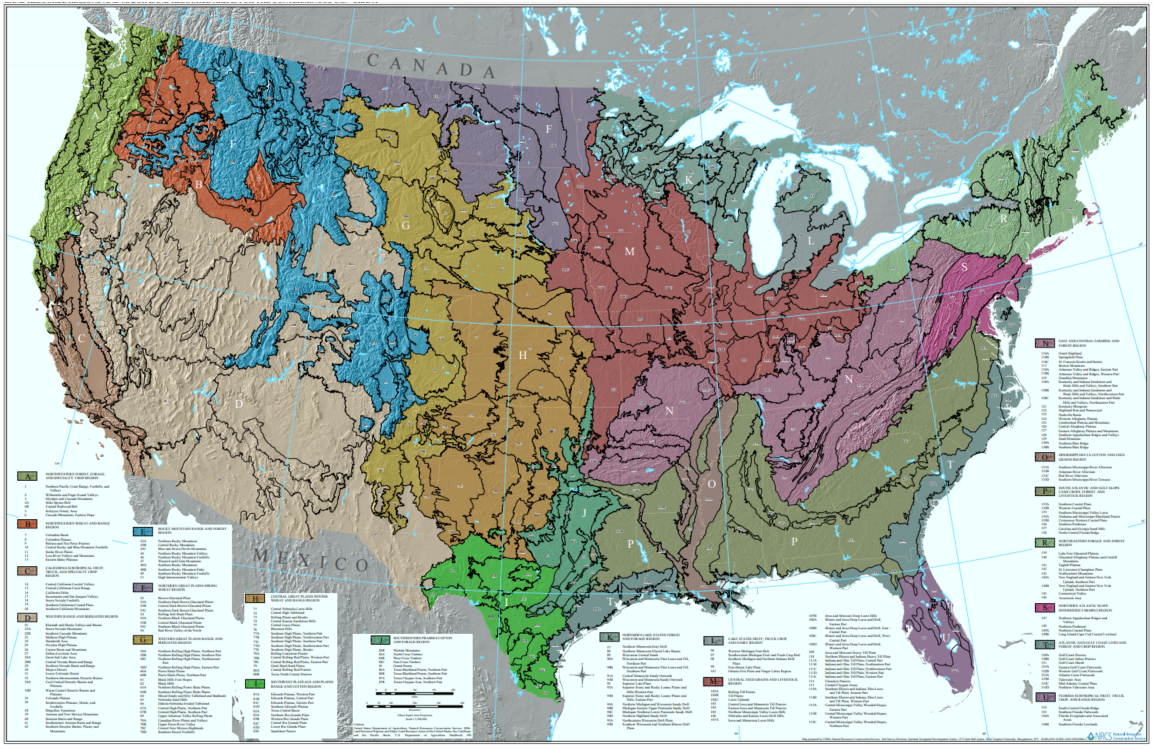
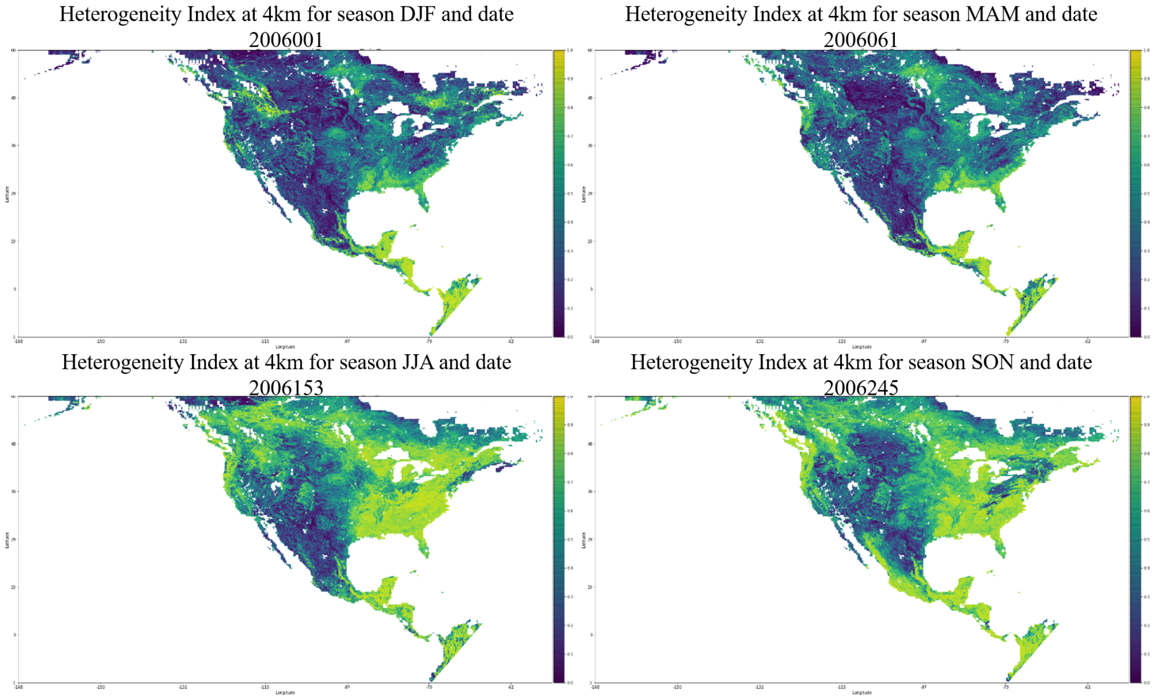


Figure 25: Major Land Resource Areas [34]



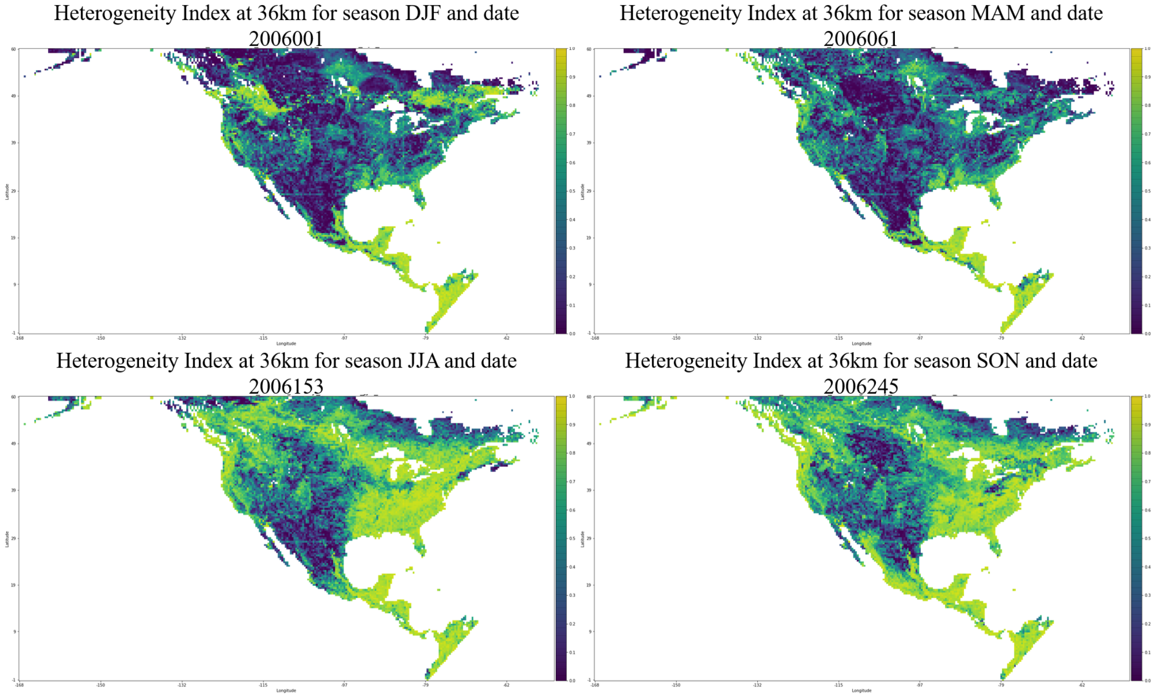
Heterogeneity Index at 4km for season DJF and date 2006001

Heterogeneity Index at 4km for season MAM and date 2006061

Heterogeneity Index at 4km for season JJA and date 2006153

Heterogeneity Index at 4km for season SON and date 2006245

Figure 26: North America (USConus): Heterogeneity Index at 4km (Year: 2006), top left is winter (DJF), top-right is spring(MAM), bottom-left is summer(JJA) and bottom-right is Fall (SON)



Heterogeneity Index at 36km for season DJF and date 2006001

Heterogeneity Index at 36km for season MAM and date 2006061

Heterogeneity Index at 36km for season JJA and date 2006153

Heterogeneity Index at 36km for season SON and date 2006245

Figure 27: North America (USConus): Heterogeneity Index at 36km (Year: 2006), top left is winter (DJF), top-right is spring(MAM), bottom-left is summer(JJA) and bottom-right is Fall (SON)

CHAPTER 5

RESULTS AND DISCUSSION FOR CLUSTER MODELS

5.1 OVERVIEW

There is a significant seasonal variation in Soil Water Retention Parameters. *Sehgal et al., 2020* [14] have found that the majority of the dominant hydrologic regimes of soil moisture drydown can be expressed with the transition point and three effective Soil Water Retention Parameters ($SWRP_{eff}$). Thus, it becomes evident to perform a separate cluster analysis on these selected parameters. The entire analysis is divided into two sets of parameters *(All Parameters and Selected Parameters)* and four seasonal SWRPs (Winter (DJF), Spring (MAM), Summer (JJA) and Fall (SON)).

To perform a comprehensive analysis on SWRP and $SWRP_{eff}$, four different clustering algorithms were applied. To get the optimal number of clusters, multiple internal cluster validity indices (CVI) were used. The results of these clusters are then evaluated using Tukey's HSD (Honestly Significant Difference) test. The formed clusters were mapped to the Heterogeneity Index and these heterogeneity index values were used to calculate significant difference between the clusters using Tukey's HSD test.

5.2 OPTIMAL NUMBER OF CLUSTERS

In this section, results for selecting an optimal number of clusters have been discussed based on the cluster validity indices. It can be noted that, in the current analysis, the number of clusters (N) on extremus has contradicting results based on Silhouette, Davies-Bouldin (DB), and Dunn. For instance, the lower number of clusters (N=2,3) are found to be optimal with a high CVI score from Silhouette and Dunn, Davies-Bouldin also has a high score for N=2,3 that contradicts the results Similar results are present for N=18,19,20 but with a lower Silhouette, Dunn and DB score.

Therefore, N=2,3,18,19 and 20 have been excluded for the optimal number of cluster consideration.

5.2.1 K-MEANS CLUSTERING

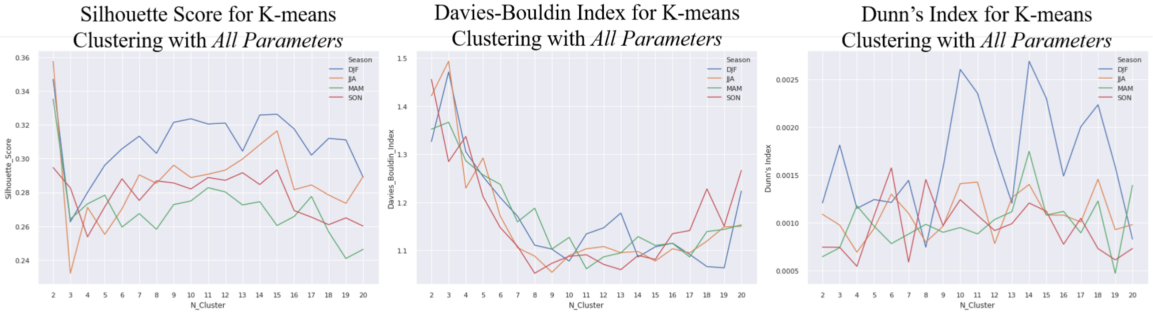


Figure 28: Cluster validity indices for K-means clustering on *All Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

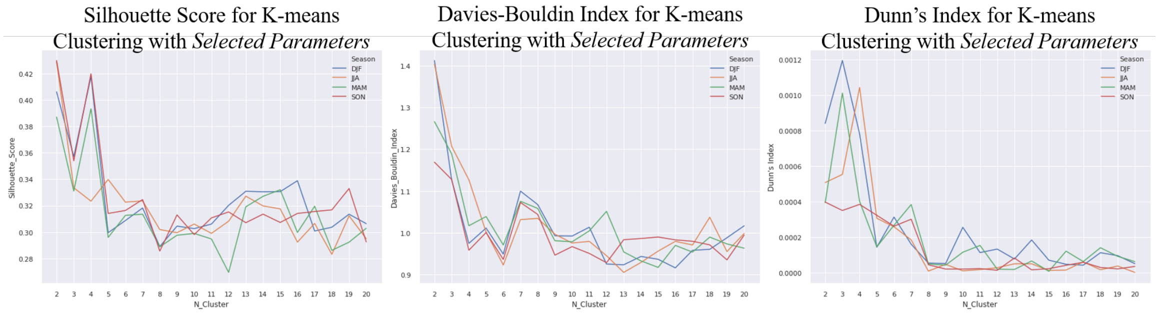
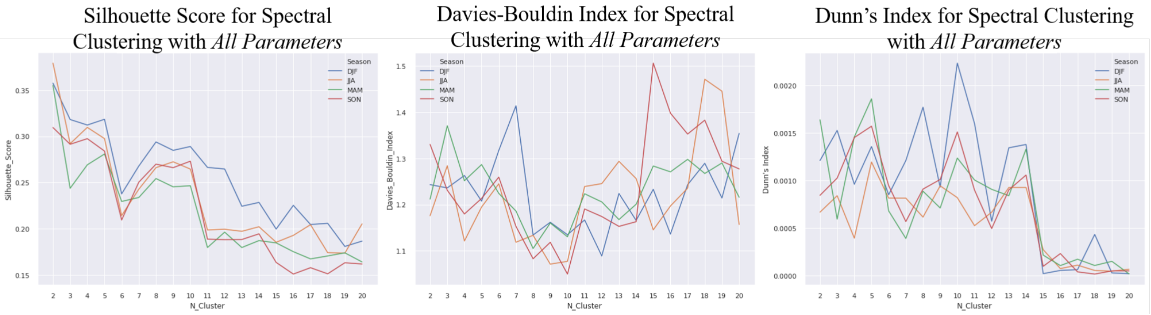


Figure 29: Cluster validity indices for K-means clustering on *Selected Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

For K-means *All Parameters* (Fig. 28), $N = 14$ is found to be optimal based to Silhouette score and Dunn's Index. Even though season JJA and SON have slightly higher score at $N = 15$, $N = 14$ is found to be more suitable for all the seasons. This is due to the fact that the season DJF and MAM at $N = 15$ have lower score compared to $N = 14$ and this indicates overall lower performance compared to $N = 14$. However, based on Davies-Bouldin, $N = 9$ is found to be optimal compared to $N = 14$. Though $N = 14$ is also in an acceptable margin. The same analysis cannot be drawn for *Selected Parameters*, where in *All Parameter* N tends to favour higher number of clusters, in *Selected Parameters* N tends to favor lower number of clusters. From Figure 29, $N = 4$ is found to be optimal based on Silhouette and Dunn's for all the seasons. For Davies-Bouldin, $N = 4$ and 6 can be consider as optimal. This experiment implies that $N = 14$ and 4 can be optimal for K-means when using *All Parameters* and *Selected Parameters* respectively.

5.2.2 SPECTRAL CLUSTERING

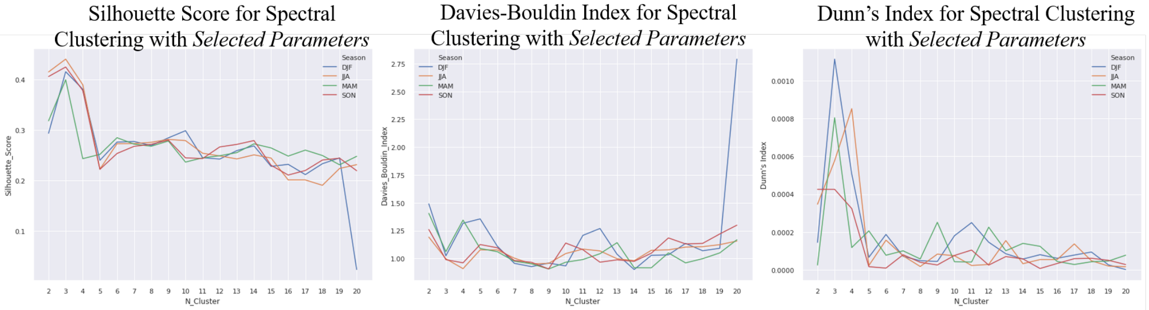

Silhouette Score for Spectral Clustering with *All Parameters*

Davies-Bouldin Index for Spectral Clustering with *All Parameters*

Dunn's Index for Spectral Clustering with *All Parameters*

Figure 30: Cluster validity indices for Spectral clustering on *All Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

For Spectral Clustering *All Parameters* (Fig. 30), $N = 4$ and 5 are found to be optimal based to Silhouette score, but for the Dunn's index $N = 4$ has much lower score compared to $N = 5$ especially in season DJF, MAM and JJA, comparatively $N = 5$ is found to be optimal for all the seasons except MAM. As for the Davies-Bouldin, $N = 8$ and 10 are found to be optimal with a higher indication of N = 10 to be more suitable for all the seasons. Due to variable optimal number of clusters from different CVI, it becomes challenging to determine optimal number of clusters for Spectral clustering using *All Parameters*. Further experiments between $N = 5$ and $N = 10$ are required to determine the optimal number of clusters.


Silhouette Score for Spectral Clustering with *Selected Parameters*

Davies-Bouldin Index for Spectral Clustering with *Selected Parameters*

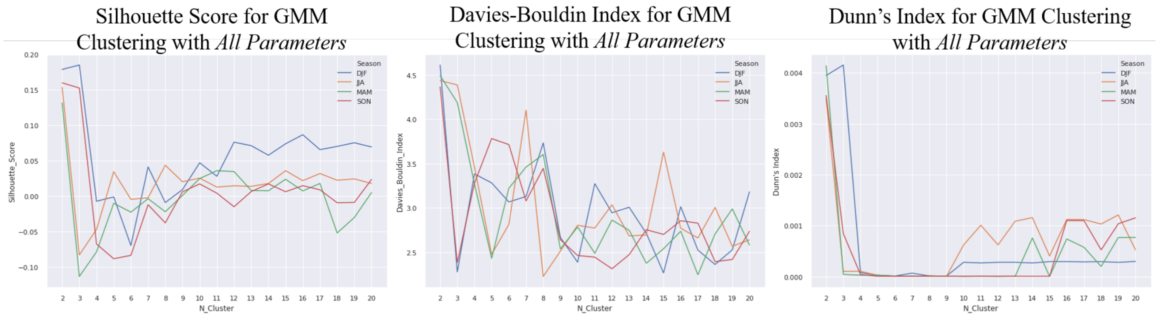
Dunn's Index for Spectral Clustering with *Selected Parameters*

Figure 31: Cluster validity indices for Spectral clustering on *Selected Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

However, this variability is not present in *Selected Parameters*. Figure 31 shows clear uniformity for all the CVI. For Silhouette and Dunn, $N = 4$ is found to be optimal for all the seasons, whereas for Davies-Bouldin $N = 4$, 8 and 9 are found to be optimal. Though at $N = 4$, DB-Index for season

DJF and MAM is higher then $N = 8$ and 9. It can be inferred that $N = 8$ or 9 are more suitable for optimal clusters, strictly based on DB-Index. But according to Silhouette and Dunn $N = 4$ is optimal, thus experiment implies that the $N = 4$ can be an optimal number of clusters for Spectral clustering when using *Selected Parameters*.

5.2.3 CLUSTERING BASED ON GAUSSIAN MIXTURE MODEL

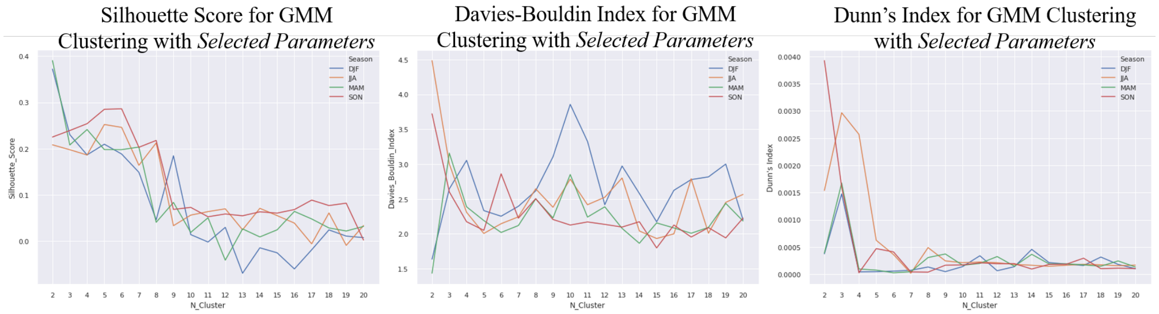


Silhouette Score for GMM Clustering with *All Parameters*

Davies-Bouldin Index for GMM Clustering with *All Parameters*

Dunn's Index for GMM Clustering with *All Parameters*

Figure 32: Cluster validity indices for GMM clustering on *All Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index



Silhouette Score for GMM Clustering with *Selected Parameters*

Davies-Bouldin Index for GMM Clustering with *Selected Parameters*

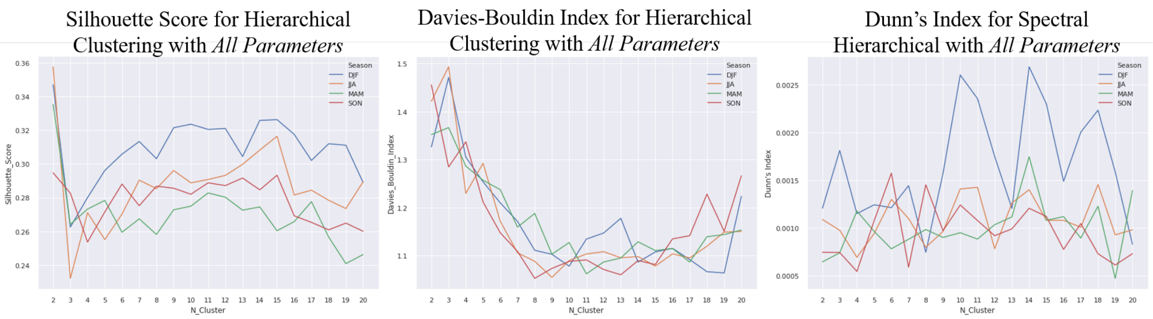
Dunn's Index for GMM Clustering with *Selected Parameters*

Figure 33: Cluster validity indices for GMM clustering on *Selected Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

Gaussian Mixture Model (GMM) has very high variability for *All Parameters* in DB-Index. This becomes challenging to determine the optimal number of clusters due to variability among the seasons (Fig. 32). For instance, based on DB-Index, optimal number of clusters for the seasons DJF, MAM, JJA and SON are $N = 15$, 8, 17 and 12 respectively. On the other hand, Silhouette and Dunn has optimal number of clusters at $N = 16$ for all the seasons. Though there is a range of low CVI (close to 0) for $N = 4$ to 10, it can be inferred that lower number of clusters are not suitable for GMM using *All Parameters*. However, it can be seen from the results for *Selected Parameters* (Fig. 33), that there is a less variability compared to *All Parameters*. For Silhouette and Dunn N

= 5 is found to be optimal for all the seasons, whereas for DB-Index $N = 5$ and 15 are found to be optimal. Although both N have comparative lower scores, $N = 15$ is slightly more suitable for all the seasons with respective to DB-Index. Thus the current experiment implies that $N = 5$ can be an optimal number of clusters for GMM clustering when using *Selected Parameters*.

5.2.4 HIERARCHICAL CLUSTERING

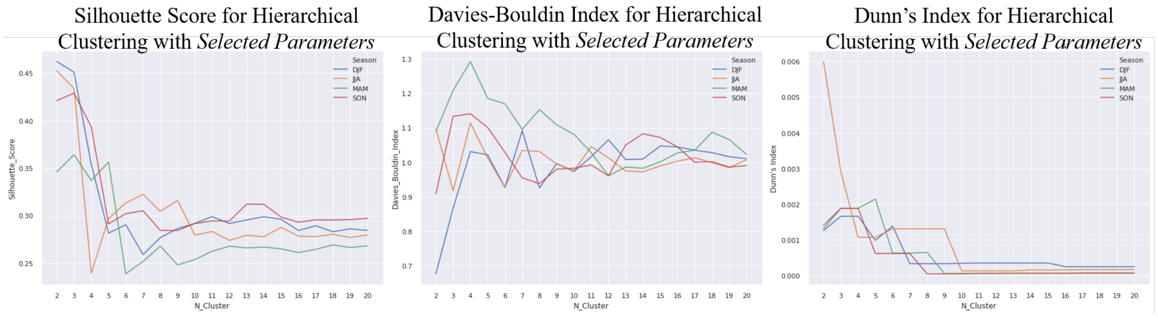

Silhouette Score for Hierarchical Clustering with *All Parameters*

Davies-Bouldin Index for Hierarchical Clustering with *All Parameters*

Dunn's Index for Spectral Hierarchical with *All Parameters*

Figure 34: Cluster validity indices for Hierarchical clustering on *All Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

For both *All Parameters and Selected Parameters* in Hierarchical clustering, Dunn's Index remains constant in a certain range, because of that it is showing as plateau on the Figure 34 and Figure 35. The difference between this range is too small, that they can be negligible. These effects are also showing up for Silhouette score in a certain cluster range. This becomes a challenge in finding out the optimal number of clusters. These plateaus might have caused due to imputation done in the data which can be a problem in a large data-set for Hierarchical clustering.


Silhouette Score for Hierarchical Clustering with *Selected Parameters*

Davies-Bouldin Index for Hierarchical Clustering with *Selected Parameters*

Dunn's Index for Hierarchical Clustering with *Selected Parameters*

Figure 35: Cluster validity indices for Hierarchical clustering on *Selected Parameters*. (from left to right) Silhouette Score, Davies-Bouldin Index and Dunn's Index

For *All Parameters*, $N = 9$ is found to be optimal based on Silhouette and Dunn for all the seasons. Whereas $N = 9$ and 13 are found to be optimal based on DB-Index for all the seasons.

Thus it implies that the $N = 9$ can be an optimal number of clusters for Hierarchical clustering using *All Parameters*. On the other hand, for *Selected Parameters* $N = 5$ and 13 are found to be optimal based on Silhouette. For Dunn $N = 4$ and 5 are found to be optimal, but for the DB-Index due to high variability in between season MAM and SON, there is no clear optimal number of clusters. The most suitable clusters that we can consider are N = 6,7,8 and 12.

5.2.5 PERFORMANCE COMPARISON

Table 2 and Table 3 summarize the cluster validity indices for *All Parameters* and *Selected Parameters*, respectively. Both the tables contains the optimal number of clusters based on CVI which were discussed in the above sections, as well as the scores for a constant number of cluster (N) for all the cluster algorithms. This is done in order to evaluate the performance between cluster algorithms and the performance for each algorithm across the seasons (*DJF, MAM, JJA* and *SON*). The number of clusters are selected after meticulously observing the CVI results.

Table 2: Summarizing the different Cluster Validity Indices for *All Parameters*: *Methodology S* - Silhouette Score; *Methodology DB* - Davies-Bouldin Index; *Methodology D* - Dunn's Index;; *Methodology N* - Optimal number of clusters; *Methodology N′* - Number of clusters to compare the performance of all clustering algorithm.

CVI	Seasons	K-means		Spectral		GMM		Hierarchical	
		$N = 14$	$N' = 4$	$N = 10$	$N' = 4$	$N = 12$	$N' = 4$	$N = 9$	$N' = 4$
S	DJF	0.3257	0.2800	0.2888	**0.3121**	0.0762	-0.0071	0.2655	0.2331
	MAM	0.2744	0.2732	0.2464	0.2690	0.0348	-0.0786	0.2046	**0.2744**
	JJA	0.3081	0.2710	0.2645	**0.3095**	0.0147	-0.0468	0.2385	0.2226
	SON	0.2846	0.2537	0.2729	**0.2973**	0.0043	-0.0675	0.2621	0.2464
DB	DJF	1.0863	1.3050	1.1345	**1.2628**	2.9443	3.3834	1.1512	1.3509
	MAM	1.1287	1.2863	1.1301	**1.2515**	2.8612	3.2676	1.1994	1.2676
	JJA	1.0980	1.2296	1.0769	**1.1210**	3.0343	3.4599	1.1541	1.3256
	SON	1.0895	1.3370	1.0490	**1.1795**	2.4423	3.2795	1.1492	1.2264
D	DJF	0.0026	0.0011	0.0022	0.0009	0.0002	7.74E-05	0.0027	**0.0019**
	MAM	0.0017	0.0011	0.0012	0.0014	1.77E-06	2.62E-05	0.0021	**0.0015**
	JJA	0.0013	0.0006	0.0008	0.0003	0.0006	0.0001	0.0025	**0.0018**
	SON	0.0012	0.0005	0.0015	0.0014	6.18E-06	3.97E-05	0.0021	**0.0015**

When comparing the $N = 4$ Spectral clustering and Hierarchical clustering are performing better than the other two algorithms for *All Parameter*(Table 2 (marked **bold**)). Spectral clustering is performing best based on Silhouette Score and Davies-Bouldin Index, whereas Hierarchical clustering is out performing other clustering algorithms based on Dunn's Index. Additionally, based on the

Dunn's Index Spectral clustering is performing as good as Hierarchical clustering for MAM and SON seasons. Thus it can be concluded that the Spectral cluster is performing *best* out of all the clustering algorithms for *All Parameters.*

Spectral clustering and Gaussian Mixture Models do not make any assumption about the structure of clusters, while clustering the data-points, due to this characteristics it is expected to have a higher performance from either one of the algorithms. But for the *Selected Parameters* it is observed that they are not the best. When comparing $N = 10$ for all the algorithms, K-means is found to be performing better for the majority of the seasons. Reasson might be due to the minimization of imputations. Although Gaussian Mixture Model(GMM) has a better performance based on Dunn's Index, K-means is performing better based on Silhouette and Davies-Bouldin Index (Table 3 (marked **bold**)). Based on the current result one may not conclude the best cluster algorithm, but it can is observed that the K-means is performing better out of all the clustering algorithms for *Selected Parameters.* The current results are then carried forward to find out the statistical significance for cluster separation.

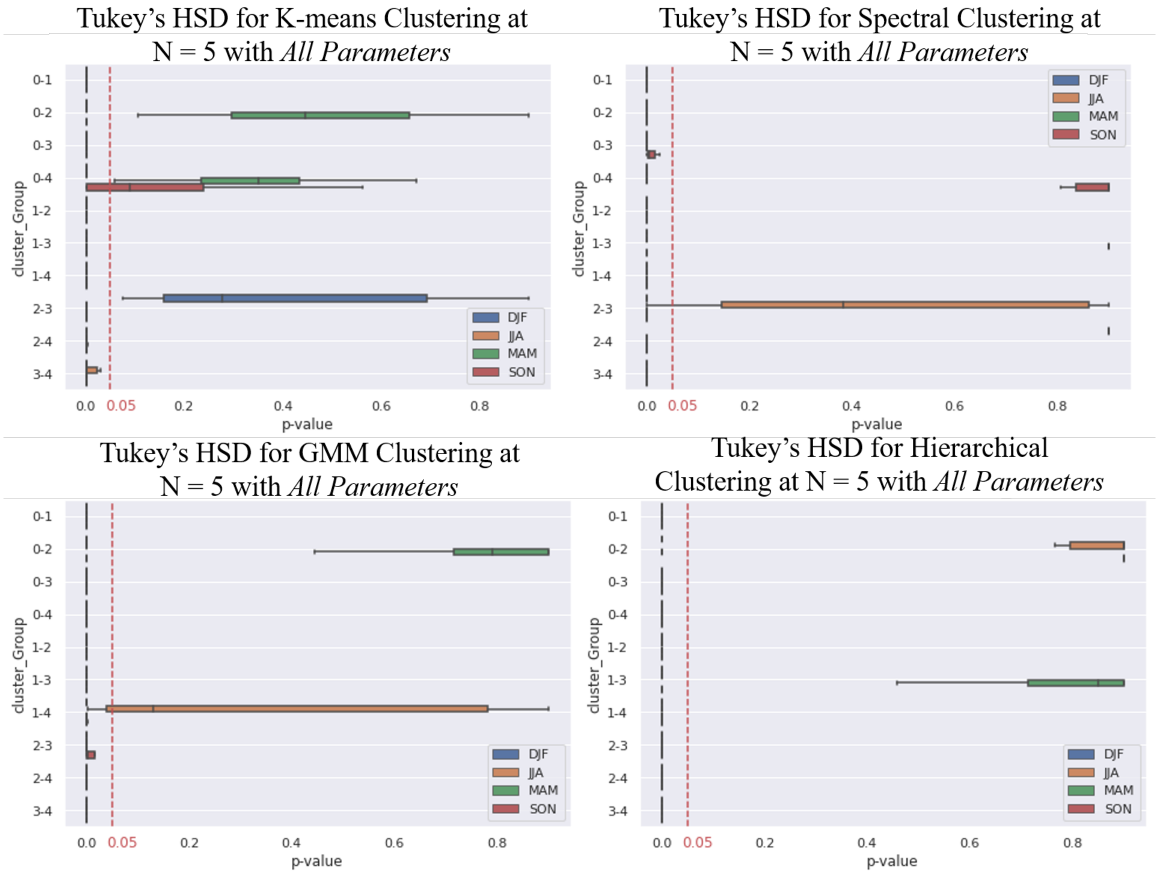
Table 3: Summarizing the different Cluster Validity Indices for *Selected Parameters*: *Methodology S* - Silhouette Score; *Methodology DB* - Davies-Bouldin Index; *Methodology D* - Dunn's Index; *Methodology N* - Optimal number of clusters; *Methodology N′* - Number of clusters to compare the performance of all clustering algorithm.

CVI	**Seasons**	**K-means**		**Spectral**		**GMM**		**Hierarchical**	
		$N = 4$	$N' = 10$	$N = 4$	$N' = 10$	$N = 5$	$N' = 10$	$N = 7$	$N' = 10$
S	DJF	0.4179	**0.3026**	0.3802	0.2983	0.2095	0.0140	0.2589	0.2917
	MAM	0.3934	**0.2989**	0.2430	0.2362	0.1979	0.0187	0.2517	0.2536
	JJA	0.3234	**0.3060**	0.3891	0.2787	0.2520	0.0561	0.3223	0.2793
	SON	0.4199	**0.2981**	0.3778	0.2446	0.2850	0.0728	0.3050	0.2915
DB	DJF	0.9744	0.9919	1.3160	**0.9333**	2.3326	3.8557	1.0916	0.9733
	MAM	1.0166	0.9782	1.3447	**0.9664**	2.1896	2.8498	1.0954	1.0811
	JJA	1.1268	**0.9752**	0.9089	1.0443	2.0053	2.7841	1.0344	**0.9752**
	SON	0.9580	**0.9666**	0.9610	1.1384	2.0495	2.1260	0.9547	0.9826
D	DJF	0.0008	**0.0003**	0.0005	0.0002	4.68E-05	0.0001	0.0003	**0.0003**
	MAM	0.0004	0.0001	0.0001	4.26E-05	7.49E-05	**0.0002**	0.0006	6.21E-05
	JJA	0.0010	1.08E-05	0.0009	7.45E-05	0.0006	**0.0002**	0.0013	0.0001
	SON	0.0004	2.10E-05	0.0003	7.56E-05	0.0005	**0.0002**	0.0006	5.02E-05

5.3 STATISTICAL SIGNIFICANCE FOR CLUSTER SEPARATION

In this series of experiments, each optimal N as well as the other number of clusters that were discussed in the previous section were used to perform Tukey's HSD test. The test is done in order to get the pair-wise significance and mean difference between the cluster groups. Heterogeneity Index used here was seasonal average of each year. For *All Parameters*, lower number of clusters were performing better than the higher number of clusters. Figure 36 represents the pair-wise box plot for all four cluster algorithms. We report the best performed algorithm with suitable N.

$N = 4$ and 5 were found to be performing better. Results for both N were similar in terms of cluster significance. For $N = 5$, K-means was having significant cluster separations for season JJA, whereas in season MAM tow pair of cluster groups were found to be insignificant, similarly in season DJF and SON a single pair of cluster group were found to be insignificant. Even though spectral clustering was able to reduce the insignificant number of pair of cluster group for season MAM, was not able to clearly separate season JJA.


Tukey's HSD for K-means Clustering at N = 5 with *All Parameters*

Tukey's HSD for Spectral Clustering at N = 5 with *All Parameters*

Tukey's HSD for GMM Clustering at N = 5 with *All Parameters*

Tukey's HSD for Hierarchical Clustering at N = 5 with *All Parameters*

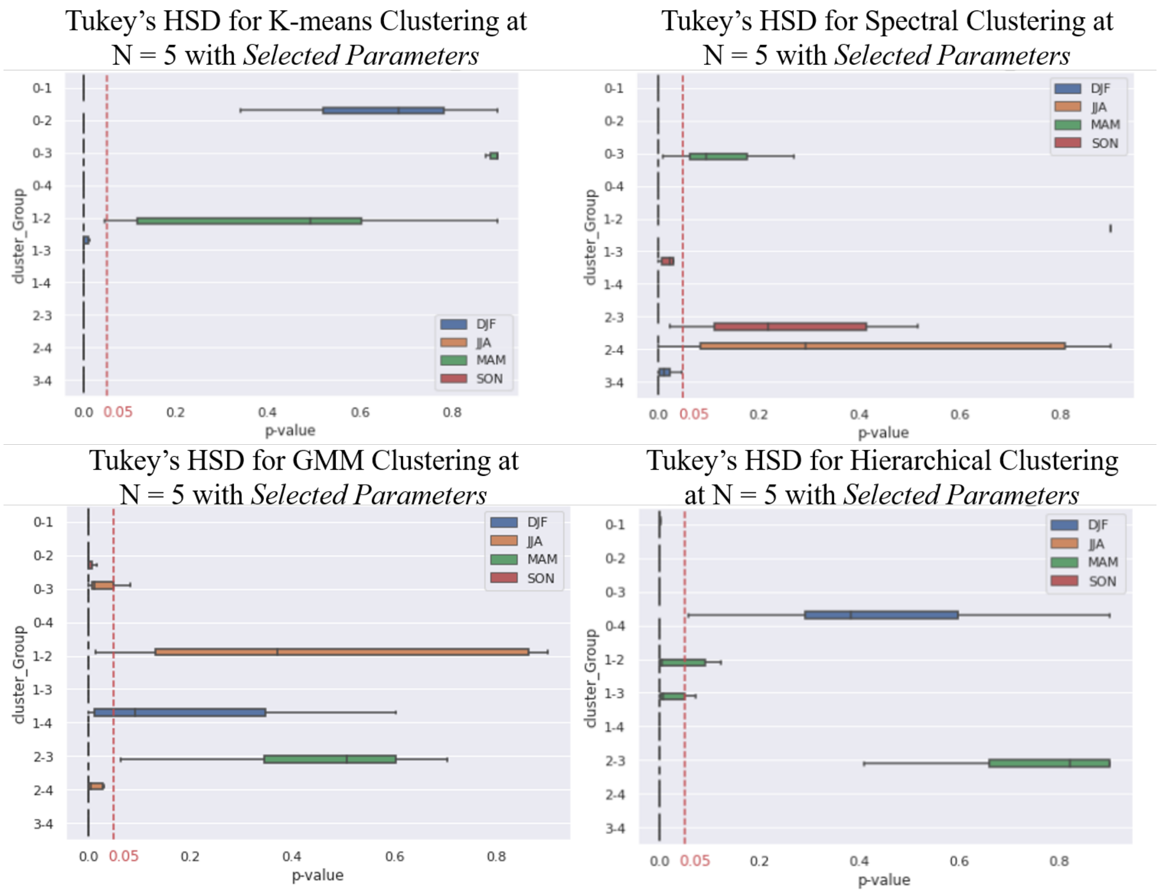
Figure 36: Tukey's HSD test for *All Parameters*. Pair-wise boxplots are for K-means (top left), Spectral Clustering (top right), Gaussian Mixture Model (bottom left) and Hierarchical clustering (bottom right) respectively.

However, Gaussian Mixture Model was found to be performing much better compare to other two algorithms. For the season DJF and SON, GMM was able to separate all the clusters, in season MAM and JJA only a single pair of cluster group was found to be insignificant for cluster separation. On the contrary, Hierarchical clustering was only able to have clear separation in season DJF. The other seasons were having insignificance for a single pair of cluster group. It is to be noted that all the algorithms were insignificant in season MAM for one or more pair of cluster groups. Season MAM is considered as leaf-on season, so we can infer that due to high temporal variability in leaf-on season clusters are not able to segregate properly.

Based of Figure 36, it can be concluded that Gaussian Mixture Model is a better suitable cluster algorithm for *All Parameters*. GMM was able to show adequate cluster separation for all the seasons. However based on the previous experiments done on CVI, it was found that GMM might not be a suitable algorithm with lower number of cluster for *All Parameters* based on Dunn's Index, but that's not the case here, even with $N = 5$ GMM is able to adequately separate the clusters. Figure 38 represents the Spatial Map of cluster Groups from Gaussian Mixture Model.

Clustering Algorithms for *Selected Parameters* were also found to be performing better with lower number of clusters. Figure 37 represents the pair-wise box plots of cluster significance for all four clusters using *Selected Parameters*. Although GMM was adequately separate clusters for *All Parameter*, when using *Selected Parameters* GMM have more insignificant clusters compare to other algorithms that are present in the season DJF, MAM and JJA for one pair of cluster groups. On the contrary, Hierarchical clustering was able to separate all the clusters for the season DJF and SON. In season MAM and JJA only a single pair of cluster group was found to be insignificant for cluster separation.

K-means was able to separate all the clusters, but for the season JJA and SON. Whereas season DJF and MAM were found to have insignificant cluster with one pair of cluster groups for DJF and two pair of cluster groups for season MAM. Moreover, in spectral clustering a single pair of cluster groups were found to have insignificant separation for all the seasons. The same pattern was repeated for spectral clustering in *All Parameters*, where all four season were found to have insignificant cluster separation. Based on the Figure 37, it can be concluded that Hierarchical clustering is a better suitable cluster algorithm for *Selected Parameters*. It was able to show adequate cluster separation for all the seasons. The high temporal variability in leaf-on season (MAM) is also being reflected here. Figure 39 shows the Spatial Map of cluster Groups from Hierarchical clustering.


Tukey’s HSD for K-means Clustering at N = 5 with *Selected Parameters*

Tukey’s HSD for Spectral Clustering at N = 5 with *Selected Parameters*

Tukey’s HSD for GMM Clustering at N = 5 with *Selected Parameters*

Tukey’s HSD for Hierarchical Clustering at N = 5 with *Selected Parameters*

Figure 37: Tukey’s HSD test for *Selected Parameters*. Pair-wise boxplots are for K-means (top left), Spectral Clustering (top right), Gaussian Mixture Model (bottom left) and Hierarchical clustering (bottom right) respectively.

Based on both sets of results we can conclude that, the Heterogeneity Index is capable of partition the soil hydraulic parameter with adequate separation. Where the difference found between clusters were highly significant for the majority part of the seasons with p-value being close to zero.

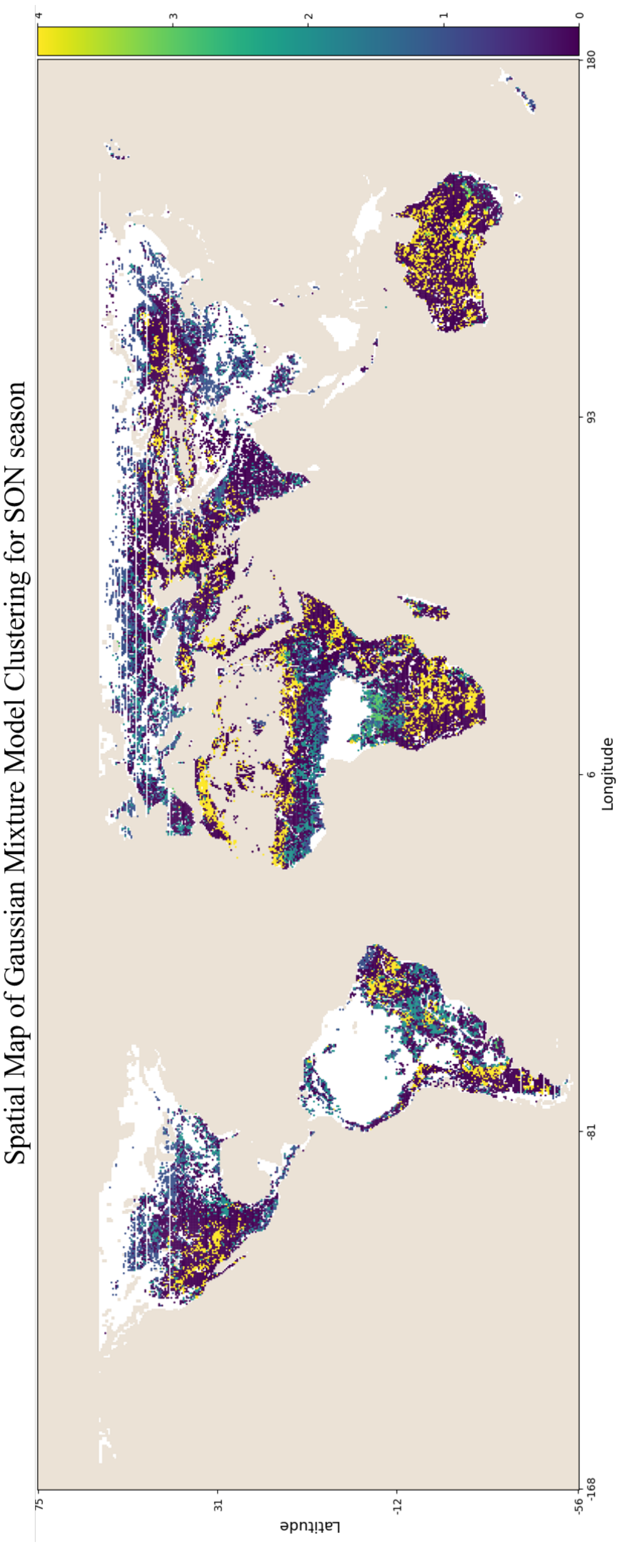

Spatial Map of Gaussian Mixture Model Clustering for SON season
Longitude
-168
-81
6
93
180
Latitude
75
31
-12
-56
0
1
2
3
4

Figure 38: Spatial Map of Gaussian Mixture Model (season SON).

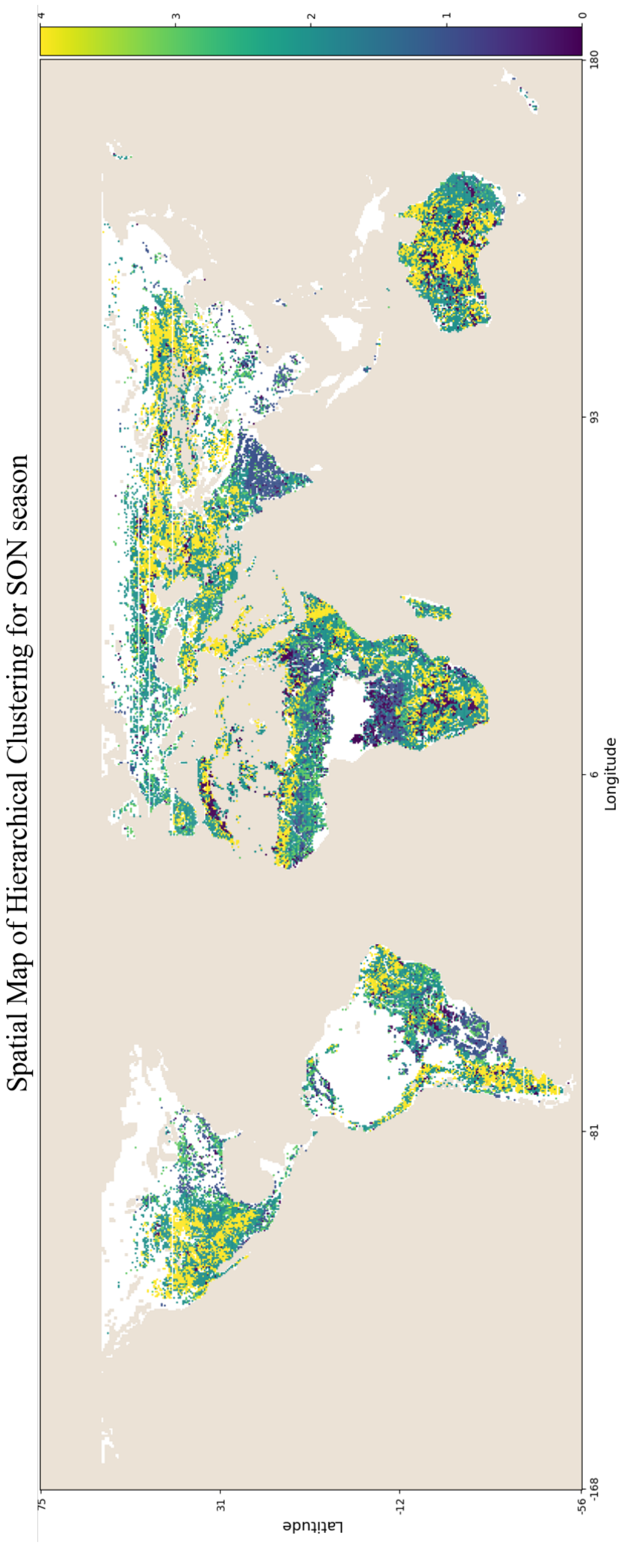


Figure 39: Spatial Map of Hierarchical Clustering. (season SON)

CHAPTER 6

CONCLUSION & FUTURE DIRECTIONS

The main purpose of this thesis was to develop a Heterogeneity Index (H-Index) that can quantify *Land-Surface Heterogeneity* at remote-sensing scale and to evaluate the utility of the developed Heterogeneity Index. Towards this end, a study was conducted where the work done by *Gaur and Mohanty, 2019* [12] on the Heterogeneity Index formulation was modified. An input-normalization schema was introduced to take into account of high variability presented in the data-sets. That helped in giving equal weight-age to %sand, Leaf Area Index and flow accumulation during the Heterogeneity Index formation which are the representative factors of soil, vegetation and topography respectively. Further more to constrain the H-Index such that the theoretical values of the final H-Index lie between 0 to 1, modifications to the previous formulation have been introduced. These modifications are introduced in a way that the new formulations are even able to incorporate the H-Index that could fall outside of observed values in the past 18 years of data-sets from 4^{th} July 2002 to 30^{th} April 2020.

The variance in land-surface heterogeneity explained by H-Index varies from 75% to as high as 99%. That means that the heterogeneity index is able to account a very high variance from the data-sets. The temporal changes in H-Index has been found to respond to changes in vegetation cover and can be used to assess the changes in land-surface heterogeneity over time. Based on the factor analysis, it was determined that spatial changes in H-Index are corresponding to the vegetation over time. These spatial heterogeneity plays a critical role in the terrestrial, water, energy, and bio-geochemical cycles from local to continental and global scales. Even though, we found that the heterogeneity responded to vegetation, our further analysis showed that there was not a high temporal correlation between heterogeneity and vegetation. They were moderately correlated to each other. The loading factor analysis showed that soils were most dominantly loaded on the H-index followed by vegetation and topography. This analysis further strengthen the validity as well as dynamicity present in the heterogeneity index. Moreover, comparison of H-Index with different hydro-climate, land cover and Major Land Resource Ares (MLRAs) showed that how well the developed H-Index is able to quantify different Eco-regions and the temporal changes happening in those regions. This work demonstrates that the combination of soil, vegetation and topography is a better explanation of effective soil hydraulic properties compare to soil beyond Representative Elementary Volume (REV) scale.

To further evaluate the utility of Heterogeneity Index as global classifier, results of unsupervised clustering were considered. It was observed that the soil hydraulic parameters can be clustered using H-Index with adequate separation whereas clustering based on Gaussian Mixture Model (GMM) and Hierarchical clustering are identified as suitable approach for spatial data-sets. This work also presented a scalabel and modular python framework called ***pyHetro*** for the Heterogeneity Index computation. The framework is scalabel enough to incorporate high computation resource requirement due to large raster image size.

Requirement of complete data matrix imposes obstacles for clustering when the missing values are the type of Missing Not At Random (MNAR). In this work, an iterative approach was used to handle the missing values, but those values still affects the results of cluster separation that can be investigated in the future. Though the cluster models were able to have adequate separation, the insignificance of inter-cluster separation still remains in some scenarios. This also leaves a scope to explore the superior techniques for clear cluster separation. It would be interesting to see if any other learning technique can be incorporated for utilizing H-Index and improve the model performance. In future, it is possible to integrate ***pyHetro*** in deep learning pipeline for predicting Heterogeneity Index. ***pyHetro*** can be used to generate ground truth for the label forecasting. The entire system can act as a look-ahead system in Land-Surface Heterogeneity.

REFERENCES

[1] Hylke E Beck et al. "Present and future Köppen-Geiger climate classification maps at 1-km resolution". In: *Scientific data* 5 (2018), p. 180214.

[2] Dino Bellugi et al. "Predicting shallow landslide size and location across a natural landscape: Application of a spectral clustering search algorithm". In: *Journal of Geophysical Research: Earth Surface* 120.12 (2015), pp. 2552–2585.

[3] *The researcher's complete guide to Leaf Area Index (LAI).* URL: `https://www.metergroup.com/environment/articles/lp80-pain-free-leaf-area-index-lai/`.

[4] Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. "k-pod: A method for k-means clustering of missing data". In: *The American Statistician* 70.1 (2016), pp. 91–99.

[5] Jon Chorover et al. "Soil biogeochemical processes within the critical zone". In: *Elements* 3.5 (2007), pp. 321–326.

[6] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[7] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57. DOI: `10.1080/01969727308546046`. eprint: `https://doi.org/10.1080/01969727308546046`. URL: `https://doi.org/10.1080/01969727308546046`.

[8] Tom G Farr et al. "The shuttle radar topography mission". In: *Reviews of geophysics* 45.2 (2007).

[9] Alberto Foni and David Seal. "Shuttle Radar Topography Mission: an innovative approach to shuttle orbital control". In: *Acta Astronautica* 54.8 (2004), pp. 565–570.

[10] Mark A Friedl, Carla E Brodley, and Alan H Strahler. "Maximizing land cover classification accuracies produced by decision trees at continental to global scales". In: *IEEE Transactions on Geoscience and Remote Sensing* 37.2 (1999), pp. 969–977.

[11] Nandita Gaur and Binayak P Mohanty. "Land-surface controls on near-surface soil moisture dynamics: Traversing remote sensing footprints". In: *Water Resources Research* 52.8 (2016), pp. 6365–6385.

[12] Nandita Gaur and Binayak P Mohanty. "A nomograph to incorporate geophysical heterogeneity in soil moisture downscaling". In: *Water Resources Research* 55.1 (2019), pp. 34–54.

[13] Tomislav Hengl et al. "SoilGrids250m: Global gridded soil information based on machine learning". In: *PLoS one* 12.2 (2017), e0169748.

[14] Vinit Sehgal, Nandita Gaur, and Binayak P. Mohanty. "Global Surface Soil Moisture Drydown Patterns". In: *Water Resources Research* (2020), e2020WR027588. DOI: https://doi.org/10.1029/2020WR027588. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR027588. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR027588.

[15] Dong Huang et al. "Ultra-scalable spectral clustering and ensemble clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 32.6 (2019), pp. 1212–1226.

[16] Yinghai Ke et al. "Development of high resolution land surface parameters for the Community Land Model". In: *Geoscientific Model Development* 5.6 (2012), p. 1341.

[17] Joseph Lee Rodgers and W Alan Nicewander. "Thirteen ways to look at the correlation coefficient". In: *The American Statistician* 42.1 (1988), pp. 59–66.

[18] Huapeng Li et al. "Performance evaluation of cluster validity indices (CVIs) on multi/hyperspectral remote sensing datasets". In: *Remote Sensing* 8.4 (2016), p. 295.

[19] Yanchi Liu et al. "Understanding of internal clustering validation measures". In: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, pp. 911–916.

[20] Javier Lozano-Parra et al. "How do soil moisture and vegetation covers influence soil temperature in drylands of Mediterranean Regions?" In: *Water* 10.12 (2018), p. 1747.

[21] M Susan Moran et al. "Estimating soil moisture at the watershed scale with satellite-based radar and land surface models". In: *Canadian journal of remote sensing* 30.5 (2004), pp. 805–826.

[22] Daniel Müllner. "Modern hierarchical, agglomerative clustering algorithms". In: *arXiv preprint arXiv:1109.2378* (2011).

[23] Virginia Murray and Kristie L Ebi. *IPCC special report on managing the risks of extreme events and disasters to advance climate change adaptation (SREX)*. 2012.

[24] *MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006*. NASA EOSDIS Land Processes DAAC, 2015. DOI: https://doi.org/10.5067/MODIS/MCD15A3H.006.

[25] Karl Pearson. "VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia". In: *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187 (1896), pp. 253–318.

[26] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *The Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[27] Carl Rasmussen. "The infinite Gaussian mixture model". In: *Advances in neural information processing systems* 12 (1999), pp. 554–560.

[28] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[29] Manish Sarkar and Tze-Yun Leong. "Fuzzy K-means clustering with missing values." In: *Proceedings of the AMIA Symposium.* American Medical Informatics Association. 2001, p. 588.

[30] Yu and Shi. "Multiclass spectral clustering". In: *Proceedings Ninth IEEE International Conference on Computer Vision.* 2003, 313–319 vol.1. DOI: 10.1109/ICCV.2003.1238361.

[31] David G Tarboton. "A new method for the determination of flow directions and upslope areas in grid digital elevation models". In: *Water resources research* 33.2 (1997), pp. 309–319.

[32] AppEEARS Team. *Application for Extracting and Exploring Analysis Ready Samples (AppEEARS). Ver. 2.46. NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, USA.* 2020.

[33] Antonio Trabucco and Robert Zomer. *Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2.* 2019. DOI: 10.6084/m9.figshare.7504448.v3. URL: https://figshare.com/articles/dataset/Global_Aridity_Index_and_Potential_Evapotranspiration_ET0_Climate_Database_v2/7504448/3.

[34] United States Department of Agriculture. *Major Land Resource Areas.* 2006. URL: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053624.

[35] Bei Zhao et al. "A spatial Gaussian mixture model for optical remote sensing image clustering". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.12 (2016), pp. 5748–5759.

[36] Robert Zomer et al. *Carbon, land and water: a global analysis of the hydrologic dimensions of climate change mitigation through afforestation / reforestation.* IWMI Research Reports H039281. International Water Management Institute, 2006. URL: `https://ideas.repec.org/p/iwt/rerpts/h039281.html`.

[37] Robert J Zomer et al. "Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation". In: *Agriculture, ecosystems & environment* 126.1-2 (2008), pp. 67–80.

APPENDIX

TERM DEFINITIONS

A.1 Landcover Classification (Definition of the IGBP classification scheme)

- Evergreen Needleleaf Forests
  - Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 meters. Almost all trees remain green all year. Canopy is never without green foliage.
- Evergreen Broadleaf Forests
  - Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 meters. Almost all trees and shrubs remain green year round. Canopy is never without green foliage.
- Deciduous Needleleaf Forests
  - Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 meters. Consists of seasonal needleleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
- Deciduous Broadleaf Forests
  - Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 meters. Consists of broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
- Mixed Forests
  - Lands dominated by trees with a percent cover >60% and height exceeding 2 meters. Consists of tree communities with interspersed mixtures or mosaics of the other four forest types. None of the forest types exceeds 60
- Closed Shrublands
  - Lands with woody vegetation less than 2 meters tall and with shrub canopy cover >60%. The shrub foliage can be either evergreen or deciduous.

- Open Shrublands
  - Lands with woody vegetation less than 2 meters tall and with shrub canopy cover between 10-60%. The shrub foliage can be either evergreen or deciduous.
- Woody Savannas
  - Lands with herbaceous and other understory systems, and with forest canopy cover between 30-60%. The forest cover height exceeds 2 meters.
- Savannas
  - Lands with herbaceous and other understory systems, and with forest canopy cover between 10-30%. The forest cover height exceeds 2 meters.
- Grasslands
  - Lands with herbaceous types of cover. Tree and shrub cover is less than 10%.
- Permanent Wetlands
  - Lands with a permanent mixture of water and herbaceous or woody vegetation. The vegetation can be present in either salt, brackish, or fresh water.
- Croplands
  - Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.
- Urban and Built-Up Lands
  - Land covered by buildings and other man-made structures.
- Cropland/Natural Vegetation Mosaics
  - Lands with a mosaic of croplands, forests, shrubland, and grasslands in which no one component comprises more than 60% of the landscape.
- Snow and Ice
  - Lands under snow/ice cover throughout the year.
- Barren

  - Lands with exposed soil, sand, rocks, or snow and never has more than 10% vegetated cover during any time of the year.

- Water Bodies
  - Oceans, seas, lakes, reservoirs, and rivers. Can be either fresh or salt-water bodies.