ANALYSIS OF RELATIONS BETWEEN SOLAR ACTIVITY, COSMIC RAYS AND THE EARTH CLIMATE USING MACHINE LEARNING TECHNIQUES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BÜKEM BELEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
EARTH SYSTEM SCIENCE

SEPTEMBER 2021

ABSTRACT

ANALYSIS OF RELATIONS BETWEEN SOLAR ACTIVITY, COSMIC RAYS AND THE EARTH CLIMATE USING MACHINE LEARNING TECHNIQUES

Belen, Bükem
Master of Science, Earth System Science
Supervisor: Assoc. Prof. Dr. Uğur Murat Leloğlu
Co-Supervisor: Prof. Dr. M. Bilge Demirköz

September 2021, 112 pages

The Earth's climate is part of a complicated system that can be affected by many different parameters, both internal and external. Important external forces on the climate are galactic cosmic rays (GCR) and the Sun. Some research has already been conducted to investigate the relationship between the climate and external forcers such as the GCR and solar activity. However, the relations are quite complicated and buried into almost chaotic meteorological measurements. This thesis looks deeper into the interactions between them. The parameters used in the correlation analysis are GCR flux, Sunspot number (SSN), total solar irradiance (TSI), UV irradiance (UVI), and the Oceanic Niño Index (ONI) as the predictor variables; total cloud amount (TCA), low cloud amount (LCA), global mean temperature anomaly (GMTA), aerosol optical depth (AOD) and precipitation as the response variables. The analysis begins with standard statistical techniques and continues with multiple regression and machine learning methods for non-linear regression, such as random forests. Both geographical and temporal patterns have been investigated. This study shows that some parameters have a weak linear correlation, while a statistically significant non-linear relationship occurs between them. It can be concluded that the

GCR-climate connection does exist, and these non-linear relations should be investigated further, specifically in certain regions of the World.

Keywords: Cosmic Rays, Solar Activity, Climate, Multiple Regression, Random Forests

ÖZ

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANARAK GÜNEŞ AKTİVİTESİ, KOZMİK IŞINLAR VE DÜNYA İKLİMİ ARASINDAKİ İLİŞKİLERİN ANALİZİ

Belen, Bükem
Yüksek Lisans, Yer Sistem Bilimleri
Tez Yöneticisi: Doç. Dr. Uğur Murat Leloğlu
Ortak Tez Yöneticisi: Prof. Dr. M. Bilge Demirköz

Eylül 2021, 112 sayfa

Dünya iklimi hem iç hem de dış birçok farklı parametreden etkilenebilen karmaşık bir sistemin parçasıdır. İklim üzerindeki önemli dış kuvvetler galaktik kozmik ışınlar (GCR) ve Güneş'tir. Şimdiye kadar iklim ile GCR ve güneş aktivitesi gibi dış etkenler arasındaki ilişkiyi araştıran bazı araştırmalar yapılmıştır. Ancak ilişkiler oldukça karmaşıktır ve neredeyse kaotik meteorolojik ölçümlerin içine gömülüdür. Bu tez, bunların arasındaki ilişkilere daha derinden bakacaktır. Korelasyon analizinde kullanılan parametreler tahmin değişkenleri olarak kozmik ışın akısı (GCR), güneş lekesi sayısı (SSN), toplam güneş ışınımı (TSI), UV ışınımı (UVI) ve okyanus Ninyo indeksi (ONI); yanıt değişkenleri olarak da toplam bulut miktarı (TCA), düşük bulut miktarı (LCA), küresel ortalama sıcaklık anomalisi (GMTA), aerosol optik derinliği (AOD) ve yağış miktarıdır. Standart istatistiksel tekniklerle başlayan analiz, rastgele ormanlar gibi doğrusal olmayan regresyon için çoklu regresyon ve makine öğrenmesi yöntemleriyle devam etmiştir. Hem coğrafi hem de zamansal örüntüler araştırılmıştır. Bu çalışmanın sonuçları, bazı parametrelerin zayıf bir doğrusal korelasyona sahip olduğunu göstermiştir ve aralarında istatistiksel olarak anlamlı doğrusal olmayan bir ilişki olduğunu göstermektedir. GCR-iklim bağlantısının var olduğu ve bu doğrusal olmayan ilişkilerin özellikle Dünya'nın belirli bölgelerinde daha fazla araştırılması gerektiği sonucuna varılabilir.

Anahtar Kelimeler: Kozmik Işınlar, Güneş Aktivitesi, İklim, Çoklu Regresyon, Rastgele Ormanlar

To my dearest family and beloved cats, Susam and Safran

ACKNOWLEDGMENTS

I, first of all, would like to thank my thesis advisor Assoc. Prof. Uğur Murat Leloğlu and express how grateful I am to have worked with him, thank him for introducing me to machine learning and mentoring me throughout my thesis. This accomplishment would not have been possible without his constant encouragement and respected recommendations.

I would like to acknowledge Prof. Dr. Bilge Demirköz, my co-advisor in this thesis, and thank her for her precious advice on research and ideas for this research.

Finally, I am very thankful for my parents, who have always supported my academic journey, especially throughout these challenging times in the pandemic. I want to thank my fiancé Kaan Dökmeci for always believing in me and thank my grandmother and grandfather for supporting me in writing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

AOD	Aerosol Optical Depth
CLOUD	Cosmics Leaving Outdoor Droplets
CN	Condensation Nuclei
CCN	Cloud Condensation Nuclei
ENSO	El Niño-Southern Oscillation
ESA	European Space Agency
GCR	Galactic Cosmic Ray
GLM	Generalized Linear Model
GMTA	Global Mean Temperature Anomaly
FD	Forbush Decrease
ISCCP	International Satellite Cloud Climatology Project
LLC	Low Level Cloud
LM	Linear Model
MERRA	Modern-Era Retrospective analysis for Research Applications
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	Mean Square Error
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
NES	Near Earth Space
ONI	Oceanic Niño Index
ONM	Oulu Neutron Monitor
OOB	Out-of-Bag
PRECIP	Precipitation
RF	Random Forest

RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
SSE	Sum of Squared Errors
SSN	Sunspot Number
SST	Sum of Squares Total
TC	Total Cloud
TSI	Total Solar Irradiance
UCN	Ultrafine Condensation Nuclei
UVI	Ultraviolet Irradiance

CHAPTER 1

INTRODUCTION

The Earth is part of a system that can be separated into several sub-systems, which are in continuous interaction with each other by means of chemical, physical, and biological activities (Kump, Kasting, & Crane, 2004). These sub-systems are the Hydrosphere, the Cryosphere, the Lithosphere, the Biosphere, the Atmosphere, and the Near-Earth Space (NES), as shown in Figure 1.1.

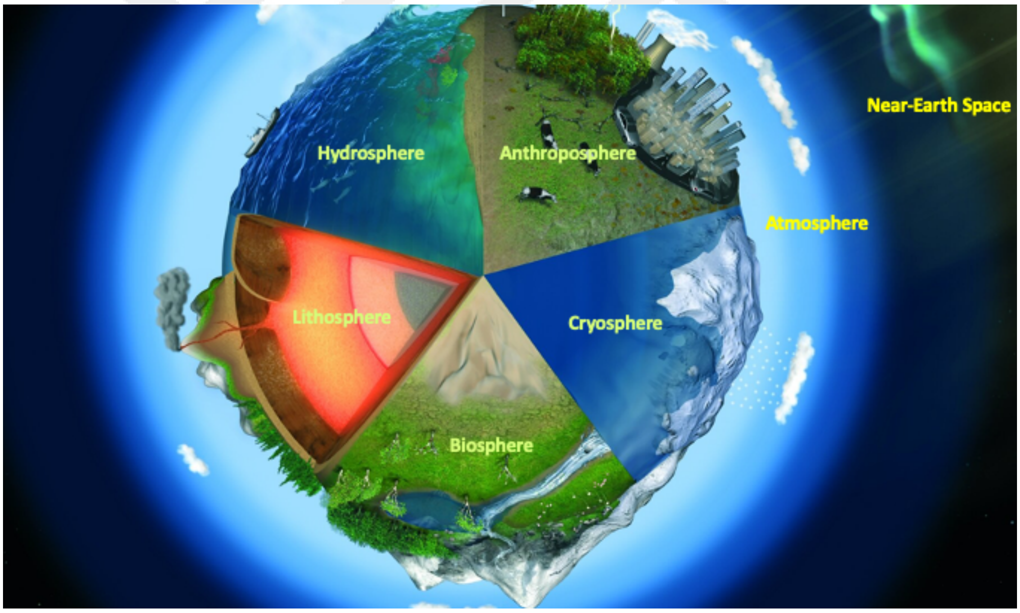


Figure 1.1. The Earth System (Image: ESA)

The hydrosphere contains water in liquid and gaseous forms, while the cryosphere contains water in the solid state, mainly ice. Oceans make a major part of the water content, while polar ice caps and glaciers are also important. Additionally, water also exists in its gas, liquid, and solid (ice) forms in the atmosphere (Stott & Huddart, 2010).

The lithosphere constitutes the solid part of the Earth. Effects on the lithosphere and its erosion shapes the land, such as mountains, plains, deserts, and valleys that provide diverse habitats for the biosphere. The biosphere includes plants, animals, humans, and other organisms that all demand nutrients from the lithosphere, water from the hydrosphere, and gas from the atmosphere for their existence and survival (Stott & Huddart, 2010).

The anthroposphere is the sub-system that humans and human activity make up. Since our economic life, industrial and technological developments have become major drivers of the environment, the anthroposphere is an essential part of the Earth System. The change in global climate thus is not only a result of the natural processes acting on the planet for millions of years but also a result of human activity, especially in the most recent years (Stott & Huddart, 2010).

The Earth is surrounded by the atmosphere, a gaseous layer receiving the solar energy coming from the Sun. The atmosphere contains essential elements to sustain life on Earth and redistributes heat and moisture from the Earth's surface (Stott & Huddart, 2010).

The NES is the region of space surrounding the atmosphere. It consists of the upper layers of the atmosphere and the Earth's magnetosphere. It is an important part of the Earth System because high energy particles constantly bombard it, and it shields the Earth from their harmful effects (Eddy, 2009).

The stability of the Earth system is maintained by its components, which are linked by feedback loops that can be either be positive or negative. Positive feedback loops amplify perturbations or forcings, while negative feedback loops reduce their effects (Stott & Huddart, 2010).

1.1 Description of the problem

The climate of the Earth can be an example of a feedback loop. The global mean temperature is maintained by various mechanisms of energy flow, including important negative feedbacks in the global climate system. The earth-atmosphere system is kept in balance by regulating the incoming and outgoing energy flows. If this stability is vaguely disturbed, the global climate will be subject to many complicated alterations (Stott & Huddart, 2010).

Earth's climate can be affected by many different parameters, both internal and external. Some of the internal forces are ocean currents, albedo, volcanic explosions, and anthropogenic activity. Humans have realized how crucial it is to keep the climate system in balance for a sustainable future in recent decades. To achieve this, external parameters and their interactions with the Earth also must be well understood, apart from the anthropogenic causes. In relatively shorter timescales, important external forces affecting the climate are galactic cosmic rays (GCR) and the Sun. It is crucial to study the external forces to understand the stages that the Earth's climate has been through since its formation and to model and predict changes in the climate.

1.2 The Approach of this thesis

It is aimed to study the interaction of galactic cosmic rays with solar activity and its effect directly or indirectly on the Earth's climate and to define the relationship pattern between them as a function of location by using machine learning. Accordingly, the goals are as follows:

1. To define the effect the galactic cosmic rays and the solar activity have on Earth's climate.

2. To reveal the temporal and geographic relationship patterns between the external forcers and climate parameters by examining the existing complex data using machine learning methods.

1.3 Contributions of this thesis

This thesis uses machine learning for the first time to look at the solar influence on climate. Many of the previous studies have used standard statistical methods such as linear correlation analysis on similar datasets with each other. In this thesis, different data sources of climate parameters such as the total cloud amount (TCA), low cloud amount (LCA), global mean temperature anomaly (GMTA), aerosol optical depth (AOD), and precipitation are used. The cloud data is obtained from two different sources, namely the ISCCP and MERRA2 projects. The period of the datasets is extended up to 2017, and multiple regression models are used as the primary investigation methods.

1.4 Organization of this thesis

This thesis is organized as follows:

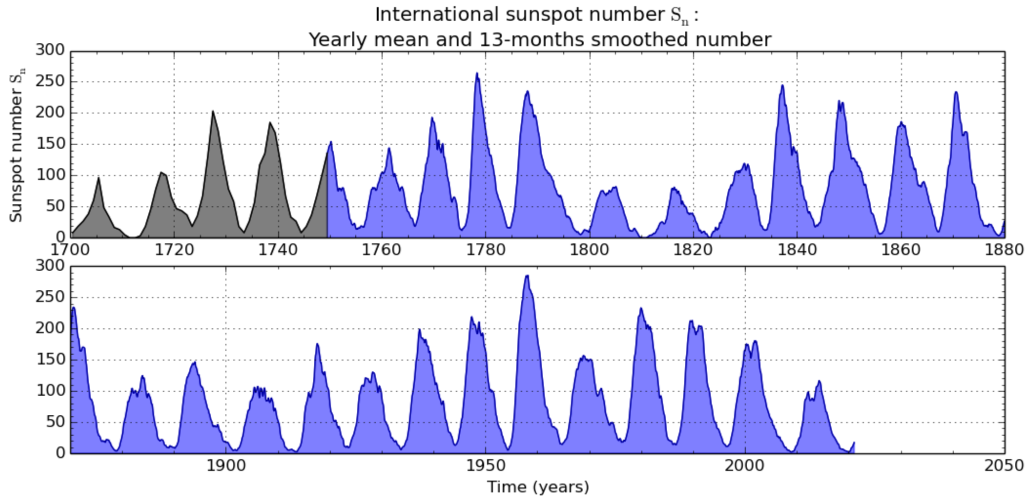
Chapter 2 provides the theoretical background that is necessary to understand the physical processes in this study. Solar activity, cosmic rays, and the properties of clouds are introduced, and an introduction to machine learning methods is given. Chapter 3 is a review of the previous research about the solar-climate connection, summarizing the papers that are for and against the hypothesis. Chapter 4 explains how the data used in the analysis were chosen and the sources of each dataset. Also, the methodology of the study is given and is detailed. The results are shown in Chapter 5 and are discussed thoroughly. Finally, the thesis is summarized and concluded in Chapter 6.

CHAPTER 2

BACKGROUND INFORMATION

2.1 Solar Activity

To explain the nature and extent of solar magnetic fields, a common term called solar activity is used. The most generally known index by which it has been specified is the number of sunspots visible in the Sun's disk at any given time. Sunspots are observed as the dark spots seen on the surface of the Sun, and they can last between days to weeks (Gray, et al., 2010). Astronomers across the globe have used a collectively described index, named the Wolf sunspot number, or more generally the sunspot number (SSN) since 1848. This relies on the effort to correct foreseeable changes in solar telescopes, observation conditions, and human observers. Solar cycles have a normal period of around eleven years. The duration of the minimum and maximum sunspot counts are referred to as solar minimum and solar maximum, respectively (Eddy, 2009).



SILSO graphics (http://sidc.be/silso) Royal Observatory of Belgium 2021 August 1

Figure 2.1. The solar activity cycles from 1700 to the present (source: SILSO data/image, Royal Observatory of Belgium, Brussels).

Figure 2.1 shows the yearly mean SSN values on a temporal scale starting from the year 1700. Up to 1749, the SSN values are annual mean values and from 1749 on (the blue part), the SSN values are 13-month smoothed values. From this figure, the periodicity of the full solar cycle of ~22 years, which can be measured from one peak to the next peak, is seen.

There are also other parameters to measure solar activity, and the two that will be used as solar proxies in this thesis are the total solar irradiance (TSI) and the ultraviolet irradiance (UVI). The Sun brings in all wavelengths to Earth's upper atmosphere. W/m^2 is the unit of total solar irradiance. At a standard range of an Astronomical Unit from the Sun, the wavelength-dependent energy response to the top of the Earth's atmosphere is called solar spectral irradiance. W/m^2-nm is the unit of solar spectral irradiance (Eddy, 2009).

The dynamics of the atmosphere and the movement in the troposphere can similarly be affected by stratospheric ozone. To explain relations between climate and solar activity, a possible mechanism could be the process that connects the stratospheric ozone to UV radiation. Another possible mechanism could be through the variations in total solar radiation leading to direct solar heating. With feedback mechanisms in the Earth system, each can be strengthened or reduced via the other (Benestad, 2006).

2.1.1 Global Energy Balance

Earth's temperature is increased by absorbing the radiation from the Sun and is decreased by emitting infrared radiation back to space. Most of the infrared radiation emitted by the Earth's surface is absorbed and reemitted by atmospheric gases, resulting in a natural greenhouse effect that warms the surface. The greenhouse effect is significantly impacted by atmospheric gases, particularly H_20 and CO_2, which absorb infrared radiation. Clouds likewise contribute to the greenhouse effect as low, thick clouds reflect sunlight and tend to cool the surface of Earth, while high, thin clouds tend to warm it (Kump, Kasting, & Crane, 2004). Earth's radiation balance is shown in the diagram in Figure 2.2.

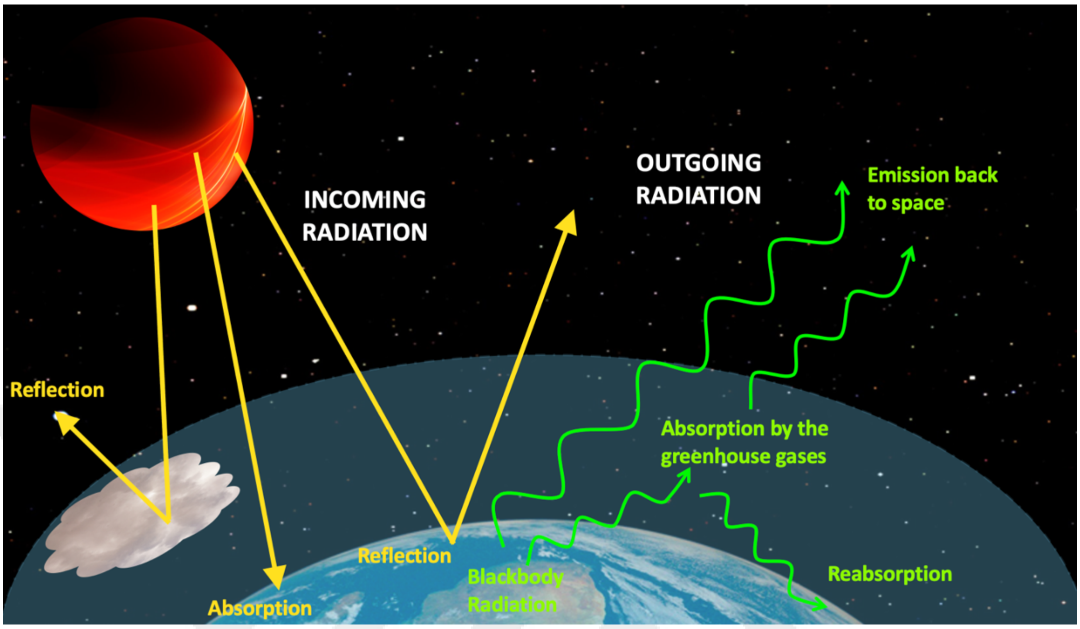


Figure 2.2. Earth's Radiation Balance

The temperature of the Earth's surface is built upon three aspects. The first one is Earth's reflectivity, while the second one is related to the amount of warming produced by the atmosphere, such as in the example of the greenhouse effect. The third and most important aspect is the solar flux available at the Earth's orbit. Solar flux is described as the amount of solar energy reaching the top of the Earth's atmosphere. Around a third of the incident energy is mainly reflected by clouds back to space; hence the whole energy is not absorbed (Kump, Kasting, & Crane, 2004).

Reflected incident sunlight by clouds causes the Earth to cool throughout the daytime. There is a direct relation between clouds and albedo, which is defined as the reflection of solar radiation, and it takes a value from zero to unity. A large fraction of current planetary albedo, which is around 0.3, is led by clouds. (Kump, Kasting, & Crane, 2004).

2.2 Cosmic Rays

Cosmic rays are very high-energy charged particles, which mainly originate from outside the Solar System; that is why they are mostly referred to as galactic cosmic rays (GCR). About 90% of all cosmic rays are protons, and the rest are electrons and the charged nuclei of heavier elements. They enter the Earth's atmosphere from every direction and collide with atoms and molecules in the air. The Earth's atmosphere shields the surface from the most energetic cosmic rays. The cosmic rays that make it to the surface are limited to the weaker, secondary particles (Eddy, 2009).

The GCR collided atoms generate other subatomic particles that are less energetic and considered the second generation. Neutrons and protons are highly energetic and heaviest among the debris. The produced electrons are called muons and pions. However, these electrons are considered short-lived particles since they cannot maintain their existence more than some microseconds (Eddy, 2009).

The so-called secondary cosmic ray particles lengthen the chain of collisions to the middle atmosphere; eventually, a few make it to the ground. In the continuation of this journey down, a third generation of subatomic particles is produced by the additional collisions of neutrons and protons with the other atoms of air. (Eddy, 2009).

A cascade of cosmic ray shower, shown in Figure 2.3, occurs with this activity of energy reduction by recurrent divisions and collisions. It continues until the number of subatomic particles which are generated reaches a maximum. With repeated cascades going further down into the denser and deeper layers of the atmosphere, it begins to decrease (Eddy, 2009).

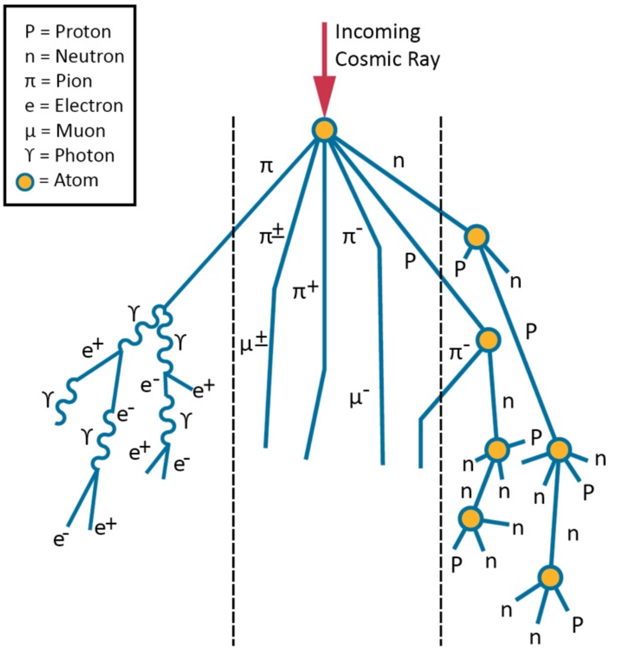


Figure 2.3. Cosmic ray cascade (Redrawn from (Eddy, 2009))

The intensity of GCRs is not uniformly spread across the globe. The intensity at the poles is nearly a factor of four higher than the equator (Kirkby & Carslaw, 2006).

Comparing solar activity with the GCR flux, one can see that they periodically act oppositely to one another, as when one is down, the other moves up. The anticorrelation between the two parameters can be seen in the time series comparison shown in Figure 2.4. Around 20% of the secondary GCR neutrons entering the Earth reduce when the Sun is more spotted and active. However, while the Sun is less active, much more GCRs can be seen. A twenty-two-year pattern acting together with the solar magnetic cycle is evident in the Earth's change in receiving GCRs (Eddy, 2009).

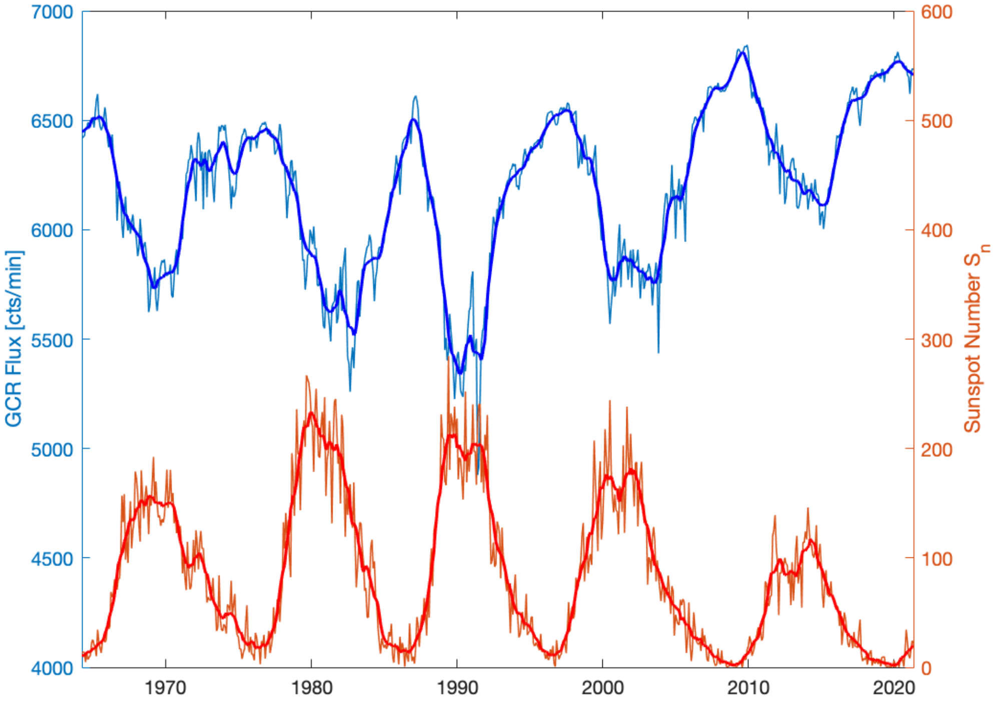


Figure 2.4. Yearly averages of SSN (Data source: WDC-SILSO, Royal Observatory of Belgium, Brussels) vs GCR flux (Data source: Oulu Cosmic Ray Monitor)

The weak energy input from cosmic rays accounts for about one billionth of the solar irradiance, but it is the most critical source of ionization in the atmosphere. The ionization in the atmosphere below 35 km is produced by cosmic rays. They interact with air molecules to form light radioisotopes (such as ^{14}C, ^{10}Be) (Carslaw, Harrison, & Kirkby, 2002). In addition, cosmic rays have two important effects on the global atmospheric electric cycle:

1. Cosmic rays are the main source of ions formed in the depths of the atmosphere far from the Earth's surface.

2. Cosmic rays have a direct effect on the electrical cycle of the atmosphere (affecting conductivity due to ionization).

While the lower levels of the atmosphere are bombarded with the ionization created by cosmic rays, it is perfectly natural to expect a relationship between cosmic rays and clouds (Carslaw, Harrison, & Kirkby, 2002). Today, when the important effect of ions on aerosols and cloud processes is well known, it becomes necessary to determine the relationship between cosmic rays and clouds.

2.3 Clouds

The effect of the change in the amount of cloudiness on the climate is significant because clouds play an active role in the control of the radiative balance of the world. The radiative effect of a cloud depends on the height of the cloud from the surface and the optical depth of the cloud. The effect of clouds on the radiative properties of the atmosphere occurs in two ways (Hartmann, 1993). These are:

1. Cooling, as a result of the reflection of short-wave infrared radiation entering the atmosphere,
2. Warming, as a result of blocking long-wave infrared radiation emitted from the ground.

Therefore, it is essential to examine and determine the relationship between cloudiness and cosmic rays because the net radiative power of global cloudiness is crucial to the Earth’s radiation budget. A global yearly average of about 65% of the Earth’s surface is covered by clouds, and in total, they apply a net cooling effect of 28 Wm^{-2} (Kirkby, 2007). Thus, small changes in the global cloud cover can have a significant effect on climate. Low-level clouds usually have high optical thickness and they act as reflectors of the sunlight, making a negative contribution to the Earth's radiation budget by cooling the surface temperature (Hartmann, 1993).

The optical depth of clouds depends on factors influencing cloud droplet size and distribution and on cloud thickness, which is affected by the atmospheric vertical temperature profile. The dimensional distribution of particles in clouds containing liquid water particles (low-level clouds) depends on the atmospheric aerosol density, acting as cloud condensation nuclei (CCN). The amount of CCN is related to the number of aerosols present in the atmosphere (Hobbs, 1993).

In cloud research, the primary focus is on the effect of pollution on clouds. The connection between cosmic rays and clouds is provided by the effect of ions on cloud microphysics. Cloud particles begin to grow by collecting small aerosol particles. These aerosol particles are formed because of natural events and pollution and enter the cloud with air movements. The number of cloud droplets growing depends on the number and properties of the aerosols (Hobbs, 1993).

2.4 Relationship Between Clouds and Cosmic Rays

Cloud formations can be predicted by many meteorological parameters (such as temperature, humidity, and atmospheric processes). In contrast, cloud properties like lifetime and reflectivity are affected by fine but significant phenomena. These phenomena taking place at the aerosol and cloud condensation particle level are commonly referred to as "microphysical properties". The connection between cosmic rays and clouds is provided by the effect of ions on cloud microphysics (Carslaw, Harrison, & Kirkby, 2002).

Two mechanisms describe the effect of cosmic rays on CCN concentration, namely, the "Ion-Aerosol Clear-Air" mechanism and the "Ion-Aerosol Near-Cloud" mechanism.

2.4.1 "Ion-Aerosol-Clear Air" Mechanism

This mechanism can be defined as the cosmic ray-created ions combining with aerosols in the atmosphere, causing the growth of CCNs. CCN are aerosols with a diameter of about 100 nm, and they act as nuclei in cloud droplet formation. The diagram in Figure 2.5. shows the growth process of cloud droplets. The main source of newly formed aerosol particles is their decomposition into ultrafine condensation nuclei (UCN) composed of condensable vapors such as sulfuric acid (H_2SO_4) (Carslaw, Harrison, & Kirkby, 2002).

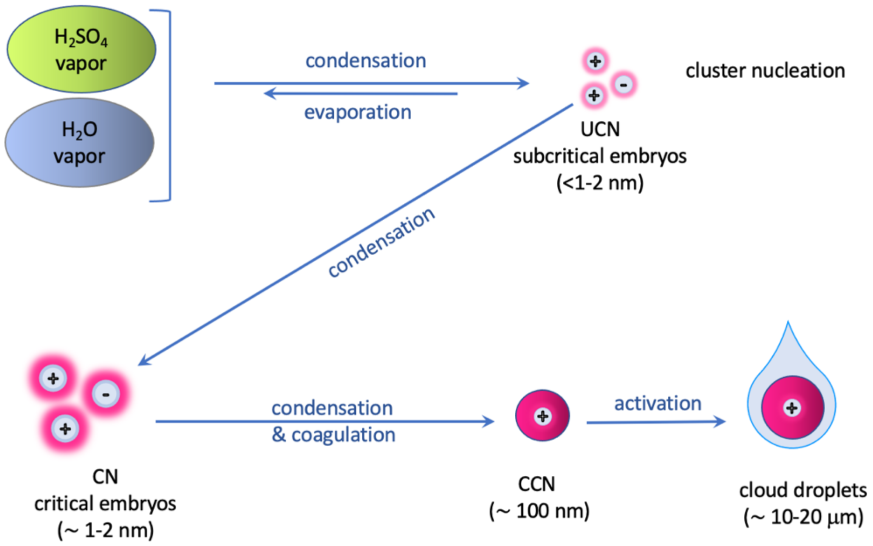


Figure 2.5. Growth of cloud droplets. (Redrawn from (Carslaw, Harrison, & Kirkby, 2002).

Model studies showed that the presence of an electrical charge made nucleation faster by reducing the nucleation threshold. This indicates that nucleation can occur at much lower ambient vapor concentrations than in the non-ionized atmosphere. The results of these models displayed that the nucleation rate of newly formed aerosols in the clear-air regions of the atmosphere is limited by the cosmic ray ionizing rate (Carslaw, Harrison, & Kirkby, 2002).

It is stated that there are two possible ways in which the ionization of cosmic rays can enhance nucleation (Kirkby, 2007). They first induce the formation of particles by stabilizing H_2SO_4 and H_2O molecule clusters through Coulomb attraction; second, they mediate the process of formation by affecting the condensation rates of molecules.

The effect that the changes in cosmic ray intensity have on clouds is similar to the indirect effect of aerosols on clouds. The concentration of aerosols increases with human-induced activities and causes the number of cloud droplets to increase in the same environment. As a result, while the number of droplets in the cloud will increase, their size will decrease. The increase in droplet concentrations results in an increase in cloud reflectivity and repression of precipitation and an increase in cloud lifetime. This is a simple explanation of the indirect aerosol effect on clouds (Carslaw, Harrison, & Kirkby, 2002). The two effects are compared in Table 2.1.

Table 2.1. Comparison of the indirect effect of aerosols and the cosmic rays on clouds (Carslaw, Harrison, & Kirkby, 2002).

	Aerosol Indirect Effect	**Cosmic Ray-Cloud Effect**
Cause	Changes in total aerosol loading.	Changes in microphysical processes by ions.
Effect on Clouds	Increase in cloud cover, cloud lifetime, cloud reflectivity. Decrease in rainfall.	Similar, but some mechanisms are unknown.
Extent of Effect	Effect in spatially limited, polluted atmosphere regions.	Most possibly global effects, but with a clean atmosphere.
Implications	Global radiative cooling that is comparable to global warming.	Potentially sizable modification of the global energy balance on longer time scales.

The cosmic ray effect on clouds and the (indirect) effect of aerosols on clouds are both related to the variation in the number of aerosols. Nevertheless, they are different from each other in three significant ways:

1. The indirect effect of aerosols on clouds is related to the change in the concentration of the total condensable vapors (mainly SO_2). In contrast, the influence of cosmic rays is associated with changes in the rates of microphysical processes.
2. The effect of cosmic rays on clouds can make small changes in the global aerosol numbers. In contrast, aerosol concentration increase due to pollution is restricted to only certain regions of the globe.

3. Pollution effects are stronger in high population areas; thus, the indirect aerosol mechanism is more effective in populated regions. Conversely, cosmic ray effects are stronger in regions with low aerosol concentrations, i.e., with low population (such as clear air above the oceans).

The "ion-aerosol-clear air" mechanism is summarized in Figure 2.6.

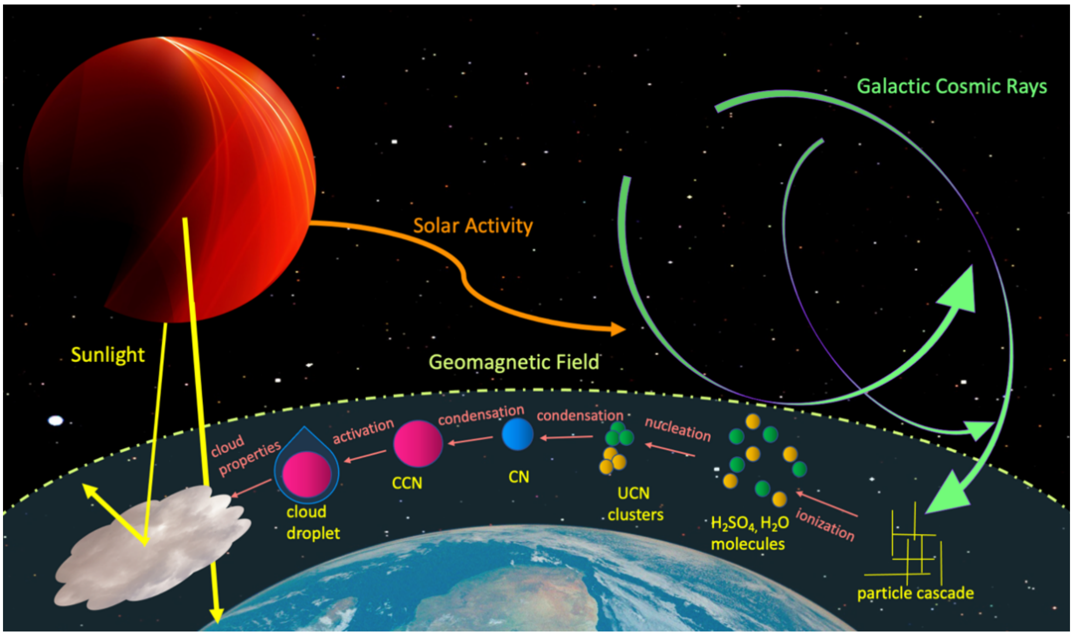


Figure 2.6. The "Ion-Aerosol Clear-Air" mechanism. (Redrawn from (Čalogović & Laken, 2015))

In summary, galactic cosmic rays, modulated by the solar activity, generate particle cascades in the atmosphere and lead to ionization. Under the right atmospheric circumstances, this ionization causes an increase in the nucleation of aerosols and leads to the growth of CCN, which has an influence on cloud properties (Čalogović & Laken, 2015).

2.4.2 "Ion-Aerosol Near-Cloud" Mechanism

This mechanism is based on the differences in aerosol electrical charge near clouds compared to the electrical charge in clear air. Changes in the clear-air electric field cause the thin clouds to be much more positively charged than the clear air above it, and these electric fields are modulated by cosmic rays (Carslaw, Harrison, & Kirkby, 2002).

It has been suggested by theoretical and experimental studies that the efficiency of aerosols is enhanced with this electrification, which is modulated by cosmic rays, yet the exact mechanism is not that well understood. It is thought that a decrease in GCR flux will lead to a decrease in electrically enhanced ice-particle formation, which in turn will decrease the amount of ice clouds and rainfall (Carslaw, Harrison, & Kirkby, 2002).

Of the two mechanisms mentioned, this thesis will be focusing on the “ion-aerosol-open air” mechanism of cosmic rays affecting cloud formation.

2.5 Cosmic Rays, Aerosols, Clouds and Climate

Aerosols play an essential role in the global climate balance and, therefore, are an important phenomenon in climate change. Changes in aerosol concentration from natural events (such as large volcanic eruptions) affect the Earth's radiation balance and thus cause global temperature changes. The impact of aerosols on climate occurs in two ways (Carslaw, Harrison, & Kirkby, 2002):

1. The direct radiative effect: with the absorption and scattering of incoming solar radiation and the radiation emitted from the ground,
2. The indirect radiative effect: with effect on cloud microphysics properties.

Thus, the theory for the effects of aerosols on clouds is that as the aerosol concentration increases, the size of the droplets forming the cloud decreases, and the cloud consisting of small droplets reaches a higher albedo. As a result, cooling occurs.

To conclude this section, the solar activity-cosmic ray-climate relationship can be summarized in the diagram in Figure 2.7.

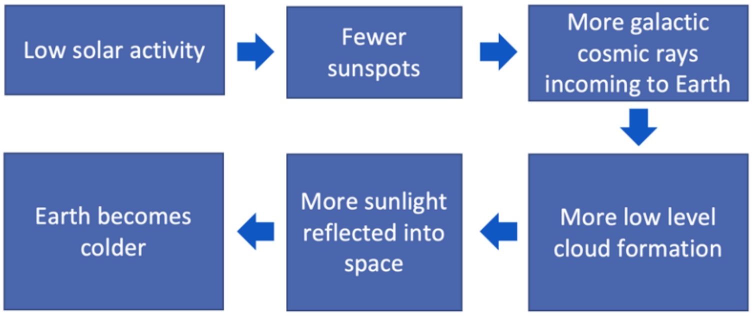

Low solar activity
Fewer sunspots
More galactic cosmic rays incoming to Earth
More low level cloud formation
More sunlight reflected into space
Earth becomes colder

Figure 2.7. The solar-climate relationship (redrawn from (Easterbrook, 2016))

The period of weak solar activity, observed from the fewer sunspots on the Sun, leads to more cosmic radiation reaching the Earth. This cosmic radiation creates an increase in the low-level cloud formation and cloudiness, which leads to more sunlight being reflected and less being absorbed. Hence, the atmosphere cools down, and the Earth becomes colder.

2.6 Machine Learning

2.6.1 Overview of Machine Learning

Machine learning (ML) forms the basis of the data analysis in this thesis; therefore, it is essential to introduce and explain some parts of ML. ML may be used in data science to predict future data from existing data. Such predictions are made using models, which are the mathematical frameworks for ML (Paluszek & Thomas, 2016). The input data in models are used to predict the value of the outputs; this is why the input data are also called the predictor variables, and the output data is called the response variable or the predictand. In statistics, the former is also called independent variables and the latter as dependent variables (Friedman, Hastie, & Tibshirani, 2001).

The learning process in ML can be separated into two main categories, namely, supervised and unsupervised. One very basic difference is that in supervised learning, training and test data sets, which will be explained in more detail in the following section, are used in the ML models. Meanwhile, unsupervised ML does not use training sets, and it is mainly used to find hidden patterns with no "correct" answer inside the data. Thus, there is no specific output variable in unsupervised ML, as the main goal of the model is to understand the data structure. On the other hand, in supervised ML, the output variable is the main learning task of models, and the differences in the type of output lead to different tasks of prediction. When quantitative output variables are predicted, the models are called regression models, and when qualitative models are predicted, they are called classification models (Friedman, Hastie, & Tibshirani, 2001).

An overview of the machine learning processes is summarized in the diagram in Figure 2.8.

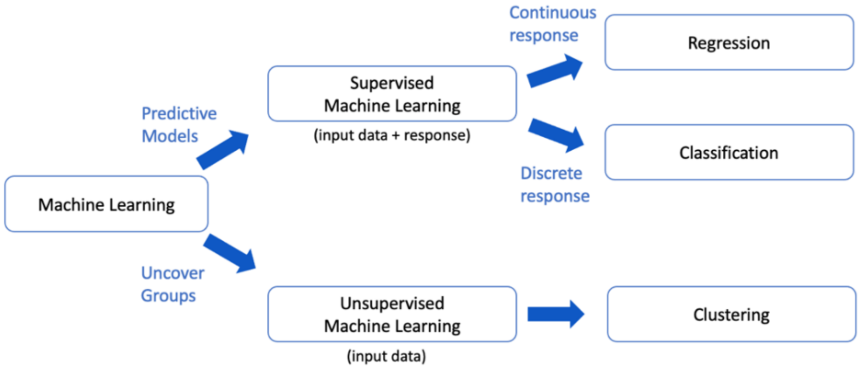


Figure 2.8. Diagram of the machine learning processes (Mathworks, 2021)

2.6.2 Supervised Machine Learning and Regression Models

Supervised ML is the learning process used in this thesis because one of the thesis goals is to analyze the effect of external forces on climate to see how they affect meteorological parameters. This can be achieved by creating a model with multiple input parameters to predict such outputs using regression (Paluszek & Thomas, 2016).

There are also two parameter types commonly referred to in ML; model parameters and hyperparameters. Model parameters are the parameters used when creating the models. Such parameters can be set before training a model, or they can be learned during the training process by the algorithm. Hyperparameters, on the other hand, are not learned by the ML algorithm in training. They are determined by the user with hyperparameter optimization (Paluszek & Thomas, 2016).

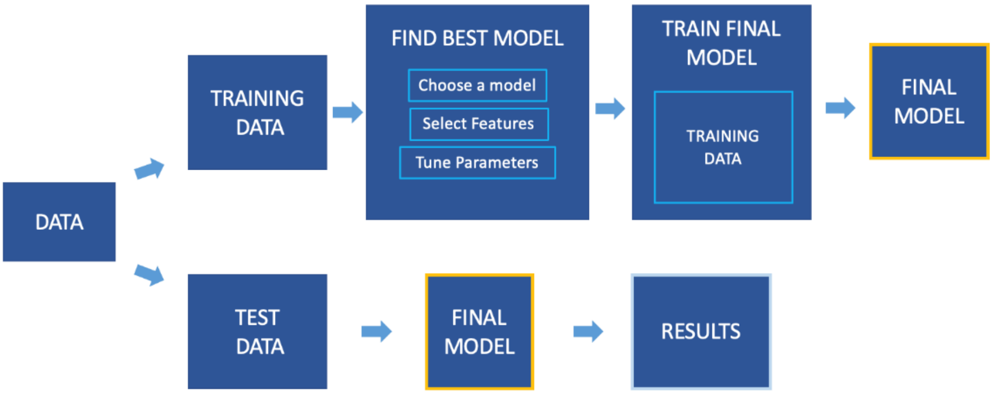


Figure 2.9. Diagram of the train-test processes (Mathworks, 2021)

Training the model is the essential part of supervised ML, and it is followed by testing the model to validate results. The complete data can be divided into two parts: training data and test data. Training data is used to construct the ML model, while the test data helps to evaluate the performance of the final model. (Paluszek & Thomas, 2016). This process is summarized as a diagram in Figure 2.9.

As mentioned above, the method used in this thesis to fit data into a model is regression. The simplest form of regression is the linear model (LM), which predicts the response variable by multiplying the predictor variables with coefficients. More complex regression models are discussed in Section 2.6.5.

The linear model can be formulized as in Equation 1;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad (1)$$

where y is the response variable, x_i are the predictors, β_i are the regression coefficients and β_0 is the intercept (bias) term (Friedman, Hastie, & Tibshirani, 2001).

2.6.3 Metrics Used in Regression Models:

Here the most important metrics used for understanding regression models are listed. Firstly, residuals (r_i) are the errors for every data point in the model (Paluszek & Thomas, 2016).

$$r_i = y_i - \hat{y}_i \qquad (2)$$

Here y_i stands for the actual response and $\hat{y}_i$ is the predicted response.

Figure 2.10. shows the residual definition graphically for one predictor variable that is displayed on the x-axis and one response variable displayed on the y-axis. The red line is the regression line that best fits the data, and the residuals are the difference between the observed and predicted values.

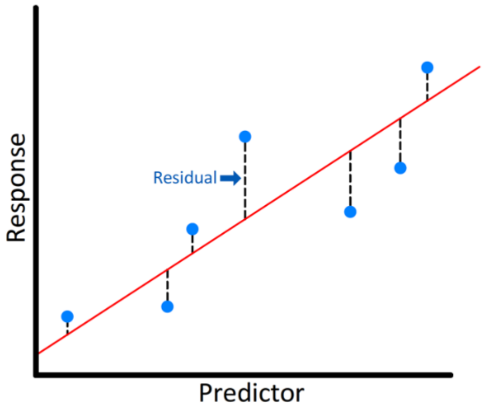


Figure 2.10. Graphical representation of the residuals (Mathworks, 2021)

The average of the squared residuals of a model is called the Mean Square Error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Most regression models aim to minimize this term during training (Friedman, Hastie, & Tibshirani, 2001).

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ , \tag{4}$$

is the square root of MSE, and it is a useful term in detecting outliers and errors (Friedman, Hastie, & Tibshirani, 2001). The Sum of Squared Errors (SSE, Eq. 5) and the Sum of Squares Total (SST, Eq. 6) are both error metrics used in calculating R^2 (Friedman, Hastie, & Tibshirani, 2001).

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (6)$$

where $\bar{y}$ is the mean response variable.

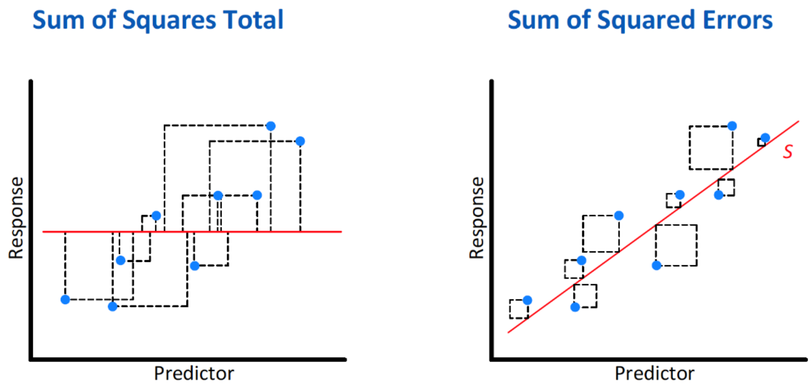


Figure 2.11. Graphical representation of the difference between the SST and SSE (Mathworks, 2021)

The Coefficient of Determination (R^2),

$$R^2 = \frac{SST - SSE}{SST}, \qquad (7)$$

is the difference in total error, calculated from the fitted model, and it theoretically has a value between 0 and 1. The value is closer to 0 when the error is large, and the model does not fit the data well, while the value is closer to 1 when the error of the model is small (Friedman, Hastie, & Tibshirani, 2001). However, a model worse than the simplest model, which is simply the mean of the predictand values, can lead to negative R^2 values. The more predictors are added to the regression equation, the more opportunities there are for random events to reduce SSE and thus result in a better R^2 value. As a result of this, in computer models, an adjusted R^2 value is more often reported, and it scales the R^2 by the number of predictors used (James, Witten, Hastie, & Tibshirani, 2013).

These metrics are used to evaluate regression models and to quantify their performance. When comparing models with each other, the error metrics MSE and RMSE are helpful, while the R^2 value is a good measure of the model fit (Paluszek & Thomas, 2016).

2.6.4 Validation of Machine Learning Models

Validation is a crucial part of ML methods, and to understand it, two terms, namely bias, and variance must be introduced. The inability of a ML method to capture the true relationship of a model is called bias, while how much the ML method can adjust the model to the differences in data is called variance (James, Witten, Hastie, & Tibshirani, 2013).

The change seen in training and test errors as the complexity of a model changes is plotted in Figure 2.12. When the complexity of a model increases, the fit of data is higher, and the error of the training set decreases. In contrast, the error of the test data increases because the model only fits the training set well and cannot generalize. Thus, when the model has high complexity, it will overfit, and the model predictions will have low bias and high variance. In contrast, it will underfit when the model has low complexity and the predictions have a high bias with low variance (Friedman, Hastie, & Tibshirani, 2001).

Hence, in ML, the ideal algorithm would have a low bias to model the true relationship between predictors and the response accurately, and it would have low variability by producing consistent predictions across different datasets (James, Witten, Hastie, & Tibshirani, 2013).

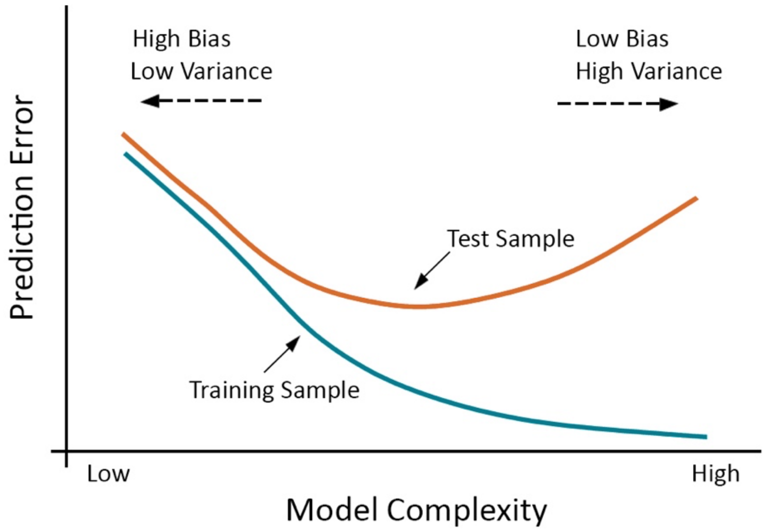


Figure 2.12. Bias- variance tradeoff (Redrawn from (Friedman, Hastie, & Tibshirani, 2001))

The bias-variance tradeoff is the tradeoff in the complexity of the model, and a good balance between the bias and variance of a model should be found, in which there should be no overfitting or underfitting (Friedman, Hastie, & Tibshirani, 2001).

Validation is what helps to prevent overfitting. With the initial raw data, the algorithm first needs to be trained to estimate parameters for the ML methods and then tested to evaluate how well the ML method works. Re-using the exact data for both training and testing would result in overfitting because the performance on data the model was not trained on should be tested (James, Witten, Hastie, & Tibshirani, 2013).

The two most used validation methods in ML are hold-out validation and k-fold cross-validation. The hold-out validation method is performed by holding out a certain percentage of the original data set and using it as the test dataset while using the remaining part as the train data. This method is the process shown in the diagram in Figure 2.9. It is a simple method; however, it has two disadvantages. Firstly, the validation of the test set is highly variable as it depends on which observations are divided into the train set and which go into the test set. Second, the error on the test set might be overestimated due to not all the observations going into the train set, and the model is trained with fewer data points (James, Witten, Hastie, & Tibshirani, 2013)

On the other hand, k-fold cross validation equally divides the original dataset randomly into *k* number of groups, takes one group, and validates the remaining data. This is done using each group one by one, *k* times, and then the results are averaged. (James, Witten, Hastie, & Tibshirani, 2013). The methods that can be used to validate data are summarized and compared in Table 2.2.

Table 2.2. Validation Methods (Paluszek & Thomas, 2016)

Holdout Validation	**K-Fold Cross-Validation**
Validates the data once	Splits the data into k subsets and validates the data k times
Better for large data sets	Better for small data sets
Faster method	Slower method

This thesis will use hold-out validation in the ML part of the analysis because the datasets are large.

2.6.5 Multiple Regression Models

Four types of multiple regression models will be used in this analysis: the three linear models are the linear regression model, generalized linear regression model, stepwise linear regression model, and the non-linear multiple regression model used is the random forest model.

2.6.5.1 Multiple Linear Regression

Simple linear regression is used when a response parameter is predicted using only one predictor (James, Witten, Hastie, & Tibshirani, 2013). However, when there is more than one predictor involved, it is called multiple linear regression, and it is formulized in Eq. 1.

When the predictors are dependent on each other, additional terms called “interaction” terms can be added to the model, and Eq. 1 becomes,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \ldots + \beta_{1n} x_1 x_n + \cdots + \beta_{n-1,n} x_{n-1} x_n \qquad (8)$$

where the cross-product terms quantify the interaction among predictors. With these additional terms, the non-linearities between predictors are also approximated.

Generalized linear regression and stepwise linear regression are extensions of the linear model. The main difference between the linear model (LM) and generalized linear model (GLM) is that in the LM, the response variable assumes a normal distribution for each set of the predictor values, while in the GLM, the response variable can have a distribution that can be, for example, binomial or Poisson as well as being normal (Ciaburro, 2017).

The stepwise model again creates a linear model starting from the entire dataset of predictors, and it adds or removes predictors at each step of the regression to determine a final model. The terms to add or remove are searched based on criteria determined by the user, such as to increase the value of R^2 (Ciaburro, 2017).

Hence, the above-mentioned three different models have been used to investigate the linear relationships in the multiple regression models of this thesis.

2.6.5.2 Random Forests

The random forest regression model is based on the decision tree method, a continuous process that splits the dataset into “branches” and expands these branches in each step. These branches go from the observation dataset to the response, and the model shape takes a final form similar to a tree, hence the name “decision tree”. Decision trees are a simple ML method to use, yet they have two significant drawbacks. The first is that they are mostly computationally expensive, and second the predictions might get deviated if the underlying data changes. Thus, instead of using one single decision tree, combining many trees will improve the prediction accuracy (James, Witten, Hastie, & Tibshirani, 2013).

Random forests construct multiple decision trees at once during the training time, and they combine the result of each prediction of each tree to determine a final output. Each decision tree selects a random sample from the original dataset, and this randomness greatly prevents overfitting (James, Witten, Hastie, & Tibshirani, 2013).

The advantages of using random forests are as follows: firstly, they are run efficiently on large datasets; and secondly, they are an effective method for estimating missing data, and they can maintain the accuracy even when a large part of the data is missing. A major disadvantage is that overfitting may be observed for datasets with noisy data (James, Witten, Hastie, & Tibshirani, 2013).

Random forests are used in this thesis because they work well with large datasets and are suitable for detecting non-linear relationships in regression.

CHAPTER 3

LITERATURE SURVEY

There have been many studies searching for a link between solar activity and the Earth's climate. Possible relationships were sought through investigating TSI and UVI variations. The variation in TSI causes an input of heat in the lower atmosphere; however, this heat by itself is not sufficient to affect the global climate. The variation in UVI is hypothesized to affect the absorption in the lower atmosphere (Carslaw, Harrison, & Kirkby, 2002). However, these studies lacked a direct explanation of the climate mechanisms, and more research was needed.

One possible explanation for this problem came from Svensmark and Friis-Christensen (Svensmark & Friis-Christensen, 1997) where they suggested a link between cosmic rays and global cloud cover. Correlation analysis was made between ISCCP globally averaged total cloud cover and GCR flux for the period July 1983-December 1991. The cloud data was smoothed with a 12-month moving mean filter to remove seasonal effects. It was concluded that the variation of 3–4% of the global cloud cover was positively correlated with the GCR flux. They stated that the calculated correlation coefficient increases from 0.95 to 0.97 if a 12-month moving mean filter is applied to the GCR data. In 2000, Marsh and Svensmark revised the analysis (Marsh & Svensmark, 2000) from Svensmark 1997, this time extending the correlation analysis period to 1994, and looked at the cloud cover at different altitudes: high, middle, and low. They claimed that the GCR correlation was restricted only to low clouds. In both studies, the global cloud cover was used with excluding the tropics region (25°S - 25°N). This was explained in 2000 by Svensmark (Svensmark, 2000) in two reasons. The first was that the flux of GCRs reduces towards the equator, and the second was that the cloud processes were different in the tropical regions compared to the regions at higher latitudes.

Pallé Bago and Butler (Pallé Bagó & Butler, 2000) also investigated the relations between the GCR and monthly ISCCP cloud levels for the years 1983-1994. They used regression to extend the study to the past century. They stated that, in general, the CR effect is expected with high cloud covers and not with low cloud cover. However, most studies unexpectedly detected a high correlation with low clouds. They continue to say that one reason for this discrepancy could be explained by how the data acquired from neutron monitors, which are located on the ground level, represent mainly the lower regions of the atmosphere. This is an important methodological limitation. They concluded their study by saying that the warming in the 20th century could be mathematically explained by solar activity effects without considering the most recent years.

Correlations were found between the cloud cover and GCRs, yet the physical mechanism behind these correlations was uncertain. In 2002, Carslaw et al. (Carslaw, Harrison, & Kirkby, 2002) talked about the different physical processes that could act as a link between climate and cosmic rays. These mechanisms are explained in detail in Section 2.4. Another important result of model studies mentioned in their paper was that the formation of aerosol concentrations was greatest in the lower atmosphere. They stated that the ionization rate in this region was a constraining factor for new aerosol particle formation; and that at the upper levels of the atmosphere, where the cosmic ray density is higher, the ionization rate had no such function. This explained why the relationship between cosmic rays and cloud cover could be more evident in low-level clouds.

The “ion-aerosol clear-air” mechanism, was also explored by Yu in 2002 (Yu, 2002). He found that with an increase in GCR rates, there is also an increase in CNs in the lower part of the troposphere. The troposphere is the lowest layer of the atmosphere, where all the weather events occur, including cloud formation. On the other hand, in the upper part of the troposphere, a decrease in CN production was found, with increasing GCR flux. They stated that these results showed that the “ion-aerosol clear-air” mechanism could explain the altitude dependency of the correlations reported earlier between global cloud cover and GCRs. Hence, this paper supported the GCR-climate relationship with further evidence.

In 2003, Marsh and Svensmark (Marsh & Svensmark, 2003) published another study, in which they suggested that the low cloud cover was statistically correlated to the El-Niño-Southern Oscillation (ENSO) process. They stated that there was a gap in the ISCCP data from September 1994 to January 1995 due to a satellite calibration problem, and for this reason, two periods were used for the analysis to compare the differences. The first period was from July 1983 to August 1994, and the second period was from July 1983 to September 2001. The ENSO index, which is the average sea surface temperature anomaly, used in their study covered the region between latitudes 5°N-5°S and longitudes (150°-90°) W. This region is in the Eastern Pacific Ocean and is known as the NINO3 area. The ENSO index was included in their study because they highlighted that it was known to have a strong effect on high cover cloud properties, and they wanted to see the relations with low cloud properties. They stated that the link between low cloud cover and GCRs throughout 1983 to 1994 was approved by the ENSO signal, yet the relation weakened when the whole period up to 2001 was considered. It was explained that this weakening was either due to the calibration problems of the satellites or due to changes in the physical processes of the ENSO events.

However, Laut (Laut, 2003) opposed the GCR-climate correlations in 2003 by revising the previous work conducted by Marsh and Svensmark (Marsh & Svensmark, 2000). He extended the datasets of the previous study up to 2000 and plotted the time series. He stated that the agreement between GCRs and low cloud cover was questionable starting from 1990, and after 1994, there was no agreement seen at all. He also stated that the change in cloud cover was delayed about six months compared to the GCR flux. That should not be the case because according to the “ion-aerosol clear-air” mechanism, the formation of clouds should occur within a few days at maximum, which should seem instantaneous when the plotted datasets are yearly averaged. He finally stated that the physical interpretation of the ISCCP low clouds was difficult because they depended on the IR measurements, which could cause problems when being detected behind high clouds.

Supporting the correlations between GCR-climate, a different approach was used in a Master’s Thesis by Akcan (Akcan, 2004) at İstanbul Technical University. The relationship between Earth’s cloudiness and the cosmic ray intensity during one solar cycle was investigated, and a prediction study was held using a neural network-based model. The study period was from 1983 to 2001, and an area between latitudes 21.25° - 58.75°N and longitudes 8.75°W - 48.75°E was chosen to be studied. It was stated that during a solar cycle of 11 years, the changes in cloud cover and the GCR flux were periodic and parallel to each other, and there was a phase difference of ~ two months between the monthly cloud cover and GCR flux. After conducting the study on the local scale, it was pointed out that the obtained results were not reliable because local effects were difficult to remove. It was suggested that future studies should be held in larger regions. The results of the model used were close to the observed values, and it was concluded that a statistically significant relationship was found. It was further suggested that in future studies, other atmospheric parameters that affect the amount of cloudiness should be included in the analysis to determine the order of importance of the factors.

In 2004, two important independent studies looked at the relationship between GCR ionization and the ISCCP low cloud amount and supported the hypothesis. The first, by Pallé et al. (Pallé Bagó, Butler, & O’Brien, 2004) studied the pixelwise statistical relations for the years 1983-2001 and found that there was a stronger correlation in the mid-latitude region. They detected a decrease in low cloud cover for the last century and concluded that this finding could be the explanation for a substantial part of the ongoing global warming.

In the following study, Usoskin et al. (Usoskin, Marsh, Kovaltsov, Mursula, & Gladysheva, 2004) investigated the GCR-cloud relationship by using latitudinal region averages for the first time, instead of the previously used global averaged cloud data. The three different regions that they used in their analysis were; the global average excluding the poles (60°S - 70°N), the tropics (25°S - 25°N), and the mid-latitudes ([60° - 25°]S and [25° - 70°]N). It is stated that they excluded the polar regions from the study due to the detection of the clouds over ice regions being problematic. They also mentioned the decreasing trend observed in the ISCCP data and that this trend was not uniformly distributed around the globe. Thus, they compared the correlations calculated from raw ISCCP data with the detrended ISCCP data; and found that detrending the ISCCP low could data greatly improved the correlation. They concluded that the GCRs are mostly correlated with the low cloud amount over the mid-latitude region for the years 1984 – 2000, and their results support the hypothesis of GCRs modulating the low cloud amount for the studied years. It is also stated that detrending the ISCCP low could data greatly improved the correlation.

Usoskin et al. (Usoskin, Voiculescu, Kovaltsov, & Mursula, 2006) reanalyzed the previous work in 2006 to see if the observed ISCCP instrument effects, also mentioned in Lauts's study (Laut, 2003), were a fact or an artifact. They showed that in some geographical regions, such as the South Pacific and North Eurasia, the correlation between GCRs and low cloud cover could be affected by the high cloud cover. However, they found that for larger geographical regions, such as Europe, the South Atlantic, Northwest Pacific, and the West Indian regions, the low cloud cover was not affected by the high or middle cloud cover. Thus, it was stated that it was safe to study these regions using the ISCCP data. They concluded that the results of earlier studies looking at the relationships between certain cloud types and any solar proxies could be distorted because of using global and latitudinally averaged datasets. Hence, they stated that future studies should be limited to the regions specified above that were not affected by higher-level clouds.

In the same year, Voiculescu et al. (Voiculescu, Usoskin, & Mursula, 2006) studied the correlations between UVI, GCR, and the pixelwise annual cloud amount data for the high, middle, and low-level clouds. They tried to distinguish between the UVI and GCR effects on clouds. They found that the low cloud amount was essentially driven by UVI in all the ocean regions but the Atlantic. They also stated that the observed high correlations seen in Europe and North America could be a result of the high aerosol concentrations generated in these critical industrial regions. They found that in most of the pixels, GCRs are correlated positively with low cloud amount and negatively with high cloud amount. They remarked that the solar effect on cloud cover is not only dependent on one solar parameter and that the climatic conditions in every pixel are also crucial. They also stated that the low cloud amount is mainly driven by UVI over oceans and dry lands, while it is driven by GCRs over moist land areas that have higher concentrations of aerosol.

In 2007, a study by Voiculescu et al. (Voiculescu, Usoskin, & Mursula, 2007), that looked at the correlation between clouds, solar proxies, internal ENSO effects, and volcanic eruptions showed that removing ENSO years from the analysis did not change the Sun–cloud relation statistical results. It is further stated that the correlation even improved between the different cloud covers, GCR flux, and UVI. This result is used to support the idea that clouds are directly influenced by solar activity. An important finding in the study is about an area in the eastern Pacific, where the relationship between GCR and clouds was found to be the opposite of the remaining part of the globe. In that specific region, the high-level cloud cover was positively correlated with the GCR flux while the low-level cloud cover was negatively correlated with GCR, while for the rest of the world, the low-level cloud cover was positively correlated with GCR, and the high-level cloud was negatively correlated. It was stated that removing the ENSO years made the unexpected correlation in the specified region disappear, and it was put forward that in the particular region of the Pacific Ocean, the ENSO effect prevailed over solar effects. It was concluded that even though the removal of the ENSO effect in the analysis does not change global correlation patterns, for some areas that are subject to being affected by internal climate processes (such as the mentioned Eastern Pacific region), the correlation results must be interpreted carefully.

Based on the studies mentioned above, the term called ‘cosmoclimatology’ was put forward by Svensmark in 2007 (Svensmark, 2007), by reviewing the evidence for the cosmic effect on climate. It was concluded that the Earth's climate was solar-driven, and the humans’ contribution to recent climate change was trivial. In 2008, Usoskin et al. (Usoskin & Kovaltsov, 2008) also reviewed the experiments and theory of the possible connection between GCRs and climate. They concluded that the link between the cosmic rays and climate seemed to be a possible climate driver; however, they stated that there were still questions on the physical mechanisms of the relationship.

In 2008, Kristjansson et al. (Kristjánsson, et al., 2008) approached the “cosmoclimatology” problem from a different perspective, mentioning that there were important inhomogeneities observed in the datasets of ISCCP. They instead used the Moderate-resolution Imaging Spectro-radiometer (MODIS) data to study the connections. Many different cloud and aerosol properties were used as the meteorological parameters in the correlation investigation. Their results showed that only the eastern Atlantic Ocean region showed a statistically significant correlation, and most of the other regions showed only a weak correlation.

Pierce and Adam (Pierce & Adams, 2009) tested the “ion-aerosol clear-air” mechanism using a general atmospheric circulation model to understand how the changes in GCR flux affect the formation of CCN, and they found that the concentration of CCNs is not very sensitive to the changes in the GCR flux. They concluded that their results showed that the “ion-aerosol clear-air” mechanism was too poor in explaining the correlations between solar activity and climate. Carslaw (Carslaw, 2009) reviewed their work and stated that even though the results of Pierce and Adam were against the cosmoclimatology theory, this would not stop the investigation of a link between cosmic rays and clouds. He said that this was first because scientists continue to be interested in the correlations between climate variables and solar proxies. Secondly, this relationship would stay a controversial subject in which all of the possible mechanisms should be explored.

Erlykin et al. (Erlykin, Gyalai, Kudela, Sloan, & Wolfendale, 2009) further analyzed the cloud-cosmic ray correlations previously stated by Svensmark and Friis-Christensen (1997) and Palle Bago and Butler (2000). The different altitude levels of global cloud cover were analyzed in latitude regions of 20-degree bands, using ISCCP data up to the year 2005. They mentioned that the middle-level cloud cover was strongly anti-correlated with the low cloud cover. Their analysis was concluded by stating that there is no “causal connection” between GCRs and the low cloud cover and that the two are only correlated due to their common dependence on the TSI at the Earth. Erlykin et al. (Erlykin, Sloan, & Wolfendale, 2009) also investigated the relationship between cosmic rays, solar activity, and the global mean temperature. They noted that global temperature fluctuation was in phase with the solar cycle (measured by SSN and TSI) and in antiphase with the GCR cycle. The delay of the GCR cycle by two to four years meant that the solar activity, rather than GCRs, was the most likely causative factor for the fluctuation. This study also concluded that the correlation of GCR intensity and low cloud cover could be due to their causal relationship with the TSI at the Earth.

Laken and Calogovic (Laken & Čalogović, 2011) investigated the GCR-cloud link using daily timescales. For this purpose, they investigated high amplitude short-term reductions in the flux GCRs, named Forbush Decrease (FD) events. They also went over the previous work looking at FD events and stated that most of the results conflicted with each other. A couple of possibilities for this conflict were stated: first that there could be no relationship between climate and solar activity; second, that a connection does exist yet it is hidden by the atmospheric conditions of each study; third that even reducing the timescale to daily datasets was not enough to detect subtle relations; and fourth that when dealing with daily timescales the sample size is too little, which could be limiting the detection of any significant signals. Laken and Calogovic addressed the third possible problem in their study by carefully choosing their samples to isolate specific period ranges. They used TSI, UVI, and GCR data and correlated them to ISCCP cloud amount at different altitudes for low and high latitude regions. They concluded that TSI or UVI variations during the GCR flux reductions do not illuminate the solar-cloud response at any altitudes.

Erlykin et al. (Erlykin & Wolfendale, 2011) wrote a paper which was a survey on the evidence for and against the GCR-cloud relationship. It was concluded that while for high latitudes, there is an observed effect, for specific regions, the GCRs do not affect the atmosphere at all. They stated that the average effect over the globe was small and insignificant.

In 2012, Voiculescu and Usoskin (Voiculescu & Usoskin, 2012) studied the correlation of GCR-induced ionization and UVI with the pixelwise global cloud cover between the years 1984 - 2009. They generated correlation maps globally for low, medium, and high cloud covers. All datasets were averaged annually to remove seasonal effects. A high positive correlation between low cloud and GCR was found in most regions, while a negative correlation was seen in the west of South America and the Indian Ocean. Meanwhile, the UVI was strongly anti-correlated over similar regions. Overall, the strongest correlations were observed for the low cloud amount compared to the other two types of clouds. They concluded that there were no solar effects on cloud cover at the global level, and the correlations existed mainly for certain climate-defining areas.

Following Krisjansson's work in 2008 (Kristjánsson, et al., 2008), Laken et al. (Laken, Pallé Bagó, & Miyahara, 2012) compared the MODIS data with the ISCCP data and looked at the relationship of GCR, TSI, and the ENSO index (MEI) with ISCCP and MODIS cloud over a more extended period. Comparing the two satellite datasets, they first mention that MODIS has a higher spectral resolution, which is better to study cloud properties. They also state that MODIS uses a more sensitive method in identifying cloud-top pressures, compared to ISCCP, which uses emissivity. An important difference between the two is that the MODIS cloud product is the daily mean cloud of the observed pixels at daytime, while the ISCCP cloud product adjusts nighttime IR cloud retrievals using the daytime visible and IR retrievals together. This results in a difference between the observed cloud amount in the MODIS and ISCCP data, especially for the daytime cloud data over land areas. They questioned the reliability of ISCCP data. It is suggested that the trend observed in the ISCCP cloud cover data itself could be mainly artificial and may be due to satellite viewing problems. They concluded that no globally significant correlation between GCR/TSI/MEI and cloud was found; and that the solar variability was not a contributor to MODIS cloud.

Again, using the MODIS data, Laken et al. (Laken B. A., Pallé Bagó, Čalogović, & Dunne, 2012) analyzed the solar-cosmic ray-cloud relations. Original research by Svensmark and Friis-Christensen (1997) using the ISCCP data was reanalyzed, without smoothing the datasets, it was shown that the resulting correlations were not statistically significant. Also, it was suggested that the ISCCP low cloud cover data itself could be mainly artificial. The ISCCP data over the 1983-2010 period was shown as a map, and the mentioned artificial trend can be observed as jumps in the dataset. They concluded their study by stating that there is no evidence for a solar-cloud link.

In 2013, Ahluwalia (Ahluwalia, 2013) searched for a correlation between GCRs and global surface temperature using the most extended robust datasets for the years 1900 - 2013. He stated that the correlation coefficient between the two parameters was not sustained throughout the entire period and exhibited both positive and negative values varying between 5 to 20 years. He concluded that there was no direct linear correlation between climate and GCRs, yet there might be non-linear effects in operation.

Erlykin et al. (Erlykin, Sloan, & Wolfendale, 2013) reviewed the recent results of the Cosmics Leaving Outdoor Droplets (CLOUD) experiment at CERN. CLOUD is a laboratory experiment set up to study the effect of the GCRs on clouds and aerosols and to understand the growth of CCN particles. The most relevant results are summarized in this paper. The results confirm that ions present in the atmosphere do increase the aerosol nucleation rates. It is concluded that even though there could be a correlation between the GCRs and climate through the nucleation of aerosols, it is expected to be very small; and further, the expected result would be a slight cooling, not warming because the intensity of incoming GCRs have increased.

Ahluwalia (Ahluwalia, 2014) looked at correlations between GCRs, TSI, SSN, and the surface air temperature on the Earth. It is stated that there is no sustained correlation between the global temperature and the GCR, TSI, SSN datasets. Also, Marsh and Svensmark's (2000) results were criticized, as the report of a strong correlation between low clouds and GCR was based on a limited dataset for July 1983-August 1994, and extending data further would make the results trivial. It is concluded that global temperatures are not affected by changes in the GCR, TSI, or SSN.

In 2015, Kilifarska et al. (Kilifarska, Bakhmutov, & Mel'nik, 2015) sought statistical relations between GCR intensity, SSN, surface air temperature, total ozone content, and the atmospheric CO_2 concentration. They stated the following chain of interaction for explaining the cosmoclimatologic theory: the GCR intensity and depth of penetration in the atmosphere are regulated by the geomagnetic field. GCR activity alters the ozone density in the tropopause that subsequently changes its temperature, which leads to the adjustment of the extratropical upper tropospheric stability and humidity. Ultimately the surface air temperature is altered due to the greenhouse effect of water vapor.

Tsonis et al. (Tsonis, et al., 2015) investigated the link between GCR flux and the year-to-year changes in global temperature. A coupling between the two parameters was found. However, they strongly indicated that this finding did not imply that GCRs could explain the current global warming. Instead, they stated that their results showed that GCRs could not explain the current increase in global temperature and that anthropogenic forcing had to be involved.

Calogovic and Laken (Čalogović & Laken, 2015) reviewed the work on cosmoclimatology up to 2015 and outlined the observational results. They talked about the difficulties and limitations in the analysis of climate parameters. They first mentioned that because solar parameters such as SSN and GCR are co-varying, disambiguating the forcings is almost impossible. Hence, they say that the previously observed effect of GCRs on climate could just be a result of GCRs being a proxy for solar activity rather than it being a separate mechanism. The second significant problem is with the internal climate forcings, such as volcanic activity or the El Nino Southern Oscillation, which can have similar periods of variation as solar activity. Hence, they can make it difficult to distinguish between the internal and external forcings on climate. Finally, they touched upon the subject of the most commonly used ISCCP cloud data itself having important quality problems, with the observed trends and errors in the dataset. It is mentioned that in long-term analyses, the ISCCP cloud data is not that reliable because problems arise from changes in the satellite calibration and the instrumentation degrading over time. They then talked about a way to solve the limitations with studies that are long-term, which is instead to work with short-term times scales, such as hourly or daily. However, they stated that the short-term studies are compromised as well due to the data being auto-correlated with each other, and the statistical analysis is complicated. It is concluded that these limitations are the reason for the different results in the cosmoclimatology studies, and they state that there is no “compelling evidence” to prove the cosmic ray and cloud relationship.

On the other hand, in 2017, Ogurtsov and Veretenenko (Ogurtsov & Veretenenko, 2017) stated that radiative properties of low clouds may be sufficient to explain the global warming effect for the last 30 years without requiring any human cause. They studied the relationship between GCRs and cloud using the ISCCP data for the years 1983 to 2010. The data was investigated in two different periods: 1983 – 2003 and 2004 – 2010. They stated that the two parameters are significantly positively correlated between the years 1983 to 2003. However, the sign of the correlation coefficient reverses in 2003, and the correlation weakens. The correlation coefficient between the global detrended monthly ISCCP low level cloud and the GCR flux is 0.62 for 1983 – 2003, while it is -0.38 for 2004 – 2010. It is suggested that the correlation coefficient sign reversal effect may be due to the changes in atmospheric circulation, mainly the transition of the circumpolar vortex from a strong state to a weak state.

Bhaskar et al. (Bhaskar, Ramesh, Vichare, Koganti, & Gurubaran, 2017) performed an extensive study to identify the possible drivers of global warming. The used parameters in the analysis were: greenhouse gases such as CO_2, CH_4, N_2O, aerosols, UVI, TSI, GCR flux, the ENSO index, and GMTA for 1984 – 2005. Instead of looking at the correlations between parameters, a different approach, implementing a non-parametric information technique, was used. It was found that aerosols on their own contributed to the global temperatures by about 23% and were a major driver of climate. The other parameters were found to contribute to global temperatures in the percentages of CO_2 by 24%, CH_4 by 19%, ENSO by 12%, UVI by 9%, GCR by 5%, N_2O by 5%, and TSI by 3%. They conclude that this analysis was done for globally averaged values, and the results could be different at regional levels.

In 2018, Veretenenko and Ogurtsov (Veretenenko & Ogurtsov, 2018) continued studying the possible reasons for the correlation coefficient sign reversing. The area of research was the mid-latitudes between the years 1983 - 2010. They mentioned that the ISCCP cloud data was in good agreement with other satellite data at these latitudes, such as MODIS. They concluded again that the stratospheric polar vortex was an important factor in the long-term cosmic ray effect on cloud cover. Strong vortex was present from 1980 to 2000, but a striking alteration occurred in the early 2000s on the state of the vortex in both hemispheres, which changed the nature of the effect of GCRs on the evolution of cyclones, which in return affected the cloud formation processes.

In the meantime, another study opposing the cosmoclimatology theory came from Ormes (Ormes, 2018), in which the studies up to 2017 were thoroughly reviewed. He concluded the review by stating that there is a lack of reliable and consistent correlation between any specific cloud type and location; and that the discussed GCR-climate processes are not strong enough to contribute to the recent global warming.

In 2019, Biktash (Biktash, 2019) studied the relations between SSN, TSI, GCR, and global air temperature from 1983 – 2017. It was stated that cyclical changes in global temperature were associated with solar activity through the effect of TSI on the Earth's atmosphere and GCRs, supporting the cosmoclimatologic theory. However, it was concluded that the slow increase of the Earth temperature during the recent solar cycles was attributed to the anthropogenic factor alone.

Another critical study in the same year came from Chapanov and Gorshkov (Chapanov & Gorshkov, 2019), where they studied the solar influence on climate, specifically in one region: Europe. The reason behind the selection of this region was that Europe has the longest and highest quality of meteorological parameter time series. Using such series, the connections between TSI, GCR, precipitation, and temperature variations for 1766 to 2000 were investigated. The chosen area was further divided into three latitude regions; (50° – 55°)N, (55° – 60°)N, (60° – 65°)N, to study latitudinal effects in detail. Oscillations of each time series were analyzed., and the sign reversal of the correlation coefficient between GCRs and climate parameters was shown for a more extended period. It was concluded that the effect of GCRs on climate is more significant in higher latitudes, particularly on precipitation over the (55° - 65°) N region.

Singh and Bhargawa (Singh & Bhargawa, 2020) also examined the relationships between TSI, GCR, cloud amount, and global surface temperature from 1983 to 2018. They found that the variation in cloud cover depended on TSI, GCR, and global surface temperature altogether, but the most contribution came from the global surface temperature. Also, for the GMTA, it was found that both the global cloud cover and TSI have an essential role. They also touched upon the decreasing trend in the ISCCP datasets and suggested that it is due to regular changes in the satellite's view angle.

One final study to mention is El-Borie et al. (El-Borie, Thabet, El-Mallah, Abd El-Zaher, & Bishara, 2020), in which the relationships between GCRs, TSI, low-level clouds, and the global surface temperature is investigated. They supported the cosmoclimatology theory by showing that in the last past century, there was a 19% decrease in the GCR flux; and stated that this caused a decrease in the global cloud cover, hence explaining global warming. They also reported a strong negative correlation between the low cloud cover and global surface temperature. It was concluded that between GCRs and the temperature anomaly variation, there existed a relationship cycle with a period of about 20 years.

Reviewing the mentioned studies, it can be said that there is no definite statement that can be hypothesized regarding the cosmic rays – clouds – climate relationship; and that further research is required on this controversial relationship.

CHAPTER 4

DATA AND METHODS

This chapter talks about the datasets used, and the methodology for the data analysis is explained.

4.1 Data Used

Data for the cosmic ray flux has been taken from the Oulu Cosmic Ray Station (ONM), operated by the Sodankyla Geophysical Observatory (Usoskin, Cosmic Ray Station of the University of Oulu, 2021). The ONM is a ground-based neutron monitor, which measures the number of high-energy charged particles reaching the Earth. It is located in Finland at latitude 65° N. The ONM has been in operation for more than 50 years, with its data being publicly available (Usoskin, Mursula, Kangas, & Gvozdevsky, 2001). It is a particularly useful source of data on cosmic flux because it is one of the most stable neutron monitors with a long record of measurements.

As previously mentioned in Section 2.1, three different parameters have been used as solar activity proxies. The SSN data has been taken from the World Data Center SILSO, Royal Observatory of Belgium, Brussels (SILSO, 1984-2017). The TSI and UVI data have been taken from the NOAA Climate Data Record (CDR) created with the Naval Research Laboratory model for solar spectral (Coddington, et al., NOAA Climate Data Record (CDR) of Solar Spectral Irradiance (SSI), NRLSSI Version 2, 2021) and total solar irradiance (Coddington, et al., NOAA Climate Data Record (CDR) of Total Solar Irradiance (TSI), NRLTSI Version 2, 2021). The UVI data used has a wavelength of 249.5 nm, corresponding to the midpoint of the UV region (100 - 400 nm).

There are a couple of different indices that can be used for the ENSO events; the Oceanic Niño Index (ONI) has been used in this analysis (NOAA, Historical El Nino / La Nina Episodes, 2021). ONI is defined as the sea surface temperature anomaly index for the region between the coordinates 5°N to 5°S, 170°W to 120°W, which is also called the Niño 3.4. Monthly averages between the years 1984 to 2017 have been used.

Three different sources for cloud datasets have been used in the analysis. The first is data from the International Satellite Cloud Climatology Project (ISCCP), H-Series (Young, Knapp, Inamdar, Hankins, & Rossow, 2018). The ISCCP data used in the analysis covers the years 1984-2017. Total cloud amount and IR cloud layers (low cloud amount) have been used as variables from this project. The cloud data has a spatial resolution of 1 by 1 degree.

The second dataset is taken from the Moderate Resolution Imaging Spectroradiometer (MODIS) (LAADS DAAC, 2021). MODIS is an imaging sensor on board both NASA research satellites, Aqua and Terra. The combined Aqua and Terra MODIS Cloud Properties product is used, covering the years 2003-2017. The variables from this dataset used in the analysis are cloud mask fraction and cloud mask fraction low, and the spatial resolution of the products is 1 by 1 degree.

The final dataset for clouds is taken from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) (NASA GMAO, 2015). The MERRA-2 project is a NASA atmospheric reanalysis for the satellite era that uses the Goddard Earth Observing System Model, Version 5 (GEOS-5) with its Atmospheric Data Assimilation System (ADAS), version 5.12.4. This reanalysis data uses advances made in the assimilation system, replacing the original MERRA dataset. It is an enhanced dataset compared to the former two cloud datasets. The variables used are total cloud area fraction and cloud area fraction for low clouds, while the data covers 1984-2017. The spatial resolution of the data is 0.5 by 0.625 degrees.

GMTA data is retrieved from the Global Historical Climatology Network-Monthly (GHCN-M) dataset and International Comprehensive Ocean-Atmosphere Data Set (ICOADS). These are combined into a single product that shows the global land and ocean temperature anomalies together (NOAA, Global Mean Temperature Anomaly Timeseries, 2021). The anomalies are calculated with respect to the 20th-century average.

The AOD data is from the NOAA Climate Data Record (CDR) of AVHRR Daily and Monthly Aerosol Optical Thickness (AOT) over Global Oceans, Version 3.0 (Zhao, 2017). The product is the depth at 0.63 microns, retrieved from NOAA PATMOS-x level-2B orbital radiance products. The spatial resolution of the product is 0.1 by 0.1 degrees.

Finally, the Global Precipitation Climatology Project (GPCP) Climate Data Record (CDR), Version 2.3, has been used for the precipitation data (Adler, et al., 2016). The provided data is on an equal degree grid of spatial resolution 2.5 by 2.5.

All of the data mentioned above have been used as monthly averages in this thesis, and they are summarized in Table 4.1.

Table 4.1. Summary of the data used in this thesis

Data	Variable Used	Temporal Resolution	Spatial Resolution
GCR	Neutron monitor count rate	Monthly averaged (Jan 1984 - Dec 2016)	Globally averaged
SSN	Sunspot number	Monthly averaged (Jan 1984 - Dec 2016)	Globally averaged
TSI	Total solar irradiance	Monthly averaged (Jan 1984 - Dec 2016)	Globally averaged
UVI	UV irradiance ($\lambda = 249.5$ nm)	Monthly averaged (Jan 1984 - Dec 2016)	Globally averaged
ONI	Oceanic Niño Index	Monthly averaged (Jan 1984 - Dec 2016)	Regionally averaged
ISCCP Cloud	Total cloud amount Low cloud amount	Monthly averaged (Jan 1984 - Dec 2016)	1° x 1°
MODIS Cloud	Cloud mask fraction (total and low)	Monthly averaged (Jan 2003 - Dec 2016)	1° x 1°
MERRA2 Cloud	Cloud area fraction (total and low)	Monthly averaged (Jan 1984 - Dec 2016)	0.5° x 0.625°
GMTA	Global land and ocean temperature anomaly	Monthly averaged (Jan 1984 - Dec 2016)	Globally averaged
AOD	Aerosol optical depth at 0.63 microns	Monthly averaged (Jan 1984 - Dec 2016)	0.1° x 0.1°
PRECIP	Global precipitation	Monthly averaged (Jan 1984 - Dec 2016)	2.5° x 2.5°

4.2 Methodology

The parameters used in this analysis are separated as predictors and predictands. The predictors are independent variables used to predict the outcomes; the GCR flux, SSN, TSI, UVI, and ONI are used as predictors. On the other hand, the total cloud (TC), low-level cloud (LLC), GMTA, AOD, and PRECIP are the predictands, hence outcomes.

First, the obtained raw data mentioned in Section 4.1 is explored, and all datasets are pre-processed using a 12-month moving average to remove seasonal effects. The different cloud datasets are compared on temporal and spatial scales. Additionally, the ISCCP cloud data is detrended because there is a clear decreasing trend seen in the dataset, and this trend is claimed to be artificial (Usoskin et al. 2004, Laken et al. 2012, Singh et al. 2020). This difference between the detrended and raw data is shown in more detail in Section 5.1. Hence the detrended ISCCP time series is used in the analysis.

The analysis started by trying to obtain the same results as previous studies investigating the linear correlation between GCRs and clouds (Svensmark et al. 1997, Usoskin et al. 2004) and then extending those studies to the present. So, the reanalysis has been conducted for three different periods; 1984-1994 (which was the analysis period of Marsh and Svensmark 2000), 1984-2000 (which was the analysis period of Usoskin et al. 2004), and 1984-2017, which is the whole period of the analysis in this thesis. Next, the linear correlation coefficients between all the variables were calculated for the full analysis period of 1984-2017.

The analysis continued with regression analysis using multiple linear regression models consisting of all the predictors and predictands. In the last step, machine learning methods were introduced to identify the non-linear relationships. Random forests were used for multiple non-linear regression. R-squared values were calculated, and cross-validation methods were used to evaluate the performance of all models.

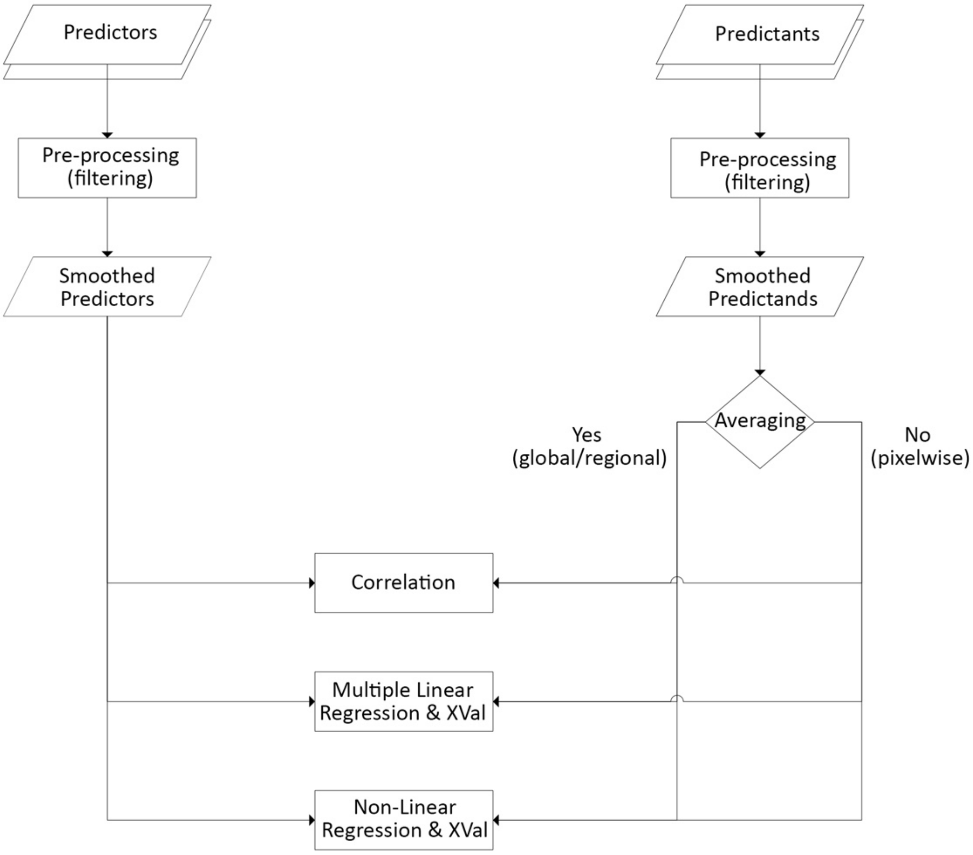


Figure 4.1. Methodology Flowchart

An overview of the methodology can be seen in the flowchart in Figure 4.1. XVal stands for the cross-validation that is used in the regression models. Validation methods are explained in detail in Section 2.6.4. The analysis in this thesis has used hold-out validation of 25%, meaning that 25% of the data has been used to test the ML models while 75% of the data has been used to train them. Holding out a certain percentage of the data can be done in two ways; first, data points can be randomly selected and held out from the dataset. Second, the data can be held out as a continuous block of data points. Such methods are investigated and compared in Section 5.4.1.

Also, both in the linear correlation and multiple regression steps, the predictand datasets are either used as global/regional averages or as pixelwise global data. The pixelwise analysis creates a different regression model for each pixel, and the results are shown as global maps. Meanwhile, there are four different latitude zones used as global or regional averages in the analysis, and they can be seen in Figure 4.2 and Table 4.1:

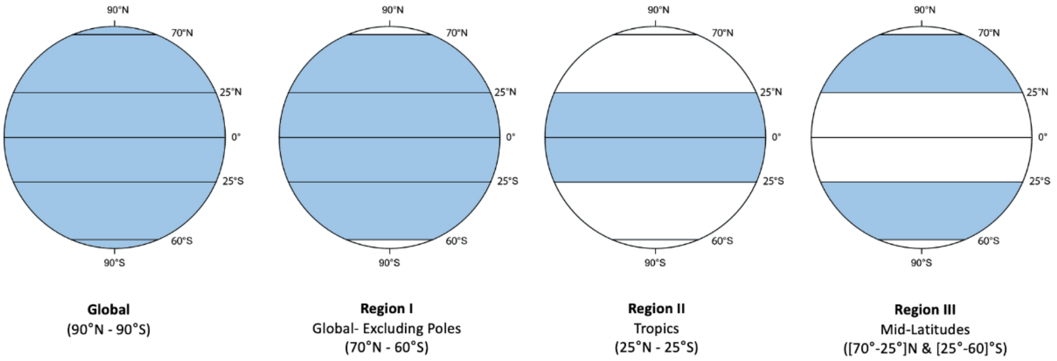


Figure 4.2. Latitudinal regions used in the analysis, graphically shown.

Table 4.2. Latitudinal regions used in the analysis

	Latitude Interval
Global	90°N - 90°S
Region I	70°N - 60°S
Region II	25°N - 25°S
Region III	(70° - 25°)N & (25° - 60°)S

The regions used in this thesis are similar to previous regional analyses ((Usoskin, Marsh, Kovaltsov, Mursula, & Gladysheva, 2004) and (Voiculescu, Usoskin, & Mursula, 2006)) and why they are used is explained in more detail in Chapter 3.

All steps of the data analysis have been conducted using the MATLAB® programming language, version R2020a. The add-on toolboxes "Statistics and Machine Leaning Toolbox" and the "Mapping Toolbox" have been used. The function *fitlm* was used to fit a linear regression model, *fitglm* was used to fit a generalized linear regression model, and *stepwiseglm* was used to fit a stepwise generalized linear model to the predictors and predictands. Additional "interaction" terms, which are the products of distinct predictors, were added to the LM and GLM regression models. Gaussian distribution was used in the GLM. The function *TreeBagger* was used to fit a random forest, non-linear model to the predictors and predictands.

Additionally, for the visualization of the linear correlation matrix between all of the parameters seen in Section 5.3.2, the programming language R has been used.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Preprocessing the Data

As mentioned previously, the first step of the analysis was to explore and preprocess the raw data. This was done by generating time series plots for each dataset and mapping the datasets available as globally gridded products.

First of all, the MODIS cloud data was compared to both the ISCCP and MERRA2 cloud products. Previous studies (Kristjánsson, et al., 2008) (Laken, Pallé Bagó, & Miyahara, 2012) have stated that the MODIS datasets are more accurate than the ISCCP products, but the MODIS data does not cover the entire time period of this analysis. So, this comparison was made to see how much the other cloud products are related to each other. The linear correlation coefficient (R) between the cloud products for January 2003 – December 2016 can be seen in Table 5.1.

Table 5.1. Linear correlation coefficients and p-values between MODIS cloud and the ISCCP- MERRA2 cloud products from 2003 to 2017

2003 - 2017	Global	Region I	Region II	Region III
MODIS vs ISCCP **TC**	0.63 p=0.0000	0.68 p=0.0000	0.72 p=0.0000	0.83 p=0.0000
MODIS vs ISCCP **LLC**	0.09 p=0.2223	-0.20 p=0.0109	-0.03 p=0.7279	-0.14 p=0.0688
MODIS vs MERRA2 **TC**	0.52 p=0.0000	0.68 p=0.0000	0.64 p=0.0000	0.60 p=0.0000
MODIS vs MERRA2 **LLC**	0.30 p=0.0001	0.56 p=0.0000	0.31 p=0.0000	0.60 p=0.0000

It can be seen from this table that, in all the defined regions, the MODIS dataset is much more correlated with the total cloud data for both cloud products. The *p*-value shows the statistical significance of the correlation coefficient value, *R*. If the *p*-value is less than 0.05, it means that the resulting *R* is statistically significant, while a *p*-value that is greater than 0.05 means it is not statistically significant. The results from Table 5.1 thus show that the correlation between MODIS and ISCCP LLC datasets is not statistically significant for all regions except Region I. Figure 5.1. shows the time series of all three cloud datasets plotted together to compare the changes in the temporal dimension.

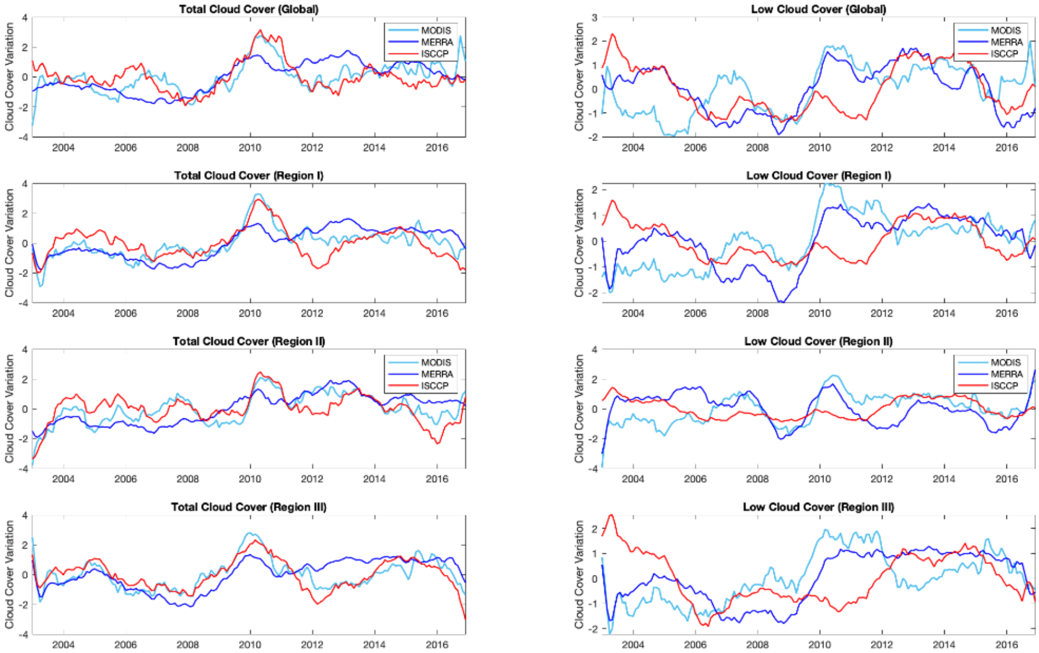


Figure 5.1. Time series of the globally and regionally averaged MODIS, ISCCP and MERRA2 cloud products for the years 2003 to 2017.

The light blue line is the MODIS cloud, the dark blue line is the MERRA2 cloud, and the red line is the ISCCP cloud. All datasets are 12-month moving averaged and normalized; the ISCCP dataset is also detrended. The detrending of ISCCP is explained further in the following paragraphs. In Figure 5.1, it can be seen that the total cloud cover for the MODIS, MERRA2, and ISCCP averaged over both global scale and regional scales are closely parallel with one another. Meanwhile, for the low cloud cover, the datasets largely differ from one another. However, comparing the three, MERRA2 and MODIS seem more correlated than the ISCCP LLC for low clouds.

Since the MODIS satellite data begins from 2003, it is also essential to check the previous years by comparing the MERRA2 and ISCCP datasets for the entire analysis period of 1984 to 2017. Table 5.2 shows the linear correlation coefficient between the globally and regionally averaged MERRA2 and ISCCP cloud datasets.

Table 5.2. Linear correlation coefficients and their p-values between MERRA2 cloud products and the ISCCP cloud products from 1984 to 2017

1984 – 2017	**Global**	**Region I**	**Region II**	**Region III**
MERRA2 vs ISCCP **TC**	0.13 p=0.0077	0.24 p=0.0000	0.45 p=0.0000	0.07 p=0.1932
MERRA2 vs ISCCP **LLC**	0.06 p=0.2341	0.12 p=0.0161	-0.11 p=0.0368	0.15 p=0.0021

It is seen that the MERRA2 and ISCCP cloud datasets are not strongly correlated in any of the regions. The total cloud cover has a relatively higher correlation coefficient than the low cloud cover; meanwhile, the two total cloud datasets are more correlated in Region II, the tropics. On the other hand, the low cloud cover has a negative correlation coefficient in Region II and a minimal positive correlation in the remaining regions. The weak relationship between the two cloud products may also be due to the low accuracy of the cloud data. The comparisons can be seen better in the temporal dimension in Figure 5.2.

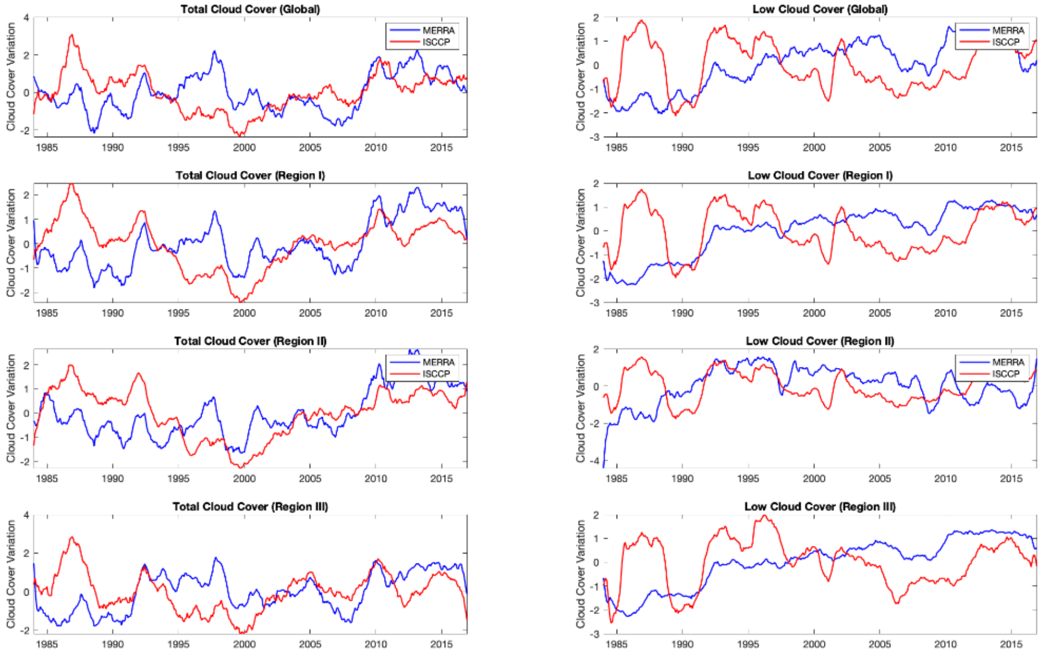


Figure 5.2. Time series of the globally and regionally averaged ISCCP and MERRA2 cloud products for the years 1984 to 2017.

From the temporal plots in Figure 5.2, it is seen that starting from around ~2003 up to 2017, trends in total cloud cover are similar between ISCCP and MERRA2 datasets, as also observed in Figure 5.1. However, there are considerable differences between 1984 to 2003; the two datasets act almost opposite to each other. As for the low cloud cover, ISCCP and MERRA2 datasets do not seem to have the same trends throughout the entire 33 years. This shows that a long time period is vital for such analyses, and using a short time frame may cause a spurious correlation.

The differences between the datasets could be due to the MERRA2 dataset being reanalysis data. Other possibilities could be because of the difference in the instrumentation used in the satellites for low cloud determination, and also due to the calibration problems of ISCCP, which is mentioned in some previous studies such as Laken et al. (Laken, Pallé Bagó, & Miyahara, 2012) in more detail in Chapter 3. Nevertheless, these differences are not enough of a reason to eliminate one of the two cloud products from the analysis. Hence the analysis will continue with both the ISCCP and MERRA2 TC and LLC products.

The time series for each of the raw datasets were plotted together with their smoothed versions. The time series plots for globally averaged data can be seen in Figure 5.3. A 12-month moving average is used to filter the data to remove the seasonal effects. The thin blue line is the raw monthly data, and the thicker red line is the 12-month moving averaged data.

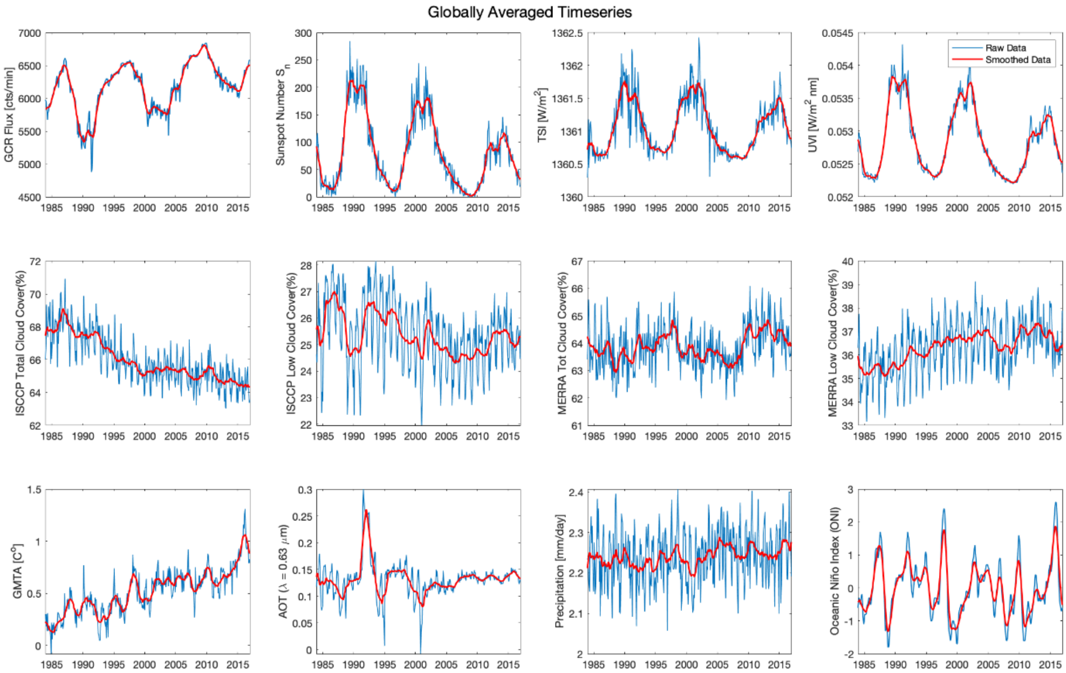


Figure 5.3. Globally averaged time series of the datasets for the years 1984 to 2017.

Looking at the figures, the decreasing trend in the ISCCP total cloud data can be seen more clearly. This trend has previously been discussed in some research ((Kristjánsson, et al., 2008), (Laken, Pallé Bagó, & Miyahara, 2012)) and is thought to be an artificial effect. That is why the ISCCP data has been detrended on top of the smoothing filter. The differences in the time series of the raw, moving averaged, and detrended ISCCP data in temporal scale can be seen better in Figure 5.4.

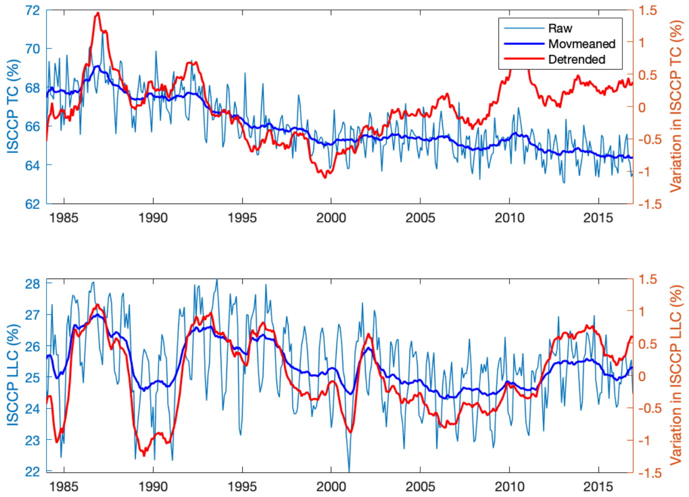


Figure 5.4. Comparison of the globally averaged raw, smoothed and detrended ISCCP time series for the years 1984 to 2017.

In the figure, the thin, light blue line is the raw ISCCP data, the dark blue line is the 12-month moving averaged (only smoothed) data, and the red line is the detrended data. The upper plot is the ISCCP total cloud cover, and the lower plot is the ISCCP low cloud cover. There is not much difference between the smoothed and detrended ISCCP low level cloud cover. However, detrending the ISCCP total cloud cover creates a significant difference. Hence, when re-analyzing the previous and for the remaining parts of the analysis, the detrended ISCCP total cloud and low cloud data have been used.

5.2 Reanalysis of Previous Studies

In this section, the linear correlation coefficient between GCR flux and cloud amount was calculated for three different periods; 1984-1994 (which was the analysis period of Marsh and Svensmark 2000), 1984-2000 (which was the analysis period of Usoskin et al. 2004), and 1984-2017, which is the entire period of the analysis in this thesis. The cloud data was averaged as in the mentioned regions in Section 4.2.

Tables 5.3 and 5.4 show the correlation coefficients and the p-values between GCR flux and ISCCP cloud cover for each period.

Table 5.3. Correlation coefficients and p-values between the smoothed GCR flux and detrended ISCCP total cloud amount

GCR-TC	Global	Region I	Region II	Region III
1984 - 1994	0.33 p=0.0002	0.52 p=0.0000	0.17 p=0.0604	0.73 p=0.0000
1984 - 2000	-0.04 p=0.5497	0.10 p=0.1505	-0.18 p=0.0068	0.42 p=0.0000
1984 - 2017	0.04 p=0.3486	0.08 p=0.0689	0.01 p=0.8718	0.20 p=0.0000

Table 5.4. Correlation coefficients and their p-values between the smoothed GCR flux and detrended ISCCP low cloud amount

GCR-LLC	Global	Region I	Region II	Region III
1984 - 1994	0.84 p=0.0000	0.84 p=0.0000	0.82 p=0.0000	0.84 p=0.0000
1984 - 2000	0.61 p=0.0000	0.60 p=0.0000	0.34 p=0.0000	0.65 p=0.0000
1984 - 2017	0.27 p=0.0000	0.23 p=0.0000	0.09 p=0.1087	0.27 p=0.0000

It is seen that just like Marsh and Svensmark suggested in 2000, the low cloud amount is highly correlated with GCRs from 1984 to 1994. Both globally and for the average cloud amounts in all regions, the correlation coefficient *R* is greater than *0.8*, which is a significantly large value. The total cloud amount does not have the same high correlations for all regions in the same period. Only region IV is highly correlated.

However, when we expand the analysis time period even by six years, there is a significant drop in the correlation. Low cloud amount is still significantly correlated in most regions, just as Usoskin et al. (Usoskin, Marsh, Kovaltsov, Mursula, & Gladysheva, 2004) stated in 2004, but the total cloud amount is not.

When the time period is expanded even further until 2017, it is seen that there is a huge drop in the *R* values for all regions and that the p-value greatly increases for all regions of the ISCCP TC and Region II of the ISCCP LC. Plotting the full period of cloud data against the GCR flux shows the temporal differences clearly in Figure 5.5. The blue line is the ISCCP cloud amount, and the red dashed line is the GCR flux. From 1984 to almost 1994, the GCR flux and the low cloud cover are highly correlated in all regions. For the total cloud cover, there is a correlation in most regions from 1984 to around 1990. However, the correlation is lost entirely after ~1990.

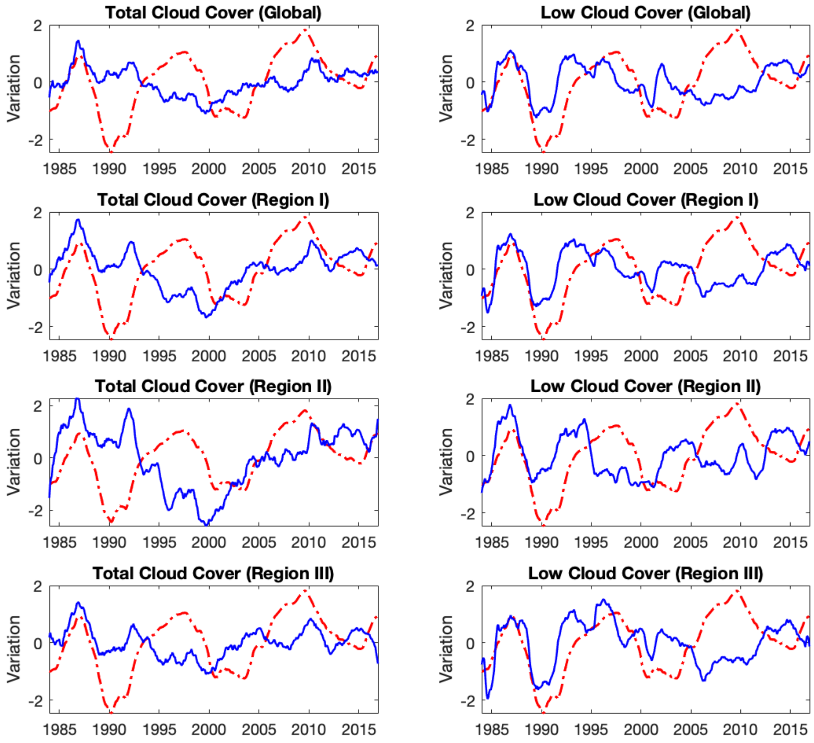


Figure 5.5. Time series of the globally and regionally averaged ISCCP cloud products compared to the smoothed GCR flux for the years 1984 to 2017.

These results show how important it is to conduct such studies on longer time scales before concluding with significant results. Even though the GCR flux and ISCCP low cloud amount appear highly correlated from 1983 to 1994, this correlation weakens after 1994 and almost completely disappears when the more recent years are included in the analysis.

Tables 5.5 and 5.6 show the correlation for the same periods between GCR flux and MERRA2 cloud datasets instead.

Table 5.5. Correlation coefficients and their p-values between the smoothed GCR flux and smoothed MERRA2 total cloud amount

GCR-TC	Global	Region I	Region II	Region III
1984 - 1994	0.21 p=0.0185	0.26 p=0.0049	0.22 p=0.0148	0.10 p=0.2596
1984 - 2000	0.49 p=0.0000	0.46 p=0.0000	0.15 p=0.0338	0.45 p=0.0000
1984 - 2017	0.31 p=0.0000	0.35 p=0.0000	0.31 p=0.0000	0.31 p=0.0000

Table 5.6. Correlation coefficients and their p-values between the smoothed GCR flux and smoothed MERRA2 low cloud amount

GCR-LLC	Global	Region I	Region II	Region III
1984 - 1994	-0.10 p=0.2587	-0.18 p=0.0492	-0.27 p=0.0025	-0.15 p=0.0967
1984 - 2000	0.46 p=0.0000	0.37 p=0.0000	0.18 p=0.0122	0.39 p=0.0000
1984 - 2017	0.26 p=0.0000	0.30 p=0.0000	-0.01 p=0.5759	0.33 p=0.0000

This time, the almost opposite can be observed with the MERRA2 data. The total cloud cover is comparably more correlated with GCR flux than the low cloud cover for all regions. Another opposite feature is that the correlation between clouds and GCRs increases as the time span is expanded. Looking at the entire analysis period between 1984 to 2017, region II is the less correlated region with less statistical significance, while region I and region III are very close in the correlation coefficient value. The relationships can be observed better in the temporal dimension in Figure 5.6.

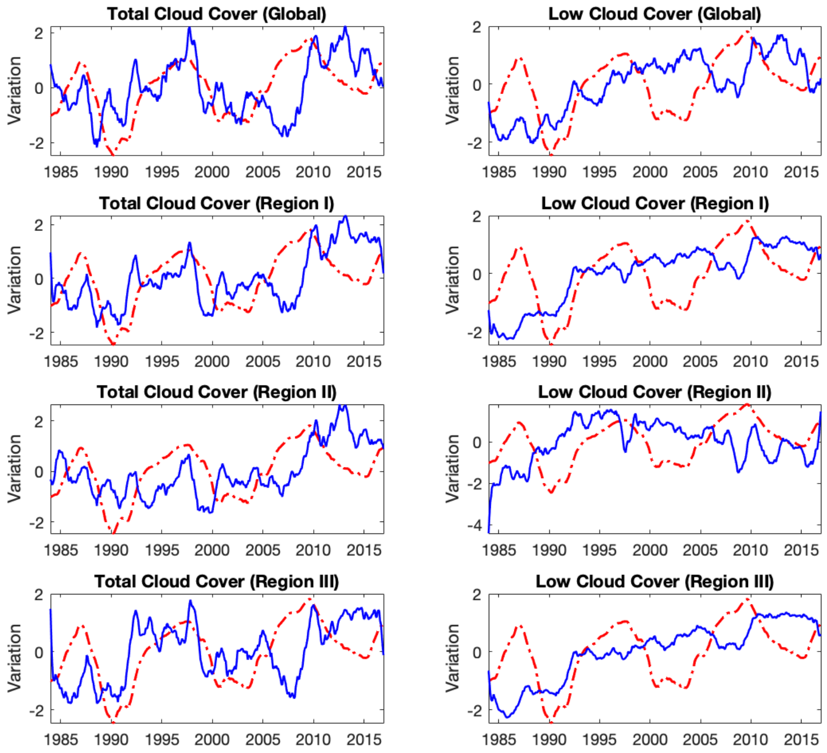


Figure 5.6. Time series of the globally and regionally averaged MERRA2 cloud products compared to the smoothed GCR flux for the years 1984 to 2017.

The globally averaged MERRA2 total cloud cover follows similar trends as GCR, both increase and decrease at almost the same time. However, the globally averaged low cloud cover and GCR flux act oppositely. As for Region III, which is expected to be the most correlated from previous literature, the low cloud cover and GCR seem parallel. Overall, MERRA2 TC and LLC are more correlated to the GCR flux compared to the ISCCP TC and LLC for the full period of 1984-2017.

The seemingly high correlations observed between certain variables for specific time periods and regions that cannot be observed for others can be interpreted differently. Firstly, they can be totally accidental. Secondly, there might be other parameters that change over time and space, and when they are not considered, we can observe the correlations only during certain times or at certain places. It is also possible that the relations are not linear, and we can only observe them when they are more or less linear under certain conditions. The noise might not be independent either.

These results show that it is tough to interpret the cosmic ray-cloud relationship based on regionally or globally averaged datasets, and it would be better to analyze the globe pixel by pixel to explore subtler correlations and to reveal location-dependent correlations that may disappear when averaged out over the globe or regions. The following section shows this pixelwise analysis in global maps.

5.3 Linear Correlation

This section shows and discusses the linear correlation with all predictors and the predictands on a regional and global scale. The globally averaged predictors are investigated with the ISCCP TC, ISCCP LLC, MERRA2 TC, and MERRA2 LLC for each pixel. The predictands GMTA, AOD, and PRECIP are investigated in regional averages.

5.3.1 Pixelwise Correlation Results

The results of the pixelwise analysis of the linear correlation between each cloud data set and the predictors are presented here in Figures 5.7 – 5.30. For each pixel, an independent Pearson correlation coefficient is calculated between the time signals.

5.3.1.1 GCR vs. cloud cover

Figures 5.7 – 5.10 show the global maps of the linear correlation coefficient between the globally averaged GCR flux and the pixelwise ISCCP and MERRA2 TC and LLC cloud products for each pixel.

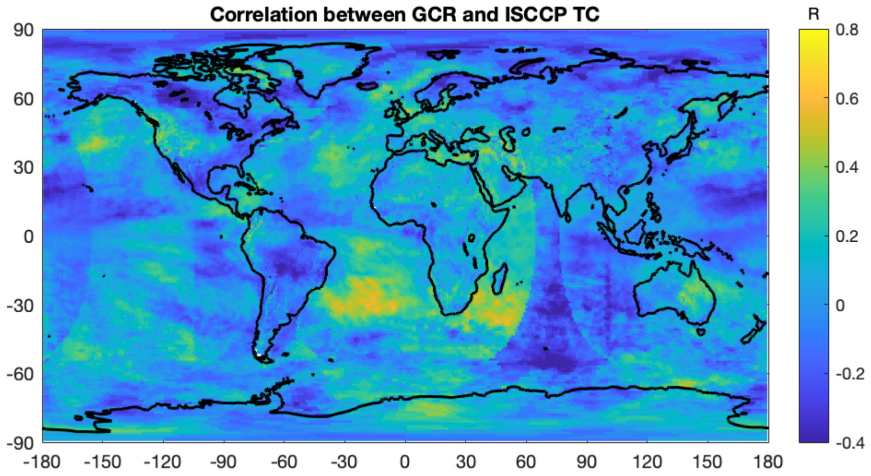


Figure 5.7. Map of the linear correlation coefficient between the 12-month moving averaged cosmic ray flux and detrended ISCCP total cloud cover for the years 1984-2017.

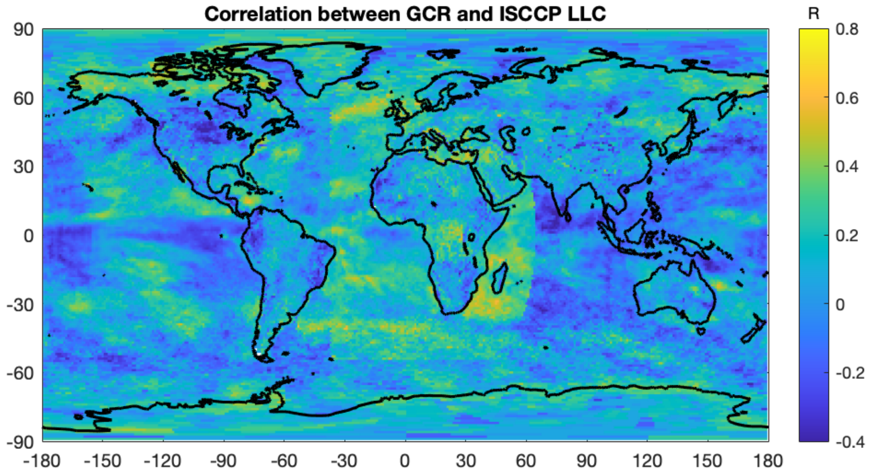


Figure 5.8. Map of the linear correlation coefficient between the 12-month moving averaged cosmic ray flux and detrended ISCCP low level cloud cover for the years 1984-2017.

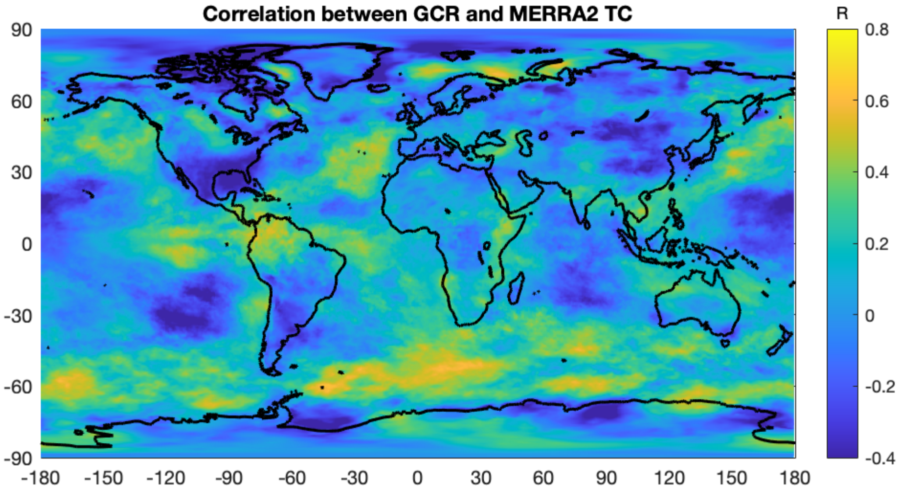


Figure 5.9. Map of the linear correlation coefficient between the 12-month moving averaged cosmic ray flux and MERRA2 total cloud cover for the years 1984-2017.

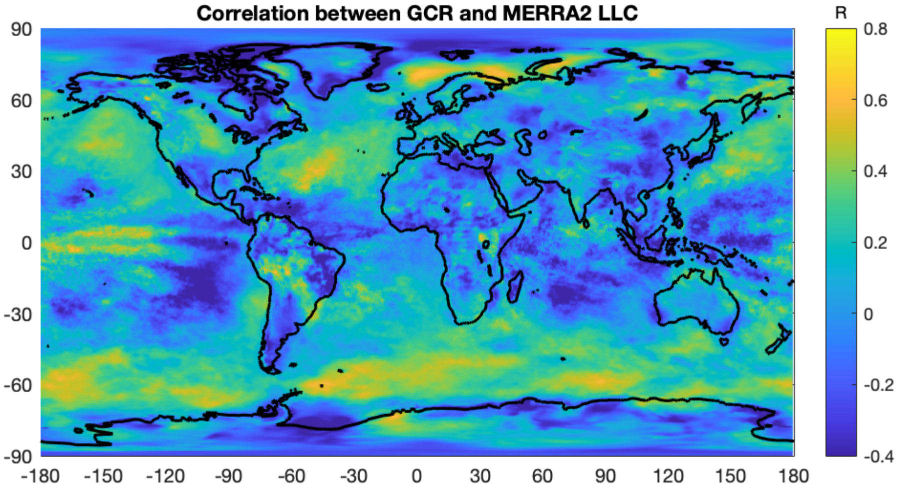


Figure 5.10. Map of the linear correlation coefficient between the 12-month moving averaged cosmic ray flux and MERRA2 low cloud cover for the years 1984-2017.

First of all, the satellite footprints mentioned in Laken et al. (Laken B. A., Pallé Bagó, Čalogović, & Dunne, 2012), can be seen in the maps using ISCCP data, especially over the Indian Ocean and slightly over the Pacific Ocean. There is also missing data in the ISCCP LLC dataset, seen as the solid indigo blue color over Asia and Antarctica in Figure 5.8.

The regions with a high positive correlation, such as the southern mid-latitudes over the Atlantic, Pacific, and Indian Oceans, can be seen in the ISCCP TC and ISCCP LLC maps. Similarly, the regions with negative correlations are parallel in both maps.

Comparing ISCCP maps with MERRA2 maps, it is seen that the correlated regions again are parallel for both TC and LLC; however, there are more pixels and areas in the MERRA2 data that have a high correlation. The region with a high correlation in the southern mid-latitudes extends over all longitudes for the MERRA2 TC and LLC data. Also, new areas with high correlation can be seen in the tropics and northern mid-latitude regions.

From these maps, it is observed that the linear correlation between GCR and both cloud products varies from each other in different regions. Even in the same latitude zones, there are both highly positive and negative correlation coefficients. Hence, taking global or regional averages lead to low *R* values.

5.3.1.2 SSN vs. cloud cover

Figures 5.11 – 5.14 show the global maps of the linear correlation coefficient between the globally averaged SSN and the pixelwise ISCCP and MERRA2 TC and LLC cloud products for each pixel.

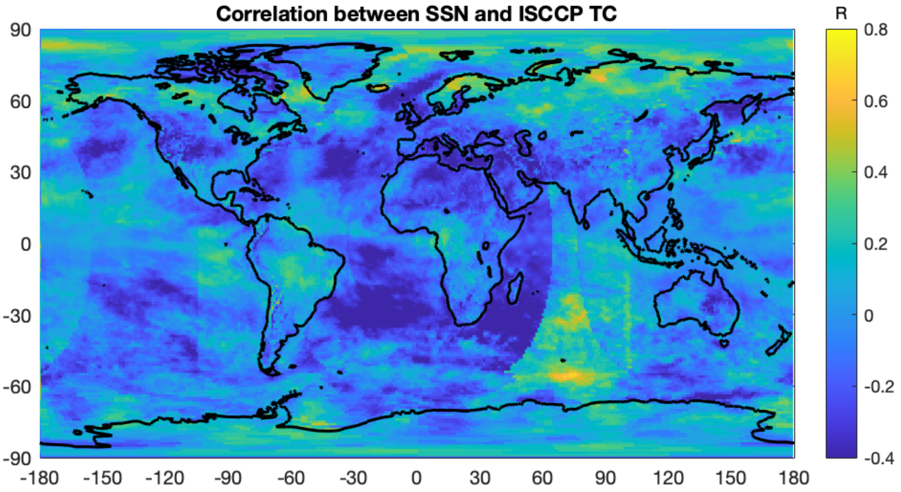


Figure 5.11. Map of the linear correlation coefficient between the 12-month moving averaged sunspot number and detrended ISCCP total cloud cover for the years 1984-2017.

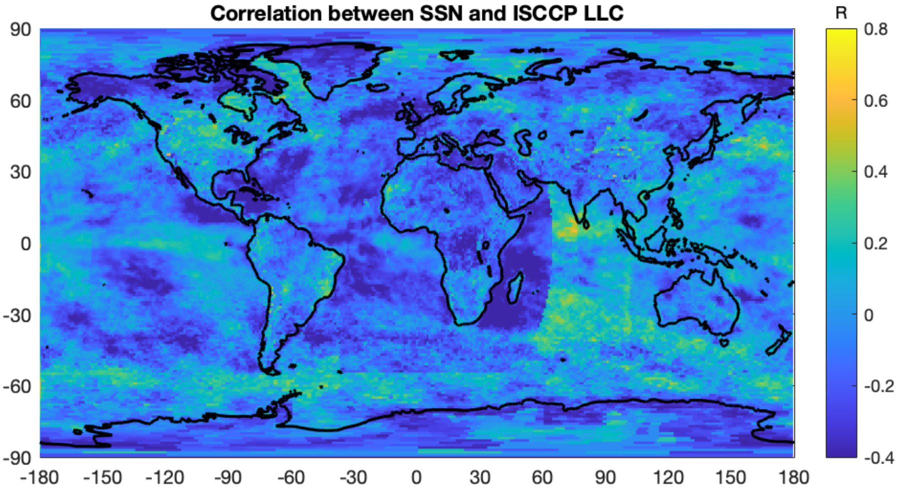


Figure 5.12. Map of the linear correlation coefficient between the 12-month moving averaged sunspot number and detrended ISCCP low cloud cover for the years 1984-2017.

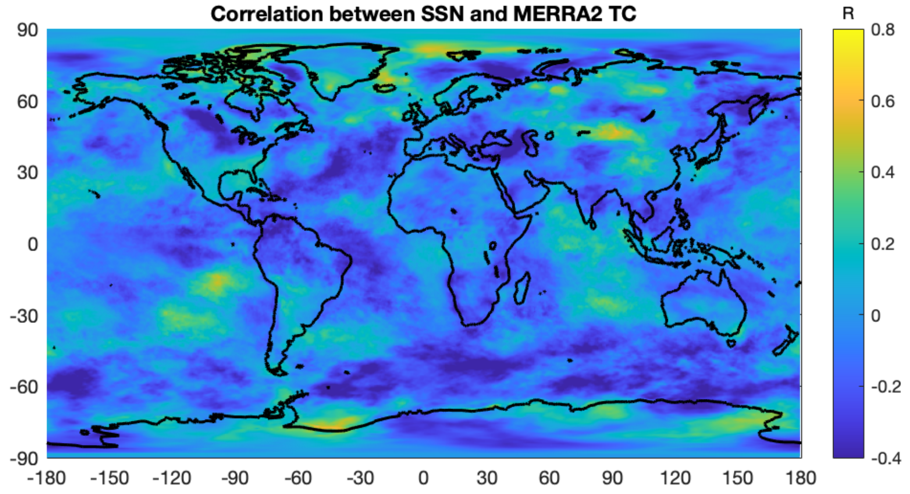


Figure 5.13. Map of the linear correlation coefficient between the 12-month moving averaged sunspot number and MERRA2 total cloud cover for the years 1984-2017.

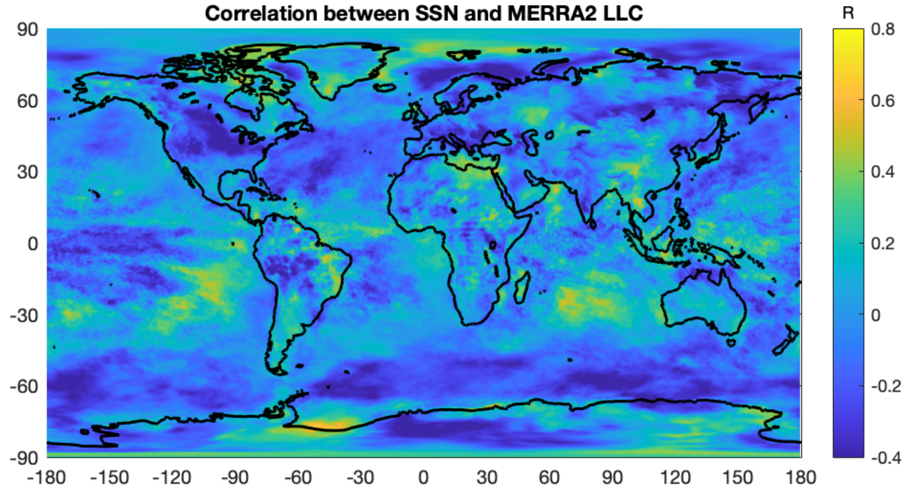


Figure 5.14. Map of the linear correlation coefficient between the 12-month moving averaged sunspot number and MERRA2 low cloud cover for the years 1984-2017.

From the ISCCP TC and LLC correlation maps with SSN, it is seen that the regions with negative correlation coefficients are the same as the positively correlated regions in the ISCCP correlation maps with GCR. Similarly, the MERRA2 maps with the correlation between GCR and SSN oppose one another for both TC and LLC. These results for both cloud products once again confirm the anti-correlation between GCR and SSN.

The TC maps for both cloud products appear to have slightly more pixels with a positive correlation compared to the LLC maps. The inverse is valid for the correlation maps between GCR and the cloud products. This could mean that SSN is less correlated with LLC while GCR is more correlated with LLC, supporting the cosmic ray – low cloud cover relationship.

5.3.1.3 TSI vs. cloud cover

Figures 5.15 – 5.18 show the global maps of the linear correlation coefficient between the globally averaged TSI and the pixelwise ISCCP and MERRA2 TC and LLC cloud products for each pixel.

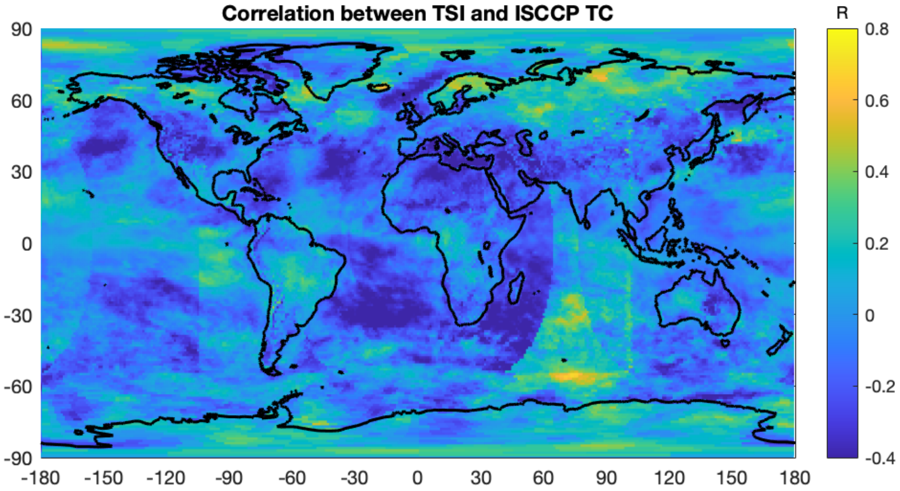


Figure 5.15. Map of the linear correlation coefficient between the 12-month moving averaged total solar irradiance and detrended ISCCP total cloud cover for the years 1984-2017.

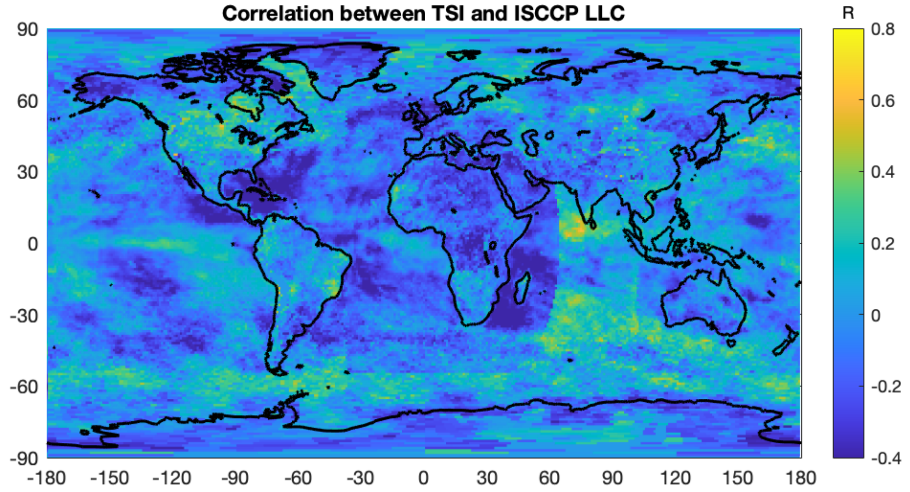


Figure 5.16. Map of the linear correlation coefficient between the 12-month moving averaged total solar irradiance and detrended ISCCP low cloud cover for the years 1984-2017.

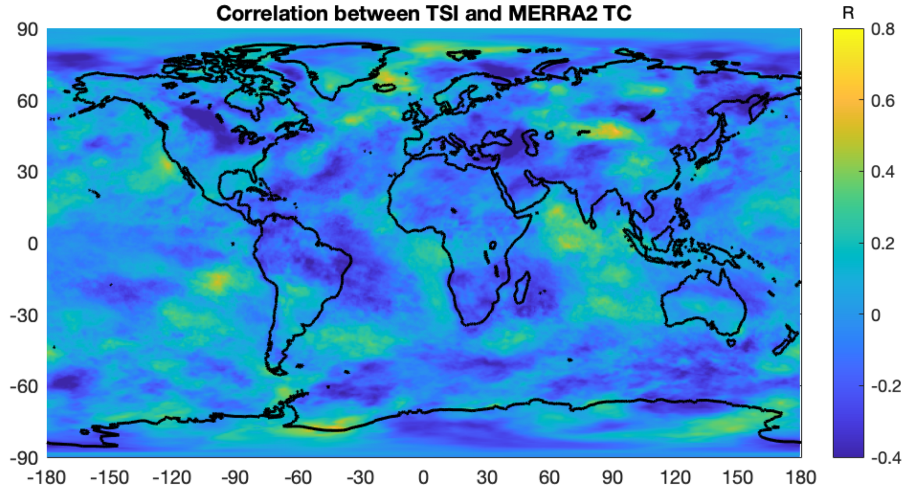


Figure 5.17. Map of the linear correlation coefficient between the 12-month moving averaged total solar irradiance and MERRA2 total cloud cover for the years 1984-2017.

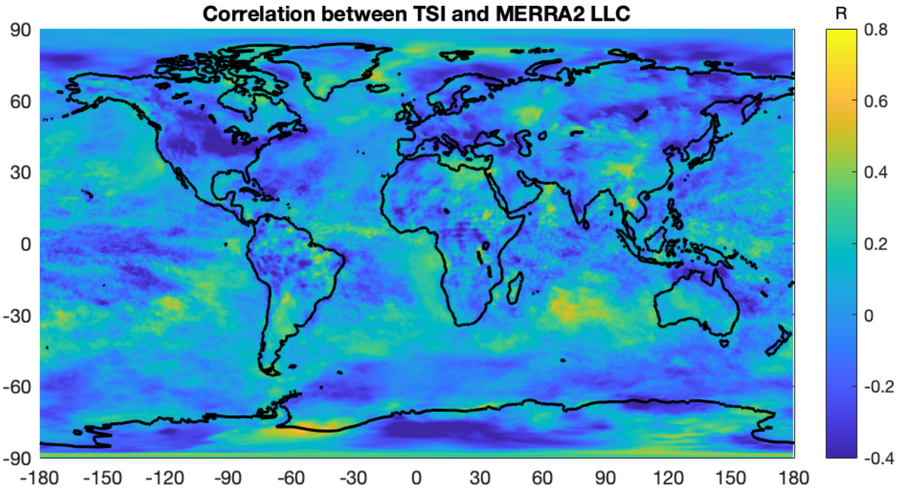


Figure 5.18. Map of the linear correlation coefficient between the 12-month moving averaged total solar irradiance and MERRA2 low cloud cover for the years 1984-2017.

Comparing the correlation maps of the cloud products with SSN and TSI, the maps appear almost the same. However, when the maps are further investigated, it can be seen that the maps with TSI for both cloud products have more pixels with higher positive correlation, in other words, more yellow pixels in more expansive areas, especially in the Indian Ocean and the Pacific Ocean.

5.3.1.4 UVI vs. cloud cover

Figures 5.19 – 5.22 show the global maps of the linear correlation coefficient between the globally averaged SSN and the pixelwise ISCCP and MERRA2 TC and LLC cloud products for each pixel.

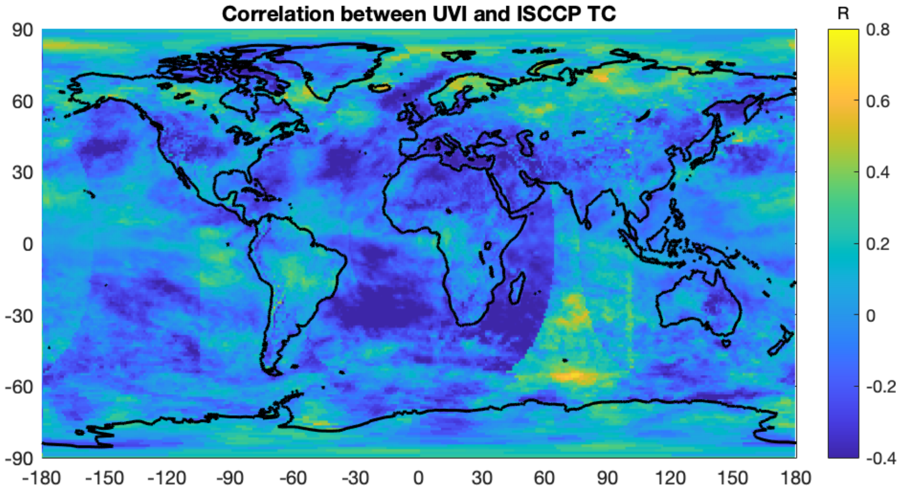


Figure 5.19. Map of the linear correlation coefficient between the 12-month moving averaged UV irradiance and detrended ISCCP total cloud cover for the years 1984-2017.

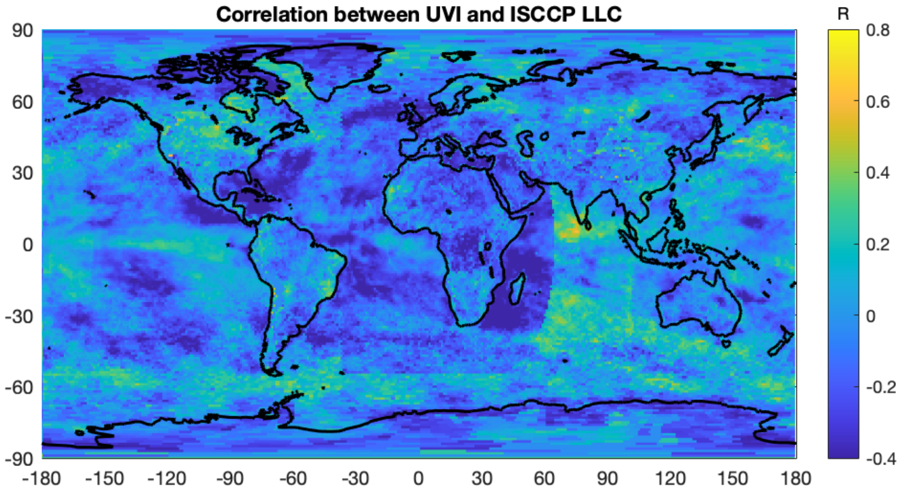


Figure 5.20. Map of the linear correlation coefficient between the 12-month moving averaged UV irradiance and detrended ISCCP low cloud cover for the years 1984-2017.

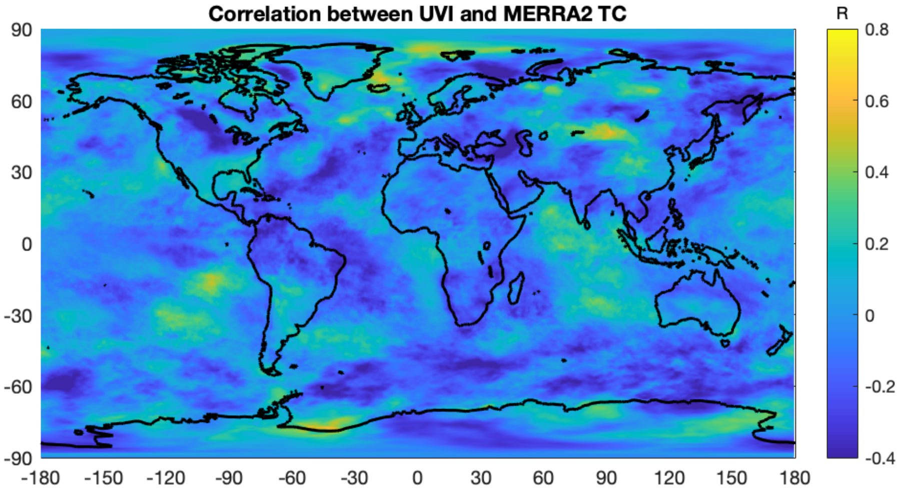


Figure 5.21. Map of the linear correlation coefficient between the 12-month moving averaged UV irradiance and MERRA2 total cloud cover for the years 1984-2017.

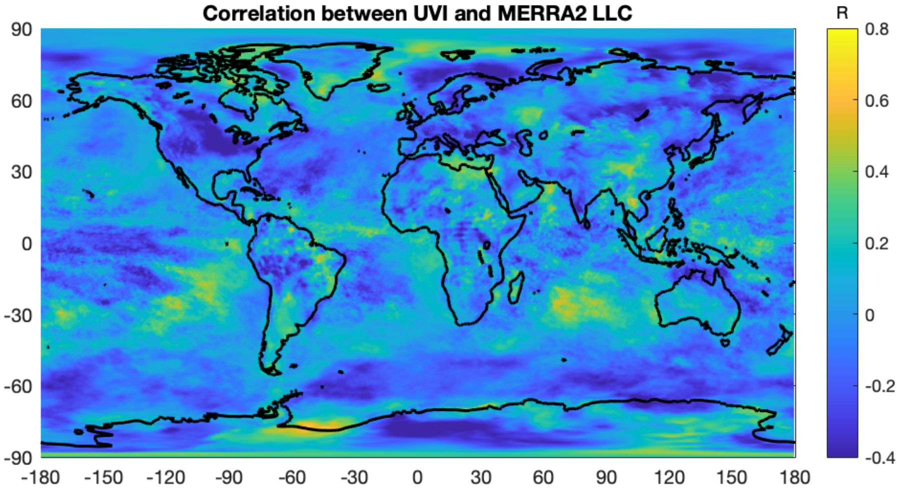


Figure 5.22. Map of the linear correlation coefficient between the 12-month moving averaged UV irradiance and MERRA2 low cloud cover for the years 1984-2017.

UVI is part of the SSI spectrum, which is the TSI measured as a function of wavelength, as mentioned in Section 2.1. Thus, the maps with UVI are expected to be significantly similar to the maps with TSI, and as assumed, the differences between maps of the two parameters are two subtle to detect by comparing maps.

5.3.1.5 ONI vs. cloud cover

Figures 5.23 – 5.26 show maps of the correlation coefficient between the globally averaged ONI and ISCCP and MERRA2 TC and LLC cloud products for each pixel.

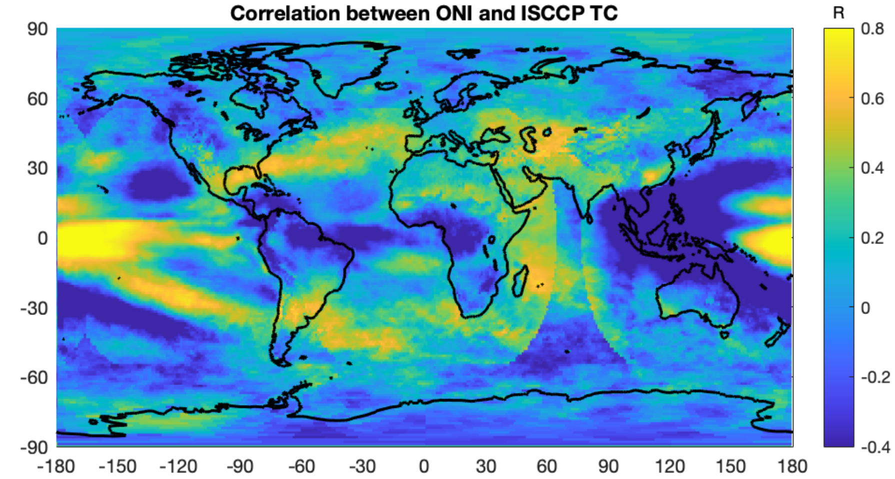


Figure 5.23. Map of the linear correlation coefficient between the 12-month moving averaged Oceanic Niño Index and the detrended ISCCP total cloud cover for the years 1984-2017.

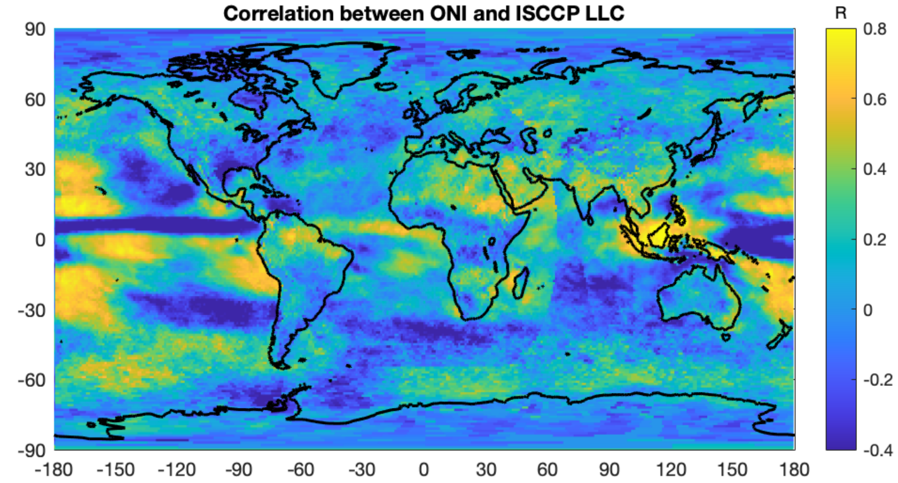


Figure 5.24. Map of the linear correlation coefficient between the 12-month moving averaged Oceanic Niño Index and the detrended ISCCP low cloud cover for the years 1984-2017.

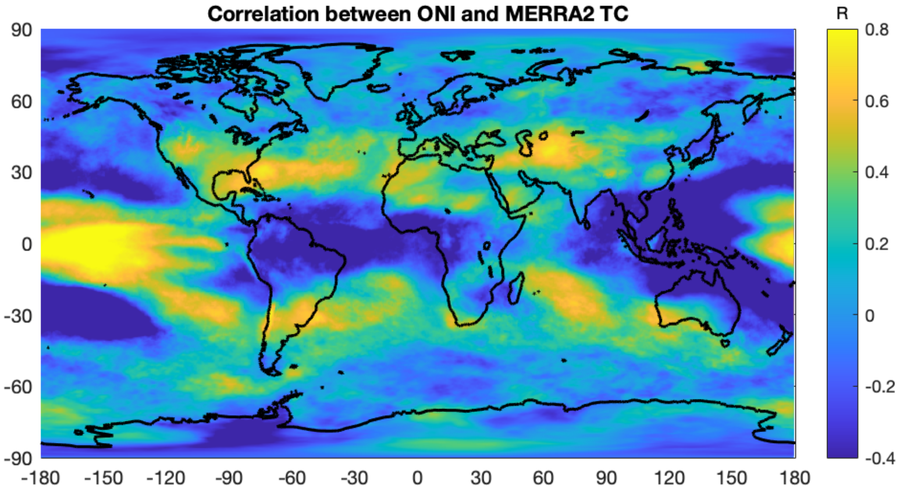


Figure 5.25. Map of the linear correlation coefficient between the 12-month moving averaged Oceanic Niño Index and the MERRA2 total cloud cover for the years 1984-2017.

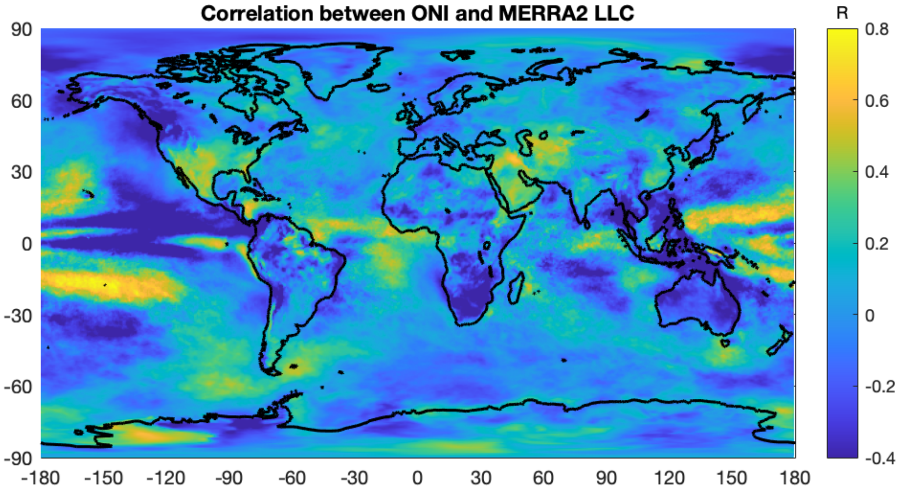


Figure 5.26. Map of the linear correlation coefficient between the 12-month moving averaged Oceanic Niño Index and the MERRA2 low cloud cover for the years 1984-2017.

The ONI maps seen in Figures 5.23 – 5.26 significantly diverge from the previous maps (Figures 5.7 - 5.22), which were produced either with proxies of solar activity (such as SSN, TSI, and UVI) or regulated by solar activity (such as GCR flux), which is an external forcer on climate. On the other hand, ONI is an index of ENSO, which is an internal event, and thus it is expected to have some different correlation pattern.

The maps using the MERRA2 cloud product and ISCCP cloud are very similar to each other. The positive correlation regions are parallel, yet the MERRA2 TC and LLC maps have more pixels with higher values of correlation coefficients. These higher correlation coefficient regions are seen in the tropics and mid-latitudes of the Pacific and Indian Oceans.

The TC and LLC maps have differences in the correlated regions for both cloud products. These differences are most evident in the equatorial region over the Pacific Ocean. TC maps show a strong positive correlation, while the LLC maps show a strong negative correlation over the equator in the Pacific. This could be because of the calibration problems of satellites, discussed in Marsh and Svensmark's 2003 paper (Marsh & Svensmark, 2003). However, it is peculiar that a similar pattern of negative correlation appears both in the ISCCP LLC map and the MERRA2 LLC.

This region is also the same region mentioned in Voiculescu et al. (Voiculescu, Usoskin, & Mursula, 2007), where it was also stated that there were odd correlations in the Pacific region, acting in opposite nature of the rest of the globe. They highlighted that the ENSO effect predominated over the solar effect in the equatorial Pacific and altered the results. Hence, it can be said that further study is required in this particular Pacific region.

5.3.2 Globally and Regionally Averaged Correlation Results

To emphasize how averaging regions once again can affect the results, here in Table 5.7, the pixels are averaged for each map in the above figures.

Table 5.7. Correlation coefficients and their p-values between the smoothed predictors and smoothed global averages of different cloud cover datasets for the years 1984-2017

	GCR	**SSN**	**TSI**	**UVI**	**ONI**
ISCCP **TC**	0.03 p=0.4950	-0.11 p=0.0320	-0.13 p=0.0088	-0.10 p=0.0462	0.36 p=0.0000
ISCCP **LLC**	0.25 p=0.0000	-0.22 p=0.0000	-0.11 p=0.0283	-0.18 p=0.0003	0.30 p=0.0000
MERRA2 **TC**	0.31 p=0.0000	-0.22 p=0.0000	-0.03 p=0.4487	-0.15 p=0.0023	0.24 p=0.0000
MERRA2 **LLC**	0.27 p=0.0000	-0.05 p=0.3629	0.14 p=0.0046	0.03 p=0.5446	-0.04 p=0.4031

The first thing that could be inferred from Table 5.7 is that globally averaged MERRA2 LLC is the cloud product with the lowest statistical significance in linear correlation, with the predictors SSN, UVI, and ONI. This means that using the globally averaged MERRA2 LLC is not very reliable in linear analysis with SSN, UVI, and ONI. The same could be said about the relationship between ISCCP TC and GCR; and MERRA2 TC and TSI. Another important relation to mention is the predictor ONI. It has the high correlation coefficient values with all globally averaged cloud products, except MERRA2 LLC, which is not a statistically significant result.

This table shows that averaging the pixels into one mean value can mislead into thinking that the parameters are not correlated at all when there are highly positive and negative pixels in certain regions. However, even if the R values are not quantitatively close, they are not wrong in qualitatively showing positive or negative correlations between parameters.

Thus, as a next step, the linear correlation between all the parameters for each geographical region was calculated as a correlation matrix. Figure 5.27 shows the correlation matrix for the parameters globally averaged and Figure 5.28 for the parameters averaged over Region III (the mid-latitudes) because that was the region with the highest overall statistically significant correlation coefficient, in Section 5.2.

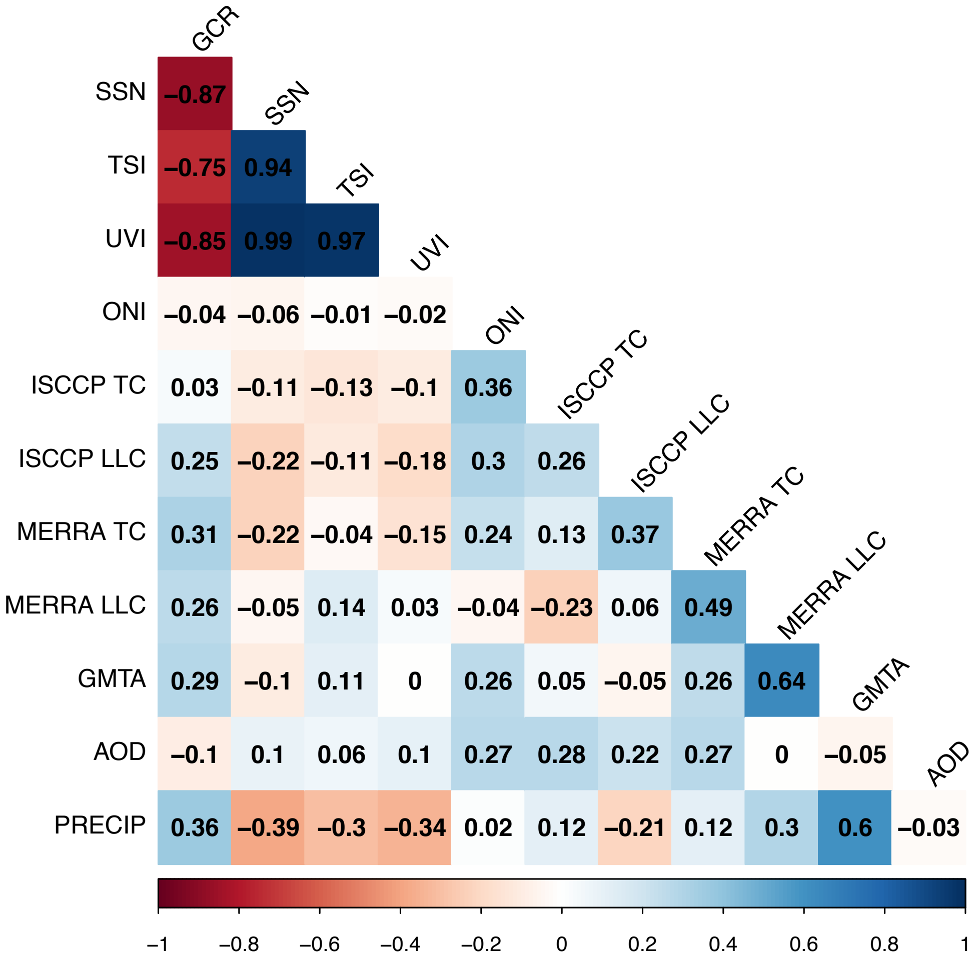
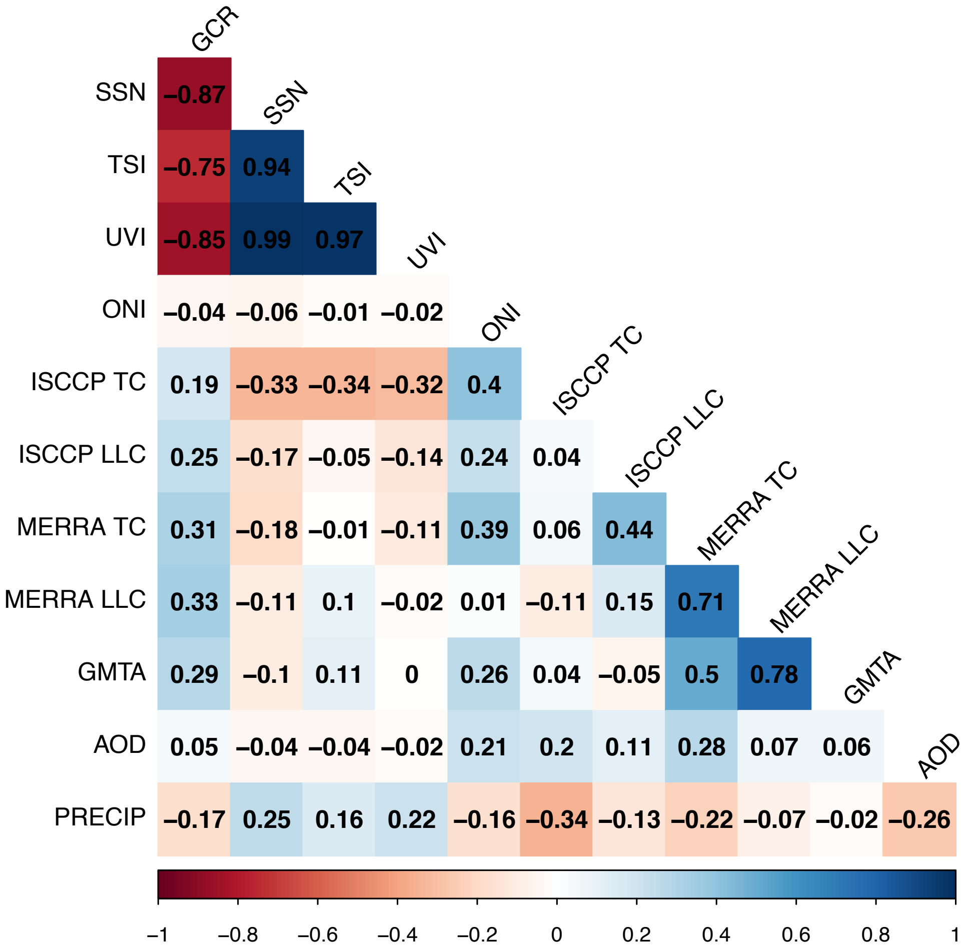


Figure 5.27. Linear correlation coefficients between the globally averaged and smoothed variables.



	GCR	SSN	TSI	UVI	ONI	ISCCP TC	ISCCP LLC	MERRA TC	MERRA LLC	GMTA	AOD
SSN	−0.87										
TSI	−0.75	0.94									
UVI	−0.85	0.99	0.97								
ONI	−0.04	−0.06	−0.01	−0.02							
ISCCP TC	0.19	−0.33	−0.34	−0.32	0.4						
ISCCP LLC	0.25	−0.17	−0.05	−0.14	0.24	0.04					
MERRA TC	0.31	−0.18	−0.01	−0.11	0.39	0.06	0.44				
MERRA LLC	0.33	−0.11	0.1	−0.02	0.01	−0.11	0.15	0.71			
GMTA	0.29	−0.1	0.11	0	0.26	0.04	−0.05	0.5	0.78		
AOD	0.05	−0.04	−0.04	−0.02	0.21	0.2	0.11	0.28	0.07	0.06	
PRECIP	−0.17	0.25	0.16	0.22	−0.16	−0.34	−0.13	−0.22	−0.07	−0.02	−0.26

Figure 5.28. Linear correlation coefficients between the smoothed variables averaged over Region III.

The predictors and predictands in Figure 5.27. and 5.28. are 12-month moving averaged, ISCCP data is also detrended. All of the inter-relationships between parameters can be seen in these correlation matrices. GCR is strongly negatively correlated with SSN, TSI, and UVI, while SSN, TSI, and UVI are strongly correlated. We also explore in further sections whether or not all predictands have novelty. Other notable correlations are the negative correlation between ISCCP TC and SSN, TSI, UVI; the negative correlation between precipitation and ISCCP TC; and the positive correlation between GMTA and MERRA TC and MERRA LLC.

From the regional analysis, it can be deduced that even though averaging the parameters can lead to missing certain smaller areas with significant correlations, they can still provide information regarding which variables are positively or negatively correlated and whether or not the correlation is high.

5.4 Multiple Regression Models

The previous sections have investigated the linear correlations between only two variables at a time. However, studying the variables together in a model is crucial to investigate the common predictive value of all the predictors and the interactions between them. This section will show the results of the multiple regression models, compare the validation methods used for the models, and discuss the importance of parameter selection.

5.4.1 Comparison of Validation Methods for Regression Models

As previously mentioned in Section 2.6.4, validating a ML model is a crucial part of the analysis to prevent overfitting and obtain more accurate results. Holdout validation of 25% is used in this analysis, yet there are different ways of implementing it to the dataset. Hence, three different validation methods were tested and compared. The first method was to hold out a randomly selected 25% inside the full dataset and test that 25%; the second method was to hold out the first 25% of the dataset, train the remaining 75%, then test the first 25% of the data points; and the third method was to hold out the final 75% of the dataset, train the model with the first 75% data points and then test the last 25% part. The last method measures how well the model can predict the future with the model trained from the past to a certain point in time; hence it is a good measure of the performance. The second model does the same backward. Although it ignores the principle of causality, it is mathematically equivalent to the last one as the whole data is already recorded.

Instead of testing validation methods for each predictand separately one by one, the best performing predictand when using global averages, which is the MERRA2 total cloud cover, was used to compare the different validation methods. The performance of each predictand will be explained in more detail in the following sections, namely, Sections 5.4.2 and 5.4.3. The comparison was first done using the smoothed and globally averaged MERRA2 TC data and the smoothed predictors GCR, SSN, TSI, UVI, and ONI for each regression model. The resulting R^2 of the models can be seen in Table 5.8.

Table 5.8. Comparison of the R^2 values for the regression models with different validation methods. GCR, SSN, TSI, UVI and ONI are used as the predictors; and globally averaged MERRA2 TC is used as the predictand

	LM	**GLM**	**Stepwise Model**	**Random Forest**
Random %25 Held Out and Tested	0.57	0.57	0.58	0.77
First 25% of Data Held Out and Tested	-3.34	-3.34	-3.43	-0.50
Final 25% of Data Held Out and Tested	-6.17	-6.17	-4.28	-5.85

It is seen that the random hold-out method works well for all the models, yet the results are suspicious. When the data is randomly selected from the entire dataset, and all data has been smoothed, it might not be that hard for the machine to predict where one data point would fit among other close data points. In other words, two successive points, which are almost identical due to the large auto-correlation functions of the predictors, may end up in the test and training sets. This is not very different from having the same points in both sets. This can be seen better in Figure 5.29, where the response data is plotted for the linear model as an example.

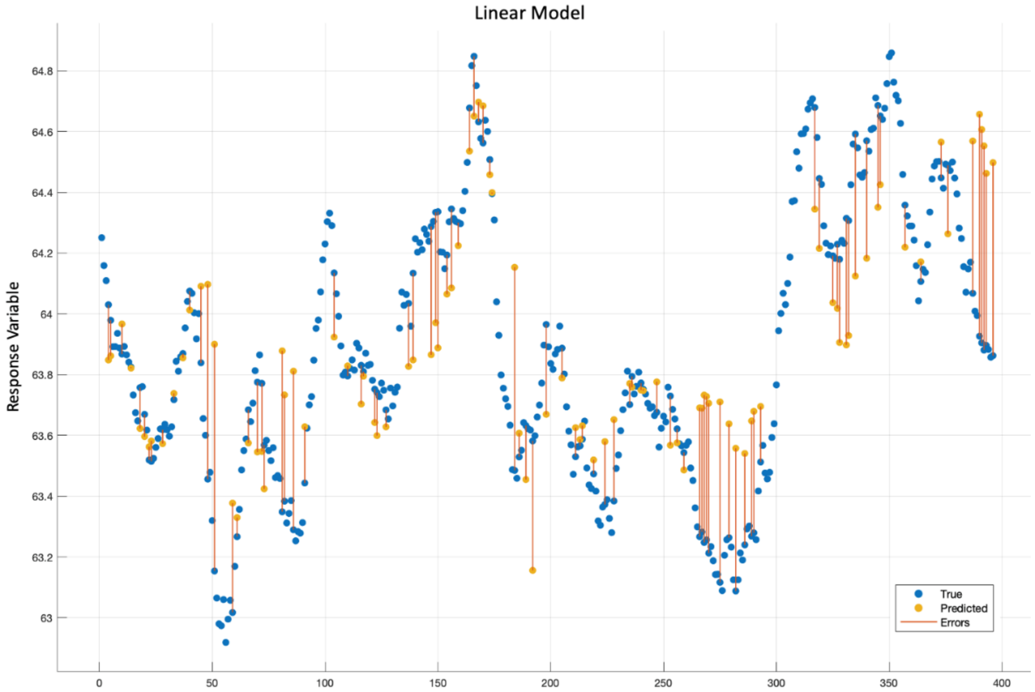


Figure 5.29. Response plot of the globally averaged MERRA2 TC, predicted by the linear model.

The blue data points in the figure are the 75% of the data used in training the model, while the yellow data points are the randomly selected 25% of the data used in testing the model. Because the data is periodic throughout the 34 years, it is not that difficult to predict the trends in the held-out data points. Thus, it is crucial to hold out continuous parts of the dataset to test and check the model's performance then.

It is seen from previous sections that it would be wrong to consider global averages without checking pixelwise results. Figures 5.30 – 5.32 show the pixelwise maps with different validation of the random forest model, using GCR, SSN, TSI, UVI, and ONI as the predictors, and globally averaged MERRA2 TC as the predictand.

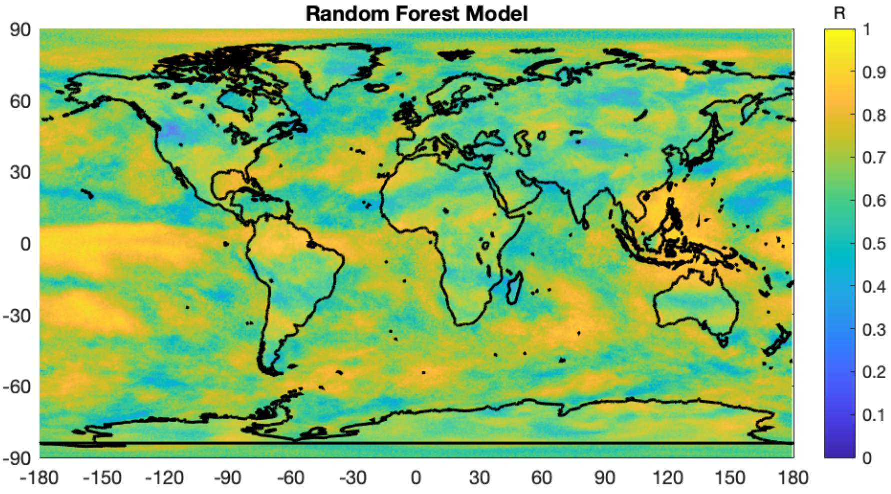


Figure 5.30. Global R^2 map of the random forest model with smoothed predictors and the pixelwise MERRA2 total cloud amount. Random 25% holdout validation is used.

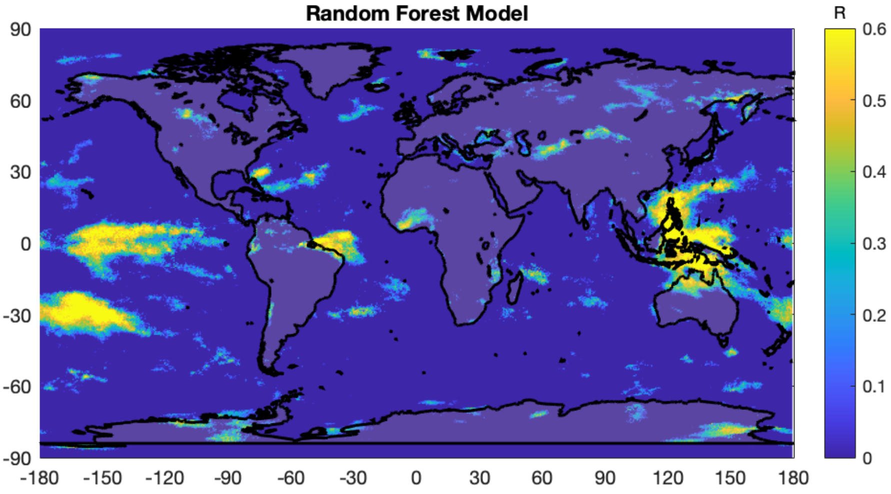


Figure 5.31. Global R^2 map of the random forest model with smoothed predictors and the pixelwise MERRA2 total cloud amount. The first 25% of the dataset is held out and tested.

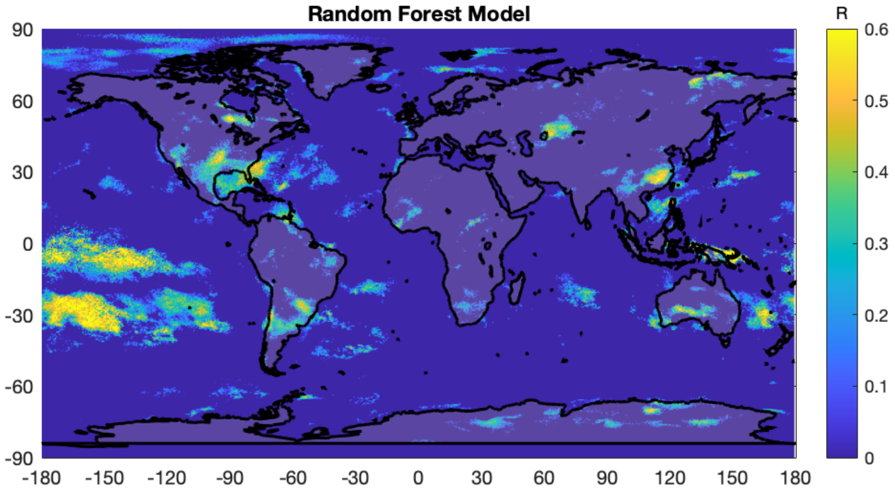


Figure 5.32. Global R^2 map of the random forest model with smoothed predictors and the pixelwise MERRA2 total cloud amount. The final 25% of the dataset is held out and tested.

Negative pixels are shown as zero in these maps in order to see the positive pixels better. Such negative pixels are the pixels where the random forest model is a poor fit. Comparing the maps with each other, it is seen that even though there are pixels that do not fit the models, similar regions have high R^2 values in the maps for all three validation methods. However, the random hold-out model creates artificially high values.

5.4.2 Globally and Regionally Averaged Model Results

The R^2 results of the models shown here are validated with random hold-out validation. Although the results are artificially high, still it can be used to assess the relative performance of methods.

Table 5.9. R^2 for the multiple regression models with smoothed GCR, SSN, TSI, UVI and ONI as predictors. The predictands are globally averaged

Globally Averaged Predictands	**LM**	**GLM**	**Stepwise Model**	**Random Forest**
ISCCP TC	0.37	0.37	0.28	0.72
ISCCP LLC	0.46	0.46	0.46	0.81
MERRA2 TC	0.60	0.60	0.58	0.83
MERRA2 LLC	0.62	0.62	0.62	0.80
GMTA	0.71	0.71	0.68	0.76
AOD	0.48	0.48	0.43	**0.78**
PRECIP	0.49	0.49	0.39	0.65

Table 5.10. R^2 for the multiple regression models with smoothed GCR, SSN, TSI, UVI and ONI as predictors. The predictands are averaged over Region I

Predictands Averaged over Region I	**LM**	**GLM**	**Stepwise Model**	**Random Forest**
ISCCP TC	0.36	0.36	0.23	0.70
ISCCP LLC	0.56	0.56	0.56	0.81
MERRA2 TC	0.73	0.73	0.70	0.80
MERRA2 LLC	0.64	0.64	0.63	**0.81**
GMTA	0.64	0.64	0.59	**0.81**
AOD	0.50	0.50	0.47	0.71
PRECIP	0.29	0.29	0.15	0.59

Table 5.11. R^2 for the multiple regression models with smoothed GCR, SSN, TSI, UVI and ONI as predictors. The predictands are averaged over Region II

Predictands Averaged over Region II	**LM**	**GLM**	**Stepwise Model**	**Random Forest**
ISCCP TC	0.41	0.41	0.42	0.72
ISCCP LLC	0.46	0.46	0.28	0.79
MERRA2 TC	0.64	0.64	0.62	**0.86**
MERRA2 LLC	0.35	0.35	0.22	0.71
GMTA	0.61	0.61	0.59	0.81
AOD	0.68	0.68	0.60	0.72
PRECIP	0.43	0.43	0.45	**0.68**

Table 5.12. R^2 for the multiple regression models with smoothed GCR, SSN, TSI, UVI and ONI as predictors. The predictands are averaged over Region III

Predictands Averaged over Region III	**LM**	**GLM**	**Stepwise Model**	**Random Forest**
ISCCP TC	0.49	0.49	0.43	**0.79**
ISCCP LLC	0.43	0.43	0.42	**0.81**
MERRA2 TC	0.59	0.59	0.63	0.80
MERRA2 LLC	0.70	0.70	0.66	**0.81**
GMTA	0.62	0.62	0.61	0.75
AOD	0.52	0.52	0.52	0.72
PRECIP	0.31	0.31	0.22	0.56

The first thing that can be said about these four tables is that the random forest model works best in all regions for all of the predictands. This shows that there indeed exists a non-linear relationship between the predictors and the predictands. Comparing the four regions, the RF results of the different region averages are very close to each other.

Only the model with PRECIP as the response variable has the highest R^2 value in Region II (the tropics), while the remaining predictands do not significantly differ from region to region. This could be because the ENSO region itself is in the tropics latitude zone, and the El-Niño warming period is known to cause more rain, especially over the tropical regions (Kump, Kasting, & Crane, 2004). Thus, it can be an expected effect for the precipitation response variable to have the highest R^2 value in Region II.

A comparison of the linear regression models with each other shows that the LM and GLM give the exact same R^2 values for all of the predictands averaged over all regions. Comparing the LM with stepwise GLM, it is seen that running the stepwise model does not change the R^2 values significantly for the LM or GLM results. For some predictands, the stepwise GLM results are even worse than the LM results. Hence, only the LM and RF models will be shown for the remaining parts of the results.

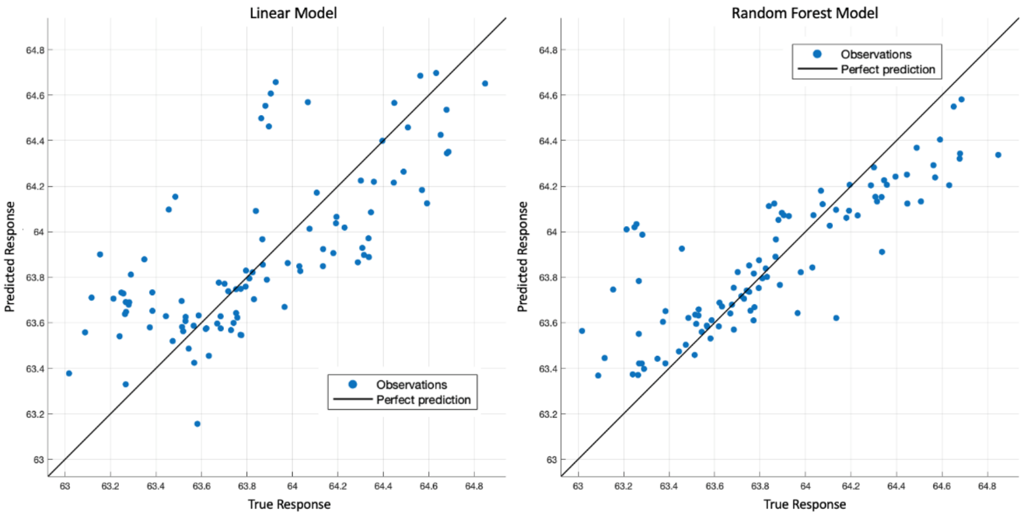


Figure 5.33. Scatter plot of the predicted response data vs the actual data for the globally averaged MERRA2 total cloud cover. The plot on the left is of the linear model, and the plot on the right is of the random forest model.

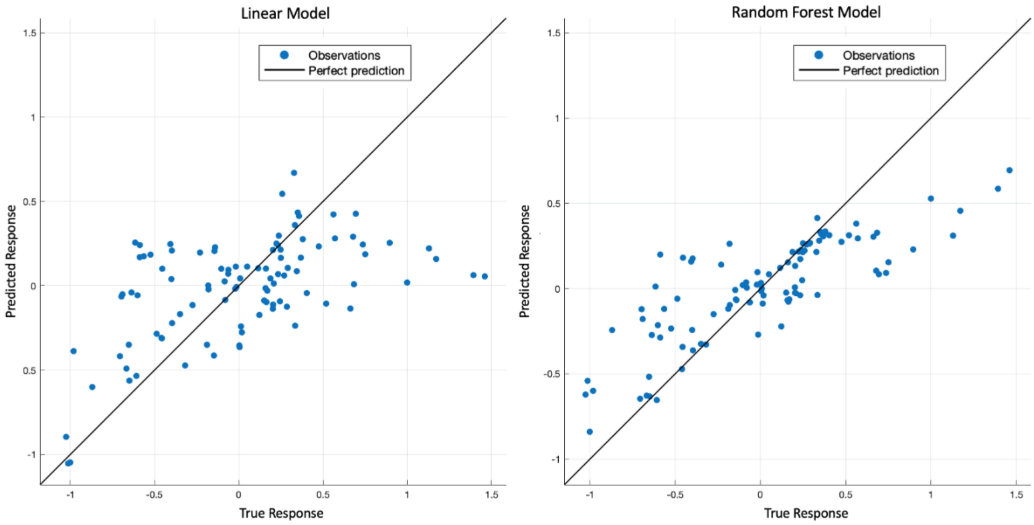


Figure 5.34. Scatter plot of the predicted response data vs the actual data for the globally averaged ISCCP total cloud cover. The plot on the left is of the linear model, and the plot on the right is of the random forest model.

Figures 5.33 and 5.34 compare the MERRA2 TC regression model results with the ISCCP TC results. This is to show how the different cloud product types respond to different models as scatter plots. MERRA2 TC has the highest values of R^2 both for LM and for RF in the globally averaged data.

The scatter plot on the left in Figure 5.33 is of the linear model, and it has a R^2 of 0.60, while the plot on the right is of the random forest model with a R^2 of 0.83. Similarly, for the scatter plots in Figure 5.34, the left is of the LM with an R2 of 0.37, while the right is of RF with an R^2 of 0.72.

Comparing both figures of the predicted versus actual responses, it is seen that MERRA2 TC data is a better fit for both the LM and the RF. This result is parallel with the *R* and *p*-values in Table 5.7, once again showing that the total cloud cover of ISCCP data is a poor fit.

5.4.3 Pixelwise Model Results

The pixelwise maps in this section will be shown for the linear model and random forest model for all cloud datasets, using the third validation method (prediction of future) only.

5.4.3.1 Linear Model Maps

This section shows maps using the linear model with the predictors and different cloud products in Figures 5.35 – 5.38.

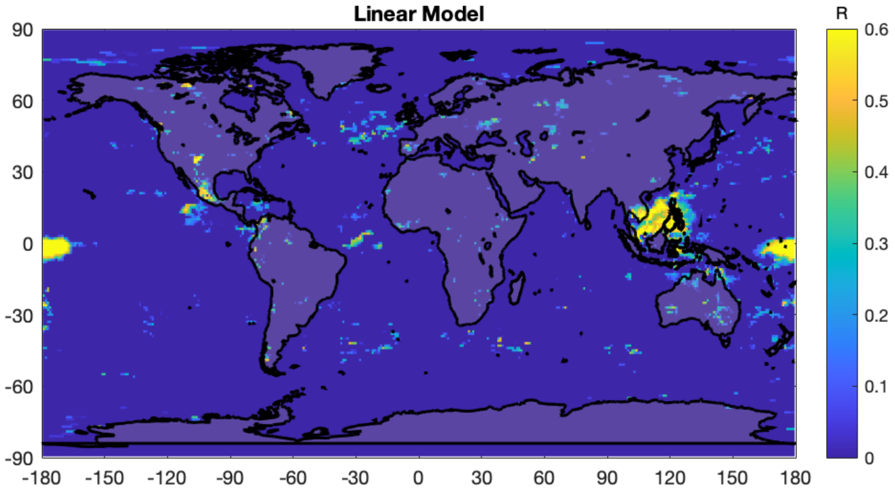


Figure 5.35. Global R^2 map of the linear model with smoothed predictors and the pixelwise ISCCP total cloud amount. The final 25% of the dataset is held out and tested.

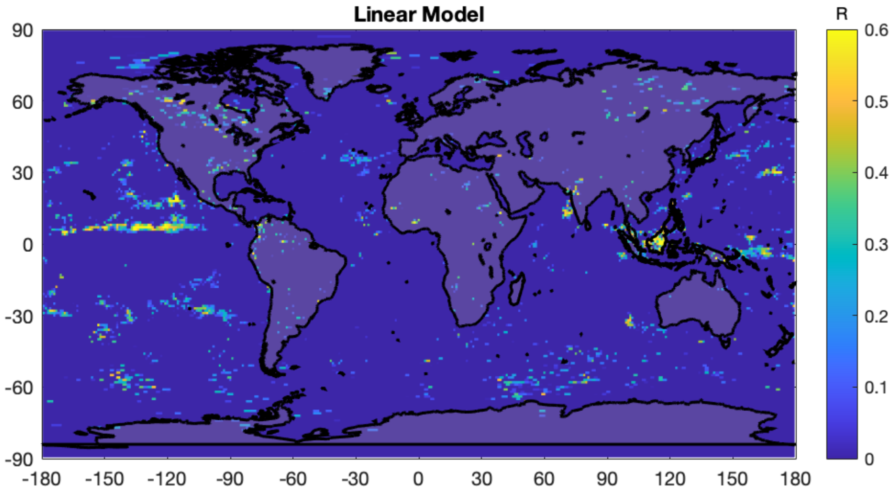


Figure 5.36. Global R^2 map of the linear model with smoothed predictors and the pixelwise ISCCP low cloud amount. The final 25% of the dataset is held out and tested.

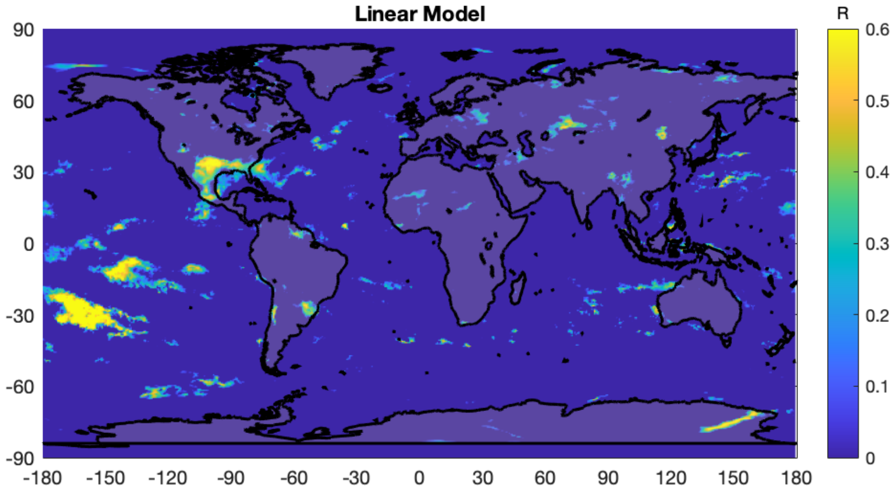


Figure 5.37. Global R^2 map of the linear model with smoothed predictors and the pixelwise MERRA2 total cloud amount. The final 25% of the dataset is held out and tested.

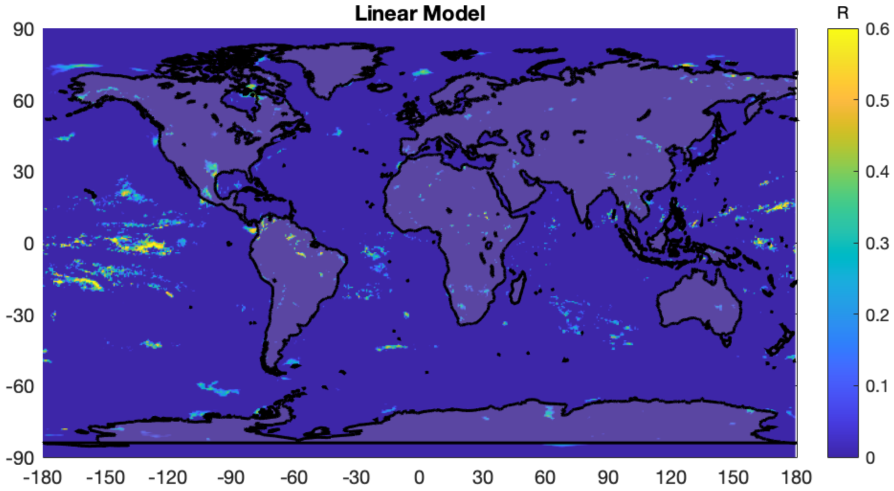


Figure 5.38. Global R^2 map of the linear model with smoothed predictors and the pixelwise MERRA2 low cloud amount. The final 25% of the dataset is held out and tested.

Firstly, the majority of pixels have a negative R^2 value in all four figures (shown as zero), meaning that the performance of the linear models is very poor for these sets of predictors and predictands. However, there are similarities in the regions with strong positive pixels. These regions are mainly in the equatorial and tropics regions in the Pacific Ocean. Some narrower strong R^2 areas can also be seen in the mid-latitudes in both hemispheres.

It is seen from Figures 5.35 to 5.38 that using TC as a predictand fits the linear model better for both cloud products. This could possibly be because of the ENSO index ONI being more correlated with the total cloud cover than the low cloud cover, as mentioned in Marsh and Svensmark (Marsh & Svensmark, 2003). Yet, this relation is considered to falsely affect the results (Voiculescu, Usoskin, & Mursula, 2007).

5.4.3.2 Random Forest Model Maps

This section shows maps using the random forest model with the predictors and different cloud products in Figures 5.39 – 5.42.

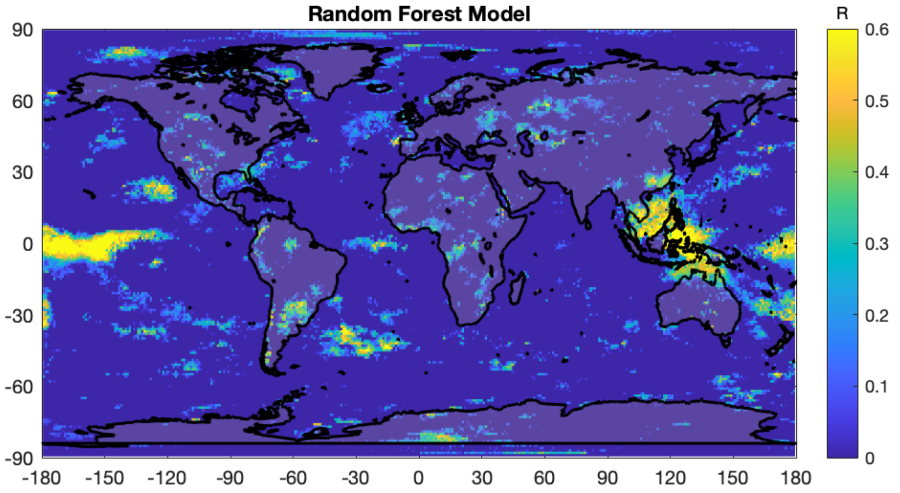


Figure 5.39. Global R^2 map of the random forest model with smoothed predictors and the pixelwise ISCCP total cloud amount. The final 25% of the dataset is held out and tested.

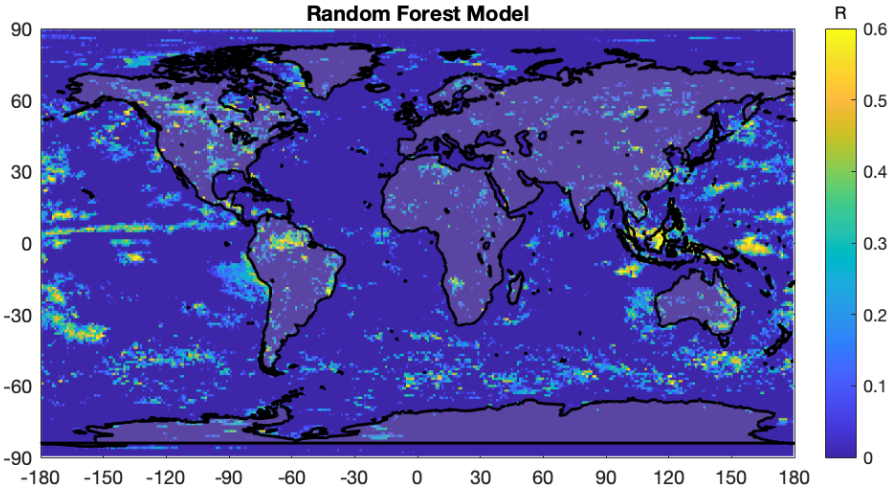


Figure 5.40. Global R^2 map of the random forest model with smoothed predictors and the pixelwise ISCCP low cloud amount. The final 25% of the dataset is held out and tested.

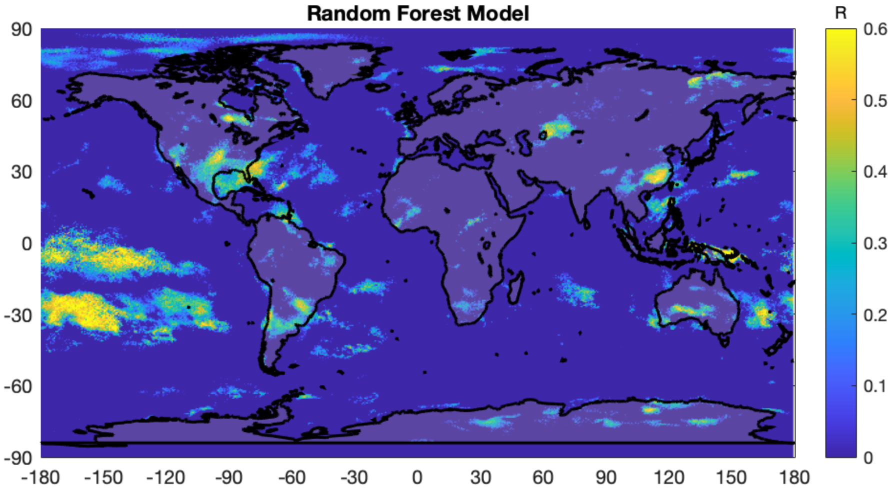


Figure 5.41. Global R^2 map of the random forest model with smoothed predictors and the pixelwise MERRA2 total cloud amount. The final 25% of the dataset is held out and tested (same figure as Figure 5.32).

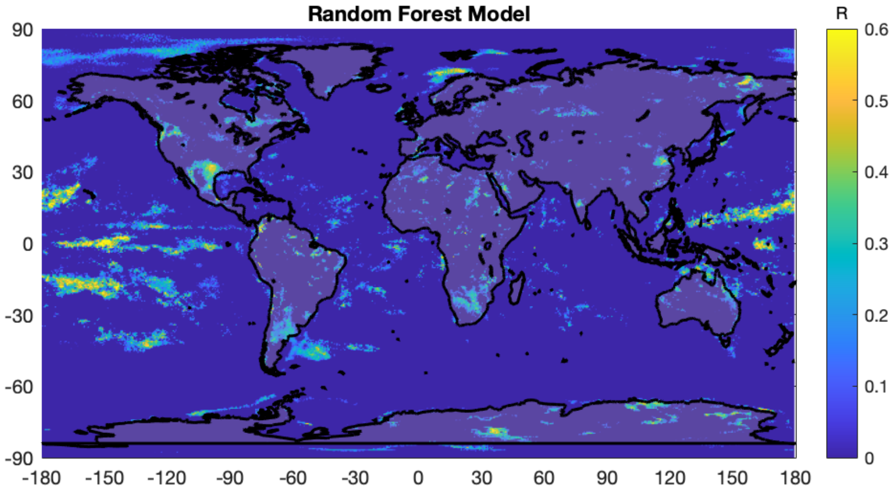


Figure 5.42. Global R^2 map of the random forest model with smoothed predictors and the pixelwise MERRA2 low cloud amount. The final 25% of the dataset is held out and tested.

The random forest model works much better than the linear model for all cloud products. The pixels with high R^2 values are more spread out over the globe, compared to the linear model, which had limited regions with positive R^2 values. This shows that the random forest model is a better fit for these sets of predictors and predictands, meaning a non-linear relationship exists between the predictors and predictands.

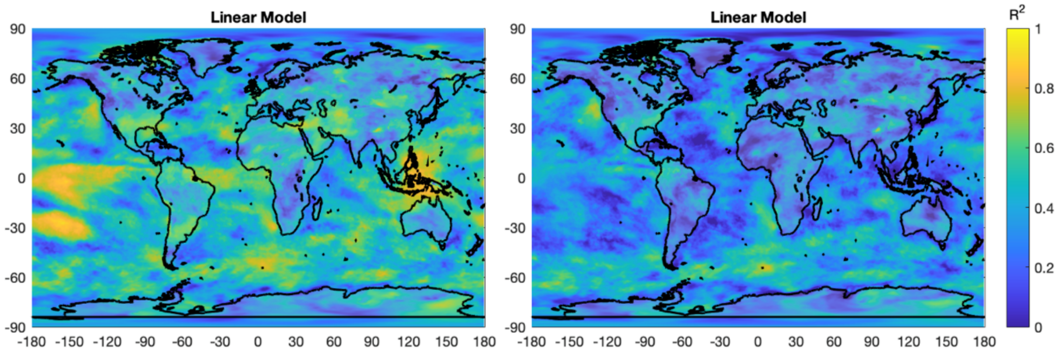
The regions with positive R^2 values are parallel in both the ISCCP and MERRA2 cloud data. However, just like in the linear model results, the TC cover of both cloud products has stronger R^2 pixels in the random forest model maps. Again, this could be because of the reported (Voiculescu, Usoskin, & Mursula, 2007) false ENSO index – high cloud cover correlation, as the most visibly strong pixels, are in the tropical Pacific Ocean.

Another reason for these visible areas along the equator could be due to sulfate (SO_4) and sea-salt (NaCl) masses observed in the Equatorial Pacific regions. The global maps of these masses are given in Mann et al. (Mann, Carslaw, Ridley, Spracklen, & others, 2012), and the areas with high concentration are parallel to the high positive R^2 pixels in the regression model maps of this thesis. The study by Mann et al. evaluates a global aerosol microphysics module and generates maps using these molecules. Both the sulfate and the sea-salt maps display similar patterns in the Pacific. The shapes for sulfate masses are also shown in Hommel et al. (Hommel, Timmreck, & Graf, 2011); and for sea-salt masses, they are shown in the research by Wang et al. (Wang, et al., 2017). However, the stated relations are beyond the scope of this study and could be part of future research.

5.4.4 The Effect of the ONI Parameter

First, the global pixelwise R^2 maps of the LM and random forest models are plotted using five predictors (GCR, SSN, TSI, UVI, ONI) and four predictors (GCR, SSN, TSI, UVI) to compare the parameter importance values with and without the ENSO index. This is to see if the observed patterns in the Pacific are related to the ONI parameter, as mentioned in Voiculescu et al. (Voiculescu, Usoskin, & Mursula, 2007). The comparison is made both for the LM and the RF model, using both random holdout validation and holding out the final 25% of the dataset. MERRA2 TC is used as the predictand in this section.

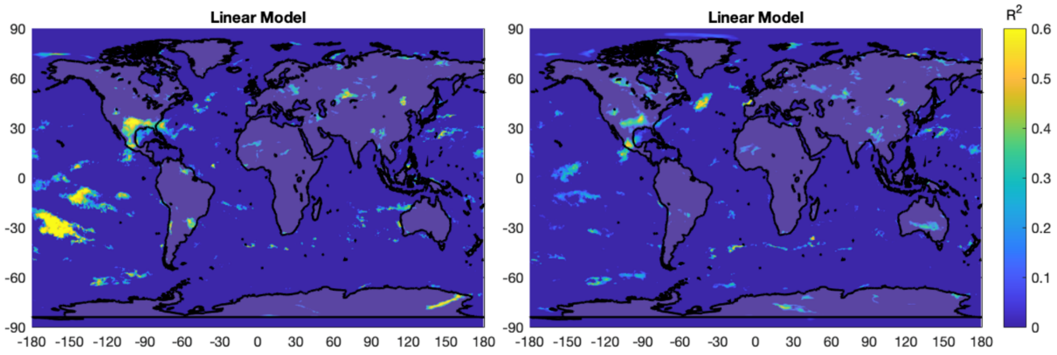
Figures 5.43 and 5.44 are the linear model maps with different validation methods used, and Figures 5.45 and 5.46 are the random forest model maps with different validation methods used, respectively.



(a) Predictors **with** ONI

(b) Predictors **without** ONI

Figure 5.43. Global R^2 map of the linear model with the pixelwise MERRA2 total cloud amount as the predictand. The predictors are a) GCR, SSN, TSI, UVI and ONI; b) GCR, SSN, TSI, UVI. Random 25% holdout validation is used.

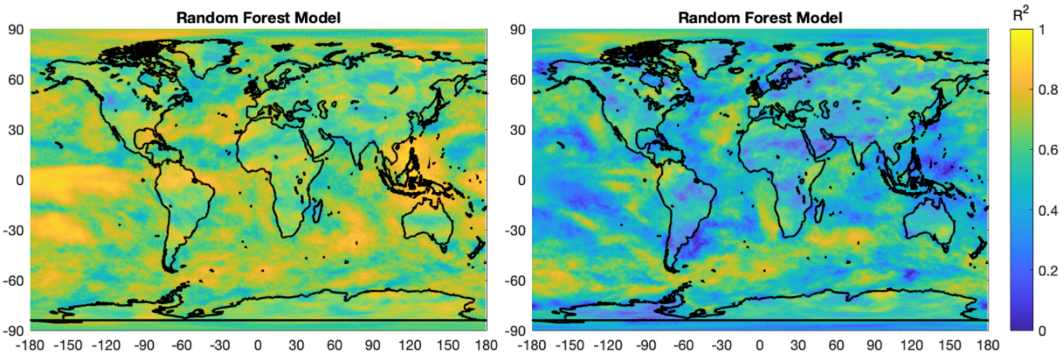


(a) Predictors **with** ONI

(b) Predictors **without** ONI

Figure 5.44. Global R^2 map of the linear model with the pixelwise MERRA2 total cloud amount as the predictand. The predictors are a) GCR, SSN, TSI, UVI and ONI; b) GCR, SSN, TSI, UVI. The final 25% of the dataset is held out and tested.

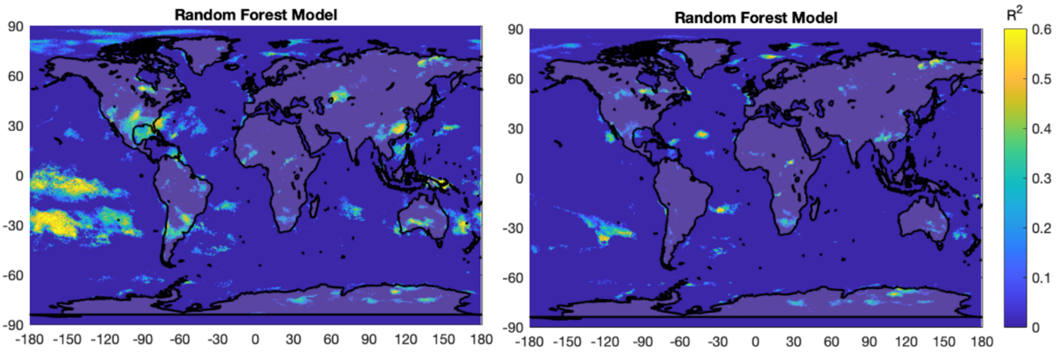
In the linear models using both validation methods, the high R^2 patterns in the Equatorial Pacific almost entirely disappear when ONI is removed as a predictor. This suggests that ONI was an essential parameter in the model.



(a) Predictors **with** ONI

(b) Predictors **without** ONI

Figure 5.45. Global R^2 map of the random forest model with the pixelwise MERRA2 total cloud amount as the predictand. The predictors are a) GCR, SSN, TSI, UVI and ONI; b) GCR, SSN, TSI, UVI. Random 25% holdout validation is used.



(a) Predictors **with** ONI

(b) Predictors **without** ONI

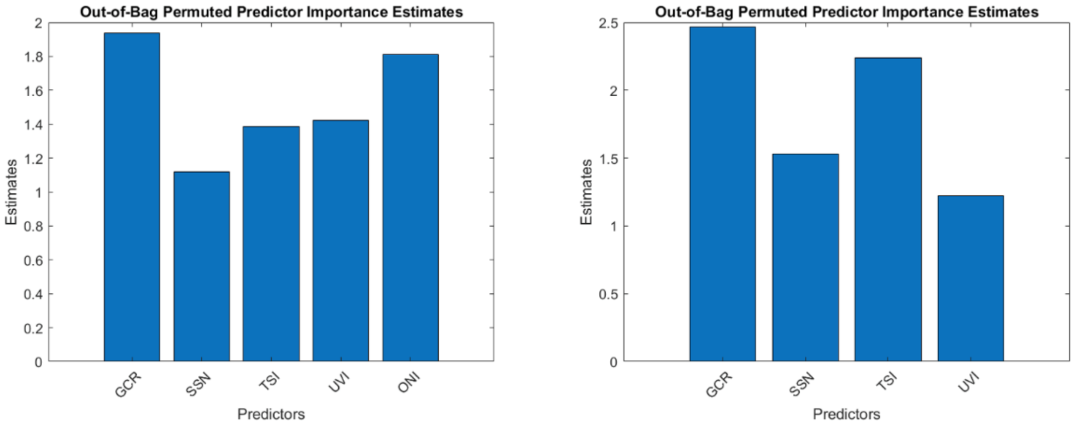
Figure 5.46. Global R^2 map of the random forest model with the pixelwise MERRA2 total cloud amount as the predictand. The predictors are a) GCR, SSN, TSI, UVI and ONI; b) GCR, SSN, TSI, UVI. The final 25% of the dataset is held out and tested.

The same is true for the RF model maps. The pixels in the Equatorial Pacific disappear after removing the ONI as a predictor, implying the importance of the parameter in that specific region.

5.4.5 Parameter Importance

The parameter importance of regression models is investigated to compare the relative effect on the model of each predictor parameter. Figures 5.47 and 5.48 show the out-of-bag (OOB) permuted predictor importance plots for the RF model maps in Figure 5.44 and Figure 5.46, with the ONI as a predictor and without the ONI parameter, respectively.

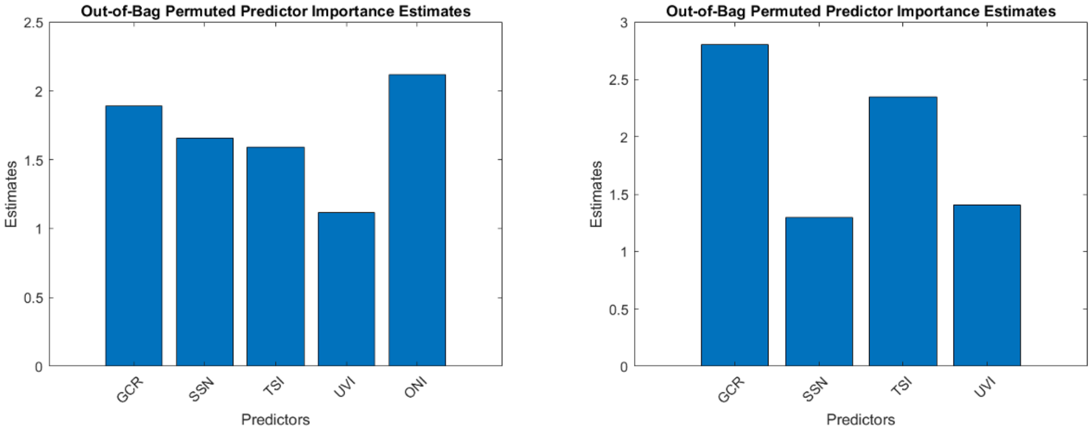
The OOB estimates are measured in non-linear regression models such as random forests, and they show how much the predictors are effective at predicting the response in the models. Permuting the values of a predictor should be affecting the error of the model if the predictor is influential in the prediction.



(a) Predictors **with** ONI

(b) Predictors **without** ONI

Figure 5.47. Out-of-Bag predictor importance plots of the random forest model with the pixelwise MERRA2 total cloud amount as the predictand. The predictors are a) GCR, SSN, TSI, UVI and ONI; b) GCR, SSN, TSI, UVI. Random 25% holdout validation is used.



(a) Predictors **with** ONI

(b) Predictors **without** ONI

Figure 5.48. Out-of-Bag predictor importance plots of the random forest model with the pixelwise MERRA2 total cloud amount as the predictand. The predictors are a) GCR, SSN, TSI, UVI and ONI; b) GCR, SSN, TSI, UVI. The final 25% of the dataset is held out and tested.

Comparing the OOB estimate figures for both with and without the ONI index, it is seen that all of the predictors appear important. Nevertheless, since the predictor parameters are correlated between themselves, these results can be misleading.

A process called Recursive Feature Elimination (RFE) is conducted to clarify this, and it is explained in the following section.

5.4.5.1 Recursive Feature Elimination

RFE is a method used to select features, and it simply removes the weakest predictors from the model to compare how the performance of the model changes. Here the method begins with all five predictors, and the results of the removal process can be seen in Table 5.13.

Table 5.13. R^2 values of the recursive feature elimination for the models, using the globally averaged MERRA2 TC as the predictand. Random 25% holdout validation is used.

Predictors	**Linear Model**	**Stepwise Model**	**Random Forest**
GCR, SSN, TSI, UVI, ONI	0.57	0.58	0.77
GCR, SSN, TSI, UVI	0.42	0.41	0.70
GCR, TSI, ONI	0.34	0.34	0.61
GCR, TSI	0.25	0.25	0.53

It is seen from this table that even though the predictor parameters are strongly correlated with each other, the R^2 value decreases every time a predictor is removed from the model. This indicates that each predictor parameter in each model has an important novelty that differs from one another.

CHAPTER 6

CONCLUSION

6.1 Summary and Conclusions

This thesis aimed to study the relationships between external forces on Earth's climate and the meteorological parameters. Using machine learning techniques, inter-relationships among cosmic rays, solar activity, and the Earth's climate in different geographical regions and temporal periods were investigated. The parameters used in the correlation analysis were galactic cosmic ray (GCR) flux, Sunspot number (SSN), total solar irradiance (TSI), UV irradiance (UVI), and the Oceanic Niño Index (ONI) as the predictor variables; total cloud (TC) amount, low-level cloud (LLC) amount, global mean temperature anomaly (GMTA), aerosol optical depth (AOD) and precipitation (PRECIP) as the response variables. All monthly averaged datasets were preprocessed, smoothed, and analyzed for the years 1984-2017. After standard statistical techniques were applied, multiple linear regression models and machine learning methods for non-linear regression were used. The findings are listed as follows:

1. Previous studies reporting a strong linear correlation between low cloud cover and cosmic ray flux were reanalyzed, and their study period was expanded up to the year 2017. It was seen that the correlation coefficient values weakened as the study period extended, and the reported correlation disappeared completely after ~2003. The MERRA2 TC and LLC were found to be more correlated to the GCR flux compared to the ISCCP TC and LLC for the years 1984-2017.

2. Pixelwise correlation analysis was conducted for the full analysis period (1984-2017) between the predictors and cloud data. A strong positive correlation was observed between GCR-cloud in the mid-latitude oceans at both hemispheres, consistent with previous work. The anti-correlation between GCR and solar activity was confirmed once again with these maps.
3. Different hold-out validation methods for the ML models were compared. It was seen that even though random hold-out validation worked well for all ML models, it was thought to create artificially high R^2 values. A hold-out validation method that holds out the final parts of a dataset and tests them was thus used, which in a time-dependent dataset means that the model uses past data to predict future data. This validation method uses predictive power, which means that the values found are less likely to be coincidental.
4. A significant non-linear relationship between the predictors and predictands has been found using random forest regression models, which has never been used before.
5. Both cloud products showed overall higher R^2 valued pixels for the TC amount data. Most of these pixels are in the Equatorial Pacific region and are suggested to be because of the reported effect of the ENSO event. The pixelwise maps were constructed without using ONI as a predictor parameter, and the high R^2 valued pixels in the Pacific vanished. This showed that the effect observed there was due to the ONI parameter.
6. The parameter importance of the predictors used in the RF models was investigated, and recursive feature elimination was conducted. The R^2 value was observed to decrease each time a predictor was removed from the analysis, confirming that each predictor parameter used in the analysis had an essential addition to the model.

It can be concluded that the GCR-climate connection does exist, and these non-linear relations should be investigated further, specifically in certain regions of the World.

6.2 Future Work and Suggestions

This analysis was mainly conducted to search for relations between the parameters, and the reasons for the relationships have not been investigated. Finding where and how much connection there is can help enlighten the mechanisms behind the relations. Based on the results of this thesis, further research should focus on the regions found with high R^2 values using the RF models in the Indian Ocean and Pacific Ocean mid-latitudes. The relationship of the predictors and predictands with the ENSO event should be investigated in-depth, specifically for the mentioned region in the Equatorial Pacific. The parameters sulfate and sea-salt masses could be added to the analysis to see how the relations change.

Finally, using the predictors as gridded datasets instead of globally averaged data would help to understand relationships for each pixel further. Such changes could be, for example, to calculate and use cosmic ray-induced ionization in the atmosphere at certain altitudes, for each pixel, instead of the neutron monitor count rates as the GCR data.

REFERENCES

Adler, R., Wang, J.-J., Sapiano, M., Huffman, G., Chiu, L., & others. (2016). *Global Precipitation Climatology Project (GPCP) Climate Data Record (CDR), Version 2.3 (Monthly)*. doi:10.7289/V56971M6

Ahluwalia, H. S. (2013). No direct correlation between galactic cosmic rays and Earth surface temperature. *Advances in Space Research, 52*, 2119–2121.

Ahluwalia, H. S. (2014). Galactic cosmic rays, total solar irradiance, sunspots, Earth surface air temperature: Correlations. *Indian Journal of Radio & Space Physics, 43*, 141–150.

Akcan, M. E. (2004). *Güneş Lekesi Çevrim Süreci İçerisinde Kozmik Işınların Yeryüzü İklim-bulut Örtüsü İle Olan İlişkisinin İncelenmesi.* Master's thesis, İTÜ Fen Bilimleri Enstitüsü.

Benestad, R. E. (2006). *Solar activity and Earth's climate.* Springer Science & Business Media.

Bhaskar, A., Ramesh, D. S., Vichare, G., Koganti, T., & Gurubaran, S. (2017). Quantitative assessment of drivers of recent global temperature variability: an information theoretic approach. *Climate Dynamics, 49*, 3877–3886.

Biktash, L. Z. (2019). Influence of Total Solar Irradiance on the Earth’s Climate. *Geomagnetism and Aeronomy, 59*, 368–373.

Čalogović, J., & Laken, B. A. (2015). Reflections on the late Cosmoclimatology. *Central European Astrophysical Bulletin, 39*, 145–160.

Carslaw, K. S. (2009). Cosmic rays, clouds and climate. *Nature, 460*, 332–333.

Carslaw, K. S., Harrison, R. G., & Kirkby, J. (2002). Cosmic rays, clouds, and climate. *Science, 298*, 1732–1737.

Chapanov, Y., & Gorshkov, V. (2019). Solar Activity and Cosmic Ray Influence on the Climate. *Geomagnetism and Aeronomy, 59*, 942–949.

Ciaburro, G. (2017). *MATLAB for Machine Learning.* Packt Publishing Ltd.

Coddington, O., Lean, J. L., Lindholm, D., Pilewskie, P., Snow, M., & (2015), N. C. (2021, January 17). *NOAA Climate Data Record (CDR) of Solar Spectral Irradiance (SSI), NRLSSI Version 2*. doi:10.7289/V51J97P6

Coddington, O., Lean, J. L., Lindholm, D., Pilewskie, P., Snow, M., & (2015), N. C. (2021, January 17). *NOAA Climate Data Record (CDR) of Total Solar Irradiance (TSI), NRLTSI Version 2*. doi:10.7289/V55B00C1

Easterbrook, D. J. (2016). Chapter 14 - Cause of Global Climate Changes: Correlation of Global Temperature, Sunspots, Solar Irradiance, Cosmic Rays, and Radiocarbon and Berylium Production Rates. In *Evidence-Based Climate Science (Second Edition)* (pp. 245-262). Elsevier.

Eddy, J. A. (2009). *The Sun, the Earth, and Near-Earth Space: A Guide to the Sun-Earth System.* Government Printing Office.

El-Borie, M. A., Thabet, A. A., El-Mallah, E. S., Abd El-Zaher, M., & Bishara, A. A. (2020). Possible effects of galactic cosmic ray flux and low-cloud amounts on global surface temperature. *Pramana, 94*, 1–10.

Erlykin, A. D., & Wolfendale, A. W. (2011). Cosmic ray effects on cloud cover and their relevance to climate change. *Journal of Atmospheric and Solar-Terrestrial Physics, 73*, 1681–1686.

Erlykin, A. D., Gyalai, G., Kudela, K., Sloan, T., & Wolfendale, A. W. (2009). On the correlation between cosmic ray intensity and cloud cover. *Journal of Atmospheric and Solar-Terrestrial Physics, 71*, 1794–1806.

Erlykin, A. D., Sloan, T., & Wolfendale, A. W. (2009). Solar activity and the mean global temperature. *Environmental Research Letters, 4*, 014006.

Erlykin, A. D., Sloan, T., & Wolfendale, A. W. (2013). A review of the relevance of the 'CLOUD'results and other recent observations to the possible effect of cosmic rays on the terrestrial climate. *Meteorology and Atmospheric Physics, 121*, 137–142.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1). Springer.

Gray, L. J., Beer, J., Geller, M., Haigh, J. D., Lockwood, & others. (2010). Solar influences on climate. *Reviews of Geophysics, 48*.

Hartmann, D. L. (1993). Radiative effects of clouds on Earth's climate. In *International geophysics* (Vol. 54, pp. 151–173). Elsevier.

Hobbs, P. V. (1993). *Aerosol-Cloud-Climate Interactions.* Academic Press.

Hommel, R., Timmreck, C., & Graf, H. (2011). The global middle-atmosphere aerosol model MAECHAM5-SAM2: comparison with satellite and in-situ observations. *Geoscientific Model Development, 4*(3), 809-834.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer.

Kilifarska, N. A., Bakhmutov, V. G., & Mel'nik, G. V. (2015). Geomagnetic field and climate: Causal relations with some atmospheric variables. *Izvestiya, Physics of the Solid Earth, 51*, 768–785.

Kirkby, J. (2007). Cosmic rays and climate. *Surveys in Geophysics, 28*, 333–375.

Kirkby, J., & Carslaw, K. S. (2006). Variations of galactic cosmic rays and the Earth's climate. In *Solar journey: The significance of our galactic environment for the heliosphere and Earth* (pp. 349–397). Springer.

Kristjánsson, J. E., Stjern, C. W., Stordal, F., Fjæraa, A. M., Myhre, G., & Jónasson, K. (2008). Cosmic rays, cloud condensation nuclei and clouds–a reassessment using MODIS data. *Atmospheric Chemistry and Physics, 8*, 7373–7387.

Kump, L. R., Kasting, J. F., & Crane, R. G. (2004). *The Earth System* (Vol. 432). Pearson Prentice Hall Upper Saddle River, NJ.

LAADS DAAC, N. L.-1. (2021). *Combined Aqua Terra MODIS Product Monthly L3, MCD06COSP_M3_MODIS.* doi:10.5067/MODIS/MCD06COSP_M3_MODIS.061

Laken, B. A., & Čalogović, J. (2011). Solar irradiance, cosmic rays and cloudiness over daily timescales. *Geophysical Research Letters, 38*.

Laken, B. A., Pallé Bagó, E., & Miyahara, H. (2012). A Decade of the Moderate Resolution Imaging Spectroradiometer: Is a Solar–Cloud Link Detectable? *Journal of Climate, 25*, 4430–4440.

Laken, B. A., Pallé Bagó, E., Čalogović, J., & Dunne, E. M. (2012). A cosmic ray-climate link and cloud observations. *Journal of Space Weather and Space Climate, 2*, A18.

Laut, P. (2003). Solar activity and terrestrial climate: an analysis of some purported correlations. *Journal of Atmospheric and Solar-Terrestrial Physics, 65*, 801–812.

Mann, G. W., Carslaw, K. S., Ridley, D. A., Spracklen, D. V., & others. (2012). Intercomparison of modal and sectional aerosol microphysics representations within the same 3-D global chemical transport model. *Atmospheric Chemistry and Physics, 12*, 4449–4476.

Marsh, N., & Svensmark, H. (2000). Cosmic rays, clouds, and climate. *Space Science Reviews, 94*, 215–230.

Marsh, N., & Svensmark, H. (2003). Galactic cosmic ray and El Niño–Southern Oscillation trends in International Satellite Cloud Climatology Project D2 low-cloud properties. *Journal of Geophysical Research: Atmospheres, 108*(D6), 1 - 11.

Mathworks. (2021). *Predictive Modeling and Machine Learning*. Retrieved from Coursera: https://www.coursera.org/learn/predictive-modeling-machine-learning

NASA GMAO, G. M. (2015). *MERRA-2 tavgU_2d_rad_Nx: 2d,diurnal,Time-Averaged,Single-Level,Assimilation,Radiation Diagnostics V5.12.4*. doi:10.5067/4SDCJYK8P9QU

NOAA. (2021). *Global Mean Temperature Anomaly Timeseries*. Retrieved 2021, from National Centers for Environmental Information: https://www.ncdc.noaa.gov/cag/global/time-series

NOAA. (2021, January 17). *Historical El Nino / La Nina Episodes*. Retrieved from National Weather Service, Climate Prediction Center: https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php

Ogurtsov, M. G., & Veretenenko, S. V. (2017). Possible Contribution of Variations in the Galactic Cosmic Ray Flux to the Global Temperature Rise in Recent Decades. *Geomagnetism and Aeronomy, 57*, 886–890.

Ormes, J. F. (2018). Cosmic rays and climate. *Advances in Space Research, 62*, 2880–2891.

Pallé Bagó, E., & Butler, C. J. (2000). The influence of cosmic rays on terrestrial clouds and global warming. *Astronomy & Geophysics, 41*, 4–18.

Pallé Bagó, E., Butler, C. J., & O'Brien, K. (2004). The possible connection between ionization in the atmosphere by cosmic rays and low level clouds. *Journal of Atmospheric and Solar-Terrestrial Physics, 66*, 1779–1790.

Paluszek, M., & Thomas, S. (2016). *MATLAB Machine Learning.* Apress.

Pierce, J. R., & Adams, P. J. (2009). Can cosmic rays affect cloud condensation nuclei by altering new particle formation rates? *Geophysical Research Letters, 36*.

SILSO, W. D. (1984-2017). The International Sunspot Number. *International Sunspot Number Monthly Bulletin and online catalogue*. Retrieved from http://www.sidc.be/silso/

Singh, A. K., & Bhargawa, A. (2020). Delineation of possible influence of solar variability and galactic cosmic rays on terrestrial climate parameters. *Advances in Space Research, 65*, 1831–1842.

Stott, T., & Huddart, D. (2010). *Earth Environments: Past, Present and Future.* Wiley-Blackwell.

Svensmark, H. (2000). Cosmic rays and Earth's climate. *Space Science Reviews, 93*, 175–185.

Svensmark, H. (2007). Cosmoclimatology: a new theory emerges. *Astronomy & Geophysics, 48*, 1–18.

Svensmark, H., & Friis-Christensen, E. (1997). Variation of cosmic ray flux and global cloud coverage, a missing link in solar-climate relationships. *Journal of Atmospheric and Solar-Terrestrial Physics, 59*, 1225-1232.

Tsonis, A. A., Deyle, E. R., May, R. M., Sugihara, G., Swanson, K., Verbeten, J. D., & Wang, G. (2015). Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences, 112*, 3253–3256.

Usoskin, I. G. (2021). *Cosmic Ray Station of the University of Oulu*. Retrieved January 17, 2021, from The University of Oulu, Sodankyla Geophysical Observatory: http://cosmicrays.oulu.fi/

Usoskin, I. G., & Kovaltsov, G. A. (2008). Cosmic rays and climate of the Earth: Possible connection. *Comptes Rendus Geoscience, 340*, 441–450.

Usoskin, I. G., Marsh, N., Kovaltsov, G. A., Mursula, K., & Gladysheva, O. G. (2004). Latitudinal dependence of low cloud amount on cosmic ray induced ionization. *Geophysical Research Letters, 31*.

Usoskin, I. G., Mursula, K., Kangas, J., & Gvozdevsky, B. (2001). Online database of cosmic ray intensities. *International Cosmic Ray Conference*, *9*, pp. 3842–3842.

Usoskin, I. G., Voiculescu, M., Kovaltsov, G. A., & Mursula, K. (2006). Correlation between clouds at different altitudes and solar activity: Fact or Artifact? *Journal of Atmospheric and Solar-Terrestrial Physics, 68*, 2164–2172.

Veretenenko, S. V., & Ogurtsov, M. G. (2018). 60-year cycle in the Earth's climate and dynamics of correlation links between solar activity and circulation of the lower atmosphere. *Geomagnetism and Aeronomy, 58*, 973–981.

Voiculescu, M., & Usoskin, I. G. (2012). Persistent solar signatures in cloud cover: spatial and temporal analysis. *Environmental Research Letters, 7*, 044004.

Voiculescu, M., Usoskin, I. G., & Mursula, K. (2006). Different response of clouds to solar input. *Geophysical Research Letters, 33*.

Voiculescu, M., Usoskin, I. G., & Mursula, K. (2007). Effect of ENSO and volcanic events on the Sun–cloud link. *Advances in Space Research, 40*, 1140–1145.

Wang, H., Wang, X., Yang, X., Li, W., Xue, L., Wang, T., & others. (2017). Mixed chloride aerosols and their atmospheric implications: a review. *Aerosol and Air Quality Research, 17*(4), 878-887.

Young, A. H., Knapp, K. R., Inamdar, A., Hankins, W., & Rossow, W. B. (2018). The International Satellite Cloud Climatology Project H-Series climate data record product. *Earth Syst. Sci. Data,10*, 583–593. doi:10.7289/V5QZ281S

Yu, F. (2002). Altitude variations of cosmic ray induced production of aerosols: Implications for global cloudiness and climate. *Journal of Geophysical Research: Space Physics, 107*, SIA–8.

Zhao, X. (2017). *NOAA Climate Data Record (CDR) of AVHRR Daily and Monthly Aerosol Optical Thickness (AOT) over Global Oceans, Version 3.0.* doi:10.7289/V5BZ642P