Analysis of the Impact of Land Cover Changes on Climate using Machine Learning

Anastasiia Kolevatova

Thesis submitted for the degree of
Master in Informatics: Programming and Systems Architecture
60 credits

Department of Informatics
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2020

Analysis of the Impact of Land Cover Changes on Climate using Machine Learning

Anastasiia Kolevatova

Analysis of the Impact of Land Cover Changes on Climate using Machine Learning

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Acknowledgements

Many wonderful people supported me during the work on this thesis. First of all, I would like to thank my supervisors: Hugo Lewi Hammer for his practical recommendations, worthwhile discussions, and much attention to this work; Pål Halvorsen for his valuable advice and review of my work; and Michael Riegler for the inspiration to conduct this research.

I would also like to express my gratitude to my friends and family, and particular thanks are due to Alexandra Martyusheva for proofreading. Finally, I would like to deeply thank my beloved husband, Ilia Kolevatov, who believes in me, encourages my work and provides enormous support.

Abstract

Climate changes and global warming are actual and widely discussed themes. The last five years were recognized as the warmest period during the whole history of observations. Global warming has a significant influence on the environment: ice melting, sea-level rise, shifting of climate zones, changes in animal behaviour, etc. It also affects regional climate and human and leads to heat waves, droughts, natural disasters and others. Therefore, it is important to understand the reasons for the changes and possible ways to prevent them.

The Intergovernmental Panel on Climate Change were founded in 1988 to study anthropogenic climate changes and their mitigation strategies. Furthermore, they have paid special attention to land use as an important factor of climate system. On the one hand, climate changes strongly affect natural land cover, and human can mitigate such a changes. On the other hand, anthropogenic land use and land cover changes, such as deforestation and urban expansion, have a huge impact on climate. However, the impact of anthropogenic land cover change remains an unexplored problem in climate science.

Nowadays, one of the main methods in climate science is simulation performed by mathematical climate models. These models describe processes in climate system with a huge number of mathematical equations. However, the simulation results are very complex and difficult to interpret. Therefore, it can be complicated to find trends and hidden patterns linking different processes in climate using the simulation results. Machine learning is a particularly promising technology and can be an efficient tool to identify patterns in the climate simulation results.

Machine learning is widely adopted in various scientific fields. However, it has a limited application in climate science at the moment. Indeed, the standard machine learning techniques often imply an application to independent and identically distributed data, while climatic data do not meet these criteria well. Climate process typically occurs locally and affects neighbor points, which manifests itself in data dependency. In addition, the probability of climate processes on the Earth is unevenly distributed. Thus, the application of machine learning in climate science requires adaptation and verification.

One particular challenge is to understand which machine learning algorithm performs better on climate data. The goal of this thesis is to develop a test that allows defining statistically significant difference in performances on spatially dependent data. We developed the test which indicated that the random forest algorithm is the most efficient algorithm applied to spatially dependent data. This is consistent with other studies which also compared algorithms for climate science using different evaluation methods.

Finally, we applied the most efficient algorithm, the random forest, to analyze the impact of land cover changes in Europe on the regional surface temperature. Our findings are mostly consistent with another study carried out on the same initial data set. However, we revealed several new patterns that have not been detected using standard statistical methods. This fact is of potential interest to researchers and requires further investigation.

Contents

Acronyms

5x2 CV test 5x2-fold cross-validation paired t test. v, viii, ix, 7, 34, 35, 49, 50, 53–55, 58–60, 64–67, 69–72, 96–98

CV Cross-validation. viii, ix, 31–35, 47, 48, 53–55, 63–65, 67–72, 97

i.i.d. independent and identically distributed. 32, 35, 38, 39, 45, 48–50, 56, 71, 96, 98

LASSO Least absolute shrinkage and selection operator. 26, 27, 54, 57, 64–67, 69, 72

LC Land cover. vii–ix, 1–8, 10, 15–18, 20, 22, 25, 35–45, 47, 51–53, 55, 61–65, 69–71, 73–98, 100–103

ML Machine learning. v, vii, 4–8, 10, 18, 20–25, 28, 30, 33, 34, 36–40, 45–48, 51–55, 57, 59–65, 67, 69–73, 75, 85–87, 92–98

MLR Multiple linear regression. 26–29, 54, 57, 64–67, 69, 71, 72

MSE Mean squared error. 22, 31, 32, 71

$\mathbf{R^2}$ Coefficient of determination. 31, 32, 48, 64, 65, 67–71, 97

RF Random forest. 29, 30, 54, 64–67, 69–72, 94, 96, 97

RMSE Root mean squared error. 31, 35, 48, 49, 52, 53, 57, 60, 64, 67–71

SVR Support vector regression. 28, 64–67, 69–72

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

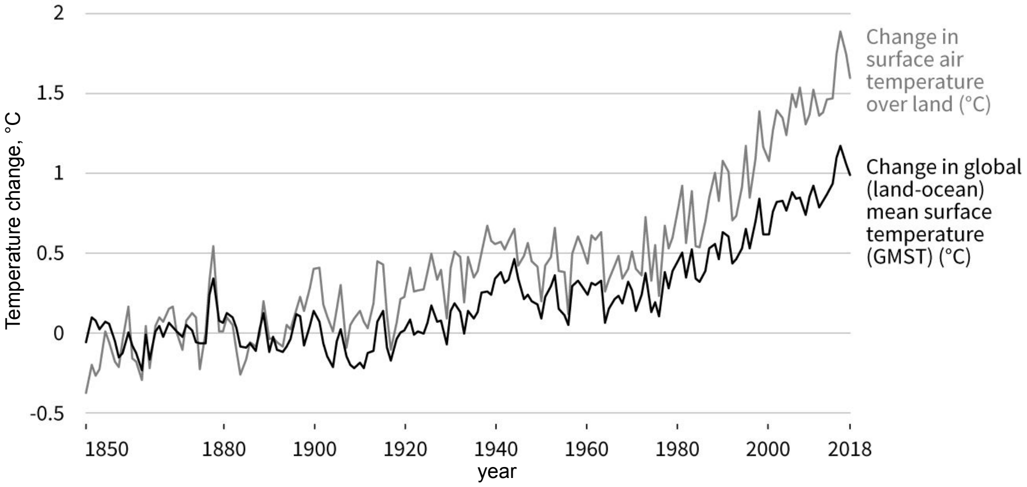
Nowadays, climate changes and global warming are indisputable facts [19, 75, 77, 91, 94]. Global surface temperature has been methodically collected since 1850. According to these records, the last 30 years exceed any previous decade in temperature. Furthermore, in some regions, the temperature has been measured over the last 1400 years, and the period between 1983 and 2012 was the warmest 30-year period during this time [19, 91, 92, 94].

The pace of global warming is constantly increasing. The global surface temperature in 2017 was the second highest annual temperature since the 1850s [43]. It was 0.38°–0.48 °C higher than the average temperature in 1981–2010 [43]. In addition, global warming was observed for all seasons in 2017 [43]. Therefore, 2017 became the warmest non-El Niño year during the entire historical period when climate data were recorded [43]. The second warmest year for the same period is 2019 [91]. Moreover, the average temperature over the last 5 years was the highest of all observations [91].

Climate changes and temperature growth have a huge impact on natural and anthropogenic systems on all continents and oceans: melting of snow and ice, sea level rise, decrease in fresh water volume and quality, changes in precipitation patterns, behavior alterations of marine organisms and animals, negative effects on agriculture and many other effects [77]. Moreover, as shown in Figure 1.1, the land surface air temperature increases twice as quick as the global (ocean and land) average temperature [94].

Anthropogenic factors, such as CO_2 emission, are considered as the main cause of global warming in the second half of the 20^{th} century [77]. Land cover (LC) transformation is distinguished among other anthropogenic factors as a cause that affects all types of climate changes. For example, it can lead to either an increase or a decrease in

Figure 1.1: Change in temperature rel. to 1850-1900 (°C). Adopted from Ref. [94]

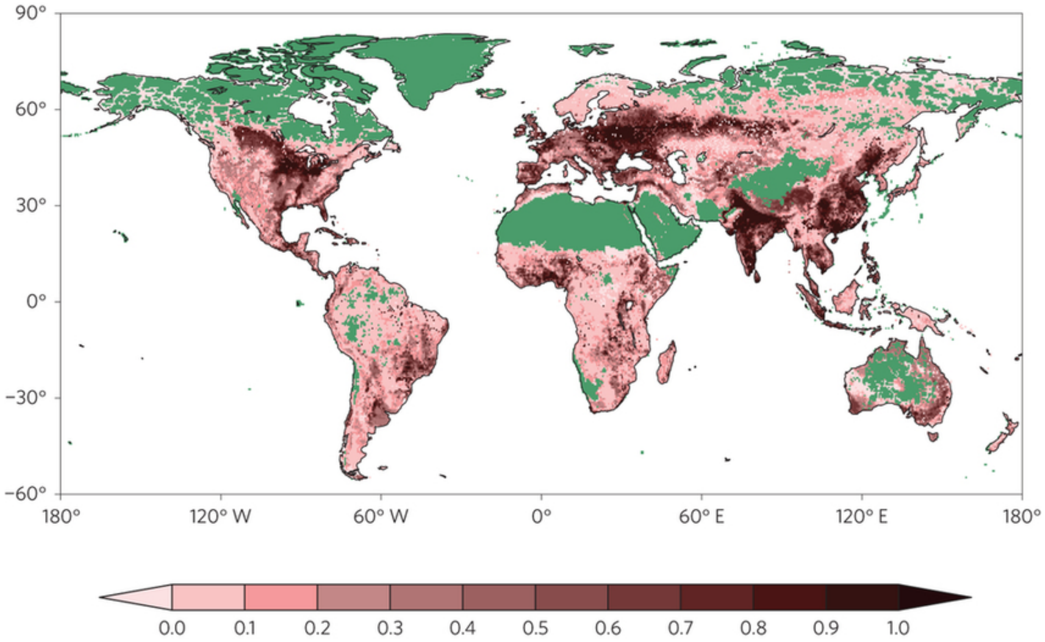


local CO_2 emissions into the atmosphere [77]. In this context, LC can be defined as a layer of soil and biomass that covers land surface and can be observed in the field and from remote sensors. For example, LC includes forests, crops, urban area, etc. [115].

The LC type transformation has various reasons. On the one hand, it can be caused by natural factors like floods, sea level rise or wildfires. On the other hand, anthropogenic factors, such as deforestation or growth of areas covered by fields, also have a significant and often dominant impact on an LC transition. Figure 1.2 presents an assessment of global LC changes in 2000 [73]. The green area in Figure 1.2 represents wilderness and non-industrial areas that have not been drastically transformed by humans. The areas colored by different shades of red represent the land used by human for hosting infrastructure, producing food, fuel or other goods. The different intensity of red color indicates the rate of changes that have taken place in LC [73]. It can be seen that changes affected a significant part of the Earth's surface.

LC plays a significant role in energy and water exchange between atmosphere and the Earth's surface. The terrestrial areas not only produce the greenhouse gases (such as CO_2), but also absorb them [94]. Therefore, sustainable land management is an important tool for climate change mitigation. The Intergovernmental Panel on Climate Change (IPCC) [94] states in a recent report that the development of appropriate policies can considerably contribute to the climate change adaptation and affect the rate of temperature rise. Some of mechanisms that have already been implemented, confirm the efficiency of this approach [94]. The good examples of this measures are sustainable food production and forest management, food waste reduction, avoidance and prevention of

Figure 1.2: Global LC changes. Adopted from Ref. [73]



deforestation and land degradation. Even more political actions can be adopted. The IPCC [94] proposes several efficient strategies for climate change mitigation such as:

1. Sustainable land use management that includes spatial planning, environmental farm planning, agricultural diversification, management of urban expansion. For instance, "green walls" can diminish the negative effect of sand and dust storms. That would lead to a better air quality and decrease the soil erosion.

2. Standardization and certification of sustainable productions. That can, for example, help consumers make a choice of products that have less impact on the environment.

3. Facilitation of transfer of knowledge and technology and its utilization.

4. Investments in the land and ecosystem restoration. The IPCC expects that return on these investments will be significant, because it will lead to more efficient production.

Nevertheless, in order to develop efficient policies, it is important to understand how different changes in LC affect local and global climate [73]. Researchers pay special attention to the importance of long-term monitoring of various types of LC transformations and their relation to climate changes [48, 77, 82, 94]. The IPCC [94] highlighted the lack of researches regarding the LC conversion and its impact. In our thesis, we aim to address this issue and consider machine learning (ML) methods as an approach to studying the impact of LC changes on climate.

One of the peculiar features of climate science is the accumulation of enormous amount of data. The estimated size of climate data exceeds ten petabytes and continues to grow exponentially [33]. Furthermore, the number of data sources also increases. Initially, information is collected by thousands of ground-based weather instruments all over the world, such as weather station, as well as by a large number of satellites that perform measurements from kilometers above the ground. Then these data are processed and transformed to the standard formats that makes gathered data comparable with each other. Some of information requires aggregation and labeling. For example, LC data can be observed from satellites and can be represented as the photograph of the surface. However, it can be hard to perform the analysis by a picture. An even more complicated task is to compare results of such an analysis from different studies.

Therefore, the standard number notions of LC types have been defined by the climate communities, for instance, the IGBP-MODIS classification system. Despite the standardization of climate information, it still remains difficult to analyze. One of the reason is the spatio-temporal dependence of data. The spatial type of dependence means that processes at a certain location also affect neighbouring locations. Events occurring in the same period of time can likely determine events in consecutive time, demonstrating temporal dependence of data. Another reason of challenges in standardization of climate information is that there are a lot of processes happening in climate systems. They affect each other, can have positive or negative feedback loops and depend on a huge number of variables.

The first attempts to describe climate data sets with mathematical equations were made at the beginning the of 20th century [74]. However, the equations were so complicated that numerical predictions of weather were impossible within a reasonable time until the computer era. The first computer-assisted weather forecast was based on the simplified atmospheric governing equations [74]. Later, an increase in computing power allowed developing mathematical models called climate models that can identify monthly and seasonal patterns. Thereafter, climate models became more and more complex and include more processes and variables.

Nowadays, simulations based on climate models are the largest source of climate information [33]. They allow researchers to model a climate response to some specific changes in a climate system. To perform an experiment, one should run a climate model with different input variables few times, and then compare the results to understand the impact of these input parameters. For instance, a climate model can demonstrate what kind of changes occurs in a climate system if the input data differs only in

LC. Nevertheless, a result obtained from the climate model can have non-linear patterns that are difficult to identify, and researchers should pay special attention to that. For example, Huang et al. [48] used a statistical method based on a ridge regression to extract from the climate model simulations the impact of precise LC transformation on temperature and precipitation. Nowadays, ML is of special interest to researchers as a powerful tool for such kind of tasks as well as for other problems within climate science [50].

1.2 Problem Statement

ML is widely used in different scientific fields, while it has a limited application within climate science [50]. One of the reason is the spatio-temporal dependence in climate information. For example, LC and temperature data are distributed in time and space. The similar changes that have occurred in different places can have diverse consequences. In addition, some processes and events are characteristic of specific areas. Whereas many of ML tools imply that observed data are independent and uniformly distributed. Thus, ML methods require adaptation or redesign to be applied to climate data. In this thesis, we study the possibility to adjust ML techniques to distinguish the impact of different LC transformations on temperature.

The main objective for this thesis is to find an evaluation method that allows comparing the performance of various ML techniques on the spatially dependent data. The found evaluation method will help to choose the most suitable ML algorithm for climate change analysis. Then we use this ML algorithm to predict the impact of LC changes on temperature changes in Europe. The following research questions should be raised:

1. *How can supervised ML techniques be applied to spatially depended data with a high variability?*

2. *Is it possible to develop a model based on an ML approach, which can predict the impact of LC changes on temperature?*

3. *How can an ML approach help to understand the effects of LC changes on surface temperature?*

Our work is divided into three stages to address the objectives listed above. The first stage implies the development of an evaluation method that allows comparing efficiency of ML approaches on spatially depended data. Practically, in order to examine the method, we design a synthetic data set that simulates data with spatial dependence.

The second stage includes verification of our hypothesis that some ML techniques perform significantly better than others. To do so, we compare performance of a few ML algorithms on the real world data with the evaluation method developed at the first stage of our research.

The third stage is based on the conclusions of the previous steps. An ML technique with the best performance will be used to make a prediction regarding an impact of LC changes on surface temperature. This part of work includes also an interpretation of the ML models because our goal is to understand the impact of LC change on temperature.

1.3 Limitations

The scope of this thesis is to develop the method for comparing the performance of different ML algorithms on spatially dependent data. To achieve this goal, we design a synthetic data set simulating spatially dependent data. On this data set, we examine the efficiency of the developed evaluation method with different data splitting strategies. We limited ourselves to a one dimensional synthetic data set because the data set of more dimensions would require much more calculation time. We also decide to limit the data splitting methods to three strategies due to time-consuming testing procedure and the overall time limitation for this master project.

The use case is to apply the developed evaluation method to four ML algorithms (random forest, least absolute shrinkage and selection operator, multiple linear regression and support vector machine) and find the one with the best performance for prediction of effect of LC change on local temperature. We limite ourselves to four ML algorithms, which are the most promising for our task. The comparison of a large number of ML algorithms can be a thesis itself, so instead we focuse on how ML approach can contribute to understanding in climate science.

In this study, we use the same data set as Huang et al. [48] to compare our findings with statistical methods. The climate model simulations of the temperature response on LC changes require a lot of computational power and time. Therefore, we are limited to LC data for the area of Europe and for two years: 1992 and 2015.

1.4 Main Contributions

This thesis is focused on study of performance of ML techniques as a tool for better understanding the impact of LC changes on surface temperature. Throughout this thesis, we have learned that ML can be an efficient approach in climate science, given the huge amount of data with a high complexity of climate information. However, standard ML tools should not be used blindly, but should be adapted to specific properties of used data.

To achieve the goal of this thesis, we have developed a method that allows evaluating and comparing performances of different ML approaches on spatially dependent data. This evaluation technique is focused on the statistical significance of a difference between ML approaches with low coefficient of determination. This approach provides a possibility to address the research objectives defined in Section 1.2.

The main contributions of this thesis are the following:

1. We develop three new methods for algorithms assessment based on *5x2-fold cross-validation paired t test* (5x2 CV test) that allows assessing performance of ML models on spatially dependent data with high variability. In contrast with other evaluation techniques, the developed methods provides statistically significant results. We evaluate the tests performances with respect to type 1 and type 2 errors and found the best one that we use to compare ML algorithms.

2. Using our technique, we compare performances of four ML algorithms: random forest, least absolute shrinkage and selection operator, multiple linear regression and support vector machine. Random forest regression possesses the superior performance over the other methods for prediction on spatially dependent data with high variability.

3. The random forest algorithm is used to predict the impact of LC transformation on the regional temperature. We find out that the ML model based on random forest regression can help to understand the effects of LC changes on surface temperature.

4. Based on our findings, we have made the predictions on regional impact of LC changes on surface temperature in Europe. To the best of our knowledge, some of these impacts have not been published previously.

1.5 Research method

The research method is a strategy to accomplish the research goal via data collection and analysis. In this thesis, we act in accordance with the design paradigm presented by the Association for Computing Machinery in the work called "Computing as a discipline". [25]. In this article, Comer et al. proposed a framework for the studies in computing. The framework consists of three major paradigms: theory, abstraction, and design.

In our research, we explore a possible application of ML techniques to climate-related tasks. To achieve our goal, we fit our work to the framework as follows:

1. *Theory* implies the study of object definitions and hypothesizes the relations between them. In this thesis, we study the specific properties of climate data and requirements for ML algorithms. We identify the lack of methods for evaluation of ML algorithm performance on spatially dependent data.

2. *Abstraction* involves the hypothesis developing, design of experiment, and analysis of experimental results. A huge part of our work belongs to this paradigm: 1) we design an experiment to test our hypothesis on different data sets, 2) we analyze data gathered from the experiment, 3) we distinguish an ML algorithm with the best performance on the spatially dependent data.

3. *Design*, in this context, means the definition of requirements, system implementation, and testing. In our research, we define the requirements for the prediction of the impact of LC change on temperature. We develop ML models to carry out predictions. Finally, we compare our results with other studies to verify our findings.

1.6 Outline

The thesis consists of the following chapters:

- Chapter 2 - Background: we introduce background information for climate science and machine learning. We pay special attention to the modern methods in climate-related studies as well as to ML algorithms and their evaluation.

- Chapter 3 - Land Cover Change Data: we describe the features of the data in climate science with a focus on the data sets used in this thesis.

- Chapter 4 - Methodology: we present our methodology by describing the possible ways to adapt the evaluating ML techniques to the spatially dependent data.

- Chapter 5 - Experiment: we demonstrate the design of the experiment that includes two experiments. Then, we analyze and compare the results of two experiments. Finally, we discuss how these findings are consistent with other studies.

- Chapter 6 - Predictions: using the results of the experiment we predict the impact of LC changes on regional temperature. We also compare these results with other studies on this subject.

- Chapter 7 - Conclusion: Finally, we provide a summary of our work and propose the directions for further studies.

Chapter 2

Background

This chapter contains the background and motivation of our thesis. In Section 2.1, we introduce the problem of climate change due to anthropogenic factors. Then we examine a special role of the impact of LC transitions on climate. In the second section, we give a brief overview of the main concepts in ML and describe the four algorithms that we use in this work. Next, we discuss different methods for the evaluation of algorithm performance. Finally, we exhibit how ML is currently used in climate science.

2.1 Climate Science

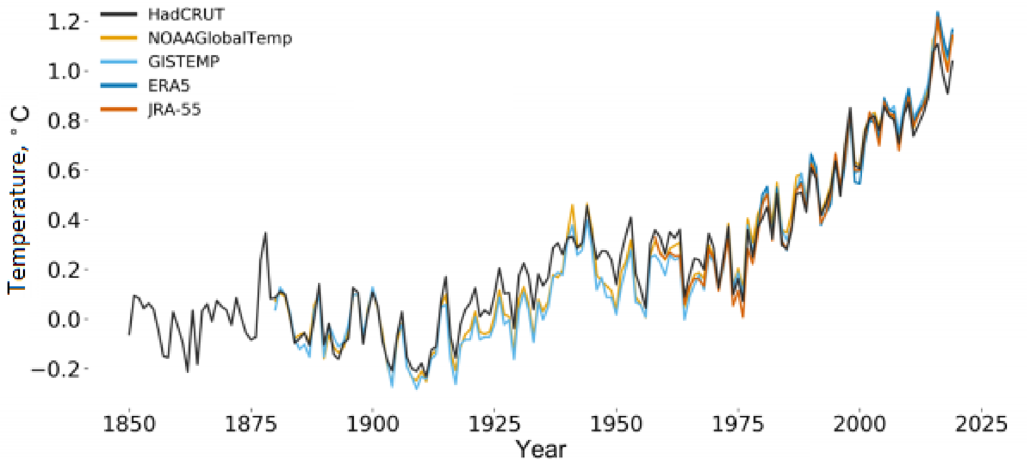
We begin this section with a background for climate science and review of the recent trends in global climate changes and especially the peculiar features of climate changes in Europe. Then we explain our choice of climate changes due to LC changes as a use case. we present the summary of recent publications on climate changes and their driven factors. Lastly, we discuss approaches for prediction of climate changes that are widely used nowadays.

2.1.1 Observed Climate Changes

Climate change has drawn the interest of researchers last hundred years [94]. The number of articles on atmospheric science per year has tripled during the period of 1965 – 1995 [38]. The number of articles per year between 1992 and 2007 has grown even more - by 4.5 times [68]. Moreover, we can expect a substantial increase in climate-related studies since climate changes are accelerating nowadays. New data and new sophisticated research methods allows distinguishing more complex patterns and revealing an impact of different factors on climate.

Starting from the 1980s, every successive decade has been warmer than any previous decade since 1850 [92]. The latest observations show that

Figure 2.1: The global annual mean temperature difference from pre-industrial conditions (1850 – 1900). The different lines in the graph correspond to several in-situ data sets (HadCRUT, NOAAGlobalTemp and GISTEMP) and two reanalysis (ERA5 and JRA55). Adopted from Ref. [92]

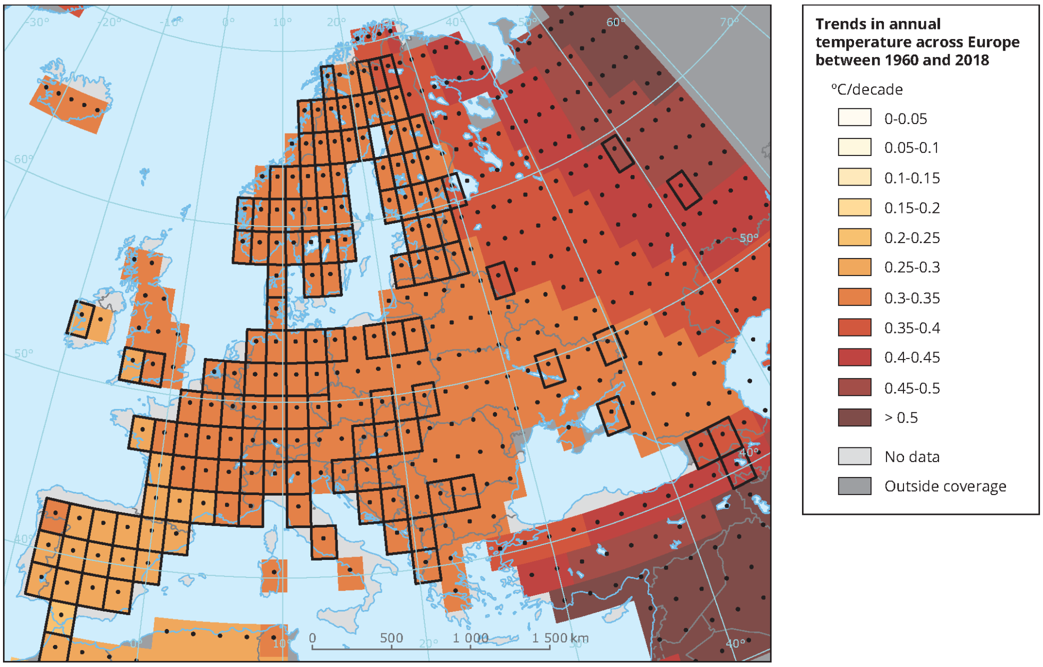


the period between 2009 and 2018 was the warmest decade through the whole observation period of the average annual surface temperature [3]. It was warmer by 0.91 – 0.96 °C than the average temperature during pre-industrial era (1850 – 1900) [3]. Based on the five data sets used by the World Meteorological Organization (WMO) [91], global mean temperature in 2019 was by 1.1 °C warmer than the pre-industrial temperature. 2017 was the warmest year, and 2019 was the second warmest year during the entire time of observation [91]. Moreover, the last 5 years (2015 – 2019) are the warmest years on record [91], indicating an acceleration of pace of global warming. The difference in global annual mean temperature between pre-industrial and industrial eras is plotted in Figure 2.1. This information is captured from several in-situ data sets (HadCRUT, NOAAGlobalTemp and GISTEMP) and two reanalysis (ERA5 and JRA55) [92]. Figure 2.1 also contains the average temperature for the first 10 months of 2019 [92].

The predictions from widely-used climate models also forecasts a further growth of global average temperature for the period of 2071 – 2100 compared to 1971 – 2000 [75]. The expected temperature increase depends on used scenarios for emissions of greenhouse gases. However, even for the lowest emission scenario, scientists estimate temperature growth of 0.3 – 1.7 °C [75]. The temperature growth should be even more prominent for a higher emission scenario. In this case, climate models predict that the increase in temperature should be between 2.6 °C and 4.8 °C [75].

The European region is especially important area for climate studies because temperature changes there exceed the global average trends [19]. During the period of 2009 – 2018, the average surface temperature in Europe was higher by approximately 1.6 - 1.7 °C than in pre-industrial period. It is much higher than the global mean temperature increase. However, changes in Europe are also not uniform. Figure 2.2 demonstrates the changes in annual surface temperature in Europe during the period of 1960 – 2018. The areas indicated by black lines are more representative because they contain three or more meteorological stations. The black dots mark areas with a significant long-term trend in temperature growth [3]. One can notice that the long-term trend in temerature growth is observed almost all over the map.

Figure 2.2: Trends in annual temperature across Europe between 1960 and 2018, °C per decade. Adopted from Ref. [3]



The climate warming in Europe has also been predicted by climate models [52]. Moreover, it is also expected that the average temperature in Europe will grow faster than global temperature [52]. The temperature in this region will grow by 2.5 – 5.5 °C in case of the highest emission scenario for the period of 2071 – 2100 compared to 1971 – 2000 [52].

Given the indisputability of climate changes and the further expected temperature growth, it is critically important to understand the main reasons for this process.

2.1.2 Climate Change Drivers

Climate is a complex system and is affected by many interrelated factors. The first trends in climate changes have been revealed many years ago [63]. However, at the beginning, researchers have mainly studied single time series – changes in global mean temperature over time [63]. The detection and attribution of these changes were main objectives of research. At that time it was impossible to distinguish anthropogenic factors from other causes of global warming.

One of the first assumptions on the human-related impact on global warming was made at the beginning of the 1980s [42, 76]. In the 1990s, it was shown that the real growth of global mean temperature in 1867 - 1982 was noticeably faster than the expected increase estimated with fluctuations in global mean temperature [109, 118]. Subsequent studies have concluded that the observed climate changes are the result of both natural factors and anthropogenic activities [46, 105, 108].

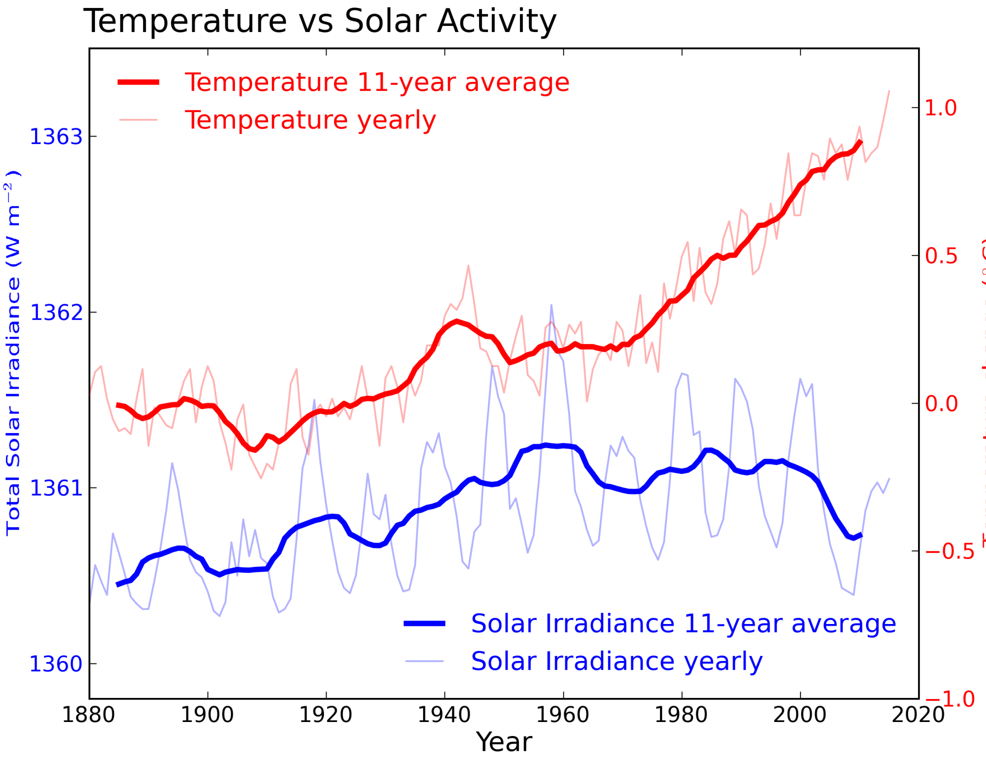
Until the middle of the 19th century, natural factors had the dominant impact on global mean surface temperature. Among them, one can distinguish the following factors:

1. **Fluctuations in solar activity** [21, 22, 45, 66]. Solar activity varies from quiet to stormy during the 11-year cycles. Moreover, solar irradiance can differ from cycle to cycle. The variation in solar activity was one of the main reasons of climate change during the pre-industrial era. However, the pace of the current temperature change cannot be explained only by solar activity. It can be clearly seen in Figure 2.3 where the red line represents global surface temperature (in degrees Celsius) and the blue line represents solar irradiance (in watts per square meter) received from the Sun. To eliminate the cyclic variability, thicker lines show average data for 11-year cycles [87].

2. **Volcanism and related aerosols** [20, 29, 66]. Volcanic eruptions lead to emission of a tremendous amount of ash and gases into the atmosphere. If eruptions are intense enough and eruption products reach the stratosphere, it can result in significant climatic cooling. Volcano-related cooling can possess a positive feedback loop with further cooling [20]. The volcanic activity was one of the main reason for a significant cold interval known as the Little Ice Age [21]. However, dissipation of eruption products usually lasts only 2 years [21]. Therefore, volcanism has a low impact on the latest climate changes and global warming.

3. **The natural greenhouse effect and water vapour** [57, 66, 114]. *Greenhouse gases*, such as H_2O, CO_2, CH_4, are characterized by high transparency in the visible range but by high absorption in the middle and far infrared ranges of electromagnetic radiation spectrum. The presence of such gases in the atmosphere captures the heat coming from the Sun and leads to the *greenhouse effect* and the Earth's surface warming [75]. Most greenhouse gases have both anthropogenic and natural origins. Among the natural greenhouse

Figure 2.3: Global surface temperature and the Sun's energy. Adopted from Ref. [87]



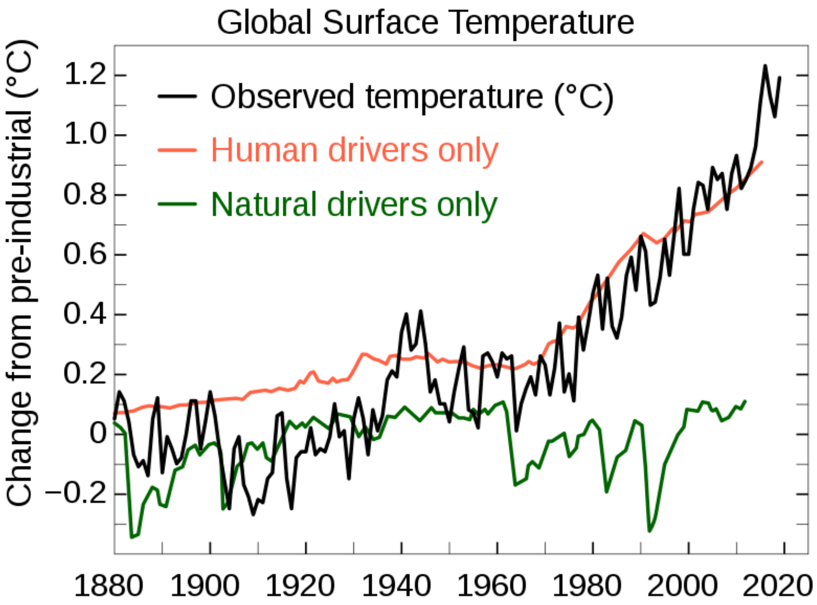
gases, water vapour has the dominant contribution (about 60 %) to the greenhouse effect on the Earth [57]. The intensity of water vapour depends on temperature that complicates its consideration in climate models [78].

The IPCC analysed a significant number of climate-related studies and estimated the different contributions to the recent climate change [19]. The IPCC concluded that the rise in global mean surface temperature is mainly driven by human activity [19]. Changes in global surface temperature caused by anthropogenic and natural drivers are presented in Figure 2.4.

Various human activities cause climate changes. Human-caused emission of greenhouse gases is the most critical anthropogenic factor in temperature trends in 1951 – 2010 [19]. The growth of greenhouse gases concentration in the atmosphere is mainly caused by industrial production and landscape change. It is shown in Figure 2.5 that land use activities have the second largest contribution to global anthropogenic greenhouse gas emissions and have already reached 25%.

In addition to the greenhouse emissions, land surface changes also facilitate other chemical and physical processes that affect climate [94]. For example, each land cover has its own surface reflectivity called

Figure 2.4: Changes in global surface temperature caused by anthropogenic and natural drivers. Adopted from Ref. [59]

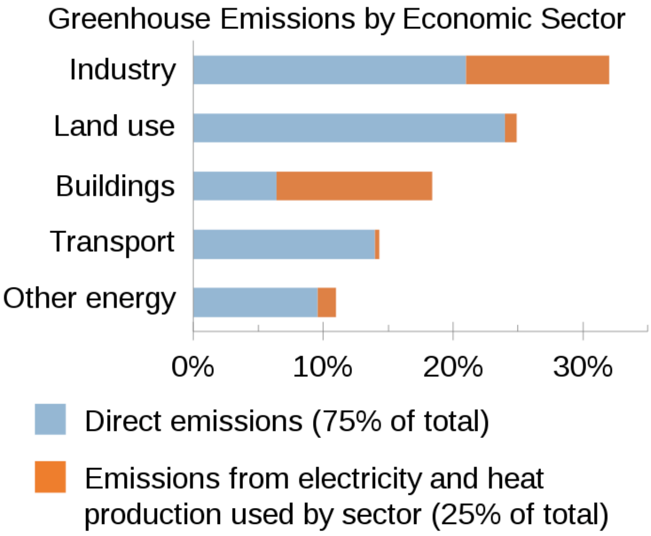


albedo [36]. On average, historical anthropogenic LC changes lead to a growth of global land surface albedo. LC with a higher albedo reflects more radiation and absorbs a smaller part of it. This leads to cooling [94]. LC transition can also affect the wind patterns because of natural obstacles created by some vegetation, for example, by trees. Deforestation also has impact on a cloud formation through the change in emission of different chemical compounds [116]. In the next section, the overall impact of LC transformation on temperature is considered in detail.

2.1.3 Impact of Land Cover Changes on Temperature

Until recently, the main measures for global warming mitigation have been focused on reduction of the fossil fuel combustion. However, the role of land use is starting to attract the scientific attention. For example, the most recent IPCC report [94] was entitled "The Climate Change and Land" which highlights the influence of land use on climate. Nowadays, around 70% of global land surface without ice is used by humankind for farming, urbanization expansion, energy production, etc. [94] The proportion of land used continues to increase following the growth of the Earth's population. It has been revealed that changes in land use can lead to local warming or cooling effect. Moreover, LC transformation can have an impact on temperature in regions located hundreds of kilometers away [94]. However, there is a two-sided process: changes in land use affect climate, but climate changes also have an impact on LC. In addition to that, some changes of LC are driven not only by anthropogenic but

Figure 2.5: Contribution from different economic sectors to emission of anthropogenic greenhouse gases. Adopted from Ref. [19]



also natural factors. As mentioned in the recent IPCC report, very few studies examine the effect of historical LC changes on seasonal climate on a regional scale [94].

In the latest report, IPCC distinguished four main trends in LC changes [94]. First of all, an increase of a cropland area by 15% since middle of the 20th century. Cropland expansion is mainly associated with the decrease of forests and leads to a significant global deforestation. Since the 1960s, forest area has reduced by 5% and continues to decline. For the same period, urban and built-up areas were doubled.

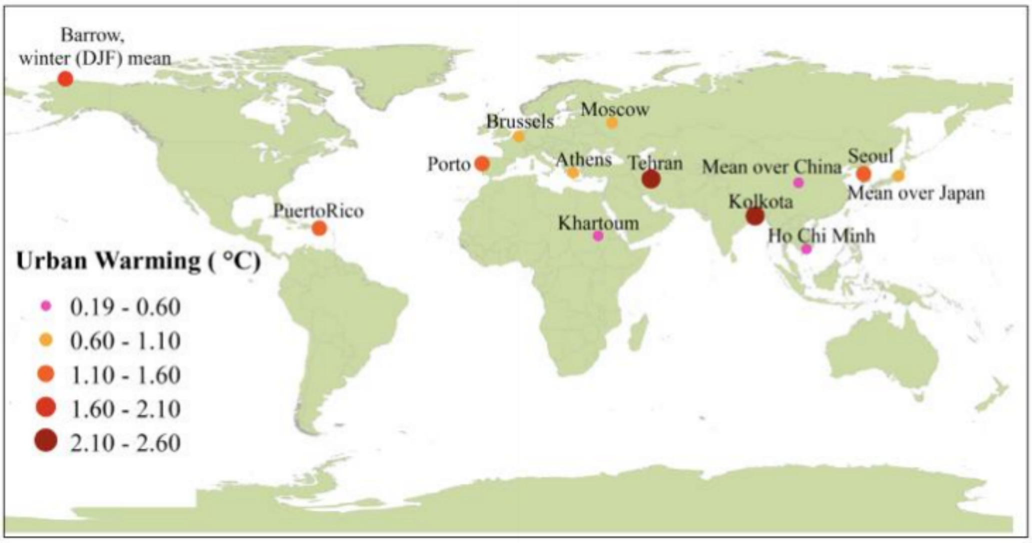
There are many ways to study the impact of LC transformations on temperature. Some works on LC changes and their impact on climate are based on directly observed data [23, 55], while others use mainly simulations and climate models [15, 35, 37]. In general, scientists agree that the climate system is very complex and depends on many factors. The impact of LC change can vary on global and regional scale. Moreover, the same transformations can lead to different consequences depending on the region where it happened. However, a few main trends in temperature change due to LC transitions can be recognized:

1. Deforestation and afforestation have different impacts on temperature depending on latitude [16, 70, 71, 97, 100, 106]. In low-latitude regions, deforestation contributes to the regional warming [16, 70, 71, 97, 100, 106]. There is no consensus regarding the impact on the mean temperature in temperate mid-latitude regions [94]. Some authors observed a low warming effect [16, 70, 71, 97, 100, 106], while others did not find any changes or even decrease in temperature [8, 26, 94]. At high-latitude, the effect of deforestation also depends on

longitude but mainly leads to local cooling [8, 16, 70, 71, 97, 100, 106]. Afforestation has an opposite effect on temperature trends [16, 71, 97, 100].

2. Urbanization is considered by many authors as a contributor to the regional and global warming [17, 48, 94, 95]. The temperature in cities and the surrounding areas grows by 0.19°C – 2.60°C per year [94]. The increase in annual mean temperature due to urbanization is shown in Figure 2.6 for different urban areas. Some authors also observed the cooling effect of urbanization in a warm dry climate [112].

Figure 2.6: Changes in annual mean surface temperature due to urbanization. Size and color intensity of the circles correspond to a degree of temperature change. Adopted from Ref. [94]



3. Abandonment of agricultural land is a huge trend typical for the European LC transformation [4, 48, 61, 67]. Recently, Huang et al. demonstrated that the decline in cropland in the Central Europe contributes to the regional cooling [48], while a similar LC change in the Eastern Europe leads to a temperature growth [48].

4. Land greening in boreal regions was firstly identified by Myneni et al. [85] in 1997 and then has attracted significant scientific interest [32, 40, 56, 94]. Shrubs and trees expansion contribute to an increase in above-ground biomass and land greening that can be observed from space [96]. Researchers mostly agree with the strong correlation between shrubbing and regional warming [13, 19, 62, 83, 84]. However, some studies consider this correlation as a result of a feedback loop [31, 83, 84]: an increase in temperature facilitates a

growth of shrub species, and then shrubbing leads to a reduction of surface albedo that contribute to warming. The studies that consider only the impact of LC transition to shrubs show that an increase of shrublands in the Arctic enhances warming [11, 13, 62].

Obviously, different LC changes have a unique effect on climate [30]. Most publications are focused on some individual LC changes, for example, deforestation [37, 64] or urbanization effects [54, 117]. However, this question is rarely studied in a broad perspective, taking into account all types of LC transitions [48].

Very recently, Huang et al. have published a study in Nature Communication [48] on the regional impact of cumulative LC changes on climate. The key point of this study is the analysis that takes into account all types of LC simultaneously and only then distinguishes the individual impact of different LC changes [48]. The LC transformations are spatially dependent, so that LC changes in one location can affect neighbouring areas. Therefore, it is worth simulating a climate response to complex LC transformations. To distinguish an individual effect of different LC changes, Huang et al. in Ref. [48] developed a new statistical method based on a ridge regression. Their promising approach was based on prediction of the impact of complex LC changes and then on identifying patterns for individual LC changes. In the this thesis, we use the same data sets as Huang et al., but we also use ML to distinguish an individual effect of different LC changes.

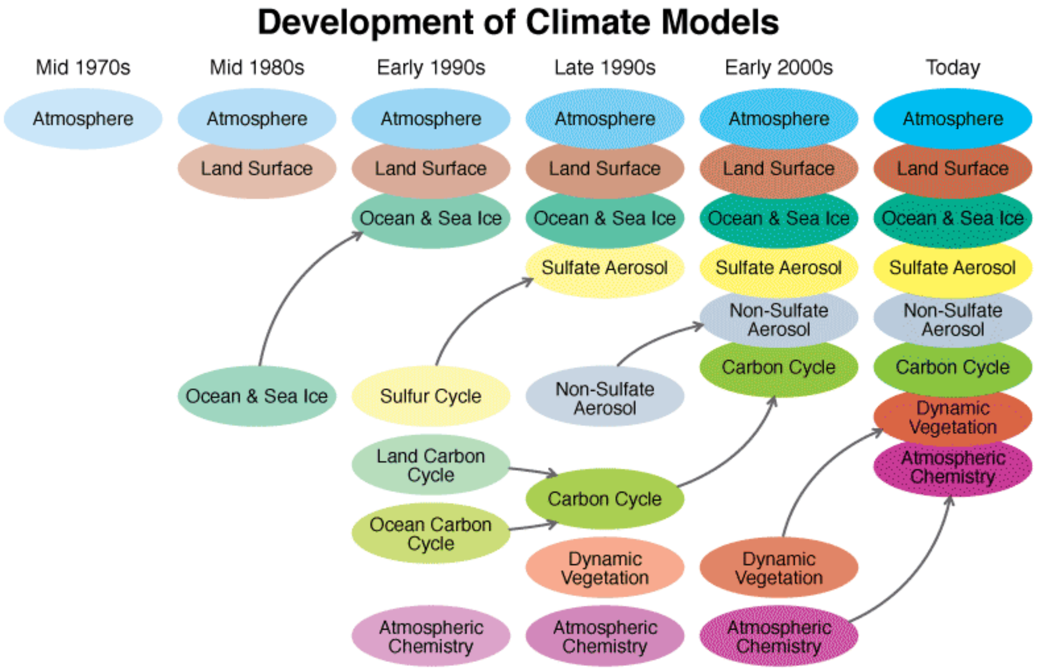
2.1.4 Climate Models

The study of the impact of LC changes on climate can be divided into two main steps. The first step is gathering and classification of data on LC change. The second step is to study the dependence and interrelation between LC changes and climate changes.

There are plenty of different climate change hypotheses proposed by scientists. However, it is challenging to unambiguously verify them since it is impossible to perform a controllable experiment on the whole planet and then to observe the results. Nevertheless, the huge number of empirical observations gathered by researchers can help in our understanding of the climate system. These data can be used for testing and verifying the climate change hypothesis [63]. Nowadays, researchers often use climate models to perform a simulation instead of experiment. In this thesis, we consider only *numerical climate models* that simulate the interaction between essential drivers of climate with the quantitative methods.

Modern climate science is mainly based on numerical weather prediction. In 1901, Cleveland was the first scientist who assumed that processes in the atmosphere are determined by thermodynamic and hydrodynamic

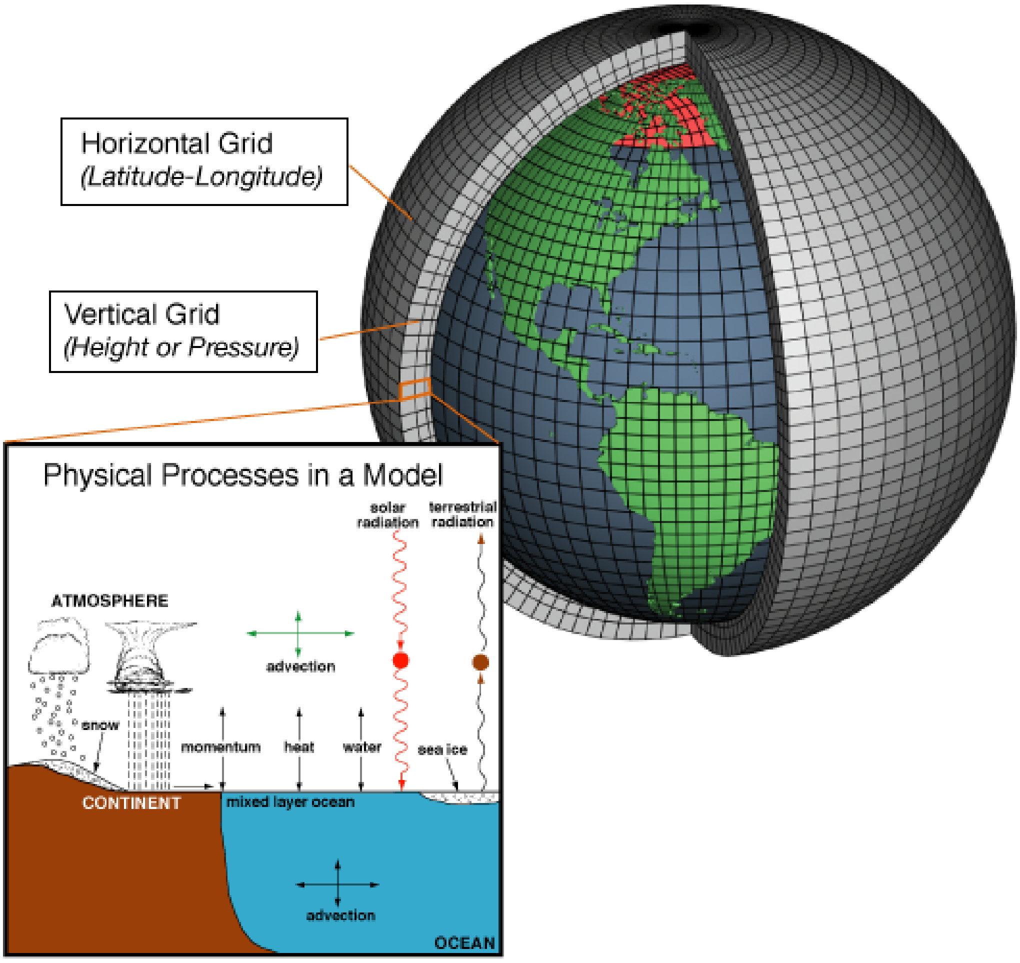
Figure 2.7: Evolution of complexity of climate models. Adopted from Ref. [86]



principles [1]. In 1956, the first climate model describing monthly and seasonal patterns in the troposphere was developed by Phillips and Norman [98]. The work mentioned above [1, 98] laid the foundation for subsequent more complex climate models. Their development was stimulated by a tremendous growth in a computation power, which allowed including much more parameters in the system and considering new processes. Figure 2.7 illustrates the development of climate model over time and shows the processes, which can be taken into account during simulation.

Nowadays, the study of climate with mathematical models is one of the main methods for climate research. The global climate models are intended to simulate the global climate of the entire planet. In simulation, the atmosphere is divided into three-dimensional grid (latitude, longitude and altitude). Mathematical equations describe the fluxes of mass and energy between the cells of this grid [41] and simulate processes listed in Figure 2.7 within the cells. The global climate model is schematically represented in Figure 2.8. The latitude-longitude resolution for different global climate models is typically between 25 km and 250 km [58]. However, Klaver et al. recently demonstrated that a 200-km grid is the most effective resolution in terms of veracity of results of global climate models [58].

Figure 2.8: Schematic representation of Global Climate Model [88]



The regional climate models function quite similar to the global climate models but are limited by a region of interest [41]. They provide more detailed information on the region scale compared to the global models, and their horizontal resolution is typically between 10 and 50 km [41].

A higher complexity of the climate models and a high number of variables allows performing more reliable and realistic simulations. On the other hand, it also makes the climate models less interpretable and also accumulates errors from different processes simulated in a system [63]. Therefore, it is necessary to carefully analyze the results of the climate model simulations to find dependencies and interrelations in these data. In this thesis, we use the climate model simulations performed by Huang et al. [48] to study the impact of LC changes on temperature. Huang et al. efficiently used an approach based on a ridge regression, and this allows us to presume that other regression approaches can also be successfully applied to this task [48]. Therefore, in this thesis, we use the ML regression algorithm to study the impact of different LC changes on surface temperature. In the next section, we introduce the application of ML for the climate change prediction.

2.2 Machine Learning

In this chapter, we present the background and motivation for using ML as a tool for climate change prediction. First of all, we give a brief overview of the major concepts in ML. Then different ML models are discussed. We also introduce how ML models can be evaluated and compared with each other. At the end of this chapter, we show various fields in climate science where ML has already been actively applied.

2.2.1 Machine Learning Algorithms

ML is a branch of AI that uses statistical learning methods to develop the ability of algorithms to "learn" from data without prior assumptions or only with few of them [80]. Probably one of the most known definition of "learning" in context of AI was made by Mitchell [79]:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

The main goal of ML is to develop a model based on the given input that will provide a required output. Models can differ from each other depending on a chosen ML algorithm and input data. Some of them can be easily interpreted, while others will work as a "black box". That means that we know only the input and output, but nothing about internal mechanics of a model. Nevertheless, all models should provide correct output for given data, but the accuracy of results should be evaluated [80].

Generally, ML can be applied to various types of tasks, and their input/output data are also quite different from each other. In the next sections, we describe the main properties and requirements of these approaches.

Main concepts in machine learning

First of all, we are going define the main notions related to ML and used in this thesis. One of the main elements of ML are *observed data*, which usually consist of pairs of *independent variables* X (also called *input variables*) and *dependent variables* Y (also called *output variables*). The observed data can be divided in two groups - labeled and unlabeled data. The *unlabeled data* consist of a raw information without explanation, gathered from the world. For instance, it can be photographs, video and audio records. The *labeled data* consist of a set of the unlabeled data with explanation, description or assigned meaningful class. An example of the labeled data is a photograph coupled with information about what is shown on it.

Let us illustrate the above definitions for our case of using the impact of LC transitions on surface temperature. Our observed data are temperature and proportion of LC in a certain cell on the grid (LC data). These data are labeled because we associate LC data with the temperature in a certain cell. Due to the problem statement, LC datum is the independent (or input) variable and temperature is the dependent (or output) variable. However, it can be reversed for some specific tasks, for instance, if we would like to study how temperature affects changes in LC.

Initially, we assume that there is a relation between X and Y, and it can be described by an unknown function usually called a *target function* $f : X \to Y$. The goal of the ML method is to find a mathematical function (also called a *model*) $g : X \to Y$ that approximates f. This can be done using an *ML algorithm* that we define as a combination of techniques and operations taken with aim to produce the g model [80]. In this thesis, we determine *training* as a process when an algorithm develops a model based on the observed data. In our use case, we consider the real process of the LC change influence on temperature as the target function f. The goal of this thesis is to develop the g ML model that will approximate the real process of the LC change influence on temperature. In our case, the training is a process where the chosen ML algorithm develops the g ML model based on the observed LC data and temperature.

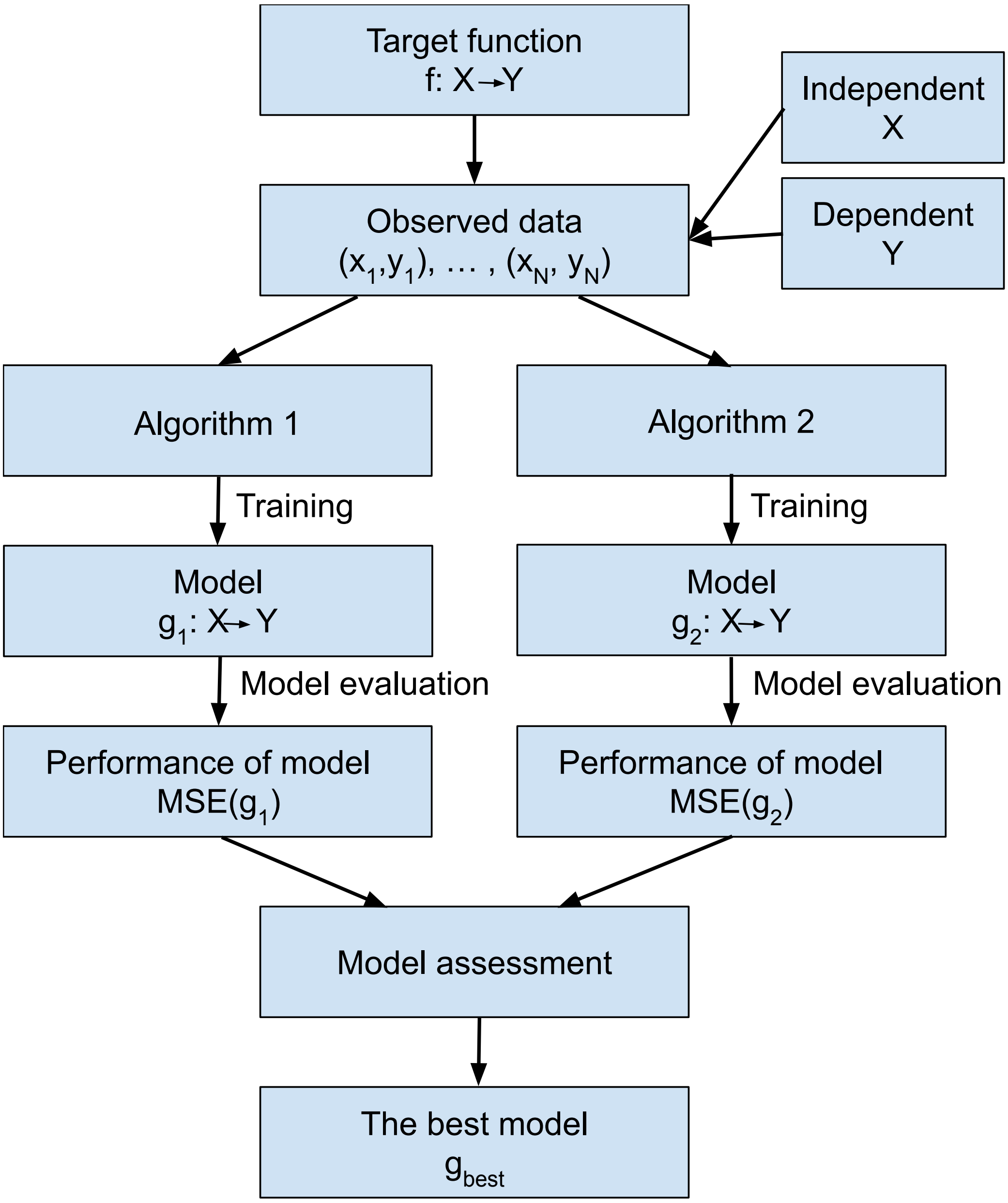
Another significant definition in ML is a *performance* of an ML model. In this thesis, we define it as a numeric representation of how good the g ML model approximates the target function f; and *model evaluation* is a technique for calculation of model performance. *Model assessment* is a process of comparing the performances of two or more models and searching for a model with the best performance. In our example of the influence of LC change on temperature, we can use Mean squared error (MSE) as a model performance, and the model evaluation process is the calculation of MSE. The model assessment can involve a comparison of two MSEs for g_1, g_2 models that are developed by two different ML algorithms. The model with the lowest MSE is characterized by the best performance. A schematic view of these relations between the described ML concepts is presented in Figure 2.9.

Machine learning paradigms

There are three main learning paradigms in ML: supervised learning, unsupervised learning and reinforcement learning.

The *supervised learning* paradigm implies a mathematical model devel-

Figure 2.9: Schematic representation of relations between the ML concepts.



opment based on the labeled data that include both input and pre-defined output. The data used for the model development are included as compounds of *training data set* that consists of an input vector X and an output vector Y. During learning, the model receives X as the input and provides some output vector Y', which is then compared with the given Y. If $Y' \neq Y$, then the model will adjust its parameters to get better performance. This procedure is repeated until the model no longer improves predictions or until it is limited by the number of possible iterations. A model can be considered as *optimal* one if it provides the correct output for an input that was not included in the training data set [80].

The supervised learning paradigm includes two large groups of algorithms: *classification* and *regression* algorithms. The goal of the first group is to categorize the input data into a limited and predefined number of classes. For example, the classification algorithm can be used to define whether a person has a decease or not, and another example is an e-mail spam detection. On the contrary, the regression algorithms have numerical output values that can be within a predefined range. These algorithms are used to calculate the output vector Y based on the information from the input vector X. The well-known examples of the regression algorithms are goods and stock price forecasts.

In the *unsupervised learning* paradigm, there are no training data, and the input data are unlabeled and unclassified. The main goal of this type of method is to determine a structure of the input data set and reveal the similarities in data [80]. *Cluster analysis* is a typical example of the unsupervised learning algorithm. The goal of cluster analysis is to group elements from the input data that have similar attributes. For instance, it can be used to distribute customers into some groups depending on their behavior.

In the *reinforcement learning* paradigm, a computer program performs actions to interact with a certain environment and collects observations and a reward. The aim is to execute actions that maximize cumulative reward [80]. The ML algorithm should explore which action will lead to the maximum reward. The reinforcement learning methods are separated from the previously mentioned paradigms. On the one hand, this is not the supervised learning since it does not use the labeled data and the training data set directly. It learns from the environment response to the performed action. Moreover, the environmental response can be non-deterministic. On the other hand, this cannot be defined as the unsupervised learning because a target reward is known. The ML model learning actions for a player in Atari games is a good example of the reinforcement learning application. The ML algorithm tries to simulate gamer behaviour and perform different actions to maximize reward. The environment is the Atari game that reacts on the actions performed by the ML algorithm. The reward is the number of points that the ML algorithm receives in the game.

Actually, all the paradigms mentioned above can be applied to climate change predictions. However, it is important to select an approach that is appropriate to a given research objective. In this thesis, we study the LC transition influence on climate, especially on surface temperature. The peculiar properties of the given data are discussed in Section 3. Looking ahead, these data are labeled and numeric. Based on the descriptions of the different ML paradigms and their particular properties given in this chapter, we chose supervised learning approach, namely, the regression algorithms as the main methods used in this thesis. However, this group includes many various algorithms, and each of them possesses some specific advantages and disadvantages. According to the "No free lunch" theorem, no algorithm can be considered as optimal for all types of supervised learning problems [119]. Therefore, we give an introduction to the main types of the regression algorithms in the next section.

Regression algorithms in supervised learning

Initially, a training data set τ is given for regression problems in the supervised learning algorithms. It contains N pairs of the input (independent) variables x^t and the output (dependent) variables d^t, where t is the number of pairs for the training set:

$$\tau = \{x^t, d^t\}_{t=1}^N \tag{2.1}$$

It is assumed that the output variables d^t depend on the input variables x^t and some unknown variables z^t. This dependence can be formally represented as a result of unknown function $f(\cdot)$:

$$d^t = f(x^t, z^t) \tag{2.2}$$

The main goal of the algorithm is to make a model $g(\cdot)$ with a parameter θ that matched the observed input x^t to the output y^t:

$$y^t = g(x^t|\theta) \tag{2.3}$$

The learning process is a search for the parameter θ of the model $g(\cdot)$ that minimizes a deviation of the predicted y^t from the pre-defined output d^t from the training set τ. This deviation is described by the loss function $L(\cdot)$ [5]:

$$\arg\min_\theta \sum_t L(d^t, y^t) = \arg\min_\theta \sum_t L(d^t, g(x^t|\theta)) \tag{2.4}$$

For the regression problem the loss function $L(\cdot)$ is often the mean squared error [5].

Many different algorithms can be used for solving the regression problems. In this thesis, we focus on four main types of algorithms:

- Multiple linear regression

- Least absolute shrinkage and selection operator (LASSO)
- Support vector regression machine
- Random forest regression

Multiple linear regression

Linear regression is the first algorithm for solving regression problems that have been deeply studied and widely adapted for various applications [101]. The linear regression model implies that relations between the input and the output variables $E(Y|X)$ are linear or that the linear relation is an acceptable approximation and can be described as follows [44]:

$$Y = f(X) + \epsilon, \tag{2.5}$$

where ϵ is the additive error term that cannot be directly observed in data. This is often a Gaussian random variable with an expectation of distribution equal to zero and a standard deviation σ: $\epsilon \sim N(0, \sigma^2)$. If we define the input variables x^t as the p-dimensional vector X, then the linear regression model is defined as follows:

$$X = (X_1, X_2, \ldots, X_p), p \in \mathbb{Z}, p > 0 \tag{2.6}$$

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.7}$$

where $\beta_0, \ldots, \beta_p$ are the unknown coefficients or parameters that should be estimated from the training data set. If $p > 1$, then the linear model is called the *Multiple linear regression (MLR)* model [44]. One of the most popular methods of fitting a model is minimizing the least squares criterion. Let us assume that we have N training pairs of (X_i, y_i), where $i = 1, \ldots, N$ and X_i is the p-dimensional vector. We can describe the model parameters as a vector $B = (\beta_0, \ldots, \beta_p)$. We also define $\hat{y}_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$ as the i^{th} prediction of Y from the model $f(\cdot)$ based on the i^{th} value of X. Then the aim is to choose B that minimizes residual sum of squares (RSS) [44]:

$$RSS = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - f(X_i))^2 \tag{2.8}$$

$$= \sum_{i=1}^{N} (y_i - (\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j))^2 \tag{2.9}$$

MLR requires special data for the input variables and the training data set. The essential properties required for MLR are the following:

1. As previously mentioned, this approach assumes linear relations the input and the output variables.

2. The training data sets should contain independent observations because a correlation between the input variables will lead to overfitting. In this thesis, we define *overfitting* as a learning error that occurs when a model learns patterns and noise in the training data so detailed that it has negative impact on a model performance on new data.

3. The input variables are precisely defined and do not imply any errors in their values.

4. As assumed in MLR, the variance of the errors ϵ_i from Equation 2.5 is constant throughout observations and is not correlated.

Despite the wide applications of MLR, there are some issues related to it. One of the issues that researchers faced is that classic MLR approach works non-efficient in situations when two or more independent variables in a multiple regression model are linearly related (see the second requirement above). This property of data is called *multicollinearity*. This is a quite common issue in real world tasks. An example of correlated input variables can be a person's height and weigh. To solve this problem, the LASSO method was developed.

Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator (LASSO) is another type of the regression models, which aims to exclude some of the input variables and hence prevent the overfitting and minimize prediction error. This is done using *L1 regularization* technique that sets some constraints on a model, reducing weight coefficients to zero for less important variables [44]. If a few independent variables are highly correlated, then only one of them is taken into account in the LASSO method, while others will be taken with zero coefficients. The L1 regularisation also helps in feature selection and makes model interpretation easier [44].

The model for LASSO is similar to that for MLR. However, the aim is to choose *B* coefficients that minimize the expression [44]:

$$\sum_{i=1}^{N}(y_i - (\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j))^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.10}$$

where λ is the regularization parameter that is defined separately and adjust the level of constraints, and $\lambda \sum_{j=1}^{p} |\beta_j|$ is the regularization term.

Support vector regression

Support vector regression (SVR) is another example of the supervised learning algorithm for regression tasks. In contrast to MLR, which aims minimizing training errors, SVR is geared to keep an error within a predefined threshold. The main idea is to define the loss function that ignores errors that are less than ε. Therefore, the ML algorithm learns only from a subset of the training data set [9]. In case of a linear target function $f(X)$, the algorithm aim is to select B coefficients in such a way as to minimize theexpression:

$$\sum_{i=1}^{N} V(y_i - f(X_i)) + \frac{\lambda}{2}||B||^2, \tag{2.11}$$

where λ is the tuning parameter and $\lambda \geq 0$, $V(\cdot)$ is the ε-intensive loss given by:

$$V(r) = \begin{cases} 0, & \text{if } |r| < \varepsilon \\ |r| - \varepsilon, & \text{otherwise} \end{cases} \tag{2.12}$$

In case of non-linear $f(X)$, SVR uses mathematical functions called a kernel. Generally, the *kernel functions* convert the given data to the desirable format. In case of SVR, the kernel expands the dimensions of the input variables. For example, let us consider only two-dimensional input variables X_1 and X_2. However, an expression including X_1 and X_2, for instance, $\frac{X_1}{X_2}$ can be more valuable for some predictions. Then the kernels expand the input variables to three-dimensional space containing X_1, X_2 and $\frac{X_1}{X_2}$. Due to the new variables, we can "include" all the non-linearity into $h_m(X)$ functions. Then we can finally define approximation of non-linear regression function $f(X)$ as a linear set of basis functions $\{h_m(X)\}$, where $m = 1, 2, \ldots, M$:

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X) + \beta_0 \tag{2.13}$$

The aim is to choose the coefficients $B = (\beta_0, \ldots, \beta_m)$ that minimize the expression given by:

$$\sum_{i=1}^{N} V(y_i - f(X_i)) + \frac{\lambda}{2} \sum \beta_m^2 \tag{2.14}$$

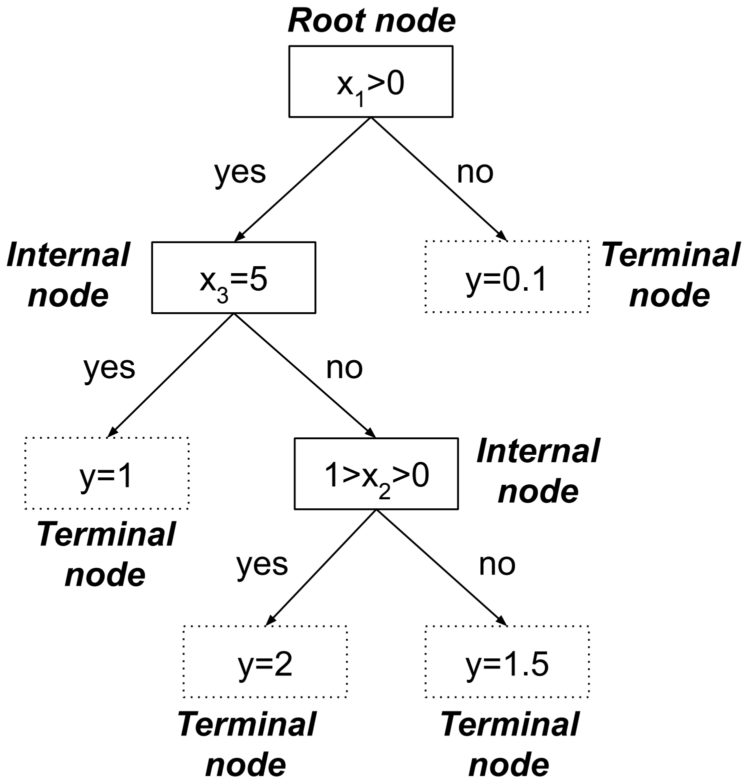
where $V(r)$ is some general measure of error [53].

Random forest regression

Random forest (RF) regression is another example of the supervised learning model. In compare with MLR, RF models are more efficient in capture of non-linearity in data but are harder to interpret.

RF algorithms are based on the decision tree learning approach. In general, a *decision tree algorithm* produces a model, which finds the output values based on a sequential set of rules applied to the input values. Figure 2.10 shows an example of a decision tree that is applied to the input variables (x_1, x_2, x_3) and predicts a value for the dependent variable y. Rules are shown in the root node and the internal node. Terminal node shows the predicted values for the output variable y, which corresponds to the input variables matching rules in the previous nodes. If a rule splits data into only two nodes, then a decision tree is called a *binary decision tree*. During training, the decision tree learns the rules and their order.

Figure 2.10: An example of a decision tree for regression.



The main problem of the decision tree algorithm is a high probability of overfitting [44]. To overcome this issue, RF implies the creation of a bunch of binary decision trees during training and output the mean prediction of different trees. Another important peculiar property of the RF algorithm is randomness. On the one hand, it randomly divides the training data into subsets in such a way that each data set corresponds to a certain decision tree. On the other hand, there are also random subsets of input variables used in building of different trees.

The method of model construction for the RF algorithm is quite different from the previously described algorithms [44]. Therefore, we use an algorithm description to provide a proper description of the RF algorithm. Let us assume B as a tuning parameter that represents the number of decision trees in the RF model. Then we can describe model as follows:

1. For $b = 1$ to B:
   (a) Choose a randomly sampled subset S from the training data set.
   (b) Create a decision tree T_b based on the subset S with a minimum node size n_{min}. To build a tree, the following operations should be done for each terminal node until n_{min} is reached:
      i. Choose randomly m variables from the p-dimensional input vector.
      ii. Choose the best variable from m variables for splitting. To do so, all of m variables should be iteratively chosen as the splitting point, and the loss function should be calculated for each of them. The variable with the lowest value of the loss function will be chosen. For example, it can be defined using RSS:
      $$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.15}$$
      where y_i is the actual value and $\hat{y}_i$ is the prediction from the model.
      iii. Create 2 new daughter nodes.
2. Output the model that will consist of B trees T_b:

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{2.16}$$

2.2.2 Model Evaluation

There is no ML method that can be considered as the optimal approach for all kinds of supervised learning problems [119]. Therefore, it is important to find out the way to evaluate different models and compare their performance.

A determination the best performing model for a given task is an important part of the model development and is usually called as a model evaluation. The main goal is to assess the model performance on future data, i.e. on data out of the given data set. In this section, we introduce two frequently used metrics for measuring the model performance: mean squared error and Coefficient of determination (also known as R^2 score). In addition, we also present one of the highly used validation techniques - Cross-validation (CV).

One of the simplest way to evaluate an algorithm performance is to test the model on data out of its training set. In this case, the algorithm considers this data as "unknown". We can pick these data from the initially given observation. Hence we divide the initial observation data set into 2 subsets: "training data set" and "test data set". The first of them uses for a model learning while the second is for evaluation of the model performance. This approach with the one split of data is called "hold-out" or "validation".

Mean squared error

Mean squared error (MSE) is one of the most popular metric for evaluation of regression models. These metrics show how well the evaluated model fits to the real data through a difference between the predicted and observed values:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2, \tag{2.17}$$

where N is the number of pairs (x_i, y_i) in the test data set, $f(x_i)$ is the prediction given by $f(\cdot)$ model for the i^{th} input variables, and y_i is the observed value. A lower value for MSE indicates a better accuracy of the estimated model [44].

Root mean squared error (RMSE) is a metric similar to MSE, but in contrast to MSE, RMSE allows providing errors with the same units as the original output. RMSE is just a square root of MSE and is given by:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2} \tag{2.18}$$

Coefficient of determination

Coefficient of determination also called as R^2 (R-squared), shows how well a model explains the observed data through a finding the percentage of the dependent variables that are explained by the model. R^2 can be described as follows:

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}, \tag{2.19}$$

where "Total variance" means the variation in the observed data. R^2 can be formulated mathematically as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}, \tag{2.20}$$

where $\overline{y}$ is the mean value of the observed data.

Usually, a higher value of R^2 corresponds to a case when a model replicates observations better. Hence if $R^2 = 0$, the evaluated model predicts worse or similarly as a model that for any input variables predicts the mean value of the output variables in the training data set: $f(x_i) = \overline{y}$. However, the use of R^2 alone is not enough to evaluate the model, and even models with lower R^2 can make meaningful predictions in some cases [81].

For example, R^2 is typically low in behavioral studies because the models implemented there often use only some of the input variables to predict an output. Falk and Miller stated that the models with $R^2 \geq 0.1$ can be considered as acceptable [34]. Cohen [18] distinguished $R^2 \leq 0.12$ as a low value, R^2 between 0.13 and 0.25 as a medium value, and $R^2 \geq 0.25$ as a high. Thus, we can conclude that even a comparison of models with a relatively low R^2 may be reasonable. Therefore, it is important to use these metrics together with others, for instance, with MSE.

The metrics discussed above do not estimate models unambiguously because different ways of splitting the observed data into the training and test data sets can greatly affect estimations. One of the most common technique to solve this issue is CV.

Cross-validation

CV is a framework for evaluating how a model generalizes to the independent data out of its training data set. This technique is especially often used to assess the accuracy of models aimed to perform some predictions. CV implies that the test and training data sets are independent, and the data are identically distributed. Variables are *independent and identically distributed (i.i.d.)* if all variables are independent from each other and possess the same probability distribution. CV is often considered as an efficient tool in evaluation of regression models [6].

The main idea of CV is to form subsets from the given data and then iteratively use one of them at a time as the test set and the others as the training set. For each iteration, evaluation metrics are estimated. At the end, a combined value of iteration metrics, for example, the average value, is calculated.

There are a few different ways to apply CV. We introduce one of the most popular technique, namely, *K-fold CV*. The main idea of K-fold CV is as follows:

1. Split the given data to K subsets of data also called *folds*
2. For $i = 1$ to K:
   (a) Choose i^{th} fold as the test data set
   (b) Use the remaining $K - 1$ folds for model training
   (c) Evaluate the model with a metric using i^{th} fold as the test data set
3. Calculate the mean value for the metric based on all found K values.

CV is not only useful for a more generalized model evaluation. This method can help to recognize overfitting problems that can occur when data are simply divided for once to the training and test data sets. In addition, this technique allows using data more efficiently, which means that even the small data sets may be enough for proper machine learning. K-fold CV can also help to evaluate the uniformity of the data distribution. This can be done through the analysis of validation metrics for different folds. If some folds perform significantly better or worse than others, this may indicate the non-uniformity of the distributed data. We introduce the peculiar properties of climate data in the Section 3 where we also shed light on the CV application in climate science.

2.2.3 Hypothesis Testing

As discussed in the previous section, the evaluation methods such as CV can estimate algorithms performances. However, it can be complicated to verify whether a difference in the performances of algorithms is statistically significant. The significance in this context means that the difference happened not by chance and we perform a *hypothesis testing* to achieve this goal. A *null hypothesis* (H_0) takes place when there is no difference in performance of two ML algorithms, while an *alternative hypothesis* (H_1) arises when two algorithms significantly differ from each other in performance. In other words, for H_1, two regression models learned on the same randomly chosen training data set should have different error rate on the same randomly chosen test data.

An important part of the hypothesis testing is the concept of statistical errors: type I error and type II error. A *type I error* appears when true H_0 is rejected during testing. In our case, it means that the test indicates that one algorithm performs significantly better than the other, but in fact both algorithms have the same performance. The *type II error* implies that false H_0 is accepted. In our context, this means that the test evaluates that the two algorithms perform equally, while in fact one of them is significantly better than the other. The relation between the type I and type II errors is shown in Table 2.1.

	H_0 is True	H_1 is True
Accept H_0	Right decision	Wrong decision - Type II Error
Reject H_0	Wrong decision - Type I Error	Right decision

Table 2.1: Type I and type II errors.

The type I error is associated with a confidence and significance levels. *The significance level* (SL) shows the probability (in percent) that the test result is the type I error. *The confidence level* (CL) is defined as $CL = (100\% - SL)$ and represents the probability (in percent) that the test result will be correct. The type II error is associated with a *power of a test* that is equal to $(1 - \text{type II error rate})$. Power of a test is a probability of rejecting H_0 when it is in fact wrong. The first possible way to reduce the type II error rate is to extend training data set. The second method is to increase the significance level that, however, leads to a rise of type I error rate. In thesis, we have the training data limited by the given initial data set (see Section 3) and prefer to have an acceptable low type I error rate, because our goal is to find the statistically significant difference between the algorithms. Thus, we mainly pay attention to the techniques for minimizing the type I error rate.

5x2-Fold Cross-Validation Paired *t* Test

Statistical testing of difference between ML methods in a CV method is not straight forward. The reason is that data are reused for each fold introducing complex dependencies that disagrees with the assumptions of traditional statistical *t* tests. The *paired t test* is a special *t* test performed on the dependent samples. In our case, it means that two different algorithms will be trained on the same training subset of the observed data and then tested on the test data subset, different from the training one. To reject or accept H_0, the paired *t* test calculates *t*-statistics, and then it is compared with a known critical value for certain distributions and a certain confident level given in Ref. [47]. The *t-statistics* for two algorithms is a ratio of the mean difference in performance to the variance of this difference. To compare the performance of two algorithms on spatially dependent data, we use a validation method the *5x2 CV test* presented in Ref. [28]. This test is specifically developed for comparison of ML algorithms. In addition, it is one of the most powerful tests with an acceptable Type I error rate [28].

The critical value for the *t*-statistics has already been calculated and published [47]. However, it is necessary to know the distributions and the confident level to find out the critical *t*-statistics, and we use 95% of the confident level in the present work. Dietterich [28] claimed that the *t*-statistics calculated for the 5x2 CV test and denoted as $\tilde{t}$ have a *t*-distribution with five degrees of freedom. Therefore, to reject the null hypothesis, the t-statistics $\bar{t}$ for the 5x2 CV test should be higher than 2.571 in accordance to the t-distribution table [47].

To perform the 5x2 CV test, the following operations should be done:

1. Define two algorithms as A and B
2. For $i = 1$ to 5
   (a) Initially observed data should be randomly divided into two subsets S_1 and S_2 of the same size.
   (b) Both learning algorithm A and B should be trained on S_1 and then tested on S_2. Thus, there are two RMSE estimations: $p_A^{(1)}$ and $p_B^{(1)}$
   (c) Then both learning algorithms A and B should be trained on S_2 and then tested on S_1. So, there are two more RMSE estimations: $p_A^{(2)}$ and $p_B^{(2)}$
   (d) Next, one calculate differences between the RMSE estimations of both algorithms A and B for an i^{th} iteration:
   $$p_i^{(1)} = p_A^{(1)} - p_B^{(1)} \text{ and } p_i^{(2)} = p_A^{(2)} - p_B^{(2)} \tag{2.21}$$
   (e) Using these differences, one calculate the variance s_i for an i^{th} replication as:
   $$s_i^2 = (p_i^{(1)} - \overline{p})^2 + (p_i^{(2)} - \overline{p})^2 \text{ where } \overline{p} = (p_i^{(1)} + p_i^{(2)})/2 \tag{2.22}$$
3. The statistics $\bar{t}$ is calculated for algorithms A and B as:
   $$\bar{t} = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} s_i^2}} \tag{2.23}$$
4. Finally, one verify that the $|\bar{t}|$ value is below the given confidence level ($|\bar{t}| \leq 2.571$), then one accept H_0 as a correct hypothesis. Otherwise, H_0 is incorrect and H_1 is correct.

We believe that the 5x2 CV test is a proper framework for the data testing with high variability. However, this method has a few limitations. First of all, CV approaches are based on the assumption that the observed data is i.i.d., which often is not the case and for example not for the spatio-temporal dependent LC and temperature data considered in this thesis. Another limitation is that the 5x2 CV test can fail if the error rates measured in the various 2-fold CV vary greatly [28]. In the chapter 4, we discuss the possible strategies to adapt the 5x2 CV test framework to the spatio-temporal dependent data.

2.2.4 Machine Learning in Climate Science

Nowadays, ML has applications in many different scientific fields. Climate scientists also pay their attention to the artificial intelligence (AI) and ML as useful tools in climate analysis. Nevertheless, ML and AI are not yet often applied in this field [50]. Traditionally, physical modelling based on theory-driven approach has been the main technique in climate science. On the contrary, ML approach has a data-driven nature and hence can be considered as method opposite to physical modelling. Previously, computer architecture and poor performance in data-intensive tasks limited an application of ML. However, this is not an issue because of the constant growth in computational power so physical modelling and ML are mostly considered as two scientific approaches complementary to each other [102].

Studies in climate science can be characterized by the following peculiar properties:

1. **Data intensity**. Climate science has collected an enormous amount of data over a long period of time. Often, the climate change tasks take into consideration hundreds of years of observations of temperature, air humidity, precipitation, LC change etc. More than ten petabyte of climate data have already been collected and are available for analysis. In addition to that, more than 400 terabytes of new data are gathered by sensors every year [2].

2. **Interdependence complexity**. Climate is a system that is determined by many interconnected processes. The main issue is that not all the processes can be described in detail, many climate processes are interconnected, forming a positive feedback loop. Therefore, it happens that modern climate models are so complicated that it is very difficult to explain their results theoretically.

3. **Non-linear behaviour**. Some processes in climate system cannot be described as linear functions of variables. The revealing of non-linear processes is a difficult task.

All the properties mentioned above characterize climate science as a field, which is suitable for studying with the ML tools. In general, we can define ML as a part of AI that allows computers to learn dependencies and relations between different parameters from given data. However, ML can be applied to different types of tasks. In a recent article, Huntingford et al. distinguished three main types of problems that can be solved with ML and AI [50]:

1. Dimension reduction in equations used in climate models can be done with ML. This can help to identify the main interactions and dominant impacts.

2. Consideration factors that should have an impact on climate but do not yet have mathematical representations. Sometimes researchers may simply assume that specific events or parameters affect climate system. However, it can be difficult to formulate them in mathematical form describing events properly. ML is able to solve this issue.

3. Identification of unknown factors, hidden patterns and dependencies within the known dimensions. This task is especially actual because ML can reveal much more complex patterns than a human can observe in data sets [102]. In particular, this task is actual for detection of patterns for the tipping points - the part of a system where small changes in some parameters significantly affect the behaviour of entire system. LC transformations is one of the tipping points for climate system [65]. In our thesis, we try to apply ML to solve such a problem.

The amount of climate data available for analysis has grown drastically during last few decades. However, predictive ability did not grow proportionally to the amount of data. In this case, ML seems be a powerful tool for improving predictions [102]. Therefore, recent works are often focused on an application of ML and AI for different subjects in climate science. For instance, in the Earth system modeling, Rodriguez-Galiano et al. use ML to perform an LC classification of data observed from satellites [104]. ML is also used for detection of past and future consequences of climate changes. For example, Wu et al. applied recently ML for studying the estimated impact of climate changes on forest aboveground biomass [120]. Liu et al. used ML for detection of extreme events based on patterns extracted from labeled historical data [72].

Probably one of the most power ML application in climate science is studying *teleconnections*, which are defined as hidden connections between components of climate models. That is challenging to extract them with standard approaches because of multidimensional nature of climate models. Here, ML can help in improving fundamental understanding in cases when output of climate models is difficult or impossible to interpret with classical approaches [102]. Boers et al. applied ML to find out the global patterns of extreme events such as extreme-rainfall [10]. Another example of this use if ML is the work by Yang et al. where they used ML to study dependencies between surface temperature and the Pacific Decadal Oscillation [121].

Despite the fact that scientists pay more and more attention to ML and AI as the tools for detection of teleconnections, there is a lack of studies dedicated to the impact of LC changes on climate change [94]. Climate models are very complicated, and it can be difficult to make dimension reductions to identify, for example, the impact of such a specific factor as LC changes on another parameter like temperature [50]. We believe that ML can be an efficient way for this kind of tasks. However, there are plenty of different AI and ML methods and algorithms that can solve the same problem. Some methods may be ineffective in tasks depended on spatial or temporal data. It is often required to adjust the standard ML tool to climate analysis [102]. Therefore, it is critically important to choose an appropriate solution that would perform most efficiently.

2.3 Summary

In summary, the LC change and its impact on climate is an important research task in the frame of climate science. There is a huge amount of climate data available on global and regional scale, and a climate model can simulate the effects of LC changes on temperature. However, the results of climate model simulations are often so complicated that it becomes difficult for human-experts to reveal patterns there. Nowadays, there is a lack of studies that considered this issue in a broad sense, i.e. the cumulative impact of all LC on climate. Thus, it is worth exploring and approving new methods and techniques to this research task.

The ML techniques seem like an appropriate tool for the climate change studies because they help to find out the hidden patterns in complex data. It is important to emphasize that most of the standard ML techniques require i.i.d. data. Supervised ML performs most effectively if the models are trained on large data-sets, which should consist of labeled data. In turn, regression algorithms require numeric data. In the next chapter, we consider the peculiar properties of climate data and discuss how climate data are consistent with these requirements.

Chapter 3

Climate Data

In this chapter, we introduce the major properties of the LC data and the climate model simulations that are used in our experiment.

Climate science has a lot of different data sources and produces dozen of petabytes of data. Moreover, the amount of data grows exponentially and is expected to exceed hundreds of petabytes already at 2030 [93]. However, not only the amount and complexity of the available data are the challenges for a researcher.

One of the biggest issues related to the climate data is spatio-temporal dependencies in these data. This property is simply described by the first law of geography [111]:

> Everything is related to everything else, but near things are more related than distant things.

Spatio-temporal dependencies imply auto-correlation and cross-correlation within the input variables. However, not only variables depend on the surrounding area and time points, but also patterns and phenomena evolve over space and time. Moreover, some events happen only in a specific region and/or time period, while the standard methods of the supervised ML require i.i.d. input variables. Hence this property of data limits the possibilities of blind application of the standard ML methods to research tasks in the field of climate science [33].

Another peculiar property of climate data is high variability and uncertainty. First of all, variability is an integral part of climate, because of natural fluctuations. Secondly, the huge number of sensors and weather stations located in different places have different measurement errors that make a contribution to the data uncertainty. In addition, simulations made by climate models are the dominant source of climate data nowadays. However, climate models also introduce some additional uncertainties due to errors in the model simulation results [33]. The variability and uncertainty of climate data limit the possible ways to use the standard ML techniques. In addition, this complicates the comparison of the different

ML techniques, since the evaluation metrics highly depend on how the data is split into the test and training data sets.

3.1 Land Cover Data

The LC data is collected by various observation systems. However, the available LC data sets from different systems are often incompatible and have limited observation periods. To solve this issue, the European Space Agency (ESA) has produced the detailed global LC maps for the period from 1992 to 2015 as a part of the Climate Change Initiative (CCI) [99]. These maps have a spatial resolution of 300 m and contain 37 LC classes from the United Nations LC Classification System (UNLCCS) [27].

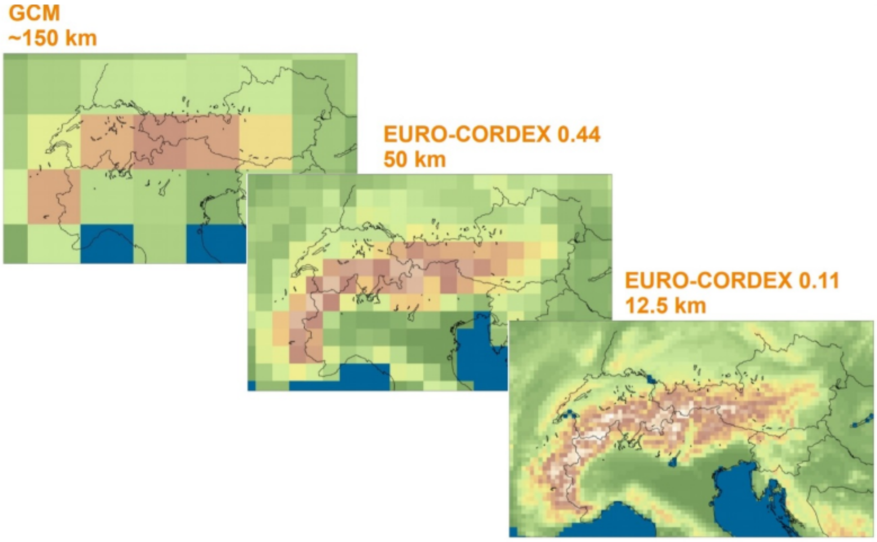
The results of our work should be comparable with the results obtained by other scientists. Huang et al. [48] transformed 37 UNLCCS LC classes to the more commonly used IPCC LC classes - the IGBP-MODIS classification system, which used in this thesis. The IGBP-MODIS system consists of 21 categories that are described in Table 3.1.

Land Cover Category, L-parameter	Land Use Description
1	Evergreen Needleleaf Forest
2	Evergreen Broadleaf Forest
3	Deciduous Needleleaf Forest
4	Deciduous Broadleaf Forest
5	Mixed Forest
6	Closed Shrublands
7	Open Shrublands
8	Woody Savannas
9	Savannas
10	Grassland
11	Permanent Wetland
12	Cropland
13	Urban and Built-Up
14	Cropland/Natural Vegetation Mosaic
15	Snow and Ice
16	Barren or Sparsely Vegetated
17	Water
18	Wooden Tundra
19	Mixed Tundra
20	Barren Tundra
21	Lake

Table 3.1: IGBP-MODIS classification system

These LC data were used to perform the simulations on a regional climate model. Due to limitations related to computational time required for simulations, only the LC data for 1992 and 2015 years were considered in this thesis.

Figure 3.1: Cordex-EU resolution in compares with Global climate model grid resolution. Adopted from Ref. [39]

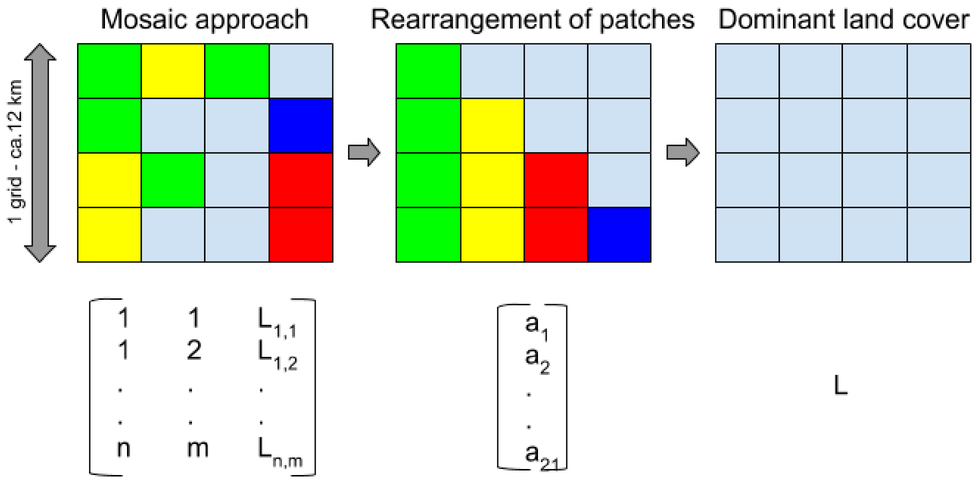


In compare with the global climate model, regional climate models have more detailed grid resolutions. The grid resolutions of the global climate models are about 150 km, while the grid resolutions for regional climate models are presented in 2 dimensions: about 50 km and about 11 km. The comparison of the different resolutions is shown in Figure 3.1. The highest resolution for the chosen regional climate model is horizontal resolution of 0.11° [48]. Therefore, the LC maps from ESA were aggregated to this scale. As a result, the data appear as a grid with a resolution of about 12 km and with a few types of LC in one grid cell. There are a few approaches to present the LC information in the grid. The three main strategies are the dominant LC strategy, rearrangement of patches, and the mosaic approach [82]. The schematic representations of these approaches are shown in Figure 3.2.

The main basic approach is the strategy of *dominant LC* where it is assumed that the whole cell is completely filled with the LC type, which is dominant in this cell. Hence each cell is described by only one parameter $L \in \{1, 2, 3, .., 21\}$ from Table 3.1.

The *patch rearrangement* strategy provides more information on LC in each cell because it records a percentage of each LC type in the cell. Thus, this type of data is stored as 21-dimension vector with components a_i representing a percentage of the area occupied by LC with number i from Table 3.1:

Figure 3.2: Schematic view of different strategies for treating LC information



$$a_i \in \{a_1, a_2, a_3, .., a_{21}\}, \quad 0 \leq a_{\mathrm{i}} \leq 1 \quad \text{and} \quad \sum_{i=1}^{21} a_{\mathrm{i}} = 1 \tag{3.1}$$

The data, which are stored with the patch rearrangement strategy can be transformed to the dominant LC type of data by choosing the L-parameter with the highest percentage, simplifying the data complexity.

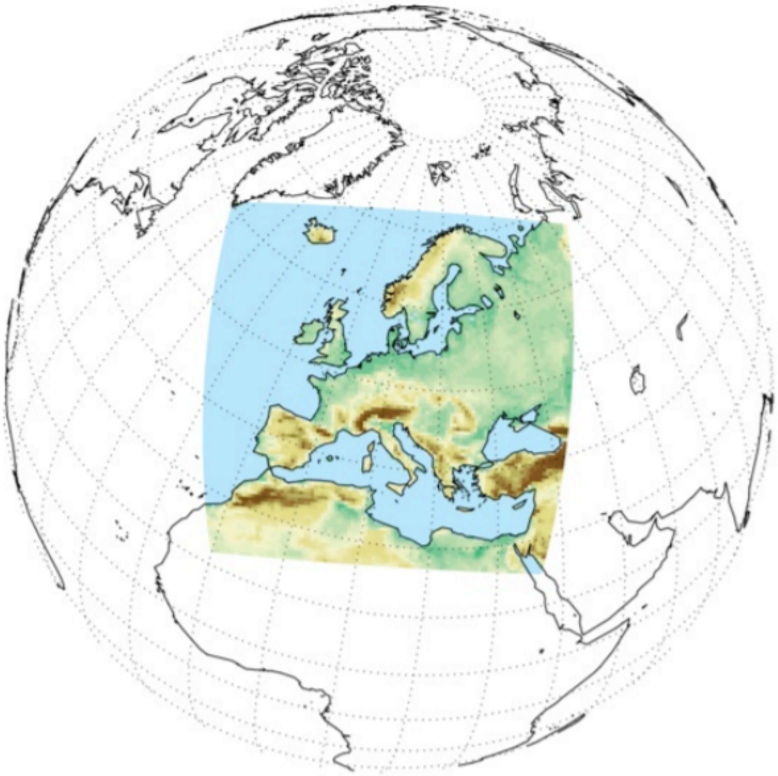
The *mosaic approach* is the most detailed strategy when each cell is split into sub-grids with the LC data resolution. Thus, we know not only a percentage of each LC in the cell but also its location. The data can be represented as a matrix with information about the location and the LC type $L_{i,j} \in \{1, 2, 3, .., 21\}$, where i, j are the numbers of rows and columns in matrix. The data presented with the mosaic approach can be transformed into both previous types of data storage reducing the data complexity.

The data used in this thesis store the LC information using the patch rearrangement strategy. Therefore, we have detailed information about the proportion of the different LC types in each cell.

In this thesis, we limit ourselves to studying the climate of Europe and cover approximately the region from about 22°W to 45°E longitude and from 27°N to 72°N latitude [60].

Surely, in the period from 1992 to 2015, some categories of LC underwent more substantial changes than others. The maps in Figure 3.4 (as well as in Appendix A) demonstrate the grade of LC changes for the most prominent LC changes within Europe such as the expansion of urban and built-up cover and changes in evergreen needleleaf forest. Different

Figure 3.3: The Cordex-EU analysis domain [39]



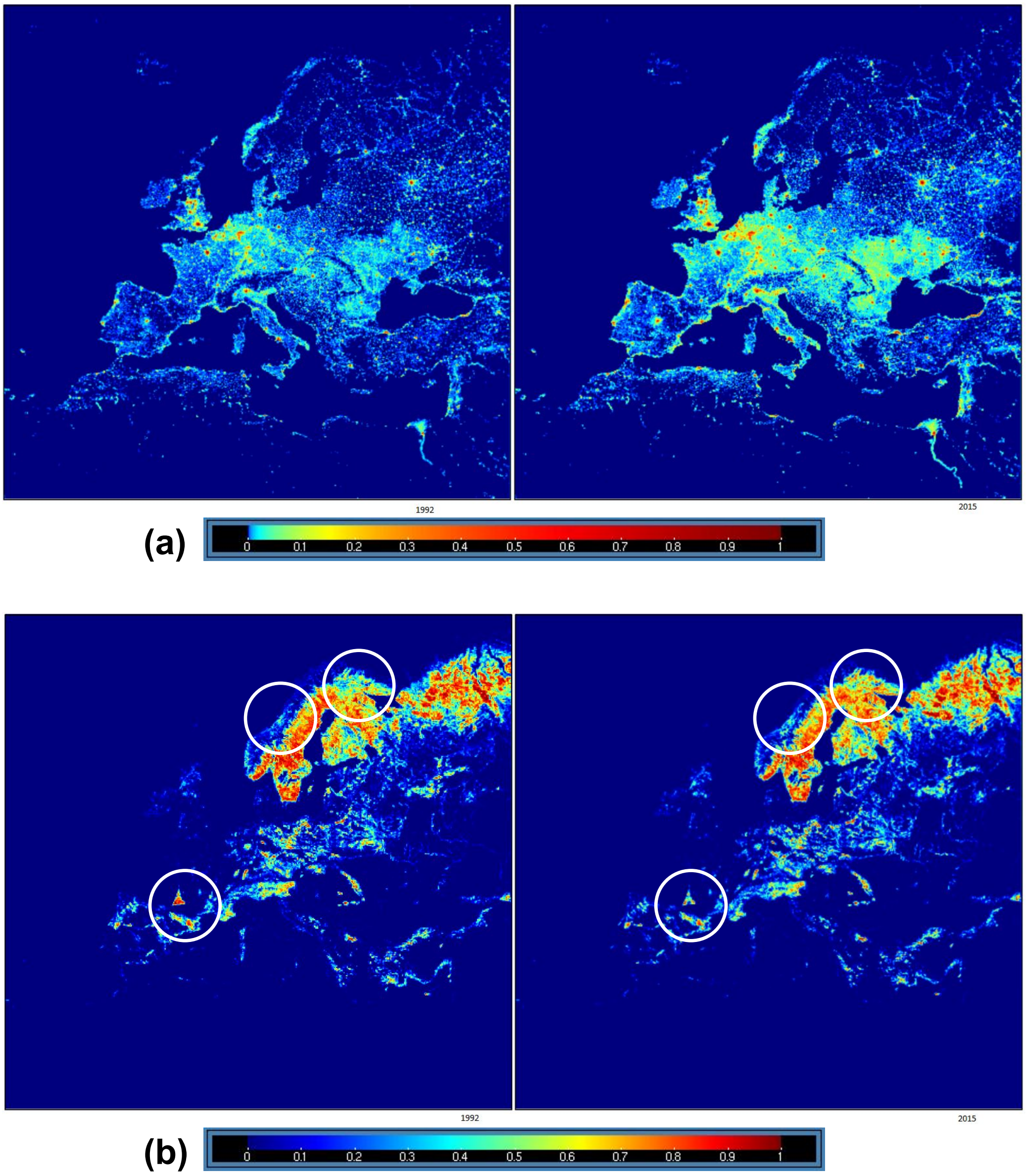
colors represent the proportion of a certain LC in each cell on the grid.

3.2 Artificial Temperature Data

We used data sets including atmospheric and surface variables from regional climate simulations of the Weather Research and Forecasting (WRF), model version 3.9.1. The WRF model made a simulations based on the input data that include the LC data for 1992 and 2015 and the settings of the international Coordinated Regional Climate Downscaling Experiment (CORDEX) initiative (EURO-CORDEX) [48]. The result of the WRF model simulations is the temperature in degrees Celsius for each day in the period between 1992-01-01 and 1992-12-31 and for each day in the period between 2015-01-01 and 2015-12-31.

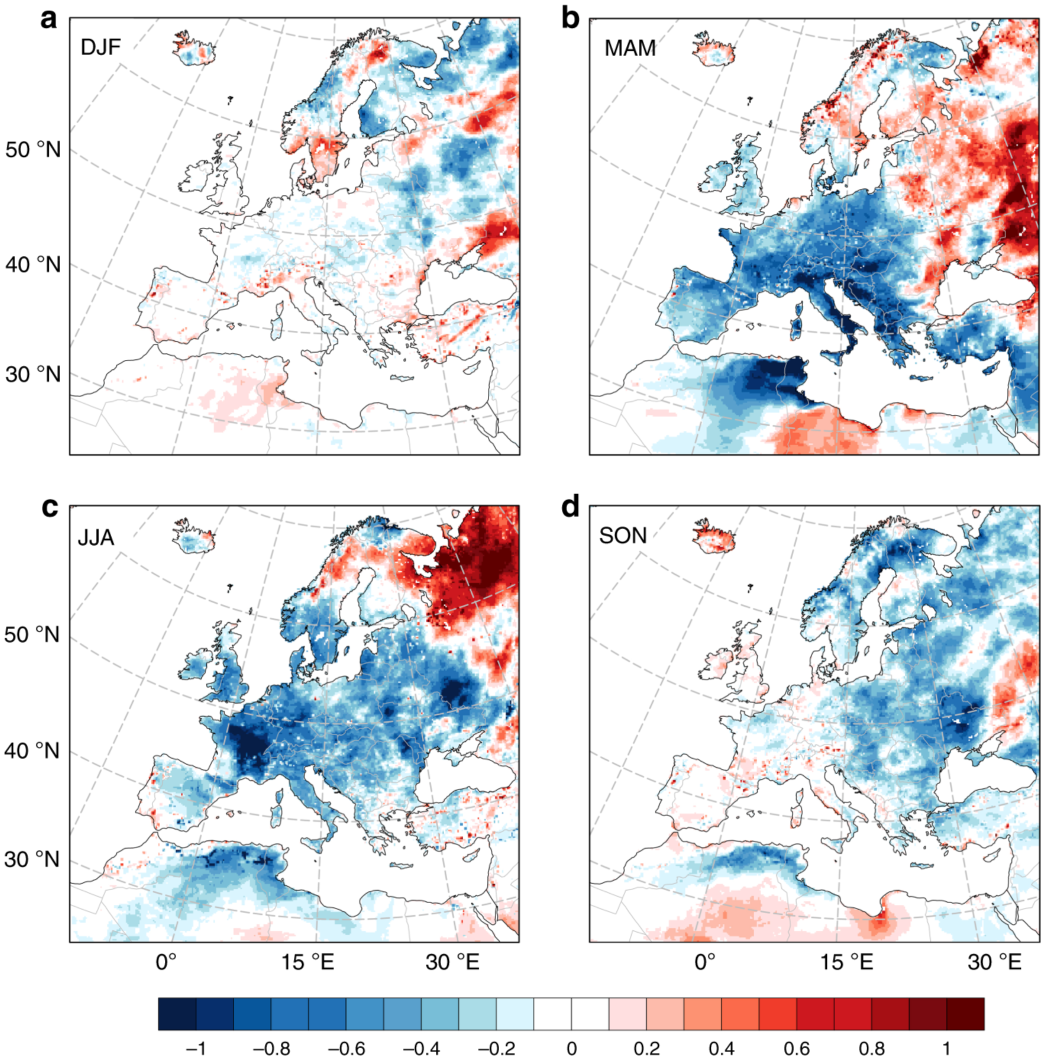
This artificial temperature is a result of two runs of the regional climate model simulation: one is with the input data of LC in 1992 and the other is with the input data of LC in 2015. These results of simulation of the regional climate model illustrate the case of how the temperature would change if only LC changes. Therefore, the absolute artificial temperature differs from the actual temperature observations for 1992 and 2015. The temperature is modeled daily for each cell during the given years. This

Figure 3.4: (a) Urban and Built-Up LC in 1992 (left) and 2015 (right), (b) Evergreen Needleleaf Forest LC in 1992 (left) and 2015 (right). White circles point at the regions with the biggest changes in LC.



allows finding the average temperature for time periods, for example, years or seasons. We can also calculate the temperature changes by simple counting the difference between the average artificial temperature in 2015 and 1992. The variation in average temperature per season for each cell is shown in Figure 3.5. The seasons are divided as follows: winter includes December, January, February (DJF), spring includes March, April, May (MAM), summer includes June, July, August (JJA) and autumn includes September, October, November (SON).

Figure 3.5: Changes of average temperature in °C by season: (a) winter, (b) spring, (c) summer, (d) autumn



3.3 Summary

A huge amount of structured and labeled LC data can be considered as prerequisites for using the supervised ML methods. However, most of the standard ML techniques assume that the observed data are i.i.d., while spatio-temporal dependence is an essential property of climate data. Therefore, the standard ML techniques cannot be directly applied to this type of tasks. In addition to that, the high variability in climate data complicates a comparison of performance for different ML methods.

Despite the challenges mentioned above, we believe that standard ML methods can be adjusted and applied to the data with spatio-temporal dependencies. In Section 1.2, we stated that our goal is to understand how supervised learning can be applied to tasks in climate science.

Therefore, in the next chapter, we strive to develop a framework that allows evaluating predictive validity and performance of the ML models on the spatio-temporal dependent data with high variability.

Chapter 4

Methodology

In Section 1.2, we defined three major objectives of our work. The first goal is to study how ML can be applied to the climate-related tasks which are typically presented by spatially dependent data with high variability. The second one is to verify whether it is possible to develop a reliable ML model that efficiently predicts the impact of LC changes on temperature. In this chapter, we are going to address these objectives.

Firstly, we start with the development of an evaluation method to compare the performance of ML algorithms on the spatially dependent data. The theoretical analysis of the possible ways to approach this problem will be considered in Section 4.1. Next, we will define requirements for ML models for the prediction of the impact of LC change on the temperature in Section 4.2.

4.1 Assessment of Algorithm Performances on Spatially Dependent Data

As discussed in Chapter 2, ML regression algorithms is a promising tool to study the impact of LC changes on temperature, and the CV framework can be used for the evaluation of these algorithms. However, we should verify the applicability of these tools to spatially dependent data. Therefore, we carry out an experiment on a synthetic data set that simulates spatially dependent data. In this section, we describe how the experiment was designed.

The CV framework is often used for evaluation of a model performance on the new data (see also CV in Section 2.2.2). However, this framework is formulated for independent and uniformly distributed data, while our LC and temperature data do not fit these requirements, being spatially dependent and non-uniformly distributed. Therefore, standard CV framework with random data partitioning can provide different results depending on how the initial data set is split into the test and training data sets. This can lead to the wrong results in model comparison. To overcome this issue, we study the statistical significance of the difference in performance between two ML algorithms. In this section, we introduce the main notions of the statistical significance testing and its application to ML.

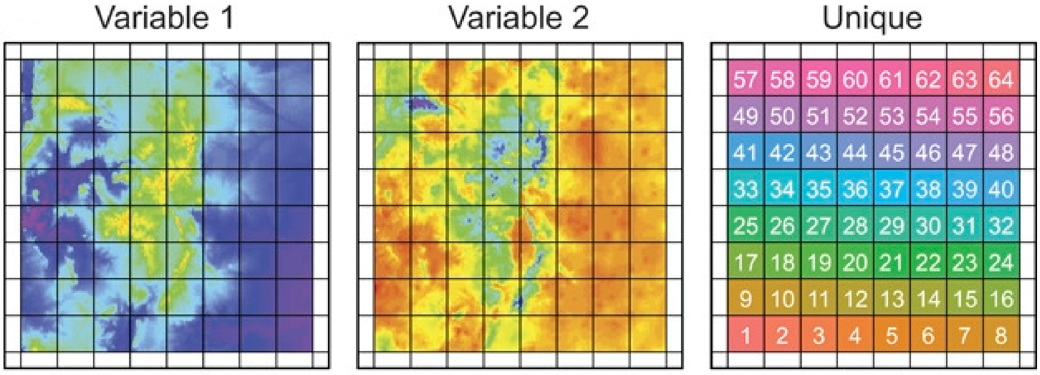
4.1.1 Spatial Cross-Validation

We use spatial CV as a reference method because it is an easy and extensively used technique. *Spatial CV* is based on the standard approach discussed in Section 2.2.2. First, we divide the data into K different folds and then reject one fold and train a model on the remaining $(K-1)$ folds. The main difference is that the initial data set should not be split randomly, but into K spatial sectors as presented in Figure 4.1.

This technique was described by Roberts et al. [103] as a way to reduce spatially dependence and avoid too optimistic assessment, which is typical for the standard CV. As mentioned in Section 2.2.2, the CV framework on the data that are not i.i.d. can show high dispersion in error estimation. Therefore, we can expect that the accuracy of predictions is non-uniformly distributed within the research area. However, we believe that this can still be helpful for comparing different models because we average K-folds evaluation metrics over the whole map, reducing dispersion.

In Section 5, we assess two algorithms A and B with the spatial CV using the evaluation metrics R^2 and RMSE. We consider that the two algorithms have the equal performance if the average R^2 and the average RMSE are equal.

Figure 4.1: Data splitting into 64 spatial sectors. The figure "Variable 1" shows a split of the input variables, while "Variable 2" demonstrates a split of the output variables within the same grid. The "Unique" figure represents a split of the data into unique 64-folds. Adopted from Ref. [103].



4.1.2 5x2 CV Test for Spatially Dependent Data

In Section 5.3.1, we will describe how the temporal dependence was reduced through the temperature averaging over a year or a season. However, the data used in this thesis are characterized by a significant spatial dependence. Therefore, the reliability of the 5x2 CV test results should be verified because this method originally implies only i.i.d. data.. To address this issue, we suggest three possible ways to split the initially observed data to the test and training data sets and, hence, to increase the reliability of the 5x2 CV test assessments. We perform an experiment that allows choosing the best strategy for the data splitting.

Bahn and McGill [7] evaluated four different approaches, shown in Figure 4.2, for dividing of spatially dependent data:

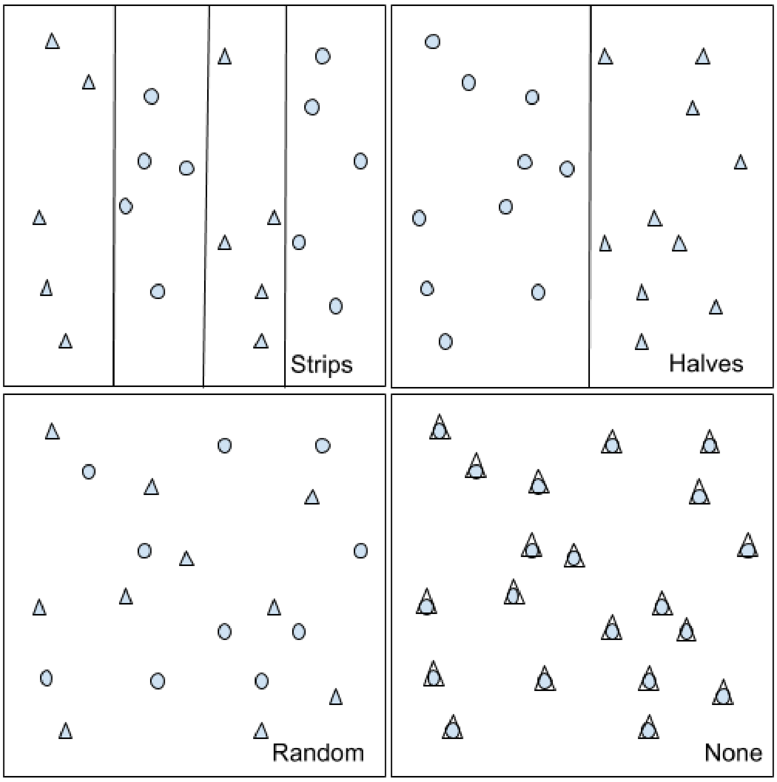
1. **"no splitting"** when the same data set is used for both testing and training;
2. **"random split"** of the data points into two groups;
3. **"strips"** divide in 4 quarters along 3 longitudinal lines where 2 parts are used in test and 2 others - in training;
4. **"halves"** partition along the longitudinal line.

The first two splitting strategies are straightforward and easy to understand. The motivation for considering "strips" and "halves" strategies is that they introduce more independence in data because there are fewer neighboring data points. However, these strategies can reduce the power of the test.

Bahn and McGill [7] found out that the last partition, the splitting into halves, has the lowest error rate in compare with the other partition strategies. The explanation is that this splitting type has the lowest dependence between the data points. Indeed, the neighboring points are spatially dependent, and the smaller the number of the neighboring points in two sets is, the less the spatial dependence is. Based on this idea, we suggest three methods for data dividing in our study.

1. The first and most obvious way to split up the data into the test and training sets is to randomly select the data points. This strategy is used in the original 5x2 CV test on the i.i.d. data. This approach, however, leads to an underestimation of the predictive error, for example, RMSE, if it is applied to the spatially dependent data [7]. Nevertheless, our main goal is to assess a performance of different algorithms but not to evaluate them numerically. Thus, we can assume that errors are equally underestimated for all algorithms, enabling the possible comparison with random data splitting.

Figure 4.2: Data splitting approaches. The circles indicate the data points used for training and the triangles are used for testing.

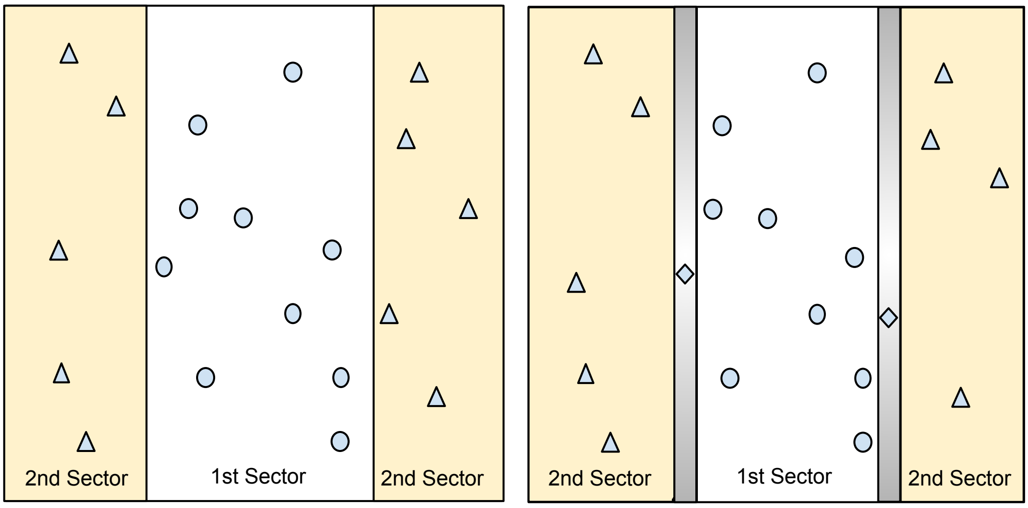


2. The second possible approach is dividing the data points into sectors similar to the "halves" strategy described by Bahn and McGill [7]. The main difference is that the 5x2 CV test implies that the data set should be differently divided five times. To obtain half of the data set, we randomly choose a sector containing a half of all data points (see Figure 4.3 (left)). The rest of the data set contains two "stripes" which we then combine with another half of the data set.

3. The third approach is a development of the previous method and is illustrated in Figure 4.3 (right). It implies excluding a part of the data points between the sectors to further reduce the spatial dependence between the test and training data sets. With this splitting strategy we aim to be in agreement with the assumptions of i.i.d. data in the original 5x2 CV test.

Later, in Section 5, we describe an experiment that we perform to evaluate which of the discussed splitting methods is the most suitable to use in the 5x2 CV test technique on the spatially dependent data.

Figure 4.3: Sector splitting approaches. The circles and triangles indicate the data points used for the training and testing data sets, respectively. The rhombuses are excluded from the data sets.



4.2 Climate Pattern Detection

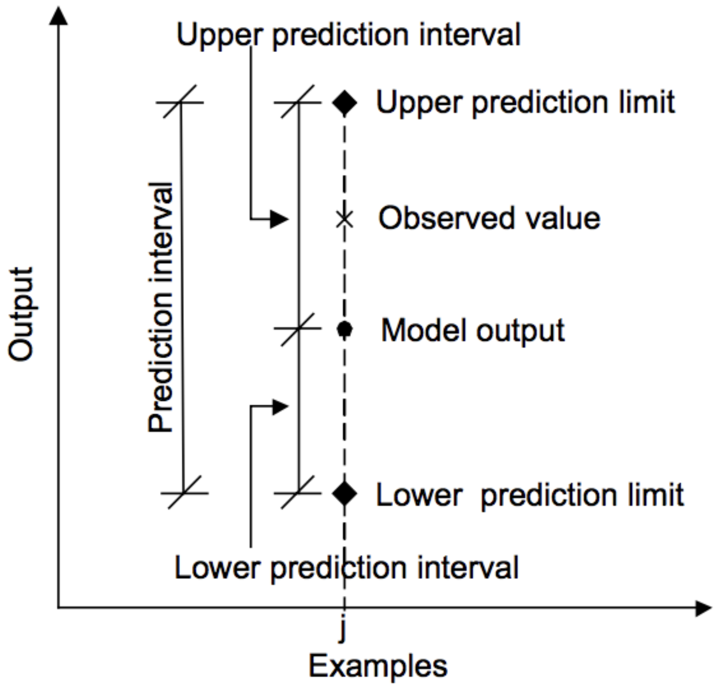
One of the goals of this thesis is to identify how LC transformation affects surface temperature. The observed LC data used in our work represent the ratio between different LC types within one cell. To understand climate patterns better, we make a prediction only for extreme cases when cells initially covered by only one LC type were completely converted into another LC type. There are 21 types of LC, and hence the total number of possible extreme cases is 420. However, an inherent property of ML algorithms is that they can only effectively recognize the patterns that were probed during training. Said in another way, predictions should be limited to the training domain. We however consider extreme cases in order to identify the impact of a particular change, i.e. when a cell initially has only one LC type, which is then completely replaced by another. But in real-world data, extreme LC transformations rarely happen. Let us call these partial LC changes as non-extreme ones. Then, in predicting the effect of LC on temperature, we use in further ML only those extreme LC changes that are often found in training data even in a non-extreme form.

In our study, we use the data sets from Ref. [48] published in Nature Communication. In this article Huang et al. also considered the impact of only extreme LC changes. This makes us believe that such an approach is reliable from both methodological and climate perspectives. In addition to that, the consideration of only extreme LC changes allows us to compare our results with the findings by Huang et al. [48].

We split the whole area into three sub-areas: Northern Europe, Central Europe and Southern Europe. For each of this region we develop the separate ML model and make a prediction for the most frequent LC transformations. This allows us to focus on the regional effects. The ML algorithm for this task will be chosen on the basis of the results of the experiment for searching the algorithm that has the best performance on the spatially dependent data.

It is not enough to simply receive the model output and then make a conclusion based on it because model prediction contains some uncertainties. Thus, we should calculate a prediction interval for each model output. The *prediction interval* is a range of values, which includes all future observations with a pre-defined probability.

Figure 4.4: Schematic explanation of the prediction interval. Adopted from Ref. [107]



A schematic explanation of the prediction interval is shown in Figure 4.4. The upper and lower prediction limits can be found with the following formula:

$$\hat{y} \pm c\sigma, \tag{4.1}$$

where $\hat{y}$ is the model output for the given input, c is the factor depending on the size of the prediction interval, and σ is the estimated standard deviation. In our case, we use $c = 1.96$ that corresponds to a 95% prediction interval [51]. We also use $\sigma = RMSE$ as the estimated standard deviation (see Equation 2.2.2 for RMSE). Since we cannot directly calculate

RMSE for the model output because we have nothing to compare it with, we calculate RMSE by splitting the initial data set in proportion 90% for training and 10% for testing. This approach is equivalent to the traditional error estimation in statistical regression analysis. Therefore, it allows us to directly compare our findings with the results obtained with the statistical method in Ref. [48] .

4.3 Summary

In Chapter 4, we considered the theoretical approach to the application of the standard ML techniques to the prediction of impact of LC changes on regional temperature. The two main aspects of this problem were examined.

The first aspect was shown in Section 4.1. We focused on the method to compare the performance of ML algorithms on spatially dependent data. The issue is that tools for standard hypothesis testing imply i.i.d data and this requirement is not fulfilled in our data sets. Therefore, the main goal was to adopt the 5x2 CV test to provide reliable results on spatial dependent and uneven distribution of data. To reach our objective, we developed three possible strategies for data splitting: random, sector, and sector with buffer. In the Chapter 5, we will evaluate the performance of the described data splitting strategies. In Section 4.1, we also described a reference method for model evaluation - spatial CV. The comparison of the results of the 5x2 CV test and reference method will be presented in Chapter 5.

The second aspect is a theoretical foundation for climate pattern identification which was discussed in Section 4.2. We considered the methodology for the prediction of the impact of LC transformation on temperature. In our study, we strive to obtain experimental results comparable to Huang et al. [48], since the data sets are the same in both studies. Therefore, we decided to consider only the extreme cases of LC changes. The results of our predictions will be demonstrated in the Chapter 6.

Chapter 5

Experiment

In Chapter 4, we introduced our new methodology, namely, the methods for performance assessment of algorithms on the spatial data and discuss preliminary data processing. In this chapter, we present the detailed description of our experiment and review the results. Afterwards, we suggest the best CV framework for our task and the best ML approach to the spatially dependent data. Finally, we discuss the consistency of our findings with other studies.

5.1 Experiment Design

The main goal of experiment is to find the best method to assess algorithms. In this thesis, we compare four regression algorithms described in Section 2.2: MLR, SVM, LASSO and RF.

The design of our experiment involves the implementation of two separated experiments: the one the synthetic data and the other on the climate data. The results of the first experiment are used as a basis for the second one. The experiment can be shortly described as follows:

1. The first experiment will be presented in Section 5.2 and will test different splitting strategies the synthetic data sets. This allows us to verify the reliability of the evaluation methods. Firstly, we will introduce the structure of the synthetic data set. Secondly, we will describe the design of the experiment. And finally, we will show the results. Based on our findings we will define the best data splitting strategy for the 5x2 CV test.

2. The second experiment will be presented in Section 5.3. There, we will compare the performances of the four ML algorithms (MLR, SVM, LASSO, and RF) with the 5x2 CV test on the real-world spatially dependent data. We will also estimate the performance of the four ML algorithms with the spatial CV method, which is used as a reference evaluation technique. At the end of this Chapter, we

compare the spatial CV evaluation of the four ML algorithms with the results of the 5x2 CV test.

5.2 Experiment 1 on the Synthetic Data

5.2.1 Experiment Design

The experiment 1 implies the calculation of the probability of detection the difference in algorithms performance for the 5x2 CV test, depending on data splitting strategies. In case of equally performing algorithms, the probability of detection the difference in algorithms performance corresponds to the type I error rate. We divide this experiment 1 to three stages:

1. Step 1 is to develop a synthetic data set that simulates the spatially dependent data. This allows us verify an efficiency of the 5x2 CV test for such kind of data. The data set is split according to one of the data splitting strategy: random, sector or sector with buffer (see also Section 4.1).

2. Step 2 is to design a couple of ML algorithms that should perform equally on the synthetic data set. This allows us to verify the type 1 error rate when the null hypothesis is true. A similar approach for determination of type 1 error rate was described and applied in Ref. [28]. We will develop several couples of ML algorithms with pre-defined performance differences. Thus, we will find out the probability of detection the difference in algorithms performance for a range of performance ratios.

3. Step 3 is to estimate the probability of detection the difference in algorithms performance with the 5x2 CV test for the three data splitting strategies (random, sector or sector with buffer) and different performance ratios of the ML algorithm designed at Step 2.

Synthetic Data Sets

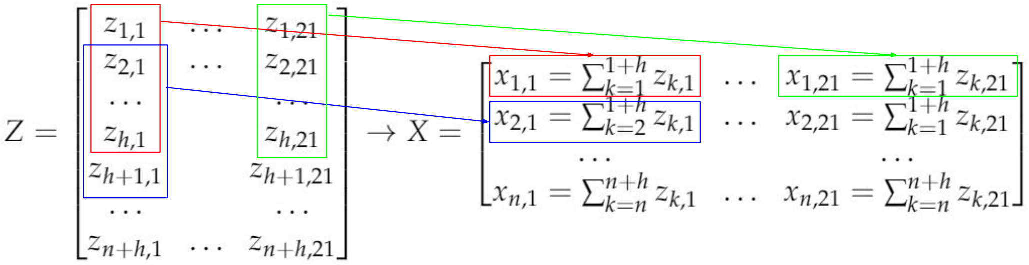
At Step 1, we design the synthetic data set $T_{all} = (X, Y)$ that simulates the spatially dependant data. To add spatial dependence in these data, we define the input variables X as a one-dimensional array that is build on randomly generated data Z. Let us define the size of the one-dimensional array X as $1 \times n$, where $n = 450$. The array consists of the p-dimensional vectors with $p = 21$ that includes information about changes in different LC types from Table 3.1. The degree of dependence in the data is determined by the parameter h equal to 20. We also generate output variables Y, which are a linear function of the input variables and spatially dependent noise data. Eventually, the synthetic data can be described by the following rules:

1. Let us generate a matrix Z of $(n+h) \times p$ size, where all elements are random and have the same normal distribution $z_{i,j} \sim \mathcal{N}(0, \sigma_z)$ for $i = 1, \dots, (n+h)$ and $j = 1, \dots, p$. Thus, the elements in Z are i.i.d. The matrix Z will be an origin of the spatially dependent input data.

2. The matrix X with the input variables and a size of $n \times p = 450 \times 21$ is constructed from the matrix Z, as presented in Equation 5.1. In Figure 5.1, one can see the spatial dependence between $x_{1,1}$ and $x_{2,1}$, which have $h-1$ common summands.

$$x_{i,j} = \sum_{k=i}^{i+h} z_{k,j} \tag{5.1}$$

Figure 5.1: Relation between Z and X matrices.



$Z =$
$z_{1,1}$
$z_{2,1}$
$z_{h,1}$
$z_{h+1,1}$
$z_{n+h,1}$
$z_{1,21}$
$z_{2,21}$
$z_{h,21}$
$z_{h+1,21}$
$z_{n+h,21}$
$\rightarrow X =$
$x_{1,1} = \sum_{k=1}^{1+h} z_{k,1}$
$x_{2,1} = \sum_{k=2}^{1+h} z_{k,1}$
$x_{n,1} = \sum_{k=n}^{n+h} z_{k,1}$
$x_{1,21} = \sum_{k=1}^{1+h} z_{k,21}$
$x_{2,21} = \sum_{k=1}^{1+h} z_{k,21}$
$x_{n,21} = \sum_{k=n}^{n+h} z_{k,21}$

3. Let us generate a vector δ of $(n+h)$ size, where elements are random, and each δ_i is normally distributed as $\delta_i \sim \mathcal{N}(0, \sigma_\varepsilon)$, where $i = 1, \dots, (n+h)$. This will be an origin of the spatially dependent noise in the output variables.

4. The noise vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ of n size is based on the vector δ, as shown in Equation 5.2, and its components are spatial dependent:

$$\varepsilon_i = \sum_{k=i}^{i+h} \delta_k \tag{5.2}$$

5. We define the output variables $Y = (y_1, \dots, y_n)$ of $n = 450$ size as a linear function of the input variables X and noise vector ε, as determined in Equation 5.3:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i, \tag{5.3}$$

where $\beta_0, \dots, \beta_p$ are some coefficients.

The final synthetic data set T_{all} consists of the spatially dependent input variables X and output variables Y that are linear functions of X. Then we divide the synthetic $T_{all} = (X, Y)$ into two sub-sets of the equal size: $T_{train} = (X_{train}, Y_{train})$ and $T_{test} = (X_{test}, Y_{test})$ using the three splitting approaches described in Section 4.

Verification of 5x2 CV Test for Spatially Dependent Data

At Step 2, we develop two ML algorithms M_1 and M_2 with equal performances, which are evaluated by RMSE. Let us define a ratio between the averaged RMSE for models based on M_1 and M_2 as a *performance ratio* $R_{perf} = \frac{RMSE_{M1}}{RMSE_{M2}}$ of the ML algorithms. For example, if the performance ratio $R_{perf} = 0.5$, then the performance of M_1 is twice better than that of M_2. M_1 and M_2 perform approximately equal if $R_{perf} = 1$, when we train and test both M_1 and M_2 enough times on the randomly generated synthetic data sets. In this thesis, we state that *approximately equal* means that $R_{perf} = 1.00 \pm 0.001$. Let us denote K as the number of data sets on which we should run M_1 and M_2 to find the correct average RMSE. Empirically, we found that $K = 1000$ is the optimal number with respect to a balance between accuracy and computational time. The design of the synthetic data implies the linear dependence between the input and output variables. Thus, we can assume that MLR and LASSO should have a proper performance on such a data set. However, their performances are not exactly the same. Let us assume that models based on M_1 have lower average RMSE and hence higher performance than M_2 models. To make them equal, we decrease the efficiency of M_1 models adding a constant c to the predicted output values and, therefore, increasing the average RMSE for M_1. Ref. [48] denoted a model with reduced performance as a *"damaged" model*. We denote a model that is produced by M_1 algorithm with the synthetic data set as $f(\cdot)$. Then, the "damaged" model $\hat{f}(\cdot)$ can be define as follows:

$$\hat{f}(X) = f(X) + c = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + c \tag{5.4}$$

To reach the desirable performance ratio $R_{perf} = P$, we perform the following operations:

1. Firstly, c is zeroed: $c = 0$
2. For $i = 1$ to K:
   (a) Generate a synthetic data set T^i_{all}
   (b) Divide T^i_{all} into T^i_{train} and T^i_{test} of the equal size in according to the splitting strategies.

(c) Train M_2 on T^i_{train}. Test the model on T^i_{test} and calculate $RMSE^i_2$ using the standard Equation 2.2.2

(d) Train M_1 on T^i_{train} and find out the model $f(\cdot)$. Then form the "damaged" model $\hat{f}(\cdot)$ according to Equation 5.4. Test the "damaged" model on T^i_{test} and calculate $RMSE^i_1$ using the following equation:

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{f}(x_j))^2} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - (f(x_j) + c))^2} \tag{5.5}$$

where c is the coefficient decreasing the performance of M_1

3. Calculate the performance ratio R_{perf} as

$$R_{perf} = \frac{\sum_{i=1}^{K} RMSE^i_1}{\sum_{i=1}^{K} RMSE^i_2} \tag{5.6}$$

4. If $R_{perf} \neq P \pm 0.001$, then choose another value for c and go to (2). Depending on whether $R_{perf} > P$ or $R_{perf} < P$, the c value should be gradually increased or decreased by a small value starting from 1. If $R_{perf} = P \pm 0.001$, then M_1 and M_2 algorithms have the performance ratio P, and we achieve our goal.

So, finally, we have the two algorithms M_1 and M_2 with the performance ratio equal to $R_{perf} = P$.

The last Step 3 is to estimate the probability of detection the difference in algorithms performance on the data splitting strategy chosen at Step 1 and the performance ratio R_{perf} chosen at Step 2. Let us denote the probability of detection the difference in algorithms performance as α. Then we can calculate α as the ratio between the number of tests, where the difference in algorithm performances is detected, and the total number of performed tests M. Empirically, we found that $M = 100000$ is the optimal number with respect to the balance between accuracy and computational time. To do so, we perform following operations:

1. For $i = 1$ to M:

   (a) Generate new synthetic data set where the input data set is the array of 1 × 450 in size, where each element of the array is a 21-dimensional vector, and the output data set is the vector of 450 in size.

   (b) Perform the 5x2 CV test on the given data set and for a model based on M_1 algorithm and and the "damaged" model based on M_2 algorithm.

(c) Conclude whether the models are recognized as equal by the 5x2 CV test.

2. Calculate the probability of detection the difference in algorithms performance α.

When we get the results of the experiment, we can evaluate whether any of the splitting data methods can provide the acceptable probability of detection the difference in algorithms performance for different performance ratios. In this thesis, we consider nine performance ratios from 1 to 0.2, where $R_{perf} = 1$ for the algorithms with the equal performance and $R_{perf} = 0.2$ for a situation when one algorithm performs five times better than another. The splitting strategy can be considered acceptable if two conditions are fulfilled:

1. For the algorithms with the equal performance, the probability of detection the difference in algorithms performance is $\alpha \leq 0.05$

2. For the algorithms with different performance, α should be as high as possible and above 0.05.

5.2.2 Experiment Results: The best data splitting strategy for 5x2 CV test

In this section, we calculate the probability of detection the difference in algorithms performance. Let us remind that the type I error rate is the probability of detection the difference in algorithms performance for algorithms with the equal performance. For the algorithms with diverse performance, the probability of detection the difference in algorithms performance shows whether the 5x2 CV test reveals their different performance. The probabilities of of detection the difference in algorithms performanc are calculated for different splitting strategies: random splitting, sector- and sector-splitting with buffer, which were described in Section 4. The experiment is performed on the synthetic data sets that simulate the spatially dependent data. The results are presented in Table 5.1 and Figure 5.2.

Figure 5.2 illustrates the dependence of the probability of detection the difference in algorithms performance on the performance ratio for different splitting strategies. The line $\alpha = 0.05$ demonstrates the maximum acceptable type I error rate for algorithms with the equal performance. This means that the performance of both algorithms is considered equal if a bar is under this line.

The performance ratio $R_{perf} = 1$ means that if algorithms perform equally, then the acceptable type I error rate should be less than 0.05. One can see from Figure 5.2 that this applies for all considered splitting strategies. As R_{perf} decreases, the ML algorithms perform more and more

Performance ratio	Random	Sector	Sector with buffer
1	0.005	0.0254	0.0236
0.95	0.0629	0.0286	0.0225
0.9	0.3256	0.0403	0.0281
0.85	0.5214	0.0594	0.0384
0.8	0.6790	0.1142	0.0671
0.7	0.8979	0.2046	0.2146
0.6	0.9850	0.2929	0.2661
0.4	0.9999	0.6388	0.5843
0.2	1	0.9805	0.9675

Table 5.1: The probability of detection the difference in algorithms performance for different data splitting strategies.

differently, and the 5x2 CV test should exhibit a sharp increase in the probability of detection the difference in algorithms performance. Ideally, the probability of detection the difference in algorithms performance should be equal to 1 for the performance ratio $R_{perf} \neq 1$. In Table 5.1 and Figure 5.2 we can see that the probability of detection the difference in algorithms performance grows most prominently for the simplest splitting strategy - the random data splitting. Indeed, already for $R_{perf} = 0.90$, the random data splitting indicates the difference in performance in 33 % of cases, while the sector splitting strategies barely detect it in only 3 - 4 % of cases. Therefore, we can conclude that the random data separation has the best sensitivity towards the performance difference of algorithms.

For example, the performance ratio of 0.4 means that RMSE for one algorithm perform in 2.5 times better than another. For this performance ratio, the sector splitting strategy detects the difference only in 64 % of cases, while the sector with buffer splitting approach detects the difference even worse - in 58 % of cases. At the same time, the random data splitting method possesses a superior effectiveness, detecting performance difference for 99.99 % of simulations.

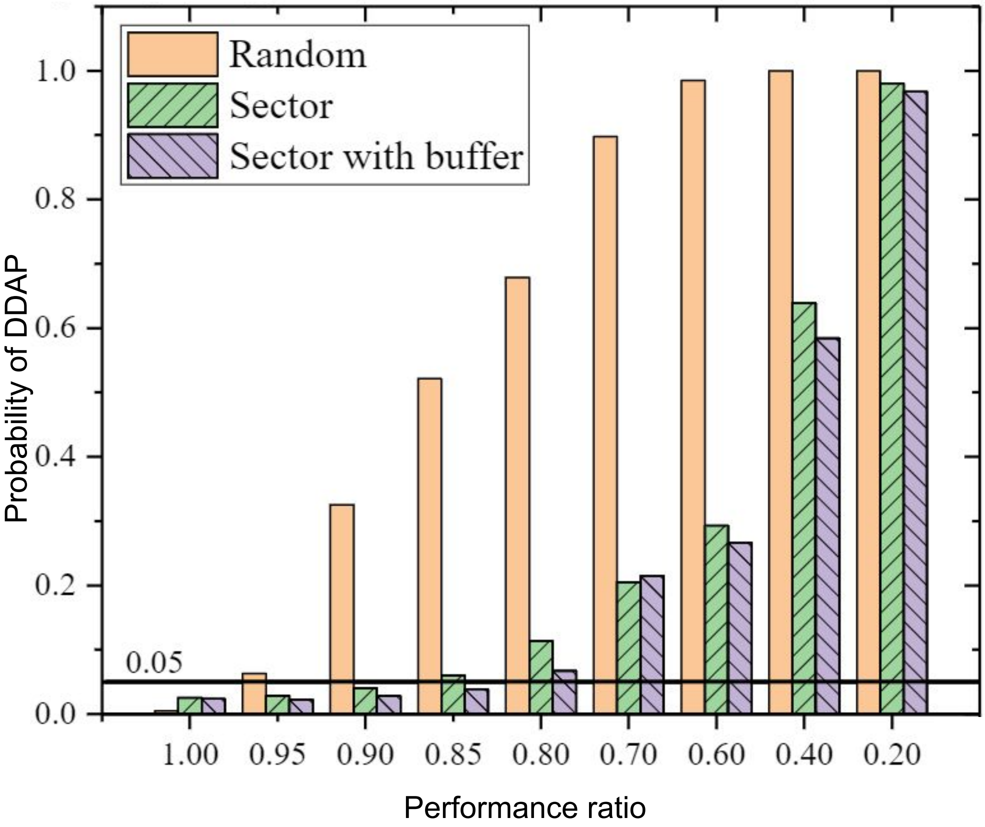
5.3 Experiment 2 on the LC Climate Data

5.3.1 Data Pre-processing

The climate data should be adapted for all tested ML algorithms. Therefore, the goal is to adjust the climate data by transforming it into a form which is the acceptable input data for ML algorithms. This task consists of the three main sub-tasks:

1. Define what data we use in ML algorithms.
2. Verify data consistency and remove suspicious data points.

Figure 5.2: The DDAP probability (probability of detection the difference in algorithms performance) depending on the performance ratio and splitting strategy.



3. Aggregate data for use in ML algorithms.

Initial files containing the LC and temperature data from the regional climate model were represented in a format called the Network Common Data Form (NetCDF). This format was developed by NASA for array-oriented scientific data [113]. We use the same data set that was used in Ref. [48] by Huang et al. In this thesis, 12 main files from the given data set were used:

1. Two files with an information on the LC data in 1992 and 2015 years.

   The research area (Europe) was divided by a grid with 467 cells in the south-north direction and 479 cells in the west-east direction, 223 693 cells are in total. There are three dimensions characterizing each data point: a cell number in the south-north direction, a cell number in the west-east direction, and time. The time dimension is an information on a year when LC information was collected. The data set used in thesis contains LC information for only two dates 1992-01-01 and 2015-01-01 which are the only possible values for a time field. Each cell (data point) contains 55 parameters with detailed description of its properties such as degrees of latitude and longitude, LC information, soil characteristics and others. In this thesis, we focus on LC and, therefore, we need only two characteristics. The first one is a *land mask* which indicates whether a given point is covered by land or water. The second is *LC category* variable that is represented by a 21-dimensional vector with components from Table 3.1. As described in Section 3, the 21-dimensional vector shows a proportion of different LC categories in a certain cell. According to the data description, we define this vector for an i^{th} cell as

$$LC_i = (lc_{i,1}, LC_{i,2}, \dots, LC_{i,21}) \text{ ,where } 0 \leq LC_{ij} \leq 1 \text{ and } \sum_{j=1}^{21} LC_{i,j} = 1 \text{ for } \forall i \quad (5.7)$$

2. Two files with the daily temperature data for the whole 1992 and 2015 years. Eight auxiliary files with the daily temperature data for four seasons in 1992 and 2015.

   The temperature data files contain five parameters. Two of them are coordinates of cells with the same dimensions as in the LC data files. The other two characteristics are representations of time dimension in two different formats. In the annual data files, the size of time dimension is 365 (both 1992 and 2015 were not leap years). In the seasonal files, the size of time dimension is between 90 and 92 days, depending on the number of days within a season. Finally, the last parameter is the daily temperature in degrees Celsius degrees for each cell.

To examine the given data quality, we check that all data points correspond to the definition given above. During testing, we found out that incorrect LC information is provided for the data points that are marked as "water" in the land mask. Therefore, we excluded 101 844 data points marked as "water" from consideration. Thus, the final number of cells used for ML is 121 849.

In this thesis, we consider changes in LC as the input variables X_i for regression algorithms and the temperature data as the output variables Y_i. If we define $X_i = (x_{i,1}, \dots, x_{i,21})$ as the input data corresponding to an i_{th} cell, then it can be described as follow:

$$X_i = LC_i^{2015} - LC_i^{1992}, \quad (5.8)$$

where LC_i^{2015} and LC_i^{1992} are the 21-dimensional vectors describing LC in 2015 and 1992, respectively (see Equation 5.7). Based on the LC data definition we can also state that:

$$\forall i \in [1, 121849] \sum_{j=1}^{21} x_{ij} = 0 \text{ and } x_{ij} \in [-1, 1], \tag{5.9}$$

where 121 849 is the total number of cells in the LC data file, excluding "water cells".

The temperature data simulated by the regional model are used as the output variables in this thesis. Therefore, we consider temperature only for cells marked as "land" in the land mask. To reduce temporal dependence [49], we consider only the average temperature for a whole year or season as dependent variables. Temperature averaging also makes our results compatible with Ref. [48]. Let us define $t_{i,j}^{year}$ as the temperature in an i^{th} cell on a j^{th} day and simulated for a certain $year = 1992$ or 2015. Then we can find the output variable Y_i as follow

$$Y_i = \frac{1}{P} \sum_{j=1}^{P} (t_{i,j}^{2015} - t_{i,j}^{1992}), \tag{5.10}$$

where P is the number of days in the given period: 365 per year and 90 – 92 per season. We label the output variables as Y_i^{year} for the average temperature for the whole year. The output variables for different seasons are labeled as Y_i^{djf} for the average temperature in December, January, February (winter, djf stands for the first letters in the names of the months); Y_i^{mam} for the average temperature in March, April, May (spring); Y_i^{jja} for the average temperature in June, July, August (summer); and Y_i^{son} for the average temperature in September, October, November (autumn).

In summary, we build five data sets with the input variables X (LC change) and the output variables Y (temperature) for use in ML algorithms. These data sets have the same input variables X_i but the different output variables Y_i corresponding to a temperature averaged over different seasons and a year. All five data sets are used to predict the impact of LC change on temperature in Section 6.1. However, in the experiment, we mainly focus on the annual temperature data set (X_i, Y_i^{year}) that is shown in Equation 5.11:

$$X = \begin{bmatrix} x_{1,1} & \dots & x_{1,21} \\ \dots & & \dots \\ x_{N,1} & \dots & x_{N,21} \end{bmatrix}, Y^{year} = \begin{bmatrix} y_1^{year} \\ \dots \\ y_N^{year} \end{bmatrix}, N = 121849 \tag{5.11}$$

5.3.2 Experiment Design: Algorithm Assessment

Algorithm Assessment with Spatial CV

The experiment 2 contains an assessment of performances of the four ML algorithms together with the base line algorithm using the spatial

CV technique described in Section 4.1. The latter is used as a reference approach for the four ML algorithms. The *base line* is an algorithm that predicts the same mean value of the output variables in a training data set for any input variable. In the base line algorithm, we split the whole map to non-overlapping sectors and each of them contains an equal number of cells. We consider three types of splitting on sectors depending on their size: 25x25, 50x50 and 75x75 cells in each of sectors. This give us an ability to observe whether the algorithm performance depends on the sector size. We train the model on all the sectors excluding one, which is later used as a test data set. This process is repeated for all the sectors containing cells with land. Then we calculate the mean values of evaluation metrics for all non-empty sectors. We apply RMSE and R^2 to assess algorithms between each other. We consider that two algorithms perform equally if the evaluation metrics have the same values on average.

Algorithm Assessment with the 5x2 CV Test

This experiment aims to reveal the superior ML algorithm among MLR, LASSO, SVR, RF with application to the spatially dependent data. To achieve this goal, we use the 5x2 CV test and a splitting strategy that is the most sensitive with respect to the difference in algorithm performance. The four algorithms MLR, LASSO, SVR, RF together with the base line are compared in pairs, where the base line is an reference algorithm that predicts the mean value of the output variables in the training data set $f(x_i) = \bar{y}$ for any input variables. If the performance of an algorithm is equal to the base line, then the algorithm considered useless.

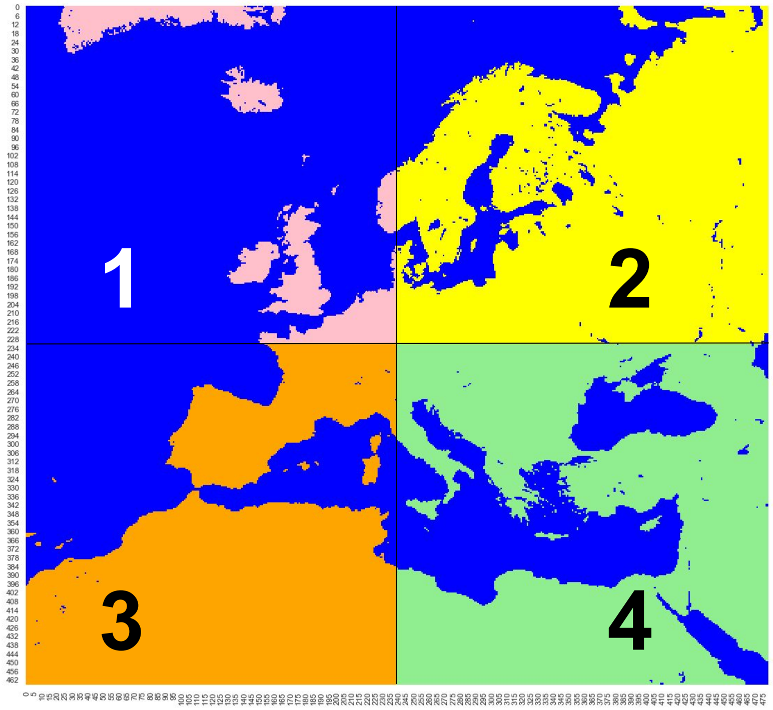
To compare two algorithms, we perform the 5x2 CV test on the real-world LC and annual temperature data to find the t value (see Section 4.1). If $|t| \leq 2.571$, then we accept the null hypothesis and consider that the two algorithms perform equally. If $|t| > 2.571$, then we consider that the two algorithms have a statistically significant difference in performance. A larger $|t|$ is associated with a larger difference. The positive t value means that M_2 performs better than M_1, while the negative one means that M_1 performs better than M_2.

In addition, we repeat this experiment on five areas: one area that includes all cells on the grid (the whole map) and four areas corresponding to the quarters of the map that are shown in Figure 5.3. This is done to verify that the performance difference between the algorithms is reliable and independent of the input data.

5.3.3 Experiment Results: Superior ML algorithm

In this section, we present the results of the experiment 2 described above. First of all, the performances of the four ML algorithms are compared pairwise with 5x2 CV test and the algorithm with the best performance is

Figure 5.3: Sectors where the 5x2 CV test was performed



.

chosen. Secondly, the estimations of the performance of ML methods are evaluated with spatial CV. Finally, we compare the results two methods.

Superior ML algorithm with the 5x2 CV test

In Section 5.2, we evaluated different tests based on 5x2 CV test. According to our estimations, 5x2 CV test with random data splitting has the type I error rate lowest among considered strategies and possesses a higher probability of detection the difference in algorithms performanc. Therefore, in this experiment 2 we used 5x2 CV test with random data splitting.

The 5x2 CV test is supposed to assess ML algorithms relative to each other meaning that it cannot provide an absolute value of an algorithm performance. The 5x2 CV test is useful when the ordinary evaluation metrics of two algorithms have nearly identical values, and R^2 is low (see Section 2.2). However, if we compare ML algorithms with a reference the base line, we can obtain an absolute evaluation of an algorithm performance for further analysis.

We run the experiment on the LC and temperature data, as described in Section 4, with the calculation of t values for pairs composed of five algorithms: base line, MLR, LASSO, RF and SVR. The t values are found for each pair of the algorithms and shown in Table 5.2. If the t value is positive, then an algorithm in a row has better performance than the corresponding algorithm in a column. For example, the RF algorithm performs better than LASSO as shown in Table 5.2. The t values are calculated only for half of the table, because the tables are symmetric with respect to the diagonals, so $t(M_1, M_2) \approx -t(M_2, M_1)$.

Full area					
	Base line	MLR	LASSO	RF	SVR
Base line	X	-	-	-	-
MLR	31.732	X	-	-	-
LASSO	0	-37.3217	X	-	-
RF	39.0171	57.5218	63.3978	X	-
SVR	20.4579	-5.4744	17.1094	-52.8771	X

Sector 1					
	Base line	MLR	LASSO	RF	SVR
Base line	X	-	-	-	-
MLR	18.3349	X	-	-	-
LASSO	1.5811	-17.94631	X	-	-
RF	35.1290	35.1453	42.8217	X	-
SVR	15.8549	-3.0303	8.5086	-18.8624	X

Sector 2					
	Base line	MLR	LASSO	RF	SVR
Base line	X	-	-	-	-
MLR	14.9467	X	-	-	-
LASSO	0	-12.2621	X	-	-
RF	47.7939	27.2626	67.8474	X	-
SVR	11.9198	-5.0461	10.5529	-17.6782	X

Sector 3					
	Base line	MLR	LASSO	RF	SVR
Base line	X	-	-	-	-
MLR	10.7008	X	-	-	-
LASSO	0	-9.2485	X	-	-
RF	27.0683	16.4368	19.1729	X	-
SVR	7.2434	-2.8546	4.6422	-16.0634	X

Sector 4					
	Base line	MLR	LASSO	RF	SVR
Base line	X	-	-	-	-
MLR	11.2362	X	-	-	-
LASSO	0	-12.0941	X	-	-
RF	32.7102	42.3702	52.0083	X	-
SVR	10.219	-5.2496	8.9192	-49.7994	X

Table 5.2: The t value calculated with the 5x2 CV test for pairs of algorithms.

Due to the use of the different data sets (the entire map and four quadrants shown in Figure 5.3), the experiment shows slightly different *t* values for the same algorithm pairs in different areas. However, the trends are identical for all the data sets. We can arrange ML algorithms according to their performance on the spatially dependent data:

1. LASSO: its performance is as bad as that of base line because the *t* values for this pair of algorithms are less than 2.571.

2. SVR: it performs significantly better than base line, but worse than MLR and RF.

3. MLR: it is the second best ML algorithm

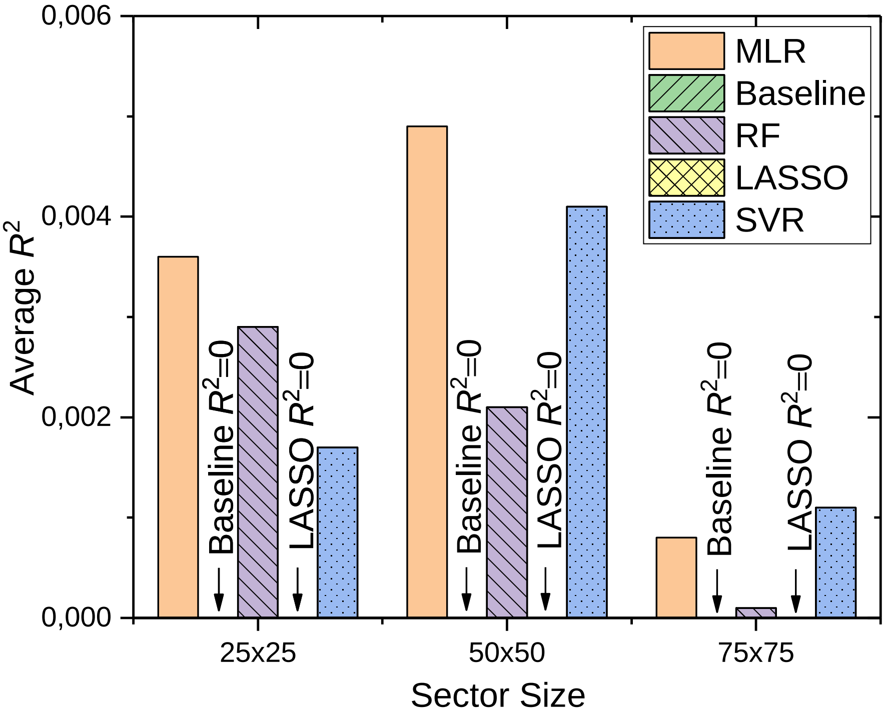
4. RF: it possesses the superior performance on the spatially dependent data since it has the highest *t* values according to the 5x2 CV test.

Superior ML algorithm with spatial CV

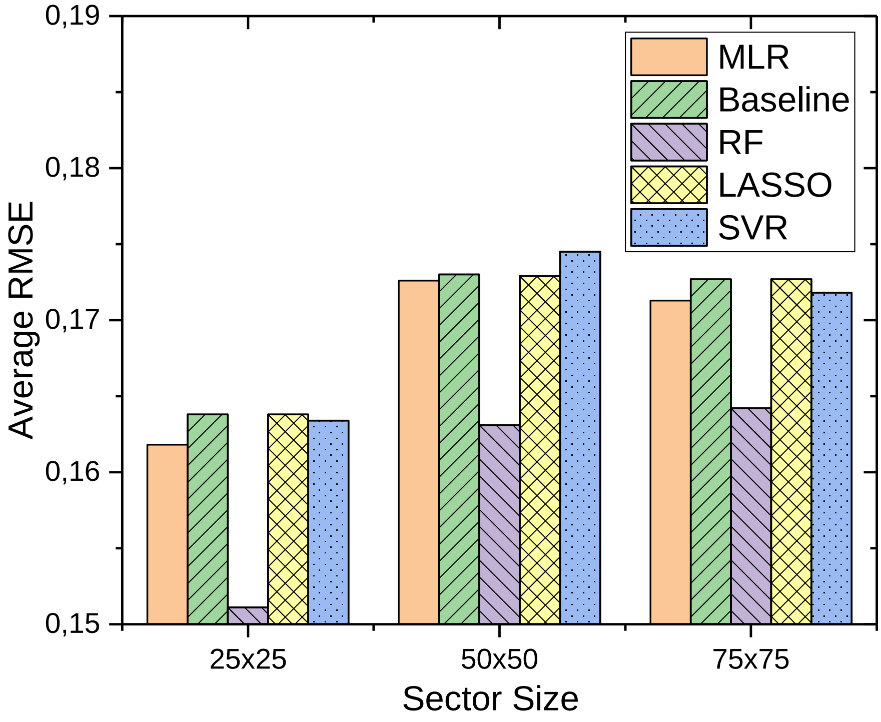
We use the spatial CV technique described in Section 4 to calculate the average R^2 and RMSE for five algorithms: base line, MLR, LASSO, RF and SVR, where base line is a reference algorithm. The whole area of the data points is divided into sectors and the evaluation metrics are calculated. The experiment is repeated several times for the different sector sizes: 25x25, 50x50 and 75x75 cells in each sector. That allows verifying that the difference between the algorithm performances is consistent. The results are presented in Table 5.3 and in Figure 5.4.

Firstly, we measure the coefficient of determination R^2, which represents how properly a model explains the relation between the input and output data. The coefficient of determination can be between 0 and 1, and higher values correspond to a better fitting of the model to the data. The experimental results demonstrate that the average R^2 is very low and almost equal to zero. This may raise doubts on the fact that ML algorithms can provide meaningful predictions.

Secondly, we measure RMSE that should be as low as possible for a model with high performance. We can see that the average RMSE is almost identical for all algorithms, including base line. This can make it hard to understand there is significant difference between the ML algorithms. In general, we can conclude that according to spatial CV, base line, LASSO and MLR have approximately the same performance, while SVR has poorer performance than these three. Finally, we can conclude that RF has noticeably better performance among the considered approaches because of the lowest RMSE.



(a) Average R^2



(b) Average RMSE

Figure 5.4: Results of spatial CV for five algorithms

Sector size 75x75		
	average R^2	average RMSE
Base line	0	0.1727
MLR	0.0008	0.1713
LASSO	0	0.1727
RF	0.0001	0.1642
SVR	0.0011	0.1718

Sector size 50x50		
	average R^2	average RMSE
Base line	0	0.1730
MLR	0.0049	0.1726
LASSO	0	0.1729
RF	0.0021	0.1631
SVR	0.0041	0.1745

Sector size 25x25		
	average R^2	average RMSE
Base line	0	0.1638
MLR	0.0036	0.1618
LASSO	0	0.1638
RF	0.0029	0.1511
SVR	0.0017	0.1634

Table 5.3: Evaluation metric calculated with spatial CV for different algorithms.

Comparison of 5x2 CV test and spatial CV

We compare the results obtained from two assessment methods that we used for evaluation of the best ML algorithm for the spatially dependent data. Based on this comparison, we conclude which algorithm should have a superior performance. Finally, we discuss why this algorithm is the most suitable for such class of tasks and how it is consistent with other publications on this subject

Both experiments, the 5x2 CV test and spatial CV assessment, have some similarities. Both techniques were applied to the same data sets of the observed LC and temperature. In both cases, five algorithms (base line, MLR, LASSO, RF, and SVR) were assessed. Both experiments were repeated several times with different parameters and data sets to demonstrate that the results were not affected by them.

The main advantage of the 5x2 CV test is that the detected difference in performances is statistically significant. In addition, the 5x2 CV test aims to assess a difference in performance between the two algorithms while the spatial CV evaluates each algorithm separately using metrics. It can be difficult to interpret the spatial CV results in case of low R^2 and low variation of RMSE for different algorithms. In addition, the spatial CV shows the ambiguous results for the same algorithms. For example, the SVR models have better performance than base line for a sector size 75x75 and 25x25 cells, but worse performance for a sector size of 50x50 cells.

Summarizing, we can conclude that the 5x2 CV test is more efficient tool to compare the algorithm performances on the spatially dependent data than the spatial CV. The 5x2 CV test can be used for evaluation of even low R^2 by the assessment with base line. This method also allows comparing algorithms with similar performance to determine the difference between them. The 5x2 CV test shows a noticeable dissimilarity between the performance of most algorithms studied in this thesis. At the same time, it is difficult to make any conclusion on the difference of the studied algorithms using the spatial CV. Based on the outcome from both experiments, we can conclude that RF demonstrates the best performance over other algorithms.

We assume that RF has better performance, because of its non-linear nature that probably better describes the dependence between LC changes and temperature. Another peculiar property of this algorithm is that it consists of several decision trees and can handle more complex patterns. For example, in our case, it can be a different temperature response for the same LC transformation depending on the region where it happens. This characteristic also facilitates the selection of the most important independent variables and the ignoring of the irrelevant variables [89]. This can be especially significant in our case because probably not all LC changes noticeably affect temperature. Breiman [12] stated that RF is quite resistant to noise that is also critical for our task, because data from the climate model simulations are noisy.

5.4 Experiment: Discussion

.

The novelty of our work is the development of the method which allows revealing a statistically significant difference in performances between ML algorithms on the spatially dependent data. The statistically significant difference is especially important in our experiment because our data sets are characterized by high noise. Other works mainly tend just to compare some specific methods on various data and do not consider the statistical significance of results. In other studies, it is common to use the K-fold CV. However, it does not always provide clear information about algorithm assessment. In this thesis, we demonstrated that the 5x2 CV test with random data splitting shows reliable results on the spatially dependent data.

In this section, we would like to discuss the outcome of the 5x2 CV test method with respect to other works on this subject. Researchers often apply the R^2 and MSE evaluation metrics and the K-fold CV with random splitting of the observed spatially dependent data (for example, Ref. [14, 69, 90]). However, this approach may have some limitations on the spatially dependent data because they are not i.i.d. Therefore, the results obtained with these approaches can overestimate a model performance that leads to the erroneous data interpretation, whereas the 5x2 CV test can be a powerful tool to reveal the performance difference between ML algorithms.

Several studies were focused on comparison of the performance of various algorithms with evaluation metrics based on the spatially dependent data, including the LC data [14, 69, 90]. Olivera et al. compared the predictive ability of MLR and RF for detection of spatial patterns of fire occurrence [90]. Olivera et al concluded that RF has much higher predictive accuracy than MLR and can identify non-linear trends. Li et al. measured the performance of twenty – three different ML methods for spatial interpolation of environmental variables such as mud content [69]. This study convincingly demonstrated using 10-fold CV that RF and various combinations of RF with other algorithms outperform other algorithms, including SVR. Li et al. also paid attention to the ability of the algorithm to reveal non-linear correlations and complex relations between variables. Chen et al. evaluated with the 5-fold CV the predictions from sixteen algorithms for the average annual fine particle and nitrogen dioxide concentrations using the spatial input variables such as the LC data [14]. In that study, RF was one of the three algorithms that show the best performance.

The research questions raised in these studies differ from those presented in this thesis. Most works also used the K-fold CV with the random data splitting and the R^2 and RMSE evaluation metrics which evaluation ability does not suit perfectly to the spatially dependent data. Therefore, we cannot directly compare our results with those in the discussed publications. We can, however, observe some common trends between these studies and our own findings. For example, these works also demonstrated that RF has a performance superior to other techniques within the application to the spatially dependent data. Thus, RF can be considered as an efficient technique for the spatially dependent data analysis.

5.5 Summary

In this Chapter 5 we discussed the design and results of the experiment. First of all, we considered the overall design and explain the motivation to carry out two experiments: one on the synthetic data and the other one on the climate data sets. Secondly, we described the design and results for each experiment. Finally, we examined how our results are comparable with the findings from other studies.

In Section 5.2, we presented the experiment 1 on the synthetic data sets. We carried out this experiment to find out the type 1 error rate and the probability of detection the difference in algorithms performance for the 5x2 CV test. We used synthetic data sets to have a non-limited amount of data with a similar level of spatial dependency. We also developed two algorithms with the pre-define performance ratio. Using these algorithms and synthetic data sets, we run the 5x2 CV test multiple times to find out the type 1 error rate and the probability of detection the difference in algorithms performance for three different splitting strategies. The experiment results allow us to conclude that a random splitting strategy has superior sensitivity to the probability of detection the difference in algorithms performance. We also came to the conclusion that the 5x2 CV test with random data splitting is an efficient and reliable tool for the assessment of the performances of the algorithms.

In Section 5.3, we used our outcome from Experiment 1 to compare performances of four ML algorithms (MLR, LASSO, RF, SVR) together with the baseline algorithm on climate data set. The algorithms were assessed by the 5x2 CV test on the five different sub-set of data to avoid any impact of data set. In Section 5.3, we also evaluated these algorithms with spatial CV. We repeated assessments three times with different sector sizes to verify that results are independent of input parameters. We found out that some results of the algorithm's assessment with spatial CV are ambiguous. However, both 5x2 CV test and spatial CV convincingly show that RF performs better than the other considered algorithms on our data set. At the end of this section, we discussed the causes of the superior RF performance. We assumed that the main reason is that RF has been developed with the focus to capture non-linearity and to cope with noise in data.

Finally, in Section 5.4 we discussed how our findings are comparable with other studies. Generally, many climate science studies agree that the RF algorithm is a suitable tool for spatially dependent data. However, many scientific papers consider only specific ML algorithms and their performance on certain tasks. In contrast, we proposed a method (the 5x2 CV test) that can evaluate the performance of any ML algorithms on different spatially dependent data.

Chapter 6

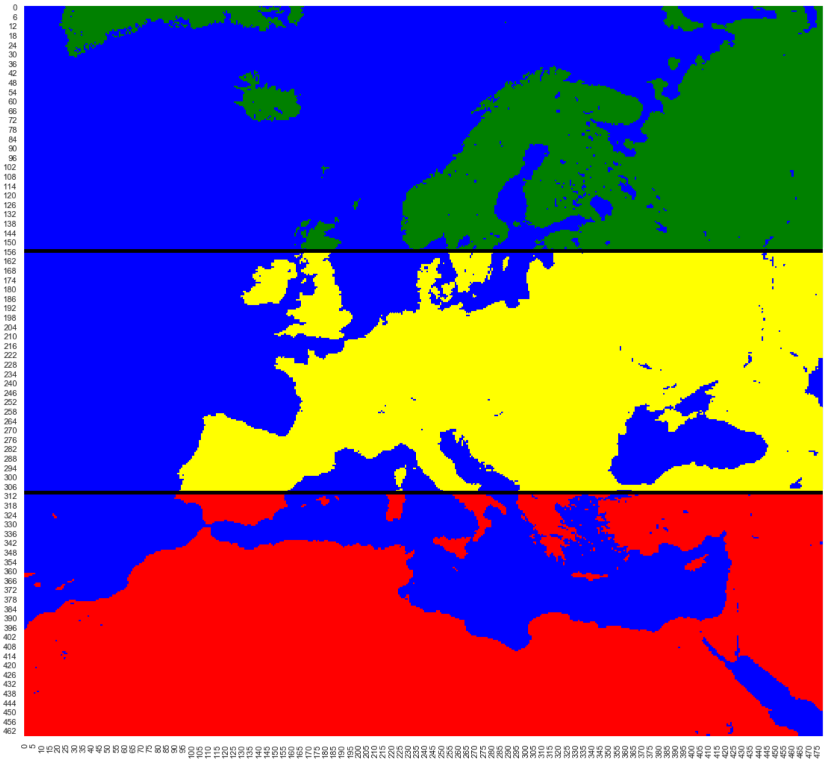
Temperature Changes Due to Land Cover Changes

In the previous chapter, we developed a method to compare the performances of the different algorithms on the spatially dependent data and the random forest method demonstrated the superior prediction ability. Now, in this chapter, we would like to test this technique on the real-world case of impact of LC transformations on surface temperature. Firstly, we build three ML models for the following regions: Northern, Central and Southern Europe (Figure 6.1). The prediction results described separately for each region. Then we build a model for the whole of Europe. The general trends for the whole of Europe are described and compared with the prediction for the three regions. Finally, we discuss the consistency of our results obtained from ML models with other studies based on statistical approaches.

6.1 Results

We make predictions of the most common extreme cases of LC transitions in different regions according to the methodology described in Section 4. The results are presented in Tables 6.1 – 6.4. The whole area and the three sub-regions were considered to reveal patterns inherent to each region. We believe that the similar extreme cases can lead to different results depending on a certain region. However, we would also like to evaluate whether this is consistent with the predictions for the whole of Europe. For each region, we identify the most characteristic LC transformations. The column "Event Rate" represents how often such LC change happens within a certain region. The ten most frequent LC transformations were chosen for Northern and Southern Europe. For Central Europe, we consider the 13 most frequent LC changes because the total number of types of LC transformations is higher there. For the whole of Europe, we considered the 20 most frequent LC changes, but in Table 6.4 we presented

Figure 6.1: Three regions used to predict the effect of LC changes on surface temperature: Northern (green), Central (yellow) and Southern (red) Europe.



only those LC transformations which have the most significant impact on temperature according to our estimation.

In addition, the prediction interval of temperature change in degrees Celsius for each season and the whole year was calculated. In this thesis, we define that the prediction interval demonstrates significant changes in temperature only if the whole interval is positive or negative. When the prediction interval is only partially positive and partially negative, then we assume that it is impossible to predict the temperature trend unambiguously. The cells with the most significant temperature growth are marked in red. The dark red corresponds to the temperate growth when the entire prediction interval is above +0.5 °C. The intense cooling is marked in blue, and the dark blue color marks the most prominent temperature drop when the whole prediction interval is below -0.5 °C.

Recall from Section 3.2 that the artificial temperature simulated by the climate model excludes the factors other than LC transformation. This allows us to pay attention to temperature trends associated exclusively with LC changes.

6.1.1 Impact of Land Cover Changes on Temperature in Northern Europe

The prediction results presented in Table 6.1 show that all ten of the most frequent LC changes can lead to a significant increase in temperature during spring and/or summer. The predictions demonstrate that 9 out of 10 LC transformations lead to a temperature growth in spring, and four of them demonstrate increase for more than +0.5 °C. Our results did not reveal any LC transition which definitely leads to any temperature decrease in Northern Europe.

In general, LC is most often transformed into open shrublands in Northern Europe. Indeed, this occurs in 6 out of 10 times of the most common LC changes. In all cases, this causes a temperature rise in spring (MAM). Moreover, a prominent temperature growth above +0.5 °C occurs when LC transforms only into open shrublands. Let us discuss the temperature patterns associated with LC change in detail.

The transition from barren or sparsely vegetated LC to open shrublands results in the most significant temperature increase in spring between +1.1 and +2.4 °C. In addition, this LC transformation is responsible for the average annual temperature rise of +0.1 – +0.7 °C.

Every change from forest to other LC types leads to a temperature growth. For example, a transformation from the different forest types (evergreen needleleaf, mixed, deciduous broadleaf) to open shrublands cuases a warming by +0.5 – +1.9 °C in spring. The transition from evergreen needleleaf forest to permanent wetland results in a temperature growth in summer by +0.2 – +1.7 °C.

Permanent wetland is the most frequently changed LC type. 75 % of cases of permanent wetland transformation is the transition to different kinds of forests (evergreen needleleaf, mixed, deciduous broadleaf). This leads to increase in temperature in spring by +0.01 – +1.7 °C, while LC change from permanent wetland to evergreen needleleaf, mixed or deciduous broadleaf forests in summer increases temperature by +0.1 – +1.7 °C.

Most of studies consider LC transformations as anti-symmetrical [48]. So, a change from LC #1 to LC #2 should have the opposite effect with respect to a change from LC #2 to LC #1. Nevertheless, our predictions detected another relation between LC changes. There are symmetrical LC transformations through the most frequent LC transition in Northern Europe: from permanent wetland to evergreen needleleaf forest and from evergreen needleleaf forest to permanent wetland. Both of these LC changes contribute to a warming during summer. At first glance, this looks like a wrong behaviour of the ML model. In this thesis, we mainly focus on the ML side of this task and do not have a strong background in climatology. However, we can try to discuss this issue.

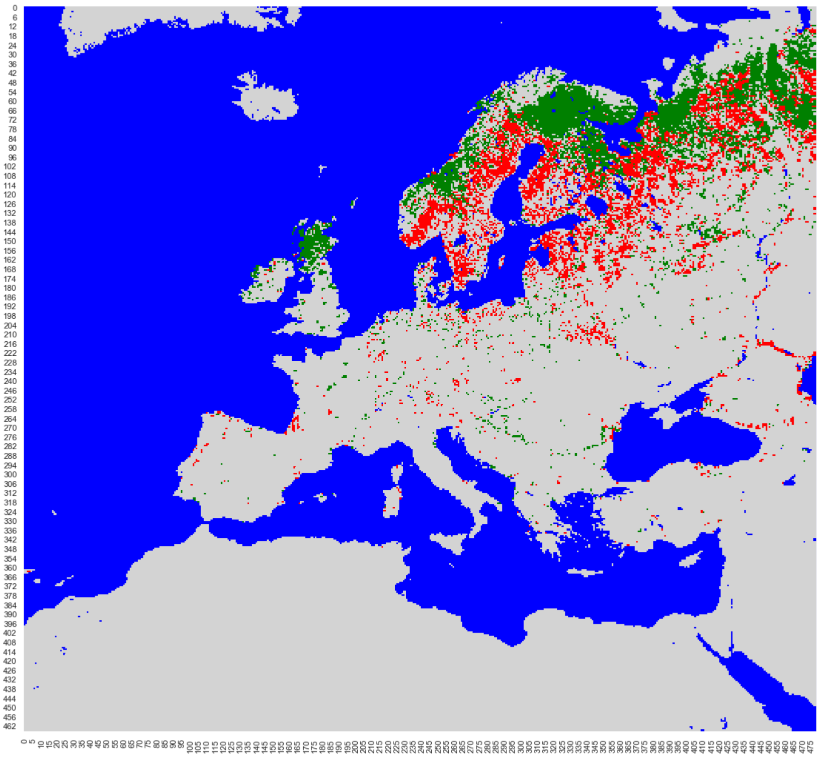
If we pay attention to the locations where such changes occured, then

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Permanent Wetland to Open Shrublands	8444	[−0.1177, 0.4338]	[−0.5883, 0.4403]	[0.3553, 1.6446]	[−0.5858, 0.9682]	[−0.7983, 0.1586]
Evergreen Needleleaf Forest to Open Shrublands	7728	[−0.1976, 0.3539]	[−0.5871, 0.4416]	[0.6168, 1.9061]	[−0.5377, 1.0164]	[−0.7309, 0.226]
Permanent Wetland to Evergreen Needleleaf Forest	6685	[−0.0996, 0.452]	[−0.5532, 0.4754]	[0.0165, 1.3058]	[0.1051, 1.6591]	[−0.6916, 0.2653]
Barren or Sparsely Vegetated to Open Shrublands	5860	[0.1398, 0.6913]	[−0.313, 0.7156]	[1.1442, 2.4334]	[−1.0866, 0.4674]	[−0.6445, 0.3124]
Permanent Wetland to Mixed Forest	5450	[−0.1356, 0.4159]	[−0.7163, 0.3123]	[0.4486, 1.7379]	[0.1063, 1.6604]	[−0.5375, 0.4194]
Evergreen Needleleaf Forest to Permanent Wetland	5191	[−0.1806, 0.3709]	[−0.775, 0.2537]	[−0.5402, 0.7491]	[0.1807, 1.7347]	[−0.661, 0.2959]
Mixed Forest to Open Shrublands	4935	[−0.2339, 0.3176]	[−0.4794, 0.5492]	[0.5399, 1.8292]	[−0.5467, 1.0074]	[−0.7691, 0.1877]
Permanent Wetland to Deciduous Broadleaf Forest	4648	[−0.166, 0.3855]	[−0.4265, 0.6021]	[0.174, 1.4633]	[−0.1017, 1.4523]	[−0.7823, 0.1746]
Deciduous Broadleaf Forest to Open Shrublands	4396	[−0.1111, 0.4405]	[−0.6223, 0.4063]	[0.5051, 1.7944]	[−0.7121, 0.842]	[−0.5298, 0.4271]
Cropland/Natural Vegetation Mosaic to Open Shrublands	4293	[−0.1426, 0.409]	[−0.4483, 0.5803]	[0.4013, 1.6906]	[−0.7162, 0.8378]	[−0.6488, 0.3081]

Table 6.1: Temperature changes in Northern Europe depending on LC transformation.

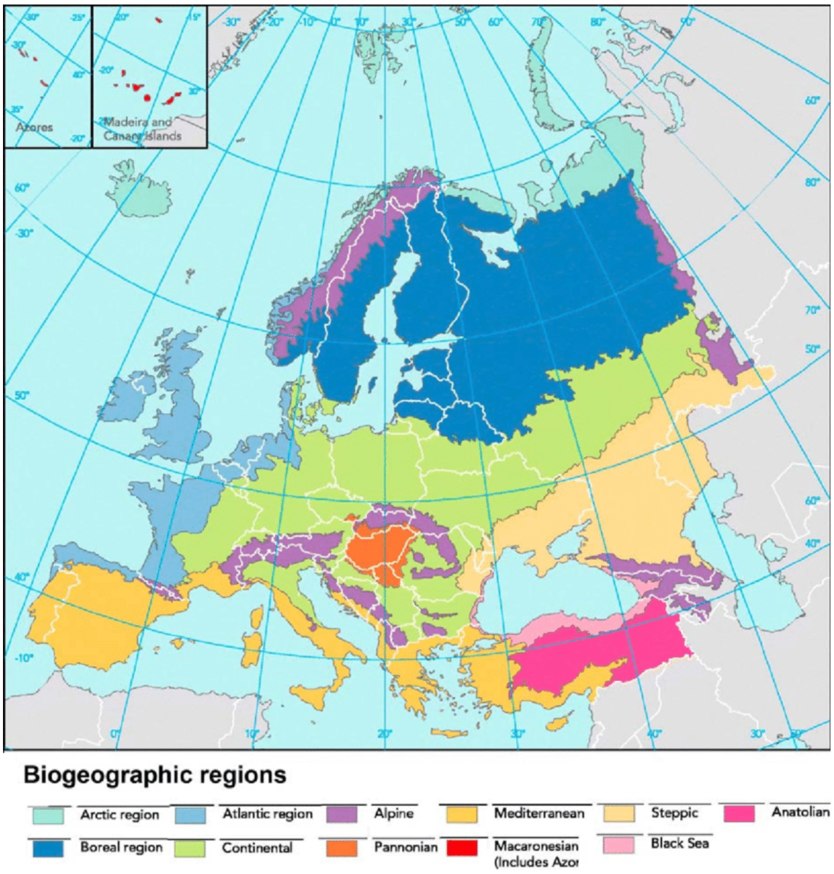
we can detect some clear patterns. In Figure 6.2, both of these types of LC changes are marked on the map. The LC transformation from permanent wetland to evergreen needleleaf forest is colored by green and located mainly along the coast line. This LC transition probably matches with the certain biogeographical regions such as Arctic, Atlantic and Alpine, which are shown in Figure 6.3. In turn, the LC change from evergreen needleleaf forest to permanent wetland (red color in Figure 6.2) occurs mainly in the boreal region (Figure 6.3). We assume that this difference in the biogeographical regions can explain the fact that the opposite LC changes result in a temperature growth.

Figure 6.2: LC changes in Northern Europe. Green color is the transformation from permanent wetland to evergreen needleleaf forest. Red color is the transition from evergreen needleleaf forest to permanent wetland.



Summarizing, we can conclude that LC changes that most often occur in Northern Europe can lead to an increase in temperature. We also detected that anti-symmetric LC transformations can contribute to a regional warming. This was rarely observed before, but we believe that it can be explained by the certain biogeographical location of these LC changes. However, this issue requires a more detailed consideration by researchers with a strong climate background.

Figure 6.3: Biogeographical regions in Europe. Adopted from Ref.[24]



6.1.2 Impact of Land Cover Changes on Temperature in Central Europe

The prediction results for Central Europe are presented in Table 6.2. The LC transitions in Central Europe occur more often than in Northern and Southern Europe. In general, despite the frequent LC transition in this region, the surface temperature is not affected critically. We can observe a significant temperature change only for 5 out of 13 most frequent LC transformations. Moreover, there is not any prediction interval demonstrating an increase or decrease above 0.5 °C. In contrast with Northern Europe, we did not detect any prominent temperature change in spring in Central Europe. However, we observe a temperature growth trend for the whole year in Central Europe more often than in the north.

All significant changes in temperature are associated with the LC transition to urban and built-up ares. Mostly it leads to a warming by +0.03 – +0.9 °C, which is observed through the whole year. The most frequent LC transformation in Central Europe is the shift from cropland to urban and built-up LC. For this LC change, we can predict the temperature growth during the whole year, in winter and summer. The LC transformation from barren or sparsely vegetated LC to urban and built-up area leads to a decrease in autumn temperature by -0.2 – -1.1 °C.

It can also be observed that 8 out of the 13 most frequent LC changes are related to the substitution of cropland or cropland/natural vegetation mosaic to other types of LC. However, our findings do not demonstrate its significant impact on temperature.

Summarizing, we can conclude that the change of LC to urban and built-up is the most frequent LC transition in Central Europe. Such LC transformations have a noticeable impact on regional temperature. In most cases, this leads to a warming, but we also observed that this can lead to a temperature decrease. One can also notice that, despite numerous transformations in LC, the impact of LC changes on temperature is mild. There is no LC transformation which results in critical temperature change when the prediction interval completely exceeds a variation of 0.5 °C.

6.1.3 Impact of Land Cover Changes on Temperature in Southern Europe

Table 6.3 shows the prediction results for extreme LC changes in Southern Europe. We observe prominent temperature changes for 9 out of 10 LC transformations. In contrast with other regions, LC transitions in Southern Europe more often lead to a temperature decrease, including a noteworthy cooling by more than -0.5 °C. Several warming trends, which take place in summer, were also identified.

Cropland and cropland/natural vegetation mosaic are most frequently replaced by another LC in Southern Europe. However, it is difficult to distinguish any temperature pattern associated with these LC changes. The transition from barren or sparsely vegetated LC to cropland leads to a temperature decrease for the whole year by -0.1 – -0.6 °C. For this LC change, we also observe a significant cooling by -0.3 – -1.0 °C during summer and by -0.1 – -0.6 °C during autumn. The shift from cropland/natural vegetation mosaic to cropland results in a prominent cooling by -0.8 – -1.5 °C in summer and by -0.1 – -1.1 °C in spring. The conversion from cropland/natural vegetation mosaic to evergreen needleleaf forest contributes to a temperature decline by -0.8 – -1.2 °C during winter and -0.03 – -1.1 °C during spring. The replacement of cropland by open shrublands causes a temperature decrease by -0.03 – -1.1 °C in spring and warming by +0.3 – +1.0 °C during summer.

A temperature growth during summer was detected for several LC transitions such as: from cropland or cropland/natural vegetation mosaic to urban and built-up with a temperature rise by +0.5 – +1.2 °C; from cropland/natural vegetation mosaic or barren/sparsely vegetated LC to open shrublands with a warming by +0.1 – +1.0 °C.

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Cropland to Urban and Built-Up	23387	[0.0373, 0.7504]	[0.0412, 0.7993]	[−0.8996, 0.7225]	[0.0409, 1.2777]	[−0.5651, 0.3416]
Cropland/Natural Vegetation Mosaic to Urban and Built-Up	18211	[−0.1335, 0.5796]	[−0.4248, 0.3333]	[−0.5393, 1.0828]	[−0.0097, 1.2272]	[−0.5714, 0.3353]
Cropland/Natural Vegetation Mosaic to Open Shrublands	14102	[−0.1972, 0.5159]	[−0.3326, 0.4255]	[−0.9703, 0.6518]	[−0.3001, 0.9367]	[−0.5031, 0.4036]
Cropland/Natural Vegetation Mosaic to Deciduous Broadleaf Forest	10880	[−0.5925, 0.1205]	[−0.3238, 0.4344]	[−0.7084, 0.9137]	[−1.0207, 0.2162]	[−0.8381, 0.0686]
Cropland to Open Shrublands	10420	[−0.2108, 0.5023]	[−0.2227, 0.5355]	[−0.492, 1.1301]	[−0.5929, 0.644]	[−0.4537, 0.453]
Cropland/Natural Vegetation Mosaic to Mixed Forest	9314	[−0.3991, 0.314]	[−0.4607, 0.2974]	[−0.3244, 1.2977]	[−0.925, 0.3118]	[−0.5842, 0.3225]
Grassland to Urban and Built-Up	8993	[0.0328, 0.7458]	[−0.1802, 0.5779]	[−0.6771, 0.945]	[−0.1435, 1.0934]	[−0.4331, 0.4736]
Cropland to Deciduous Broadleaf Forest	8816	[−0.4883, 0.2248]	[−0.3208, 0.4374]	[−0.5455, 1.0766]	[−0.9398, 0.297]	[−0.6289, 0.2778]
Deciduous Broadleaf Forest to Open Shrublands	8406	[−0.4172, 0.2959]	[−0.659, 0.0991]	[−1.2918, 0.3303]	[−0.4992, 0.7377]	[−0.2814, 0.6253]
Cropland/Natural Vegetation Mosaic to Evergreen Needleleaf Forest	8231	[−0.4309, 0.2821]	[−0.5338, 0.2243]	[−0.6073, 1.0148]	[−0.7261, 0.5107]	[−0.5276, 0.3791]
Deciduous Broadleaf Forest to Urban and Built-Up	7977	[0.051, 0.7641]	[−0.4977, 0.2604]	[−0.3922, 1.2299]	[0.1501, 1.3869]	[−0.3628, 0.5439]
Evergreen Needleleaf Forest to Urban and Built-Up	7406	[0.1921, 0.9051]	[0.1358, 0.894]	[−0.3556, 1.2665]	[−0.0935, 1.1433]	[−0.3311, 0.5756]
Barren or Sparsely Vegetated to Urban and Built-Up	5583	[−0.4757, 0.2373]	[−0.3166, 0.4415]	[−0.6784, 0.9437]	[−0.5866, 0.6502]	[−1.0893, −0.1826]

Table 6.2: Temperature changes in Central Europe depending on LC transformation.

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Barren or Sparsely Vegetated to Urban and Built-Up	5233	[−0.1294, 0.3696]	[−0.1071, 0.219]	[−1.0459, 0.0215]	[−0.0131, 0.7063]	[−0.1367, 0.4365]
Cropland to Urban and Built-Up	4473	[−0.1651, 0.3339]	[−0.0778, 0.2483]	[−0.8228, 0.2446]	[0.4718, 1.1912]	[−0.2072, 0.3661]
Barren or Sparsely Vegetated to Cropland	3455	[−0.6361, −0.1371]	[−0.1009, 0.2253]	[−0.8968, 0.1706]	[−0.9814, −0.262]	[−0.6312, −0.0579]
Cropland/Natural Vegetation Mosaic to Urban and Built-Up	3417	[−0.2611, 0.238]	[−0.1155, 0.2107]	[−0.9243, 0.1432]	[0.4579, 1.1773]	[−0.2176, 0.3557]
Cropland/Natural Vegetation Mosaic to Open Shrublands	3071	[−0.2362, 0.2628]	[−0.2812, 0.045]	[−0.7906, 0.2769]	[0.2316, 0.951]	[−0.405, 0.1683]
Cropland/Natural Vegetation Mosaic to Cropland	2029	[−0.3848, 0.1142]	[−0.1325, 0.1937]	[−1.1184, −0.051]	[−1.5166, −0.7972]	[−0.3396, 0.2337]
Barren or Sparsely Vegetated to Open Shrublands	1926	[−0.2398, 0.2593]	[−0.2134, 0.1128]	[−0.7118, 0.3556]	[0.0753, 0.7947]	[−0.3321, 0.2412]
Cropland/Natural Vegetation Mosaic to Evergreen Needleleaf Forest	1424	[−0.2675, 0.2315]	[−1.1594, −0.8333]	[−1.0964, −0.029]	[−0.2992, 0.4202]	[−0.525, 0.0483]
Cropland to Open Shrublands	1422	[−0.313, 0.1861]	[−0.2122, 0.1139]	[−1.1058, −0.0384]	[0.2722, 0.9916]	[−0.0905, 0.4828]
Cropland/Natural Vegetation Mosaic to Deciduous Broadleaf Forest	1412	[−0.4676, 0.0315]	[−0.1385, 0.1877]	[−1.1793, −0.1119]	[−0.4408, 0.2786]	[−0.4877, 0.0856]

Table 6.3: Temperature changes in Southern Europe depending on LC transformation.

Summarizing, we can conclude that the most frequent LC changes in Southern Europe mainly lead to a cooling. These patterns are in contrast with typical processes in two other regions. However, some LC transformations also result in a temperature increase during summer.

6.1.4 Impact of Land Cover Changes on Temperature in the Whole of Europe

The temperature prediction for extreme LC transitions based on the data set for the whole of Europe are shown in Table 6.4. We made predictions for the 20 most frequent LC changes. However, in the table, we present only the results by which it is possible to unambiguously determine the temperature trends.

In the Table 6.4, one can notice that most of LC transformations lead to a local warming. The most general trends are associated with the urban expansion and the cropland replacement. On the European scale, the urban expansion results in a temperature increase throughout the year. One can also observe that the cropland replacement by another LC causes a temperature increase in summer.

There are also three frequent types of LC transformation that were not observed for the three regions separately. Two of these LC replacements are related to deforestation. The transition from deciduous broadleaf forest to cropland and evergreen needleleaf forest to cropland lead to a temperature decrease through the whole year with especially significant cooling by -0.6 – -1.8 °C in summer. In turn, we predict that the LC change from cropland to mixed forest contributes to a warming during summer. One more LC change contributing to the regional cooling is the transformation from barren or sparsely vegetated LC to urban and built-up areas, which leads to a temperature decrease on -0.02 – -1.2 °C.

Summarizing, we can conclude that the most patterns predicted for the whole of Europe are also typical for the regions studied separately. However, some patterns can only be detected on a scale of the whole of Europe. Mostly, LC changes in Europe lead to a temperature increase. Nevertheless, we also found three LC transformations that contribute to the regional cooling.

6.1.5 Comparison of Land Cover and Temperature Changes in Different Regions

Let us compare the predictions for three separate regions (Northern, Central and Southern Europe) to reveal some common trends. The set of the most frequent LC changes noticeably varies depending on a region because of the typical LC that is intrinsic to a certain region. Even for the similar LC transformation we can observe different temperature

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Cropland to Urban and Built-Up	29694	[0.0074, 0.6232]	[−0.3763, 0.3403]	[−0.743, 0.6816]	[0.131, 1.3258]	[−0.5429, 0.2748]
Cropland/Natural Vegetation Mosaic to Open Shrublands	21539	[−0.2683, 0.3475]	[−0.5078, 0.2088]	[−0.9844, 0.4402]	[0.1809, 1.3757]	[−0.4492, 0.3685]
Evergreen Needleleaf Forest to Open Shrublands	16561	[−0.3856, 0.2302]	[0.2147, 0.9313]	[−0.8066, 0.6179]	[−0.3429, 0.8519]	[0.2222, 1.0399]
Cropland to Open Shrublands	13891	[−0.4164, 0.1994]	[−0.5177, 0.1988]	[−0.6879, 0.7367]	[0.1593, 1.3542]	[−0.4127, 0.405]
Cropland/Natural Vegetation Mosaic to Evergreen Needleleaf Forest	12671	[−0.3734, 0.2425]	[−0.4416, 0.275]	[−0.8785, 0.546]	[0.0558, 1.2506]	[−0.5779, 0.2398]
Barren or Sparsely Vegetated to Urban and Built-Up	12350	[−0.1896, 0.4262]	[−0.3071, 0.4095]	[−1.0188, 0.4057]	[−1.2128, −0.018]	[−0.6348, 0.1829]
Deciduous Broadleaf Forest to Urban and Built-Up	10773	[0.0911, 0.7069]	[−0.433, 0.2836]	[−0.1756, 1.2489]	[0.0826, 1.2774]	[−0.2724, 0.5453]
Evergreen Needleleaf Forest to Urban and Built-Up	10629	[0.0343, 0.6501]	[−0.1379, 0.5787]	[−0.444, 0.9805]	[−0.2144, 0.9804]	[−0.4064, 0.4113]
Deciduous Broadleaf Forest to Cropland	10587	[−0.943, −0.3271]	[−0.3909, 0.3257]	[−1.3147, 0.1098]	[−1.7565, −0.5617]	[−0.8687, −0.051]
Grassland to Urban and Built-Up	10448	[0.0068, 0.6227]	[−0.3413, 0.3753]	[−0.6201, 0.8045]	[−0.2194, 0.9754]	[−0.4165, 0.4012]
Barren or Sparsely Vegetated to Open Shrublands	10131	[0.1269, 0.7427]	[−0.1964, 0.5202]	[1.1181, 2.5427]	[−0.7157, 0.4791]	[−0.5438, 0.2739]
Evergreen Needleleaf Forest to Cropland	9983	[−0.9488, −0.3329]	[−0.1522, 0.5643]	[−1.2604, 0.1642]	[−1.7591, −0.5643]	[−0.5507, 0.267]
Permanent Wetland to Open Shrublands	9935	[−0.1733, 0.4425]	[−0.4717, 0.2449]	[0.2164, 1.6409]	[−0.3666, 0.8282]	[−0.7281, 0.0897]
Mixed Forest to Open Shrublands	9320	[−0.2837, 0.3322]	[−0.617, 0.0996]	[0.1774, 1.602]	[−0.359, 0.8359]	[−0.627, 0.1907]
Cropland to Mixed Forest	8943	[−0.2976, 0.3182]	[−0.3925, 0.3241]	[−0.7132, 0.7113]	[0.4207, 1.6156]	[−0.6425, 0.1752]

Table 6.4: Temperature changes in the whole of Europe depending on LC transformation.

patterns in these three regions. The only frequent LC transition observed in all regions is the transformation from cropland/natural vegetation mosaic to open shrubland. For this LC transition, we detected slightly different patterns in each region: when we consider the whole of Europe and Southern Europe, we obtain warming during summer; in Northern Europe – a warming during spring; while in Central Europe there are no significant temperature changes.

One can observe the different temperature behavior for a shift from deciduous broadleaf forest to open shrublands depending on the considered region. In Northern Europe, this transition leads to a significant increase of the average spring temperature by more than +0.5°C, while in Central Europe no significant changes were detected.

Northern and Southern Europe possess the common LC transition from barren or sparsely vegetated LC to open shrublands, which leads to a warming in both regions. However,a temperature growth in the north is detected not only for spring but also for the whole year, while in the south we observe a warming exclusively in spring.

Central and Southern Europe have seven similar LC transformations. However, only one of them has an unambiguous temperature pattern within the prediction interval in both regions. Hence, we can only compare the change from cropland to urban and build-up. We can notice some similarities there. First of all, this LC transformation is the most frequent in Central Europe and the second most frequent in Southern Europe. Secondly, we observe that in both regions this LC transition leads to a temperature increase by +0.04 – +1.3°C during summer.

Let us now turn to the comparison of the prediction for the whole of Europe with three separated regions. Indeed, we can notice similarities in temperature patterns, however, some trends slightly differ from each other. For example, th replacement of barren or sparsely vegetated LC by urban and built-up LC leads to a cooling in autumn in Central Europe, while for the whole of Europe we identify a cooling only in summer. The only transformation from cropland/natural vegetation mosaic to evergreen needle forest shows a significant difference for the whole of Europe and Southern Europe. This LC transformation demonstrates a cooling during winter and spring in the south, but contributes to a warming in summer on the scale of the whole of Europe.

Summarizing the conclusions through three regions, we can distinguish several common trends. First of all, all the LC transitions to open shrublands in Northern and Southern Europe lead to a temperature increase during spring or summer. Secondly, we can conclude that the most frequent changes in Southern Europe often lead to a temperature decrease, while in Northern Europe they contribute to a local warming. However, we also observe that LC changes have a different impact on temperature depending on the region. Therefore, it is important to study different areas separately to obtain a clear understanding of climate patterns.

6.2 Discussion

One of the questions raised in Section 1.2 is aimed to realize how ML techniques can improve the understanding of the climate systems. In this section, we compare our predictions of the impact of LC changes on the regional temperature with other works on this subject. In the first subsection, we discuss in general how our findings consistent with various studies. In the second subsection, we focus exclusively on the comparison of our findings based on ML with Ref. [48] where the same data set was used together with statistical analysis.

6.2.1 How Are Our Predictions Consistent With Other Studies?

The pattern detection with the ML algorithm is a quite new approach for prediction of the impact of LC changes. Thus, we would like to consider how our results are consistent with other studies based on statistical approaches and climate model simulations.

Many studies revealed a strong correlation between temperature increase and growth in shrub species [13, 19, 62, 83, 84]. Some of these researchers discussed the positive feedback loop when LC transitions affect climate, while temperature changes also influence LC transformation [31, 83, 84]. Firstly, a warming increases a spreading of shrublands. Then LC transition to shrublands influences the energy exchange, increasing the absorption of solar radiation. This, in turn, results in a temperature rise. However, it can be complicated to distinguish what is the main driver in this feedback loop. In this thesis, we study only on the impact of LC transformation on temperature growth, ignoring the effect of a warming on LC. So, we observed that the transition to open shrublands alone leads to a temperature increase in Northern and Southern Europe. Thus, one can assume the driving role of LC in the feedback loop was discussed above.

Some works demonstrate that shrubland increase in Arctic can lead to an annual temperature increase [11, 13, 62], which is consistent with our own findings. However, most articles only consider the growth of shrubs and do not pay attention to the initial cover. Therefore, our approach can help in understanding how prominent is the effect of LC transformation to shrubs depending on the initial LC. For instance, the replacement of barren or sparsely vegetated cover to shrublands causes a more significant warming than a temperature rise associated with transition from permanent wetland to open shrublands.

Urbanization and its impact on temperature is another subject which draws the interest of climate scientists. In general, researchers conclude that the transition to urban and built-up covers causes a warming [17, 48, 94, 95]. Indeed, we also observed that most of the LC changes to urban and built-up covers results in a temperature growth during the whole year as well as seasonally. However, we identify the cooling effect in Central Europe during autumn, associated with the transformation from barren or sparsely vegetated cover to urban and built-up areas. This finding is not consistent with the results presented by Huang et al. for the same climate data set [48]. We assume that it can be explained by the different subsets of data as well as different methodologies. We obtained this result from the ML model trained on the data set of Central Europe, while Haung et al. considered changes for the whole area and used a statistical approach. The effect of cooling due to urbanization is rarely observed in the climate literature but have already been reported previously. For example, Trusilova et al. [112] mentioned a reduction in temperature in cities located in a warm dry climate exactly in Central Europe, which is similar to our findings.

Deforestation and its contribution to a temperature increase are the important research subjects that have been explored by many authors [48, 71, 110]. In this thesis, we also observed the similar trend. Most LC changes associated with deforestation observed in our work lead to a significant temperature increase. These findings are consistent with the results of Huang et al., who demonstrated a warming effect of transition from forest to any LC using the same data set [48].

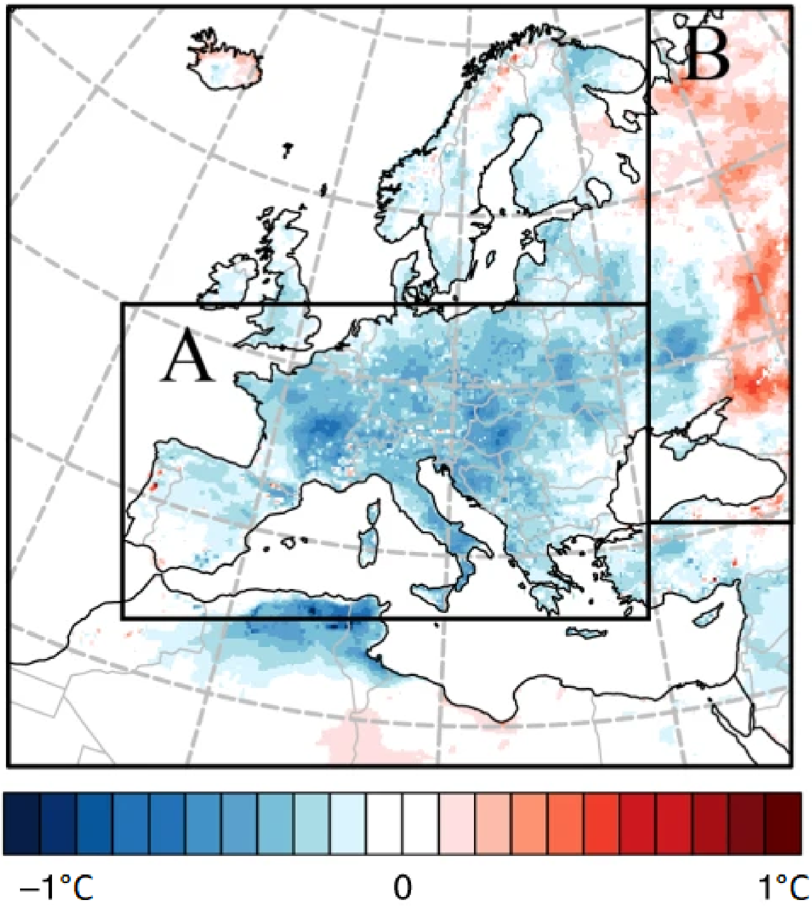
Afforestation is considered as a possible solution to the problem of the warming effect of deforestation because of its contribution to cooling [71, 94, 110]. In this thesis, we detected such a trend in Southern Europe where the shift from corpland/natural vegetation mosaic to evergreen needleleaf or deciduous broadlef forest results in a significant cooling. However, in Central Europe, we could not identify a clear pattern in temperature change associated with afforestation. Moreover, the transition from permanent wetland to any kind of forest contributes to a warming in Northern Europe. This is consistent with the results of Li et al. where a transition of any LC to forest leads to a cooling in tropical regions but to warming in high latitudes [71].

Summarizing, we can conclude that our predictions of the LC change impact on temperature are consistent with the main trends described by the IPCC [19, 94] and other studies. This supports the assumption that the ML techniques can be a powerful tool in climate science, and it is possible to develop a model that can make a meaningful prediction. In addition, our approach allows us to extract more complex patterns and have more clear understanding of the effect of different LC transitions. This demonstrates that the ML techniques can help to figure out the effect of LC changes on surface temperature.

6.2.2 How Are Our Predictions Consistent With Huang et al. [48]?

In our prediction of the LC changes impact on temperature, we use the same data set as Huang et al. in their article [48]. However, we applied the Random Forest ML algorithm to distinguish patterns, while Huang et al. used a statistical method based on ridge regression. Therefore, in this section, we compare our results with the findings from the mentioned article.

Figure 6.4: The simulated average annual temperature changes depending on LC changes between 1992 and 2015 in Sectors A and B in Europe (°C). Adopted from Ref. [48].



First of all, it is important to mention that Huang et al. consider three regions within Europe. The two regions are sector A and sector B, shown in Figure 6.4, while the third region is the whole of Europe. The second important difference is that Huang et al. excluded all data points without LC changes from the analysis. To be consistent with this article, we train new ML models for the same three regions from Figure 6.4 and limit our train data set to data points with LC changes. We made predictions for the 20 most frequent LC changes in each region. The results are presented in Tables 6.5 – 6.7, where we included only LC transformations with a significant impact on temperature. Comparing the results from Table 6.5 with our previous prediction for the whole of Europe, we can conclude that the trends and patterns are more or less similar.

Our results have a larger prediction interval, and we cannot make predictions for rare LC transformations. In general, our findings are quite similar to the results reported by Huang et al., but they are not identical. Comparing our predictions with the results from Ref. [48] for a data set of the whole of Europe, we make the following conclusions:

1. In the article [48], the LC change from forest to urban leads to a warming by +(0.27±0.06) °C. In our predictions, we have a similar trend, but the slightly different values. Namely, the transition from deciduous broadleaf forest to urban and built-up areas is associated with a temperature growth by +(0.38±0.35) °C, while from evergreen needleleaf forest to urban and built-up areas corresponds to an increase of +0.41±0.35°C. Thus, these results are consistent with Ref. [48] within the uncertainty interval.

2. It was also predicted [48] that the transformation from forests to grassland contributes to a warming by $+(0.23 \pm 0.06)$ °C. However, in our predictions, we could not detect a significant impact of such LC transition on temperature. Our model provides a value of $-(0.17 \pm 0.35)$ °C for transformation from evergreen forest to grassland and a value of $-(0.1 \pm 0.35)$ °C for transition from deciduous forest to grassland. This does not allow us to unambiguously conclude whether a warming or cooling is associated with this LC change.

3. Huang et al. stated that the impact of LC transition from forests to cropland leads to a temperature increase of $+(0.15 \pm 0.03)$ °C, but our results are not consistent with this. We observe a cooling during the whole year for transformation from evergreen needleleaf forest to cropland by -(0.67±0.35) °C and from deciduous broadleaf forest to cropland by -(0.66±0.35)°C.

4. Another statement made in the article is that the urban expansion always leads to a warming regardless of the initial LC type. Our estimates are consistent with these findings.

5. LC transformation from wetland to shrubland results in a temperature increase according to Ref. [48]. We observe the same trend in our findings using the data set for the whole of Europe.

Let us now consider the results for the sector A and B. The first sector labelled as A covers Central and Western Europe that are characterized

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Cropland to Urban and Built-Up	29694	[−0.1095, 0.5937]	[−0.4502, 0.3942]	[−0.785, 0.8301]	[0.027, 1.3695]	[−0.5228, 0.3572]
Evergreen Needleleaf Forest to Open Shrublands	16561	[−0.3181, 0.3851]	[0.2658, 1.1102]	[−0.3451, 1.27]	[−0.6863, 0.6563]	[0.2583, 1.1384]
Cropland to Deciduous Broadleaf Forest	11646	[−0.5629, 0.1403]	[−0.4679, 0.3765]	[−1.634, −0.0189]	[0.2414, 1.5839]	[−0.6733, 0.2067]
Deciduous Broadleaf Forest to Urban and Built-Up	10773	[0.0264, 0.7296]	[−0.513, 0.3314]	[−0.274, 1.341]	[0.015, 1.3575]	[−0.4198, 0.4603]
Evergreen Needleleaf Forest to Urban and Built-Up	10629	[0.0596, 0.7628]	[−0.2953, 0.549]	[−0.5673, 1.0477]	[−0.2642, 1.0784]	[−0.2066, 0.6735]
Deciduous Broadleaf Forest to Cropland	10587	[−1.0134, −0.3102]	[−0.4249, 0.4195]	[−1.1042, 0.5108]	[−1.673, −0.3305]	[−0.7769, 0.1031]
Barren or Sparsely Vegetated to Open Shrublands	10131	[0.0637, 0.7669]	[−0.2913, 0.5531]	[0.9013, 2.5163]	[−0.8623, 0.4802]	[−0.5733, 0.3067]
Evergreen Needleleaf Forest to Cropland	9983	[−1.0217, −0.3185]	[−0.2695, 0.5749]	[−1.0603, 0.5547]	[−1.6745, −0.332]	[−0.3364, 0.5437]
Permanent Wetland to Open Shrublands	9935	[−0.1182, 0.585]	[−0.5377, 0.3067]	[0.2947, 1.9098]	[−0.595, 0.7475]	[−0.8171, 0.063]

Table 6.5: Temperature changes in the whole of Europe depending on LC transformation. Predictions are made on the basis of only data points with LC changes.

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Cropland/Natural Vegetation Mosaic to Deciduous Broadleaf Forest	8976	[−0.6412, −0.0524]	[−0.2406, 0.328]	[−1.2267, 0.1087]	[−0.7084, 0.3442]	[−0.6221, 0.2074]
Cropland to Open Shrublands	8524	[−0.5578, 0.031]	[−0.298, 0.2707]	[−1.3739, −0.0384]	[−0.5809, 0.4716]	[−0.6457, 0.1838]
Cropland to Deciduous Broadleaf Forest	7122	[−0.685, −0.0962]	[−0.2119, 0.3568]	[−0.4789, 0.8565]	[−0.8909, 0.1616]	[−0.8157, 0.0138]
Cropland/Natural Vegetation Mosaic to Evergreen Needleleaf Forest	6414	[−0.4846, 0.1042]	[−0.7952, −0.2266]	[−1.1644, 0.171]	[−0.4558, 0.5967]	[−0.5663, 0.2631]
Evergreen Needleleaf Forest to Urban and Built-Up	6096	[0.1467, 0.7355]	[0.15, 0.7187]	[−0.8192, 0.5162]	[0.1045, 1.157]	[−0.4747, 0.3547]
Evergreen Needleleaf Forest to Open Shrublands	6019	[−0.1043, 0.4845]	[0.0963, 0.6649]	[−0.9576, 0.3778]	[−0.4927, 0.5598]	[0.1503, 0.9798]
Cropland to Evergreen Needleleaf Forest	5053	[−0.5695, 0.0193]	[−0.3515, 0.2171]	[−0.8173, 0.5182]	[−0.8472, 0.2053]	[−0.9201, −0.0906]
Deciduous Broadleaf Forest to Grassland	4277	[−0.5226, 0.0662]	[−0.3219, 0.2468]	[−1.347, −0.0115]	[−0.7677, 0.2848]	[−0.2892, 0.5403]

Table 6.6: Temperature changes in Section A depending on LC transformation. Predictions are made on the basis of only data points with LC changes.

LC change	Event Rate	Year, °C	Winter (DJF), °C	Spring (MAM), °C	Summer (JJA), °C	Autumn (SON), °C
Cropland/Natural Vegetation Mosaic to Open Shrublands	5145	[−0.1033, 0.6071]	[−0.4577, 0.5521]	[−0.638, 0.5391]	[0.3519, 1.6471]	[−0.5828, 0.3482]
Cropland/Natural Vegetation Mosaic to Deciduous Broadleaf Forest	4290	[−0.5308, 0.1795]	[−0.5325, 0.4773]	[−0.2372, 0.94]	[−1.3278, −0.0326]	[−0.8725, 0.0585]
Cropland/Natural Vegetation Mosaic to Mixed Forest	3870	[−0.1859, 0.5244]	[−0.5949, 0.4149]	[0.1376, 1.3148]	[−0.4581, 0.8371]	[−0.7402, 0.1908]
Evergreen Needleleaf Forest to Open Shrublands	3481	[0.0355, 0.7458]	[−0.7113, 0.2985]	[−0.4385, 0.7387]	[0.1904, 1.4856]	[−0.4755, 0.4555]
Permanent Wetland to Open Shrublands	3060	[−0.1929, 0.5174]	[−0.7506, 0.2592]	[0.1301, 1.3073]	[−0.0641, 1.2311]	[−0.7126, 0.2184]
Mixed Forest to Open Shrublands	2527	[−0.0828, 0.6275]	[−0.7151, 0.2947]	[0.0776, 1.2547]	[−0.1364, 1.1588]	[−0.7471, 0.1839]
Cropland to Mixed Forest	2483	[−0.3609, 0.3494]	[−0.4181, 0.5917]	[0.24, 1.4172]	[−0.7061, 0.5891]	[−0.3362, 0.5948]
Permanent Wetland to Evergreen Needleleaf Forest	2476	[0.1376, 0.8479]	[−0.573, 0.4368]	[0.9988, 2.176]	[0.2335, 1.5287]	[−0.7487, 0.1823]
Evergreen Needleleaf Forest to Permanent Wetland	2296	[−0.1686, 0.5417]	[−0.7761, 0.2337]	[−0.4648, 0.7124]	[0.5553, 1.8505]	[−0.7805, 0.1505]

Table 6.7: Temperature changes in Section B depending on LC transformation. Predictions are made on the basis of only data points with LC changes.

by a huge amount of abundant croplands. Comparing our results for the sector A with Huang et al., we can discuss the following phenomena:

1. Huang et al. paid special attention to an impact of deforestation. The conversion from evergreen forest to cropland leads to a temperature increase by +(0.21±0.03) °C. The transition from deciduous forest to cropland causes a temperature growth of +(0.12±0.03) °C. In our predictions, we could not detect these particular patterns in the Sector A because of limited amount of data points with such LC transitions. However, we observed a similar trend for deforestation associated with urbanisation. For example, in our study, the transformation from evergreen needleleaf forest to urban area contributes to a temperature increase of +(0.44±0.29) °C.

2. Huang et al. revealed that the transitions from cropland to another LC mostly result in a cooling, and our findings indicate the same pattern (see Table 6.6).

The sector B is Eastern Europe. In this area, Huang et al. highlighted two significant trends:

1. Huang et al. found that the conversion from cropland to evergreen or deciduous forest results in a temperature increase. We also detected that afforestation leads to a temperature rise. For example, the conversion from cropland or cropland/natural vegetation mosaic to mixed forest contributes to the regional warming by +(0.24-1.14) °C and +(0.1-1.13) °C, respectively.

2. Huang et al. also found out that LC change from evergreen to cropland leads to a temperature change by -(0.14±0.05)°C, and the transformation from deciduous forest to cropland results in a cooling by -(0.10±0.05)°C. We did not find a significant impact of these LC transformation on temperature exactly in the Sector B. However, this trend is typical for the whole of Europe where the change from evergreen needleleaf forest to cropland affects temperature by -(0.67±0.35) °C, and the transition for deciduous broadleaf forest to cropland leads to a temperature change by -(0.66±0.35)°C.

The method introduced by Huang et al. postulates anti-symmetry in the impact of LC transformations. It means that the transformation from LC #1 to LC #2 should have opposite effect on temperature with respect to the transition from LC #2 to LC #1. However, as mentioned in Section 6.1.1, our ML approach does not assume this pre-condition hence it is more flexible and may reveal some hidden patterns. In the sector B, we found out that some symmetric LC transitions lead to a similar impact on temperature. The LC change from both permanent wetland to evergreen needleleaf forest and from evergreen needleleaf forest to permanent wetland causes a warming.

Summarizing, we can conclude that our findings mainly demonstrate the same trends as in the discussed article [48]. Huang et al. used the statistical approach based on ridge regression, while we used the ML method. Nevertheless, in both studies, many similar patterns such as temperature increase due to urban expansion and regional warming due to deforestation in Central Europe, were found. In addition, both works revealed a difference in the impact of LC changes depending on the regions. The main difference is that the method of Huang et al. assumes anti-symmetry, while our ML approach does not have such an assumption. This allows us to observe new patterns that have not been observed before.

6.3 Summary

In this Chapter 6, we presented and discussed the predictions of the regional temperature changes in Europe due to LC transformations that took place between 1992 and 2015. We started from the presentation of predictions for three regions: Northern, Central, and Southern Europe. We considered the LC changes that are typical for each region and their impact on the regional temperature. Then we considered the impact of LC changes in the whole of Europe. Finally, we examined our findings in the context of other studies on this subject.

One can notice that each region can be characterized by some specific LC transformations that are rarely observed in other areas in Europe. For example, shrublands expansion is often observed in Northern Europe but rarely happens in other areas. In addition to that, the most typical LC changes in the region usually lead to similar temperature patterns. For example, most LC changes in Northern Europe result in a significant temperature increase, especially, during spring. LC transformations in Central Europe are also mainly contributing to warming and this trend is typical for the whole year. In contrast, many changes in Southern Europe facilitate significant cooling.

In Section 6.2, we compared our results with findings from the other studies. We demonstrated that our predictions are not contradicting other studies. Moreover, we also revealed some LC-associated temperature patterns which were not examined by other studies including Huang et al. [48]. We paid special attention to this publication because our works were carried out on the same data sets as Ref. [48]. Moreover, Huang et al. used a statistical technique to reveal temperature pattern, and this comparison could verify our ML approach.

Huang et al. divided Europe into three regions differently than we did in our experiment. To compare our results directly, we developed new additional ML models for these three regions. It is interesting to note that our ML predictions are similar to the "statistical" results by Huang within the given confidence intervals.

Our ML approach has also exposed some temperature trends which were not observed in Ref. [48]. For example, according to our prediction LC transformations from permanent wetland to evergreen needleleaf forest and from evergreen needleleaf forest to permanent wetland are symmetric: both LC changes lead to the warming in Northern Europe (see Section 6.1.1). We assume that this symmetric behavior is associated with different biogeographical regions within Northern Europe.

Summarizing the discussion for this chapter, we came to the conclusion that RF can be considered as an efficient technique for the climate-related tasks. It provides results consistant with other studies, and also add to the understanding of the regional temperature change due to LC transformations.

Chapter 7

Conclusion

7.1 Summary

The last hundred years, climatologists have observed the constant temperature increase. Since 1850, every decade has been characterized by a higher temperature than every previous one. Global warming leads to the significant changes in climate system and ecosystem. Nowadays, scientists agree that human-related activities are the main reasons of warming. Lately, the IPCC has paid special attention to land use as the cause of climate changes and also suggested the possible ways to mitigate its impact. The IPCC has also mentioned a lack of articles on the impact of LC transformations on regional climate. However, the study of this impact can be challenging.

The amount of climate data gathered by sensors and satellites grows exponentially, but the predictive ability does not increase at the same rate. One of the dominant reasons is the complexity of climate system and interconnections between different elements. Climate models describe processes in climate system with the set of mathematical equations. They allow performing simulations of climate changes. However, the simulation results are also very complex, and it can be challenging to reveal some non-trivial patterns. ML can be a solution to this problem and used to find out hidden patterns in the simulation results of mathematical climate models.

In this thesis, we aim to solve the issues described above. In Section 1.2, we defined three main objectives of our study. Our first goal is to find out how ML methods can be applied to climate data with spatial dependency. The second objective is to develop an ML model that can predict temperature changes due to LC transformations. The last aim is to use our ML model to improve the understanding of climate processes.

To adapt ML techniques to the spatially dependent climate data, we should develop an assessment method to figure out an ML algorithm with the best performance for the certain task. We developed an assessment method called the 5x2 CV test for ML algorithms on the spatially dependent data. We performed an experiment on the synthetic spatially dependent data to verify that the 5x2 CV test provides reasonable results. In the experiment, we examined the probability that the 5x2 CV test can detect a difference in algorithm performances. The experimental results clearly demonstrated that the 5x2 CV test method can be an efficient tool to assess ML algorithms. Moreover, ML algorithms can be compared with the reference ML technique (base line algorithm), therefore, the 5x2 CV test method can also act as an evaluation method providing the absolute metric of algorithm performance.

To test the practicability and correctness of our evaluation, we performed a case study using the RF algorithm with application to the impact of extreme LC changes on temperature. In general, our predictions follow the climate patterns revealed by statistical methods on the same initial data set. However, we were also able to detect several additional complex trends associated with the impact of LC changes on local surface temperature in Europe. Summarizing, the 5x2 CV test can be an effective tool to assess ML algorithms for tasks with spatially dependent data, and the RF algorithm can be considered as a prominent ML approach for a data set with spatial dependence.

During the work on this thesis, we have considered different aspects of climate science and ML. We believe that our findings help to apply ML in various areas of climate science. To summarize our study, we want to list major conclusions:

1. In Section 2.1, we showed that global warming is an indisputable fact and LC transformations can both contribute to temperature increase or mitigate warming.

2. We described in Section 2.1.4 that climate models provide complex simulated data that can be analyzed, for example, with ML methods. In this thesis, we aimed to use ML to study the impact of LC changes on temperature.

3. ML should not be blindly applied to any type of data because some kinds of data are not i.i.d. as it was shown in Section 2.2.

4. Our experiment shows that ML algorithms can be adapted to certain data types, for example, to spatially dependent data.

5. The efficient tool of algorithms' assessment is required to adapt ML methods to spatially dependent data. In section 5.2, we proposed 5x2 CV test for this task and verified that it provides reliable results.

6. According to our estimations in Section 5.3, models based on the RF algorithm possess the superior performance in the prediction of the temperature changes due to LC transformations.

7. Using the ML models based on RF algorithm revealed impacts of LC changes on temperature that are not only consistent with other studies, but also exhibited previously unobserved temperature patterns.

8. We performed our experiment on the same data sets as Huang et al. [48] where the impact of LC changes on temperature was studied using the statistical method. Our results are consistent with findings from Ref. [48] as it is discussed in Section 6.2.2.

7.2 Contributions

In Section 1.2, we raised three research questions that we consider in this thesis. Summarizing our work, we quote our objectives and show how this study solved the assigned issues.

1. *How can supervised ML techniques be applied to spatially depended data with a high variability?*

   This question was answered in Section 5.2 where we developed method 5x2 CV test to assess algorithms' performances and verified its reliability on synthetic data sets. We demonstrated that this method can also evaluate ML techniques on the spatially dependent data by comparing them with base line. Moreover, matching the results of 5x2 CV test with the spatial CV, we demonstrated that 5x2 CV test is efficient evaluation tool even when ML model has low R^2 scores.

2. *Is it possible to develop a model based on an ML approach, which can predict the impact of LC changes on temperature?*

   This objective was addressed in Section 5.3 where we used 5x2 CV test to compare four ML algorithms with each other and with baseline. According to our evaluations, models based on RF demonstrate superior performance in the prediction of regional temperature changes due to LC transformations. In Section 6.2, we demonstrated that the results of the models' predictions are consistent with other studies. This supports the efficiency and expediency of the models based on RF for the studied task.

3. *How can an ML approach help to understand the effects of LC changes on surface temperature?*

   We answered this question in Section 6 where we displayed the results of predictions of the impact of LC changes on surface temperature. We performed an analysis on the same data sets as Huang et al. in Ref. [48] where the statistical method was used. Mostly, our results are similar to the outcome of this article. However, we also revealed some impacts of LC changes on temperature that, to our best knowledge, have not been previously observed by other studies. Probably, such results have been achieved because of peculiar properties of ML models. For example, they can distinguish symmetrical patterns.

Concluding, ML techniques can be a useful tool in climate science. The developed 5x2 CV test can efficiently identify an ML algorithm most suitable for a certain problem. The major advantage of 5x2 CV test is that it can be applied to any tasks on spatially dependent data.

7.3 Future Research

Because of time limitations regarding the duration of master project, there are several directions in which our work can be expanded. In the current section, we describe some of the main improvements that can be made.

Certainly, a further study should be carried out to consider more different methods of data splitting for the 5x2 CV test method. Some other data splitting strategies may provide a superior probability of detection the difference in algorithms performance on the spatially dependent data. For example, a synthetic data set can be expanded to the two-dimensional spatially dependent data, and the splitting strategy can be based on the random picking of sectors. Going further, one can develop a completely other assessment method that has no assumption of i.i.d. and designed specifically for spatio-temporal dependent data.

In addition, various ML techniques and statistical methods can be compared with the 5x2 CV test to develop new tools for climate science. This will allow performing enhanced predictions and extracting more complex patterns. Another possible way to improve predictions is to compare models with another input variables. For example, in our thesis we used only the difference in LC as an input variable, but one can use both the initial and final LC as the input variables. Furthermore, not only one data point but groups of neighboring data points can be used as the input variables, which can also have some effect on the predictive validity.

As part of the case study, we considered only the simulation for two years, 1992 and 2015, while it is possible to run a climate model with LC data for several years. This is time and power-consuming, but allows better understanding of the impact of LC transitions on temperature. In addition, one can study not only the impact on temperature, but also on other climate characteristics, for example, on precipitations.

Appendices

Appendix A

Changes in Land Cover between 1992 and 2015

In Chapter 3, we described LC data where we also showed changes in urban areas and evergreen needleleaf forests between 1992 and 2015. In Appendix A, we demonstrate the grade of other LC changes within Europe. Different colors represent the proportion of a certain LC in each cell on the grid.

Figure A.1: Deciduous Broadleaf Forest LC in 1992 and 2015

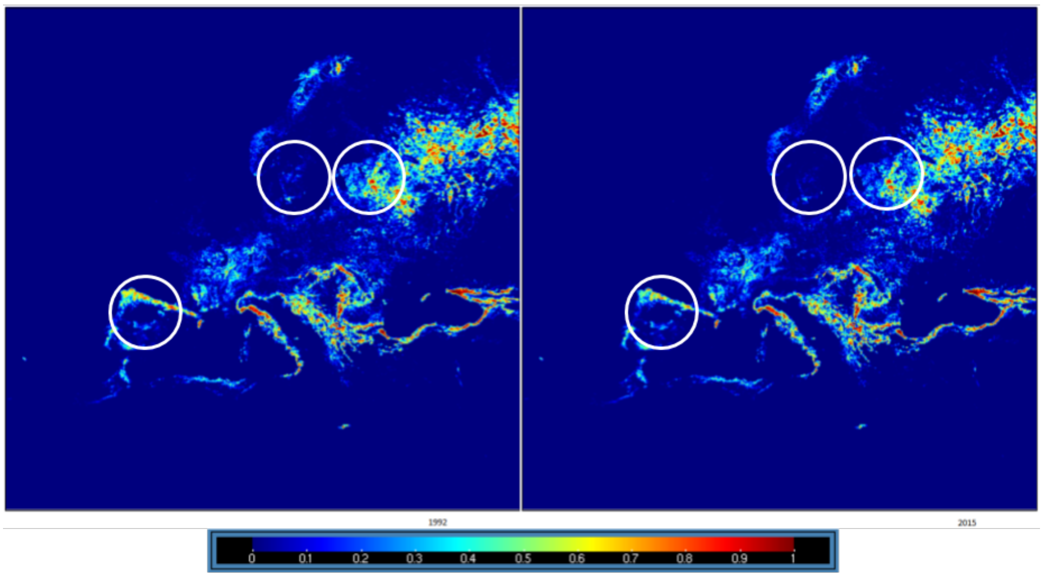


Figure A.2: Open Shrublands LC in 1992 and 2015

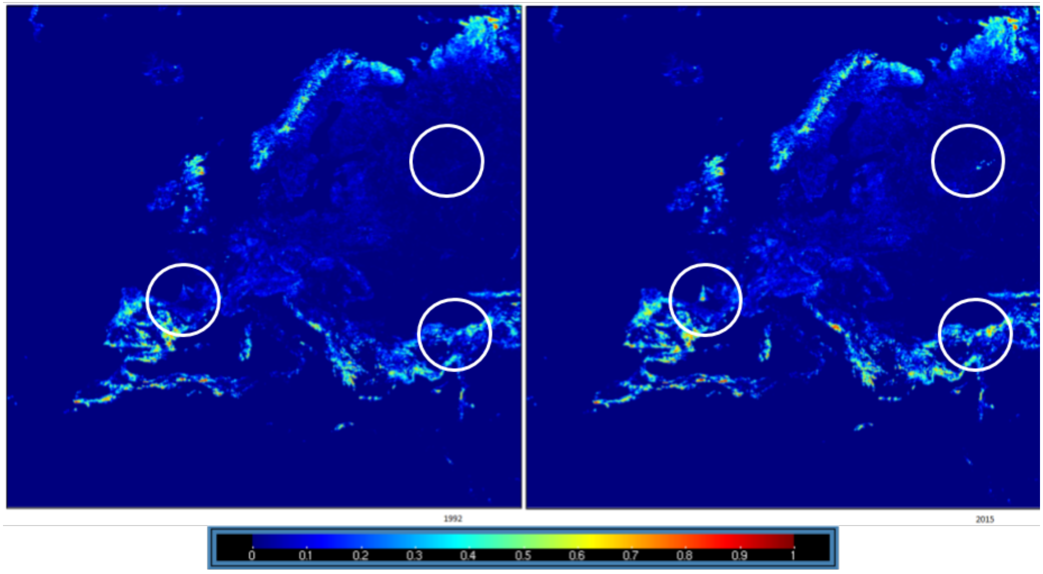


Figure A.3: Grassland LC in 1992 and 2015

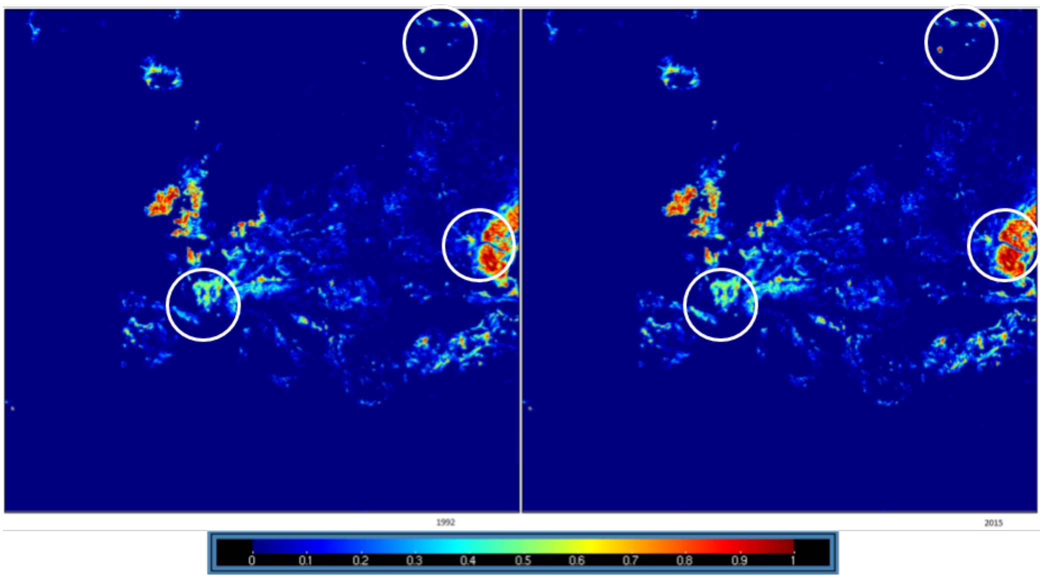


Figure A.4: Permanent WetLC in 1992 and 2015

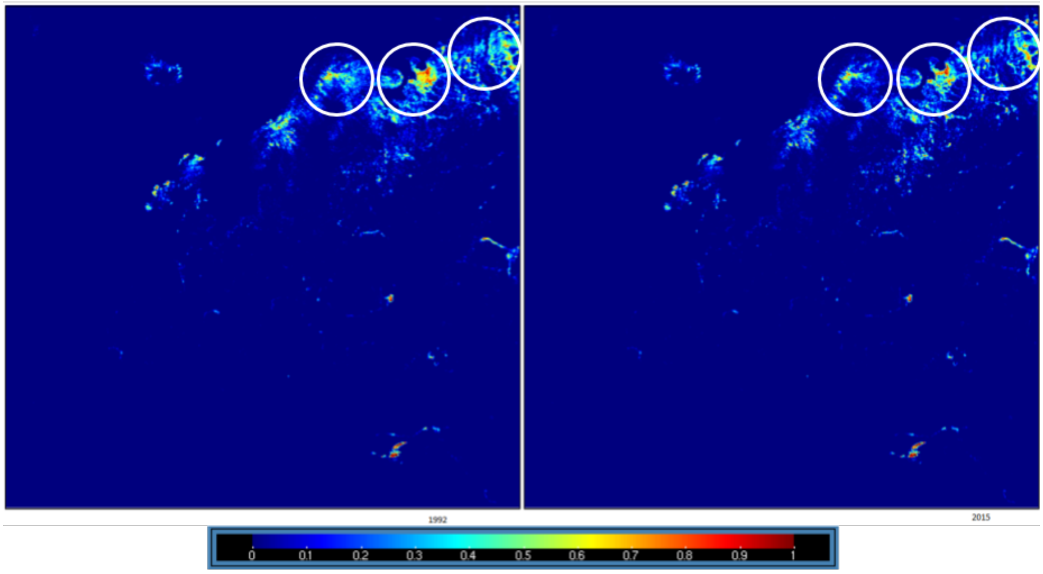


Figure A.5: Cropland LC in 1992 and 2015

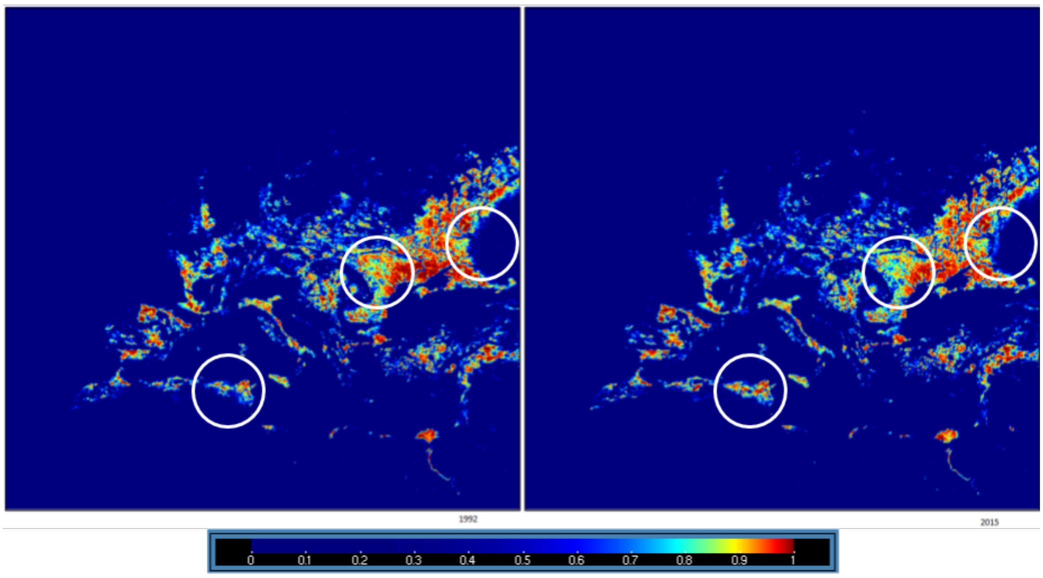


Figure A.6: Cropland/Natural Vegetation Mosaic LC in 1992 and 2015

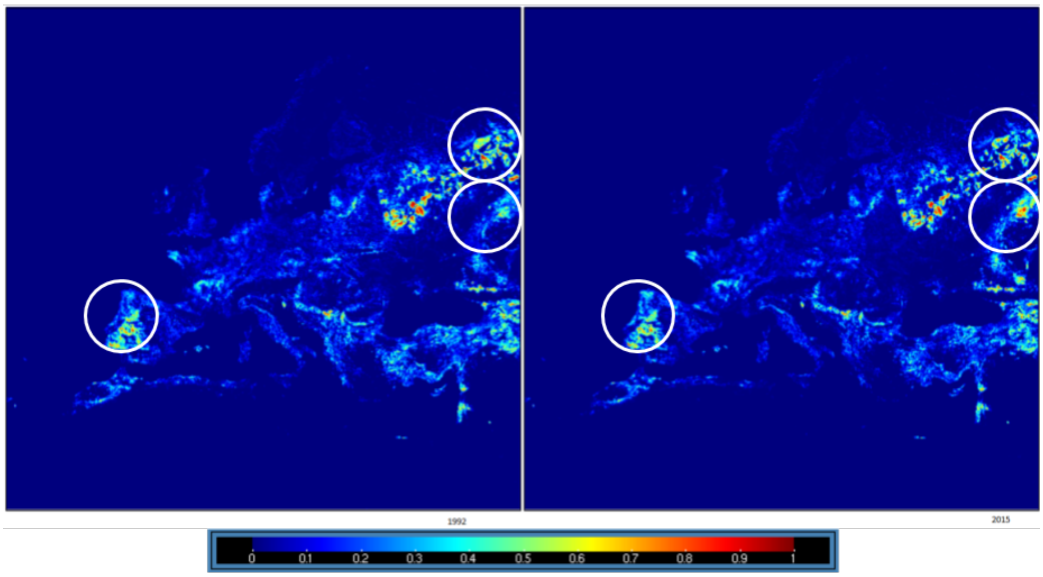
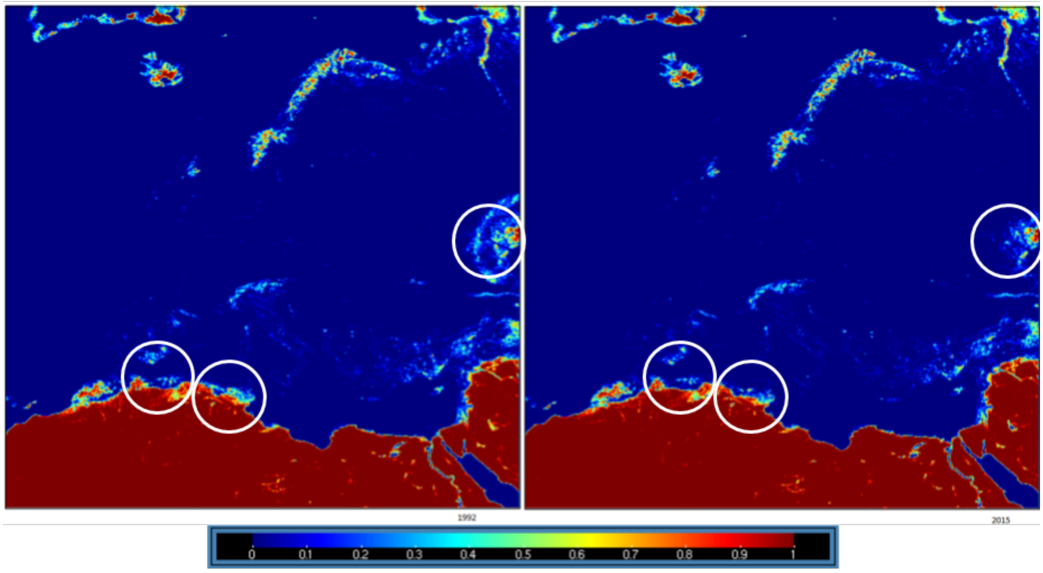


Figure A.7: Barren or Sparsely Vegetated LC in 1992 and 2015



Bibliography

[1] Cleveland Abbe. 'The Physical Basis Of Long-range Weather Forecasts'. In: *Monthly Weather Review* 29.12 (1901), pp. 551–561. DOI: 10.1175/1520-0493(1901)29[551c:TPBOLW]2.0.CO;2.

[2] Athos Agapiou. 'Remote Sensing Heritage In A Petabyte-scale: Satellite Data And Heritage Earth Engine© Applications'. In: *International Journal of Digital Earth* 10.1 (Nov. 2016), pp. 1–18. DOI: 10.1080/17538947.2016.1250829.

[3] European Environment Agency. *Global and European temperature*. Copenhagen, Denmark, 2019. URL: https://www.eea.europa.eu/data-and-maps/indicators/global-and-european-temperature-9/assessment.

[4] Camilo Alcantara et al. 'Mapping The Extent Of Abandoned Farmland In Central And Eastern Europe Using MODIS Time Series Satellite Data'. In: *Environmental Research Letters* 8.3 (2013), p. 035035. DOI: 10.1088/1748-9326/8/3/035035.

[5] Ethem Alpaydın. 'Machine Learning'. In: *WIREs Computational Statistics* 3.3 (2011), pp. 195–203. DOI: 10.1002/wics.166.

[6] Sylvain Arlot and Alain Celisse. 'A Survey of Cross Validation Procedures for Model Selection'. In: *Statistics Surveys* 4 (July 2009), pp. 40–79. DOI: 10.1214/09-SS054.

[7] Volker Bahn and Brian J. Mcgill. 'Testing The Predictive Performance Of Distribution Models'. In: *Oikos* 122.3 (2013), pp. 321–331. DOI: 10.1111/j.1600-0706.2012.00299.x.

[8] Govindasamy Bala et al. 'Combined Climate And Carbon-cycle Effects Of Large-scale Deforestation'. In: *Proceedings of the National Academy of Sciences* 104.16 (2007), pp. 6550–6555. DOI: 10.1073/pnas.0608998104.

[9] Debasish Basak, Srimanta Pal and Dipak Patranabis. 'Support Vector Regression'. In: *Neural Information Processing Letters and Reviews*. Vol. 11. 10. Nov. 2007, pp. 203–224.

[10] Niklas Boers et al. 'Complex Networks Reveal Global Pattern Of Extreme-rainfall Teleconnections'. In: *Nature* 566.7744 (Feb. 2019), pp. 373–377. DOI: 10.1038/s41586-018-0872-x.

[11] CJW Bonfils et al. ‘On The Influence Of Shrub Height And Expansion On Northern High Latitude Climate’. In: *Environmental Research Letters* 7.1 (2012), p. 015503. DOI: 10.1088/1748-9326/7/1/015503.

[12] Leo Breiman. ‘Random Forests’. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[13] F S Chapin et al. ‘Role Of Land-surface Changes In Arctic Summer Warming’. In: *Science* 310.5748 (2005), pp. 657–660. DOI: 10.1126/science.1117368.

[14] Jie Chen et al. ‘A Comparison Of Linear Regression, Regularization, And Machine Learning Algorithms To Develop Europe-wide Spatial Models Of Fine Particles And Nitrogen Dioxide’. In: *Environment International* 130 (2019), p. 104934. DOI: 10.1016/j.envint.2019.104934.

[15] Francesco Cherubini et al. ‘Quantifying The Climate Response To Extreme Land Cover Changes In Europe With A Regional Model’. In: *Environmental Research Letters* 13.7 (May 2018), p. 074002. DOI: 10.1088/1748-9326/aac794.

[16] Francesco Cherubini et al. ‘Quantifying The Climate Response To Extreme Land Cover Changes In Europe With A Regional Model’. In: *Environmental Research Letters* 13.7 (June 2018), p. 074002. DOI: 10.1088/1748-9326/aac794.

[17] A Chrysanthou et al. ‘The Effects Of Urbanization On The Rise Of The European Temperature Since 1960’. In: *Geophysical Research Letters* 41.21 (2014), pp. 7716–7722. DOI: 10.1002/2014GL061154.

[18] J Cohen. ‘A Power Primer’. In: *Psychological Bulletin* 112.1 (1992), pp. 155–159. DOI: 10.1037/0033-2909.112.1.155.

[19] Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)] *IPCC, 2014: Climate Change 2014: Synthesis Report*. Geneva, Switzerland, 151 pp.: IPCC, 2014.

[20] Claire Cooper et al. ‘Evaluating The Relationship Between Climate Change And Volcanism’. In: *Earth-Science Reviews* 177 (Nov. 2018), pp. 238–247. DOI: 10.1016/j.earscirev.2017.11.009.

[21] Thomas Crowley. ‘Causes of Climate Change Over the Past 1000 Years’. In: *Science* 289.5477 (Aug. 2000), pp. 270–277. DOI: 10.1126/science.289.5477.270.

[22] Ulrich Cubasch et al. 'Simulation of the influence of solar radiation variations on the global climate with an ocean-atmosphere general circulation model.' In: *Climate Dynamics* 13 (Jan. 1997), pp. 757–767. DOI: 10.1007/s003820050196.

[23] Virginia Dale. 'The Relationship Between Land-Use Change and Climate Change'. In: *Ecological Applications* 7.3 (Aug. 1997), pp. 753–769. DOI: 10.1890/1051-0761(1997)007[0753:TRBLUC]2.0.CO;2.

[24] Claude De Broyer et al. *Action Plan for Conservation of Groundwater Biodiversity. Fifth EU FP: Global Change, Climate and Biodiversity*. Jan. 2005. DOI: 10.13140/RG.2.1.2961.5121.

[25] Peter Denning et al. 'Computing as a discipline'. eng. In: *Communications of the ACM* 32.1 (1989), pp. 9–23. ISSN: 00010782. DOI: 10.1145/63238.63239.

[26] N Devaraju et al. 'Quantifying The Relative Importance Of Direct And Indirect Biophysical Effects Of Deforestation On Surface Temperature And Teleconnections'. In: *Journal of Climate* 31.10 (2018), pp. 3811–3829. DOI: 10.1175/JCLI-D-17-0563.1.

[27] Antonio Di Gregorio. *Land Cover Classification System: Classification Concepts And User Manual: LCCS*. Vol. 2. Food & Agriculture Org., 2005.

[28] Thomas G Dietterich. 'Approximate Statistical Tests For Comparing Supervised Classification Learning Algorithms'. In: *Neural Computation* 10.7 (1998), pp. 1895–1923. DOI: 10.1162/089976698300017197.

[29] Matthias Döörries. 'In The Public Eye: Volcanology And Climate Change Studies In The 20th Century'. In: *Historical Studies in the Natural Sciences* 37.1 (2006), pp. 87–125. DOI: 10.1525/hsps.2006.37.1.87.

[30] Gregory Duveiller, Josh Hooker and Alessandro Cescatti. 'The Mark Of Vegetation Change On Earth's Surface Energy Balance'. In: *Nature Communications* 9.1 (2018), p. 679. DOI: 10.1038/s41467-017-02810-8.

[31] Sarah C Elmendorf et al. 'Plot-scale Evidence Of Tundra Vegetation Change And Links To Recent Summer Warming'. In: *Nature Climate Change* 2.6 (2012), pp. 453–457. DOI: 10.1038/NCLIMATE1465.

[32] H Epstein et al. 'Tundra Greenness'. In: *Arctic Report Card 2018* (2018). URL: https://arctic.noaa.gov/Report-Card/Report-Card-2018/ArtMID/7878/ArticleID/777/Tundra-Greenness.

[33] Faghmous and Vipin H. Kumar James. *Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities*. 2014.

[34] R. Falk and Nancy Miller. 'A Primer for Soft Modeling'. In: *The University of Akron Press* (Jan. 1992).

[35] Johannes J. Feddema et al. 'The Importance of Land-Cover Change in Simulating Future Climates'. In: *Science* 310.5754 (2005), pp. 1674–1678. DOI: 10.1126/science.1118160.

[36] Kirsten L Findell et al. 'Modeled Impact Of Anthropogenic Land Cover Change On Climate'. In: *Journal of Climate* 20.14 (2007), pp. 3621–3634. DOI: 10.1175/JCLI4185.1.

[37] Kirsten Findell et al. 'The Impact Of Anthropogenic Land Use And Land Cover Change On Regional Climate Extremes'. In: *Nature Communications* 8.1 (Dec. 2017), pp. 1–10. DOI: 10.1038/s41467-017-01038-w.

[38] Bart Geerts. 'Trends in Atmospheric Science Journals: A Reader's Perspective'. In: *Bulletin of the American Meteorological Society* 80.4 (1999), pp. 639–652. DOI: 10.1175/1520-0477(1999)080<0639:TIASJA>2.0.CO;2.

[39] Andreas Gobiet, Daniela Jacob and Euro-Cordex Community. 'A New Generation Of Regional Climate Simulations For Europe: The EURO-CORDEX Initiative'. In: *EGU General Assembly Conference Abstracts*. Apr. 2012, p. 8211.

[40] Scott J Goetz et al. 'Satellite-observed Photosynthetic Trends Across Boreal North America Associated With Climate And Fire Disturbance'. In: *Proceedings of the National Academy of Sciences* 102.38 (2005), pp. 13521–13525. DOI: 10.1073/pnas.0506179102.

[41] Lee Hannah. *Chapter 2 - The Climate System and Climate Change*. Ed. by Lee Hannah. Second Edition. Boston: Academic Press, 2015, pp. 13–53. ISBN: 978-0-12-420218-4. DOI: 10.1016/B978-0-12-420218-4.00002-0.

[42] J. Hansen et al. 'Climate Impact of Increasing Atmospheric Carbon Dioxide'. In: *Science* 213.4511 (1981), pp. 957–966. DOI: 10.1126/science.213.4511.957.

[43] Gail Hartfield, Jessica Blunden and Derek S. Arndt. 'State of the Climate in 2017'. In: *Bulletin of the American Meteorological Society* 99.8 (2018), Si–S310. DOI: 10.1175/2018BAMSStateoftheClimate.1.

[44] Trevor Hastie. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. Springer Series in Statistics. New York, NY: Springer New York : Imprint: Springer, 2009. ISBN: 9781282827264. DOI: 10.1007/978-0-387-84858-7.

[45] Gabi Hegerl et al. 'Multi-fingerprint Detection And Attribution Analysis Of Greenhouse Gas, Greenhouse Gas-plus-aerosol And Solar Forced Climate Change'. In: *Climate Dynamics* 13.9 (1997), pp. 613–634. DOI: 10.1007/s003820050186.

[46] Gabriele Hegerl et al. 'Optimal Detection And Attribution Of Climate Change: Sensitivity Of Results To Climate Model Differences'. In: *Climate Dynamics* 16.10-11 (Oct. 2000), pp. 737–754. DOI: 10.1007/s003820000071.

[47] Tai-Wen Hsu et al. 'A Study Of Extreme Value Analysis On Typhoon Wave'. In: *Coastal Engineering Proceedings* 1.34 (Jan. 2014), p. 38. DOI: 10.9753/icce.v34.waves.38.

[48] Bo Huang et al. 'Predominant Regional Biophysical Cooling From Recent Land Cover Changes In Europe'. In: *Nature Communications* 11.1066 (Feb. 2020), pp. 1–13. DOI: 10.1038/s41467-020-14890-0.

[49] NA Hughes and A Henderson-Sellers. 'The Effect Of Spatial And Temporal Averaging On Sampling Strategies For Cloud Amount Data'. In: *Bulletin of the American Meteorological Society* 64.3 (1983), pp. 250–257. DOI: 10.1175/1520-0477(1983)064<0250:TEOSAT>2.0.CO;2.

[50] Chris Huntingford et al. 'Machine learning and artificial intelligence to aid climate change research and preparedness'. In: *Environmental Research Letters* 14.12 (Nov. 2019), p. 124007. DOI: 10.1088/1748-9326/ab4e55.

[51] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018. DOI: 10.1111/j.1467-856X.2005.00178.x.

[52] Daniela Jacob et al. 'EURO-CORDEX: New High-resolution Climate Change Projections For European Impact Research'. In: *Regional Environmental Change* 14.2 (Apr. 2014), pp. 563–578. DOI: 10.1007/s10113-013-0499-2.

[53] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. DOI: 10.1007/978-1-4614-7138-7.

[54] Phil D Jones et al. 'Assessment Of Urbanization Effects In Time Series Of Surface Air Temperature Over Land'. In: *Nature* 347.6289 (1990), pp. 169–172. DOI: 10.1038/347169a0.

[55] Eugenia Kalnay and Ming Cai. 'Impact of Urbanization and Land-Use Change on Climate'. In: *Nature* 423.6939 (June 2003), pp. 528–31. DOI: 10.1038/nature01675.

[56] TF Keenan and WJ Riley. 'Greening Of The Land Surface In The World's Cold Regions Consistent With Recent Warming'. In: *Nature Climate Change* 8.9 (2018), pp. 825–828. DOI: 10.1038/s41558-018-0258-y.

[57] J. T. Kiehl and Kevin E. Trenberth. 'Earth's Annual Global Mean Energy Budget'. In: *Bulletin of the American Meteorological Society* 78.2 (1997), pp. 197–208. DOI: 10.1175/1520-0477(1997)078<0197:EAGMEB>2.0.CO;2.

[58] Remko Klaver et al. 'Effective Resolution In High Resolution Global Atmospheric Models For Climate Studies'. In: *Atmospheric Science Letters* 21.4 (Feb. 2020), e952. DOI: 10.1002/asl.952.

[59] Thomas Knutson et al. 'Detection and Attribution of Climate Change'. In: *Climate Science Special Report: Fourth National Climate Assessment* 1 (2017), pp. 114–132. DOI: 10.7930/J01834ND.

[60] Sven Kotlarski et al. 'Regional Climate Modeling On European Scales : A Joint Standard Evaluation Of The EURO-CORDEX RCM Ensemble'. In: *Geoscientific Model Development* 7.4 (July 2014), pp. 1297–1333. DOI: 10.5194/gmd-7-1297-2014.

[61] Tobias Kuemmerle et al. 'Hotspots of Land Use Change in Europe'. In: *Environmental Research Letters* 11.6 (2016), p. 064020. DOI: 10.1088/1748-9326/11/6/064020.

[62] David M Lawrence and Sean C Swenson. 'Permafrost Response To Increasing Arctic Shrub Abundance Depends On The Relative Influence Of Shrubs On Local Soil Cooling Versus Large-scale Climate Warming'. In: *Environmental Research Letters* 6.4 (2011), p. 045504. DOI: 10.1088/1748-9326/6/4/045504.

[63] Le Treut, H., R. Somerville, U. Cubasch, Y. Ding, C. Mauritzen, A. Mokssit, T. Peterson and M. Prather. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)] *2007: Historical Overview of Climate Change. In: Climate Change 2007: The Physical Science Basis.* Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2007.

[64] Quentin Lejeune et al. 'Historical Deforestation Locally Increased The Intensity Of Hot Days In Northern Mid-latitudes'. In: *Nature Climate Change* 8.5 (2018), pp. 386–390. DOI: 10.1038/s41558-018-0131-z.

[65] Timothy Lenton et al. 'Tipping Elements in the Earth's Climate System'. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.6 (Mar. 2008), pp. 1786–93. DOI: 10.1073/pnas.0705414105.

[66] Marcel Leroux. *Global Warming — Myth or Reality? The Erring Ways of Climatology*. Berlin, Heidelberg: Springer Science & Business Media, 2005, pp. 99–122. ISBN: 978-3-540-28100-9. DOI: 10.1007/3-540-28100-2_6.

[67] Myroslava Lesiv et al. ‘Spatial Distribution Of Arable And Abandoned Land Across Former Soviet Union Countries’. In: *Scientific Data* 5 (2018), p. 180056. DOI: 10.1038/sdata.2018.56.

[68] Jinfeng Li et al. ‘Bibliometric Analysis Of Atmospheric Simulation Trends In Meteorology And Atmospheric Science Journals’. In: *Croatica Chemica Acta* 82.3 (Dec. 2009), pp. 695–705. DOI: 10.5562/cca3210.

[69] Jin Li et al. ‘Application Of Machine Learning Methods To Spatial Interpolation Of Environmental Variables’. In: *Environmental Modelling & Software* 26.12 (2011), pp. 1647–1659. DOI: 10.1016/j.envsoft.2011.07.004.

[70] Yan Li et al. ‘Local Cooling And Warming Effects Of Forests Based On Satellite Observations’. In: *Nature Communications* 6 (2015), p. 6603. DOI: 10.1038/ncomms7603.

[71] Yan Li et al. ‘Potential And Actual Impacts Of Deforestation And Afforestation On Land Surface Temperature’. In: *Journal of Geophysical Research: Atmospheres* 121.24 (2016), pp. 14–372. DOI: 10.1002/2016JD024969.

[72] Yunjie Liu et al. ‘Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets’. In: *ArXiv Preprint ArXiv:1605.01156* (May 2016).

[73] Sebastiaan Luyssaert et al. ‘Land Management And Land-cover Change Have Impacts Of Similar Magnitude On Surface Temperature’. In: *Nature Climate Change* 4.5 (Apr. 2014), pp. 389–393. DOI: 10.1038/nclimate2196.

[74] Peter Lynch. ‘The Origins Of Computer Weather Prediction And Climate Modeling’. In: *Journal of Computational Physics* 227.7 (Mar. 2008), pp. 3431–3444. DOI: 10.1016/j.jcp.2007.02.034.

[75] M. Collins et al., Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, ed. T. F. Stocker et al. *Long-Term Climate Change: Projections, Commitments and Irreversibility, in Climate Change 2013: The Physical Science Basis.* Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013.

[76] Roland A. Madden and V. Ramanathan. ‘Detecting Climate Change due to Increasing Carbon Dioxide’. In: *Science* 209.4458 (1980), pp. 763–768. DOI: 10.1126/science.209.4458.763.

[77] Gregg Marland et al. ‘The Climatic Impacts Of Land Surface Change And Carbon Management, And The Implications For Climate-change Mitigation Policy’. In: *Climate Policy* 3.2 (2003), pp. 149–157. DOI: 10.1016/S1469-3062(03)00028-7.

[78] Ahilleas Maurellis and Jonathan Tennyson. 'The Climatic Effects of Water Vapour'. In: *Physics World* 16.5 (May 2003), p. 29. DOI: 10.1088/2058-7058/16/5/33.

[79] Tom M Mitchell. *Machine learning*. McGraw-Hill series in computer science, Artificial intelligence. New York: McGraw-Hill, 1997. ISBN: 0070428077. DOI: 10.1036/0070428077.

[80] Mohssen Mohammed, Muhammad Badruddin Khan and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. 1st ed. CRC Press, 2016. DOI: 10.1201/9781315371658.

[81] Ferenc Moksony. 'Small Is Beautiful: The Use and Interpretation of R2 in Social Research'. In: *Szociológiai Szemle, Special issue* (Jan. 1990), pp. 130–138.

[82] Nicole Mölders. *Land-Use and Land-Cover Changes. Impact on Climate and Air Quality*. Dordrecht, Netherlands: Springer, 2012. DOI: 10.1007/978-94-007-1527-1.

[83] H Myers-Smith et al. 'Shrub Expansion In Tundra Ecosystems: Dynamics, Impacts And Research Priorities'. In: *Environmental Research Letters* 6.4 (Dec. 2011), p. 045509. DOI: 10.1088/1748-9326/6/4/045509.

[84] Isla H Myers-Smith et al. 'Climate Sensitivity Of Shrub Growth Across The Tundra Biome'. In: *Nature Climate Change* 5.9 (2015), pp. 887–891. DOI: 10.1038/NCLIMATE2697.

[85] Ranga B Myneni et al. 'Increased Plant Growth In The Northern High Latitudes From 1981 To 1991'. In: *Nature* 386.6626 (1997), pp. 698–702. DOI: 10.1038/386698a0.

[86] NASA. *Graphic: Development of Climate Models*. 2012. URL: https://www.giss.nasa.gov/research/briefs/puma_02/evolution.gif.

[87] California Institute of Technology NASA - Jet Propulsion Laboratory. *Graphic: Temperature vs Solar Activity*. 2019. URL: https://climate.nasa.gov/climate_resources/189/graphic-temperature-vs-solar-activity/.

[88] National Oceanic and Atmospheric Administration. *Climate Model - Atmospheric Model Schematic*. 2017. URL: https://celebrating200years.noaa.gov/breakthroughs/climate_model/AtmosphericModelSchematic.png.

[89] Oleg Okun and Helen Priisalu. 'Random Forest For Gene Expression Based Cancer Classification: Overlooked Issues'. In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer. 2007, pp. 483–490. DOI: 10.1007/978-3-540-72849-8_61.

[90] Sandra Oliveira et al. 'Modeling Spatial Patterns Of Fire Occurrence In Mediterranean Europe Using Multiple Regression And Random Forest'. In: *Forest Ecology and Management* 275 (2012), pp. 117–129. DOI: 10.1016/j.foreco.2012.03.003.

[91] World Meteorological Organization. *WMO Confirms 2019 As Second Hottest Year On Record*. 2020. URL: https://public.wmo.int/en/media/press-release/wmo-confirms-2019-second-hottest-year-record.

[92] World Meteorological Organization. *WMO Provisional Statement on the State of the Global Climate in 2019*. 2019. URL: https://library.wmo.int/doc_num.php?explnum_id=10108.

[93] Jonathan T Overpeck et al. 'Climate Data Challenges in the 21st Century'. In: *Science* 331.6018 (2011), pp. 700–702. DOI: 10.1126/science.1197869.

[94] P.R. Shukla, J. Skea, R. Slade, R. van Diemen, E. Haughey, J. Malley, M. Pathak, J. Portugal Pereira (eds.) *Technical Summary, 2019. In: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. IPCC, 2019.

[95] Roberta Paranunzio et al. 'Evaluating The Effects Of Urbanization Evolution On Air Temperature Trends Using Nightlight Satellite Data'. In: *Atmosphere* 10.3 (2019), p. 117. DOI: 10.3390/atmos10030117.

[96] Thomas C Parker et al. 'Exploring Drivers Of Litter Decomposition In A Greening Arctic: Results From A Transplant Experiment Across A Treeline'. In: *Ecology* 99.10 (2018), pp. 2284–2294. DOI: 10.1002/ecy.2442.

[97] Lucia Perugini et al. 'Biophysical Effects On Temperature And Precipitation Due To Land Cover Change'. In: *Environmental Research Letters* 12.5 (2017), p. 053002. DOI: 10.1088/1748-9326/aa6b3f.

[98] Norman A. Phillips. 'The General Circulation Of The Atmosphere: A Numerical Experiment'. In: *Quarterly Journal of the Royal Meteorological Society* 82.352 (1956), pp. 123–164. DOI: 10.1002/qj.49708235202.

[99] Ben Poulter et al. 'Plant Functional Type Classification For Earth System Models: Results From The European Space Agency's Land Cover Climate Change Initiative'. In: *Geoscientific Model Development* 8 (2015), pp. 2315–2328. DOI: 10.5194/gmd-8-2315-2015.

[100] Jayme A Prevedello et al. 'Impacts Of Forestation And Deforestation On Local Temperature Across The Globe'. In: *PloS one* 14.3 (2019). DOI: 10.1371/journal.pone.0213368.

[101] Simo Puntanen. 'Linear Regression Analysis: Theory and Computing by Xin Yan, Xiao Gang Su'. In: *International Statistical Review* 78.1 (2010), pp. 144–144. DOI: 10.1111/j.1751-5823.2010.00109_11.x.

[102] Markus Reichstein et al. 'Deep Learning And Process Understanding For Data-driven Earth System Science'. In: *Nature* 566.7743 (Feb. 2019), pp. 195–204. DOI: 10.1038/s41586-019-0912-1.

[103] David R Roberts et al. 'Cross-validation Strategies For Data With Temporal, Spatial, Hierarchical, Or Phylogenetic Structure'. In: *Ecography* 40.8 (2017), pp. 913–929. DOI: 10.1111/ecog.02881.

[104] V.F. Rodriguez-Galiano et al. 'An Assessment Of The Effectiveness Of A Random Forest Classifier For Land-cover Classification'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 67 (2012), pp. 93–104. DOI: 10.1016/j.isprsjprs.2011.11.002.

[105] Benjamin Santer, K.E. Taylor and J.E. Penner. 'A Search For Human Influences On The Thermal Structure Of The Atmosphere'. In: *Nature* 382.6586 (Aug. 1995), pp. 39–46. DOI: 10.2172/116649.

[106] Catherine E Scott et al. 'Impact On Short-lived Climate Forcers Increases Projected Warming Due To Deforestation'. In: *Nature Communications* 9.157 (2018), pp. 1–9. DOI: 10.1038/s41467-017-02412-4.

[107] Durga Shrestha and Dimitri Solomatine. 'Machine Learning Approaches For Estimation Of Prediction Interval For The Model Output'. In: *Neural Networks : The Official Journal Of The International Neural Network Society* 19.2 (Apr. 2006), pp. 225–35. DOI: 10.1016/j.neunet.2006.01.012.

[108] Peter A. Stott et al. 'External Control of 20th Century Temperature by Natural and Anthropogenic Forcings'. In: *Science* 290.5499 (2000), pp. 2133–2137. DOI: 10.1126/science.290.5499.2133.

[109] R. J. Stouffer, S. Manabe and K. Ya. Vinnikov. 'Model Assessment Of The Role Of Natural Variability In Recent Global Warming'. In: *Nature* 367.6464 (Feb. 1994), pp. 634–636. DOI: 10.1038/367634a0.

[110] Gustav Strandberg and Erik Kjellström. 'Climate Impacts From Afforestation And Deforestation In Europe'. In: *Earth Interactions* 23.1 (2019), pp. 1–27. DOI: 10.1175/EI-D-17-0033.1.

[111] Waldo Tobler. 'Cellular Geography'. In: *Philosophy in Geography* 20.1 (Jan. 1979), pp. 379–386. DOI: 10.1007/978-94-009-9394-5_18.

[112] K Trusilova et al. 'Urbanization Impacts On The Climate In Europe: Numerical Experiments by the PSU–NCAR Mesoscale Model (MM5)'. In: *Journal of Applied Meteorology and Climatology* 47.5 (2008), pp. 1442–1455. DOI: 10.1175/2007JAMC1624.1.

[113] Unidata. *NetCDF Documentation*. 2019. URL: https://www.unidata.ucar.edu/software/netcdf/docs/index.html.

[114] Roeland Van Malderen et al. 'A Multi-site Intercomparison Of Integrated Water Vapour Observations For Climate Change Analysis'. In: *Atmospheric Measurement Techniques* 7.8 (Aug. 2014), pp. 2487–2512. DOI: 10.5194/amt-7-2487-2014.

[115] Peter H Verburg et al. 'From Land Cover Change To Land Function Dynamics: A Major Challenge To Improve Land Characterization'. In: *Journal of Environmental Management* 90.3 (2009), pp. 1327–1335. DOI: 10.1016/j.jenvman.2008.08.005.

[116] Jingfeng Wang, Rafael L Bras and Elfatih AB Eltahir. 'The Impact Of Observed Deforestation On The Mesoscale Distribution Of Rainfall And Clouds In Amazonia'. In: *Journal of Hydrometeorology* 1.3 (2000), pp. 267–286. DOI: 10.1175/1525-7541(2000)001<0267:TIOODO>2.0.CO;2.

[117] Qihao Weng. 'Impacts of Urbanization on Land Surface Temperature and Water Quality'. In: *Techniques and Methods in Urban Remote Sensing*. Wiley-IEEE Press, 2020, pp. 267–306. DOI: 10.1002/9781119307303.ch11.

[118] Tom M. L. Wigley and Sarah C. B. Raper. 'Natural Variability Of The Climate System And Detection Of The Greenhouse Effect'. In: *Nature* 344.6264 (1990), pp. 324–327. DOI: 10.1038/344324a0.

[119] David Wolpert and William Macready. 'Macready, W.G.: No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1(1), 67-82'. In: *Evolutionary Computation, IEEE Transactions on* 1.1 (Apr. 1997), pp. 67–82. DOI: 10.1109/4235.585893.

[120] Chunyan Wu et al. 'Modeling And Estimating Aboveground Biomass Of Dacrydium Pierrei In China Using Machine Learning With Climate Change'. In: *Journal of Environmental Management* 234 (Mar. 2019), pp. 167–179. DOI: 10.1016/j.jenvman.2018.12.090.

[121] Hui Yang et al. 'Strong but Intermittent Spatial Covariations in Tropical Land Temperature'. In: *Geophysical Research Letters* 46.1 (Dec. 2018), pp. 356–364. DOI: 10.1029/2018GL080463.