FACULTY OF
BIOSCIENCE ENGINEERING

# UNDERSTANDING VEGETATION ANOMALIES WITH MACHINE LEARNING METHODS

Ruben Ingels
Student ID: 01202223

Promotor: Prof. Dr. Willem Waegeman
Copromotor: Dr. Matthias Demuzere
Tutor(s): Ir. Stijn Decubber

Master thesis submitted for obtaining the degree: master in Bio-ingenieurswetenschappen.
Academic year: 2017 - 2018

GHENT
UNIVERSITY

Gent, August 24, 2018

| | |
|---|---|
| The promotor, | The author, |
| Prof. Dr. Willem Waegeman | Ruben Ingels |

I would like to thank everyone who was involved in the makings of this thesis, whether it be technical, practical or mental support.

Thanks to Willem Waegeman, Stijn Decubber and Christina Papagiannopoulou for always making time to the best of your abilities to answer all the questions that I had for them. Your answers often provided the push that I needed to keep going and the discussions were of great help. Thank you for your numerous input, support and feedback.

Additional thanks to Stijn Decubber for getting me started with python, while still allowing me to code everything myself. I have learned so much from this, much more than I had ever anticipated.

Thanks to Diego Miralles and Matthias Demuzere for the genuine interest in my research and for your feedback. This motivated me to keep searching for new answers and new questions.

Additional thanks to Diego Miralles for always answering my panicky mails in a matter of hours. Your positive encouragement really helped to keep the mood light and the work going.

Thanks to everyone who helped me to gather the data and everyone who has worked on the research that precedes this thesis. Your work was of vital importance for me.

Thanks to my friends for all the motivational talks and reminding me to have fun.

To finish, a warm thanks to Bo Broddelez and my parents for putting up with me during this whole year and for providing all your love and support. You helped me keep my head up.

# CONTENTS

# ABSTRACT ENGLISH

Quantifying the causal influence of climate drivers on vegetation is not a trivial task, especially on a global scale. In this thesis, recent satellite data is used to model vegetation anomalies, using both linear ridge regression models and non-linear random forest models. This allows for causal inference within a Granger causality framework. The importance of water as a driver for vegetation anomalies is quantified on a global scale and compared to the importance of other drivers. Furthermore, the impact of new variables in the dataset is quantified. Finally, the effect of the high spatial and temporal resolution of the data is investigated within the Granger causality framework.

# ABSTRACT DUTCH

Het kwantificeren van de causale invloed die klimaat drijvers globaal hebben op vegetatie is niet triviaal. In deze thesis worden recente satelliet data gebruikt om vegetatie anomalieën te modelleren, met lineaire ridge regressie modellen en niet-lineaire random forest modellen. Deze causaliteit analyse wordt gevoerd binnen een Granger causaliteit framework. Het belang van water als een drijver voor vegetatie anomalieën wordt globaal gekwantificeerd en vergeleken met andere drijvers. De impact van nieuwe variabelen in de dataset wordt gekwantificeerd. Finaal wordt het effect van de hoge spatiale en temporele resolutie van de data binnen het Granger causaliteit framework onderzocht.

CHAPTER 1

# INTRODUCTION

Vegetation dynamics are largely influenced by external climatic drivers: air temperature, incoming radiation and water availability (Bonan, 2011; Wu et al., 2015; Papagiannopoulou et al., 2017a). These drivers have been changing with the climate and different ecosystems have shown a difference in resilience to these changes (Seddon et al., 2016). Thus, climate drivers heavily impact vegetation both by their long-term availability and their change over time.

Meanwhile, vegetation has a non-neglible impact on these climate drivers. This impact stems from effects such as changes in albedo, surface roughness and evaporation. These alterations effect the global climate system through changes in cloud formation, the CO2 balance and the energy flux partitioning (Zhang et al., 2014; Heimann and Reichstein, 2008; Papagiannopoulou et al., 2017b).

On top of this two-way interaction, some other variables influence either vegetation or climate in a significant manner. For example, human behaviour influences vegetation through deforestation, land use change and the instigation of large-scale wildfires (Andela et al., 2017; Tepley et al., 2015). Also, anthropogenic CO2 emmissions significantly and irreversibly contribute to a global rise in temperature (Solomon et al., 2009).

This all gives rise to different negative and positive feedback loops, respectively dampening or strengthening different interactions. These feedback loops combine to form a highly non-linear system of interactions, which happens on a variety of spatial scales. This scale ranges from local (land use change influences local climate (Stohlgren et al., 1998)) to regional (the 2003 and 2010 european heatwaves impacted vegetation adversily (Bastos et al., 2014)) to global (anthropogenic $CO_2$-fertilization has an impact on global primary production (Devaraju et al., 2016)).
On top of that, the temporal scale of these interactions also varies massively, due to the frequency of occurence of impacting phenomena and the latent response of natural systems. This scale ranges from direct (the impact of large-scale wildfires on vegetation (Puig-Gironès et al., 2017)) to annual (the strong seasonality of natural

systems) to interannual (El Niño southern oscillation impacts the mainland vegetation (Li and Kafatos, 2000)).

The combination of these feedback loops on a variety of spatial and temporal scales results in a complex web of multi-way interactions. The goal of this thesis is to isolate and quantify the impact of different climate variables on vegetation, with respect to the complex system they are entangled in. This should provide additional understanding in the response of global vegetation to climate drivers in recent history.

This task is referred to as causal inference and is preferably done within a solid theoretical framework. In this work, the framework proposed in Papagiannopoulou et al. (2017b) will be adopted, since it was designed for this task and has proven succesful in a very similar setting. This machine learning framework allows for using a non-linear model. This is useful for these vegetation-climate interactions, which have often been reported as non-linear. (Heimann and Reichstein, 2008; Bonan, 2011; Zhang et al., 2014; Papagiannopoulou et al., 2017b). It is also able to handle the high co-linearities between different climate variables. Furthermore, it utilizes time series decomposition and promotes the use of high-level features for modelling.

Considering the recent wealth of high-resolution climate and vegetation data (Ma et al., 2015), only a selection of sources is used to construct the big data set used in this work. Most sources consist of remote sensing data, coming from satellites and spanning most of the vegetated land. NDVI (normalized difference vegetation index) is used as a remote sensing proxy for vegetation productivity (Pinzon and Tucker, 2014). Different climate variables are used to represent the more general groups of temperature, water availability and radiation. These groups have been often investigated for their individual control over global vegetation (Nemani et al., 2003; Wu et al., 2015; Seddon et al., 2016; Papagiannopoulou et al., 2017a). The variables used are part of the aforementioned complex web of vegetation-climate interactions. To fully utilize the power of this proposed framework, the other variables in this web of interactions should be included in the analysis (Granger, 1969; Papagiannopoulou et al., 2017b).

This thesis is done as part of the SAT-EX project (http://www.sat-ex.ugent.be/). It is very similar to the work of Papagiannopoulou et al. (2017b,a), with the biggest novelty being the data, which is more recent: 2003-2015 compared to 1981 - 2010 in Papagiannopoulou et al. (2017b)). The data also has a higher spatial and temporal resolution. The coding for this work was all done from scratch in python (with initial support provided by Ir. Stijn Decubber).

## 1.1 Goal and research questions

The goal of this thesis is to gain quantitative knowledge on how climate drivers influenced global vegetation between 2003 and 2015 and compare the results to previous research (Papagiannopoulou et al., 2017b,a). This is represented by the main research questions.

1. Is water the most important climate driver for vegetation anomalies globally as presented in Papagiannopoulou et al. (2017a)?
2. What is the impact of the new variables in the dataset on the analysis?
3. What is the influence of a higher spatial and temporal resolution on the Granger Causality Framework compared to previous work of Papagiannopoulou et al. (2017a)?

## 1.2 Outline

In Chapter two, a literature review is presented that contains a selection of research papers. The aim is to represent the variety of ways these vegetation-climate interactions are researched and where this work fits into that field.

In Chapter three, the used data are discussed. This comprises the data collection procedure, the data merging and the data preprocessing that was done.

In Chapter four, the used methods are presented. First, the construction of the NDVI anomalies is explained. Then, the Granger causality theory is given together with the machine learning models that are used. The construction of model features is discussed, which consists of a time series decomposition, lagged and cumulative variables.

In Chapter five, the results are presented and discussed.

In Chapter six, some general conclusions and suggestions for further research are given.

CHAPTER 2

# LITERATURE REVIEW

Quantifying the relationship between climate and vegetation has been the topic of many research projects, on a variaty of scales and using a wide range of different approaches. From this vast amount of research, some selected papers are presented in the following sections. This is in no way a representative overview of the available literature, but merely a slice of context in which to place this work. To facilitate the comparison between literature, a short structure is presented of the common differences in scale, goal and methods of the different works.

1. **Scale**
   (a) Space
      - regional
      - global
   (b) Time
      - incidental
      - decadal
2. **Goal**
   (a) Future prediction
      - climate change
      - vegetation change
        – classification (vegetation type)
        – regression (quantity of vegetation)
   (b) Causal inference
      - influence of climate on vegetation
      - influence of vegetation on climate

3. **Method**

   (a) Mechanistical Modelling

   - earth system model
   - regional climate-vegetation model

   (b) Data-driven Modelling

   - linear models
   - non-linear models

   (c) Correlation studies

Within the presented classification, this thesis is a global study that spans thirteen years. The goal is to infer the causal impact of climate drivers on vegetation. The methods used are both linear and non-linear data-driven modelling.
In the following sections, the literature is presented according to the used method: mechanistical modelling, data-driven modelling and correlation studies.

## 2.1 Mechanistical modelling

Mechanistical modelling of vegetation-climate interactions is done through a tight coupling of dynamic vegetation models and climate models.

Climate models simulate the interactions of important climate drivers using differential equations based on the laws of physics, fluid motion, and chemistry. Included in these models are the main earth compartments including atmosphere, oceans, land surface, vegetation and ice. By capturing the interactions between these compartments, these models aim to capture the feedback loops that arise from these interactions and can thus be used to either study or predict climate change and its impact on the different compartiments.

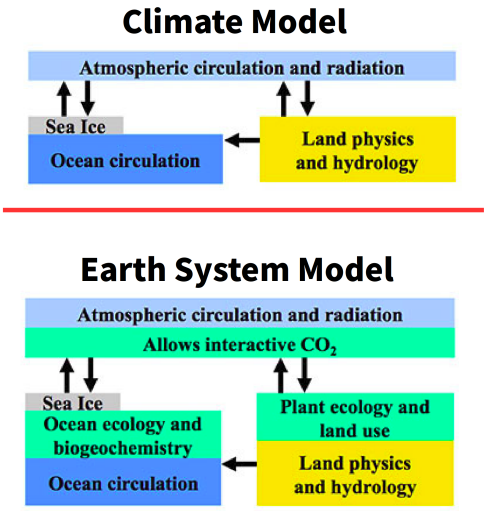



Figure 2.1: Earth System Model vs Climate Model, Adapted from SOCCOM, Princeton University, Retrieved July 3, 2018 from `https://soccom.princeton.edu/content/what-earth-system-model-esm`

Dynamic vegetation models simulate the changes in vegetation (type, biomass, behaviour) arising from climate change and the associated impact on carbon, nutrient

and hydrological cycles (Bonan et al., 2003; Krinner et al., 2005). Climate models traditionally used an over-simplified representation of vegetation, which is why the coupling of global climate models and dynamic vegetation models results in more realistic models (Foley et al., 1998).

Most of these coupled vegetation-climate models are based on global climate models and represent the entire earth system (see Figure 2.1). These models are called earth system models and they play an important role in understanding vegetation-climate interactions and predicting the long-term effects that arise from them. More recently, attempts have also been made to couple regional climate models to dynamic vegetation models. This might provide a useful tool for studying climate-ecosystem interactions on a regional scale (Wang et al., 2016)

These vegetation-climate feedback loops arise from a changing climate and the vegetation response. However, how the feedback loops themselves will change over time with an ever changing climate remains a source of uncertainty. As discussed in Heimann and Reichstein (2008), as long as there is no fundamental understanding of the processes involved, simulations of such models can only illustrate the importance of the feedbacks, but cannot present a conclusive picture.

## 2.2 Data-driven modelling

Data-driven modelling does not strive to capture the underlying physical processes that gouvern the interactions. Instead, general-purpose models are used to predict the response variable based on the values of the regressor variables (regressors). The modelling approaches originate from the fields of statistal analysis and machine learning and can require large amounts of data. Linear models are frequently used for their simplicity and robustness. However, this linearity assumption is a simplification, since the vegetation-climate interactions have often been observed as non-linear (Heimann and Reichstein, 2008; Bonan, 2011; Zhang et al., 2014; Papagiannopoulou et al., 2017b).

A balance between prediction accuracy and robustness is sought, which is called the bias-variance trade-off. When balanced correctly, the model is able predict the response variable for data that is was trained on, but also for new data that were not used for training. In data-driven modelling, the goal is not always to predict the futute. When causal inference is the goal, statistical and machine learning theory is often needed to estimate the reliability of the produced conclusions.

Chen et al. (2014) performed a regional study that aims at quantifying the impact of soil moisture on Australian mainland vegetation. The used methods are statistical, and consist of windowed cross correlation, quantile regression and piecewise linear regression. A strong positive relationship was found between soil moisture and NDVI anomalies, with the anomalies typically lagging behind soil moisture by one month. The data range from 1991 to 2009 on a monthly time scale.

Wu et al. (2015) quantifies the impact of temperature, precipitation and solar irradiation on vegetation using linear regression models. In these regression analyses, lagged versions of the climate variables were used, as to include the resilience and latent response of vegetation to perturbations. NDVI was used as a proxy for vegetation and the analyses were done on a global scale. The data range from 1982 to 2008 on a monthly time scale. Uni-variate linear regression was used to model the predictive power of each climate driver on NDVI seperately and the determination coefficient ($R^2$) is calculated each time. Next, multiple linear regression was used to model the combined predictive power on NDVI. Also, from the determination coefficients, the partial correlation coefficient of each climate driver is calculated and used as a measure for that driver's influence on vegetation growth. For the multivariate linear regression, an additive linear model was used, which automatically invokes the assumption that the combined effect of these climate drivers on vegetation is a weighted sum of their individual effects. This means that no interaction effects between these climate drivers are considered, while these effects are observed in reality (Luo et al., 2008). Also, one should keep in mind that by using the raw NDVI instead of the anomalies, the seasonal cycle of the NDVI is included (for details, see Section 4.1). Thus, part of the predictive power in these regression models results from matching the seasonal cycles of NDVI and a seasonal climate driver (f.e. temperature). This reflects reality, but is not very informative.

Seddon et al. (2016) researched the sensitivity of global terrestrial ecosystems to climate variablility. Through a combination of statistical methods, a vegetation sensitivity index is calculated. This index is a weighted sum of the vegetation sensitivity to perturbations in temperature, cloud cover and water availibility. The weights are derived from the significant coefficients ($p < 0.1$) from a principle component regression (PCR). This PCR does not use raw monthly data, but the monthly Z-scores. The vegetation sensitivity to the different perturbations is calculated from the vegetation anomalies, by removing seasonality from the raw data for months with significant PCA regression coefficients. To account for the memory effects of vegetation, a one-month lagged variable is also included. The data range from 2000 to 2013 on a monthly time scale

The entire analysis is thorough and uses a clever variety of statistical methods. However, by using the Z-score, it relies on the assumption that the variable in question is normally distributed. This assumption will be violated in some regions for some variables (f.e. precipitation often shows a skewed distribution in this thesis's dataset). Therefore, using z-scores for regions where a variable is non-normally distributed will result in the corresponding p-values losing their meaning. Another observation that could be made is that by using linear regression on principal components, the analysis also relies on a linearity assumption.

Ackerly et al. (2015) researches the effect of different projected climate change scenarios on the vegetation of the San Francisco Bay (SFB) Area. This research aims at predicting the change in vegetation type distribution that would result from these different scenario's (only current vegetation types of the SFB area are taken into account). This problem is referred to as a multi-class classification problem. Climate and hydrological variables are taken from the climate models output and combined with present-day topological and wind speed distribution data. The model of choice is a multinomial logistic regression model that classifies in a one-vs-all fashion (computationally intensive). Instead of predicting one vegetation class for every location, the model returns probabilities for the different types, which are then summated over the entire SFB area to produce a total probability distribution.

In this study, some liberties are taken but the authors themselves identify the limitations of their approach. In order to use the time series output (90 years) of the climate model simulations as an input for the logistic regression model, the time series are chopped into 30-year pieces, averaged and further treated as equivalent individual scenarios. Some uncertainty also arises from the extrapolation of present day climate-vegetation relations to future climate conditions. The models only predict steady-state vegetation distributions, which take hundreds to thousands of years to establish and ignore the transitional phase. The final vegetation distribution is thus predicted as independent from transitional phenomena such as repeated droughts, which can cause persistent shifts in vegetation (Mueller et al., 2005).

### 2.2.1 SAT-EX project

Studies of high importance for this thesis were conducted by Papagiannopoulou et al, in the framework of the SAT-EX project (http://www.sat-ex.ugent.be/). The goal of the SAT-EX project is threefold. Firstly, it wishes to provide evidence of how climate extremes have changed over the satellite era and identify the drivers behind these changes. Secondly, it strives to provide new insights into past changes in vegetation and the role of climate and climatic extremes on these changes. Thirdly, it works to

test to what extent the IPCC earth sytem models reproduce both the changes in climatic extremes and the associated response in vegetation. The project is conducted using different data-driven modelling approaches on global satellite datasets.

Initial efforts are presented in Papagiannopoulou et al. (2016), where climate-vegetation interactions are investigated using satellite data and machine learning techniques. In this work, a novel framework is presented for identifying climatic drivers that affect vegetation on a global scale. Firstly, an inclusive data-collection approach is used to incorporate all the available data for these climate drivers that meets certain requirements (span and resolution). The time series in this data are decomposed into three parts: a linear trend, a seasonal cycle and the remaining anomalies. The resulting time series are then used to look at the predictive performance of different decomposed climate drivers on vegetation anomalies. These models are ridge regression models (McDonald, 2009) and random forest models (Breiman, 2001). In this analysis, lagged variables, cumulative variables and variable extremes are used. The explained variance $R^2$ ($= 1 - \frac{RSS}{TSS}$) is used as a measure of prediction accuracy in a 5-fold cross-validation scheme.

The addition of climate variables leads to a substantial increase in model preformance, showing that these variables contain additional information about the vegetation state. In this setting, the non-linear random forest models outperform the linear ridge regression models, suggesting again the non-linearity of these interactions.

The proposed framework is further completed in Papagiannopoulou et al. (2017b) through the incorporation of a Granger causality analysis (for details see section 4.2). This allows for the isolation of the causal influence of different climate variables on vegetation anomalies. The quantification of causal influence remains an estimation, but this is arguably the best available technique at this time in this setting. It is argued that no statistical tests exist for investigating the significance of these non-linear interactions in this setting, and that the construction of such a test is not a trivial task. The machine learning models used are ridge regression and random forest. The regressor variables are constructed through time series decomposition of these climate variables to isolate the trend, seasonality and anomalies. Additional high-level features are constructed from these decomposed time series, including lagged variables, cumulative variables and the occurence of extreme events. The proposed pipeline now consists of merging data from various databases, time series decomposition, high-level feature construction, predictive modelling and a Granger causality quantification. This framework is particularly useful due two factors: the ability to desentangle the co-linearities between climate variables and the flexibility to allow for both linear and non-linear models.

This framework is further used in Papagiannopoulou et al. (2017a) to investigate the causal influence of different climate drivers on vegetation anomalies. The climate variables used are grouped into three main driving forces: temperature (surface and near-surface), water availability (precipitation including snow and soil moisture) and irradiation (incoming and net). The data spans from 1981-2010 and is rescaled to a monthly time series on a 1° x 1° spatial resolution. The influence of each of these three climate driver groups on vegetation is quantified. Different regions are classified according to which climate driver group has the most influence.

It is concluded that changes in water availability lead to a lagged response in vegetation and that its impacts are longer lasting than those of radiation and temperature. The impact of extremes is also investigated and again, extremes in water availability are the most important. Water availability is identified as the primary factor driving NDVI anomalies globally, with 61% of the continental surface vegetation being water-driven. This constrasts with earlier studies (f.e. Wu et al. (2015)) that reported a lower global importance of water availability for vegetation. This difference can be related to the different methods used. Interestingly, for most of the regions reported to be water-driven, the supply of precipitation is expected to decline following global warming. Furthermore, the reported impact of extremes in water availability underlines the gravity of the projected increase in hydro-climatic extremes.

## 2.3 Correlation studies

Correlation studies are data-driven studies that investigate the correlation of two or more variables. This can be done by calculating a simple correlation coefficient between variables within a selected period or over the whole time span. Another option is the use of event composition (also known as superposed epoch analysis) to investigate the effect of recurring phenomena, such as droughts or wildfires.

Simple correlation studies were some of the first studies to be conducted on global vegetation-climate interaction. By assessing the covariance between vegetation time series and the lagged time series of climate variables, the influence of vegetation on climate is identified in Liu et al. (2006). Similarly, the influence of climate on vegetation is derived from the covariance between climate variable time series and lagged vegetation time series. To quantify these influences, a simplified causality model is used.

Event composition is used in Nicolai-Shaw et al. (2017) to investigate the effect of soil moisture drought events during the peak of the growing season, on a global scale. First, the peak of the growing season is mapped globally. Then, the periodes are

isolated when soil moisture drought occurs during the peak of the growing season. The periods when the peak of the growing season does not coincide with drought are the reference data. For every climate and vegetation variable, the anomalies between the isolated and reference periods are calculated and the covariance between the different variables is quantified.

The climate variables and vegetation correlate with soil moisture drought as expected, while the resulting anomalies in vegetation activity are often delayed compared to the anomalies in climate variables. Forests show much less vegetation anomalies than other land cover types, likely because they have access to water in deeper layers. For some forest-covered regions, positive anomalies in vegetation activity are even identified during drought periods. This could relate to the fact that the presence of deeper soil moisture is not reflected in the utilized soil moisture dataset (ESA-CCI). Drought events in this analysis only reflect upper soil dryness, which often relates to warmer-than-average conditions. These conditions can lead to higher fotosynthesis in energy-limited regions and thus positive vegetation anomalies.

CHAPTER 3

# DATASET CONSTRUCTION

The outline for this chapter is as follows. Firstly, the data collection procedure is discussed. Secondly, a description of the datasets is given. Thirdly, the data merging and preprocessing methods are discussed.

## 3.1 Data collection

In this thesis, the data units are often omitted. In most research settings, this is considered very bad practice. In this machine learning setting however, all data is either fed to a scale-invariant algoritm or rescaled before analysis. Thus, unless units are of particular relevance for their physical impications, they are omitted from figures and text. The data were provided by courtesy of ir. Stijn Decubber[1], ir. Christina Papagiannopoulou[1], dr. Matthias Demuzere[2], ir. Brianna Pagan[2], ir. Brecht Martens[2] and prof. dr. Diego Miralles[2].

A selective data collection approach was used in this work, in contrast to the inclusive data collection approach proposed in Papagiannopoulou et al. (2017b). For every considered variable only one data source is used. The reasoning here is that due to the higher resolution of this analysis, a dataset for this analysis contains much more data points than the dataset for the same variable in the previous studies. Thus, in order to keep the total data size reasonable, only a selection of data sources can be used. Furthermore, the scope of this thesis should be smaller than the scope of Papagiannopoulou et al. (2017b), so an inclusive data selection procedure is not desired.

The used climate variables are grouped in three main groups as proposed in Papagiannopoulou et al. (2017b): water availability, temperature and irradiation. Additional variables that were not yet included in the works of Papagiannopoulou et al. (2017b) are vapour pressure deficit (VPD), irrigation and large-scale wildfires. A short

[1]KERMIT, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering
[2]LHWM, Department of Water and Forest Managment, Faculty of Bioscience Engineering

structure of the variable groups is presented below and more details are presented in the following sections.

1. Vegetation
    - Normalized difference vegetation index (= NDVI)
2. Water availability
    - Precipitation
    - Snow water equivalent (= SWE)
    - Near-surface soil moisture
    - Root-zone soil moisture
    - Vapour pressure deficit (= VPD)
3. Temperature
    - Near-surface air temperature at midday
    - Near-surface air temperature at midnight
    - Difference between day and night air temperatures
4. Radiation
    - Net radiation
5. Other
    - Burned area (Wildfire)
    - Irrigation

### 3.1.1 Vegetation

The response variable for this study is vegetation productivity. This means that the models will try to predict the vegetation productivity, based on the other variables. To be fully correct, the response variable is not the NDVI as such, but rather the NDVI *anomalies*. The construction and meaning of these anomalies are discussed in Section 4.1. Different remote sensing proxies are available for vegetation productivity, of which the Normalized Difference Vegetation Index (NDVI) is the most commonly used (Pinzon and Tucker, 2014). This proxy shows a near-linear relation with photosynthetically active radiation absorbed by vegetation, and therefore with light-dependent photosynthetic activities in the upper canopy. NDVI is near-linearly related to various productivity variables, such as area-averaged net carbon assimilation and transpiration. An important drawback of NDVI is the well-known saturation effect that occurs in

dense vegetation with high biomass values (Glenn et al., 2008). Data from the Global Inventory Modeling and Mapping Studies (GIMMS) 3g version 1.0 is used (Pinzon and Tucker, 2014).
» https://ecocast.arc.nasa.gov/data/pub/gimms/3g.v1/

### 3.1.2 Water availability group

The water availability group contains precipitation, snow water equivalent, soil moisture at the surface and root zone, irrigation and vapour pressure deficit.

Precipitation data comes from the Multi-Source Weighted-Ensemble Precipitation (MSWEP) dataset, which combines a variaty of remote sensing and other sources. The data is taken from Beck et al. (2017).
» http://www.gloh2o.org/

Snow water equivalent is used to approximate water in the form of snow cover and is a combination. The data is taken from Metsämäki et al. (2015).
» http://www.globsnow.info/swe/

Soil moisture is a property that goes several meters below the earth surface, which makes it inheritably more difficult for remote sensing. A first data source is the European Space Agency Climate Change Intiative (ESA CCI) combined soil moisture product (Dorigo et al., 2017). This dataset is based on remote sensing alone and is thus limited to surface moisture.
» http://www.esa-soilmoisture-cci.org/node/145

A second soil moisture data source is the Global Land Evaporation Amsterdam Model (GLEAM v3.1), which includes deeper soil moisture. This model estimates terrestrial evaporation and root-zone soil moisture from satellite data observations. The data is taken from Martens et al. (2017).
» https://www.gleam.eu/

Vapour pressure deficit (VPD) is the difference between the saturated vapour pressure (the amount of moisture the air can hold when it is saturated) and the vapour pressure (the amount of moisture in the air). VPD is a measure for air drought and is of importance for plant transpiration, due to its regulatory effects on stomatal conductance (McAdam and Brodribb, 2015). Therefore, it is included in the water availablility group. The VPD is calculated from its definition formula ($VPD = P_v^{sat} - P_v$). The data for the vapour pressure is taken from the Atmospheric Infrared Sounder (AIRS) (Tian, 2016b). The saturated vapour pressure is function of the air temperature ($°C$) and is calculated as $P_v^{sat} = 611 * exp\left[\frac{19.65 * T_a}{T_a + 273}\right]$. These calculations were performed by ir.

Brianna Pagan.
» https://airs.jpl.nasa.gov/data/physical_retrievals

### 3.1.3 Temperature group

Temperature is the second climate driver group. It is a another important driver, especially for certain vegetation types (Williams et al., 2013). The near-surface air temperature at midday and midnight are used to produce two seperate datasets. The difference between the day and night time temperatures is also included as a variable. The data is taken from the Atmospheric Infrared Sounder (AIRS) (Tian, 2016a)
» https://airs.jpl.nasa.gov/data/physical_retrievals

### 3.1.4 Radiation group

The third climate driver group of importance is radiation. The variable used is net radiation, which is the difference between incoming and outgoing energy at the top of the atmosphere. This net radiation is the part of the incoming solar energy that is available to influence the climate. This is calculated through a radiation balance. The resulting data is sourced from the Clouds and Earth's Radiation Energy System (CERES) (Loeb et al., 2016)
» https://icdc.cen.uni-hamburg.de/1/daten/atmosphere/ceres-radiation.html

### 3.1.5 Other

Large-scale wildfires are included in the analysis. They arise from an interplay between vegetation, temperature, drought, lightning and human interference. Climate is a dominant control on wildfire activity, regulating vegetation productivity and plant moisture content. Over short time scales, rainfall during the dry season suppresses fire activity, whereas over longer time scales, fuel build-up during wet years in more arid ecosystems can increase burned area in the subsequent years (Andela et al., 2017). The vegetation recovery after a wildfire varies wildly depending on vegetation type, fuel build-up and climate conditions such as drought and water availability (Gouveia et al., 2012; Lu and He, 2014). The remote sensing product for the total burned area is used as a proxy and is sourced from the $4^{th}$ generation of the Global Fire Emissions Database (GFED4) (Giglio et al., 2013)
» http://www.globalfiredata.org/data.html

Irrigation is the anthropogenic addition of water to vegetation, often for agricultural purposes. Although this variable is definitely of importance for the local water availability, it is not a climate variable. Thus it is not included in the water availability group, which should only contain climate variables. This is directly related to research question one, which questions the impact of water availability as a *climate* driver. Irrigation is included in the analysis to produce a more correct Granger causality analysis (see Section 4.2). The data is taken from the global irrigation model (GIM) described in appendix C of Müller Schmied et al. (2014)
» *Data requested by dr. Matthias Demuzere*

## 3.2 Description of the data

### 3.2.1 Data dimensions

Collecting the data from the aforementioned sources results in twelve datasets, one with "raw" data for every variable. These datasets all have three dimensions (two in space and one in time), hence they are named datacubes.

The two spatial dimensions of each dataset are the Latitude and Longitude, and together they describe the place of a data point on the globe. Latitudes range from -90° at the South Pole to 90° at the North Pole and give your position compared to the Equator (see Figure 3.1).

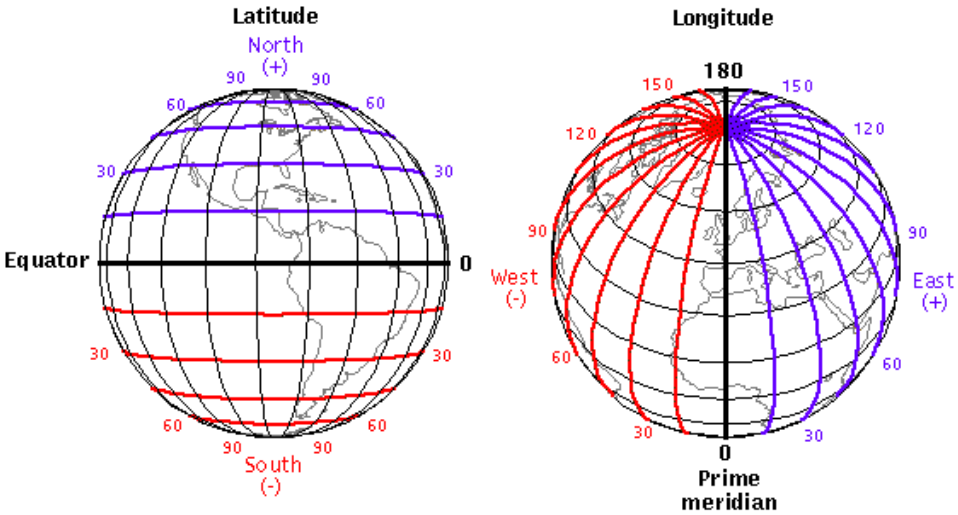



Figure 3.1: Latitude and Longitude, Slawitsky K., Hicksville High School, Retrieved July 14, 2018 from `https://www.hicksvillepublicschools.org/Page/11725`

Longitudes range from -180° to 180° and give your position compared to the Prime meridian in Greenwich. On a rectangular map of the globe, Latitude corresponds to

the vertical position (the y coordinate on a graph) and Longitude corresponds to the horizontal position (the x coordinate). These spatial coordinates divide the globe into a orthogonal grid. The spatial resolution of a dataset determines on what scale the measurements are done. In remote sensing, this scale is often expressed in degrees. A spatial resolution of 0.5° x 0.5° means that measurements are made for every square in a grid of 0.5° longitude on 0.5° latitude.

Such squares are further referred to as pixels and they are the smallest space over which an individual measurement is made. The total number of pixels is dependent on the spatial resolution and is calculated from $\#pixels = \frac{180° * 360°}{(resolution)^2}$. A 0.5° x 0.5° spatial resolution corresponds to a grid of 259200 pixels if the entire globe is covered. It also follows that doubling the spatial resolution corresponds to a quadrupling of the available data (see Figure 3.2). Some attention should also be paid to naming conventions: Changing the resolution from 0.5° x 0.5° to 0.25° x 0.25° is an upscaling to a *higher* resolution.

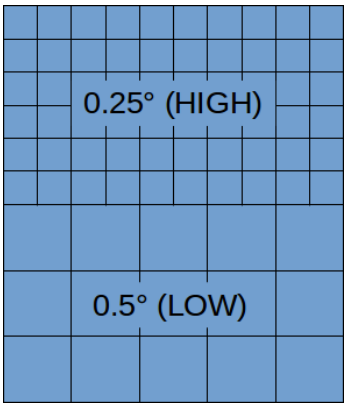


Figure 3.2: spatial resolution

The pixels are (quasi-)square in shape and are identified by the coordinates of the central point. For each pixel, a variable is represented by a series of measurements over time, called a time series. Each variable has a limited time span over which data is available. Each variable also has its own temporal resolution, which is how often a measurement is made. For example, a variable might be measured weekly and hence the temporal resolution is seven days. There is a subtle nuance in the representativeness of a remote sensing data point. The data point is a snap-shot in time and is thus often not representative for that entire time step. On the contrary, a data point is representative in space as it represents the average state of the entire pixel.

### 3.2.2 Data merging

The twelve raw datasets differ in the aforementioned properties, which prevents them from being merged straightaway. They are recorded over different time spans and may miss data in different areas. They also have different native resolutions, both in time and space. To ease the further handling of these datasets, a common time span, spatial span and resolution are defined.

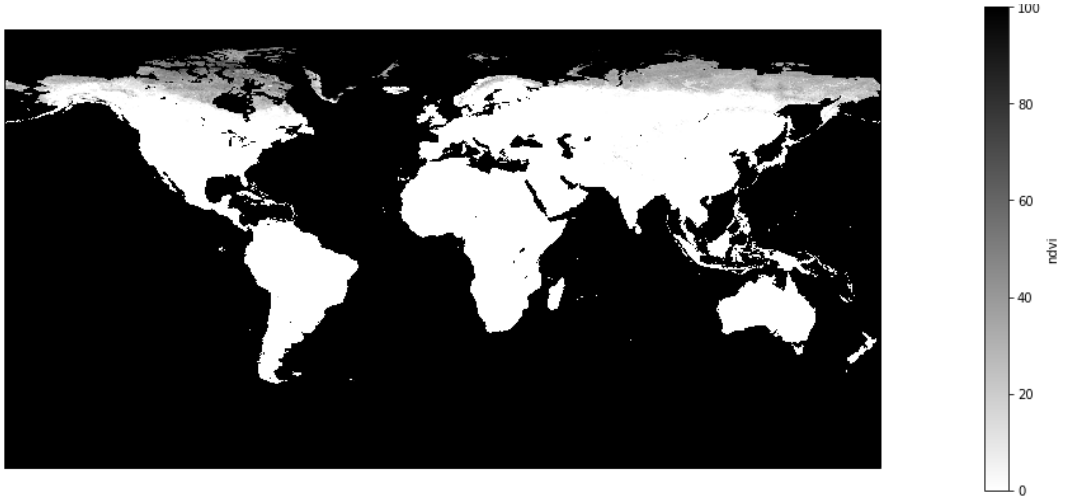



Figure 3.3: NDVI: percentage of missing values

The common spatial span for this study is chosen from the spatial span of the NDVI dataset, since this is the response variable. Ocean and freshwater pixels don't have values for NDVI. In Figure 3.3, the percentage of missing NDVI values in each pixel is represented for that location. If a pixel has more then 50% missing values in its time series, it is part of a water body or an area with limited vegetation and is not included. This results in a land mass similar to that of Papagiannopoulou et al. (2017b). However, in that study large land areas exist for which no data was available, because the NDVI was so low that no reliable measurements were possible. These areas are mainly deserts such as the Sahara region and large parts of Saoudi Arabia and Mongolia. In this study, the total NDVI is summed up over time and pixels with a low total NDVI are excluded from the analysis. The NDVI time series in these regions often have very low variance and do not neccesarily reflect the actual vegetation dynamics. The low variance also proved problematic during the construction of the NDVI anomalies later on. Pixels with a total NDVI below 400000 were exluded (different cut-offs were tested). The excluded regions are shown on Figure A.1 in Appendix A, together with a map of missing data from Papagiannopoulou et al. (2017a) in Figure A.2. The paper from Seddon et al. (2016) reviewed in Section 2.2 shows areas of excluded data similar to what is used in this thesis (see Figure A.3 in Appendix A).

The spatial resolution adopted in this thesis is 0.25° x 0.25°, which corresponds to sixteen pixels per patch of 1° x 1°. These pixels are (quasi-)squares of roughly 27.8km x 27.8km around the equator and get smaller towards the poles. The temporal resolution is bi-weekly, which corresponds to two data points every month, or one data point roughly every 15 days. For comparison, the resolution in the works of Papagiannopoulou et al. (2017b) is 1° x 1° and monthly (one data point every month). The time span of these datasets is chosen from the $1^{st}$ of january 2003 to the $15^{th}$ of december 2015. Thus every time series includes 13 years, which corresponds to

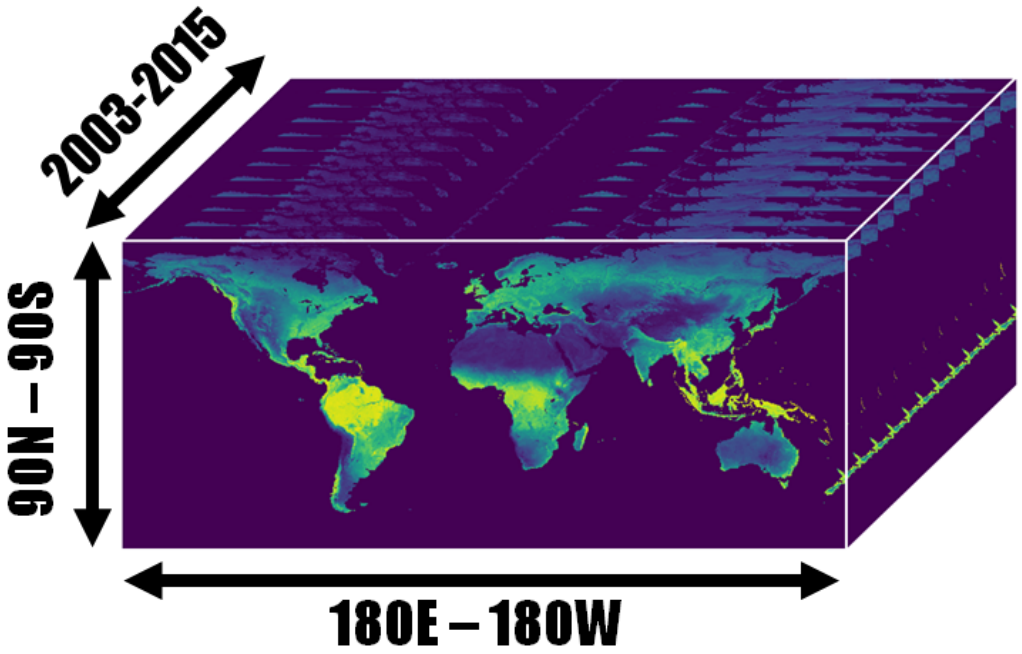



Figure 3.4: Dimensions of the datacubes

312 data points. This limited time span is due to the high desired resolution: remote sensing data records extend way earlier than 2003, but for most variables not in the required resolution. In the work of Papagiannopoulou et al. (2017b), the include data ranges from 1981 to 2010, an extended period of 30 years. The dimensions of the resulting data cubes are presented in Figure 3.4. Each dataset is 2.6 GB and contains 13.86 times more data points then similar datasets in Papagiannopoulou et al. (2017b) due to the higher resolution. Despite their large size and high resolution, the limited time span may prove to be a strong limitation of these datasets.

## 3.3 Dataset Preprocessing

Now that the properties are known, each of the raw datasets is to be transformed into a final datacube. This consists of filling the gaps and adopting the desired resolution in space and time. If the native resolution of a dataset is higher then 0.25° x 0.25° or bi-weekly, downsampling is required. If the native resolution is to lower, upsampling is required. For most datasets, the preprocessing was performed by ir Stijn Decubber at the start of this thesis. For four datasets, this was done by myself and the process is described in more detail.

Finding the optimal approach for gap-filling and resampling these datasets in three dimensions is not as straightforward as it might seem. The ideal method is tailored to exploit the specific properties of each dataset and is robust and fast. Some advanced

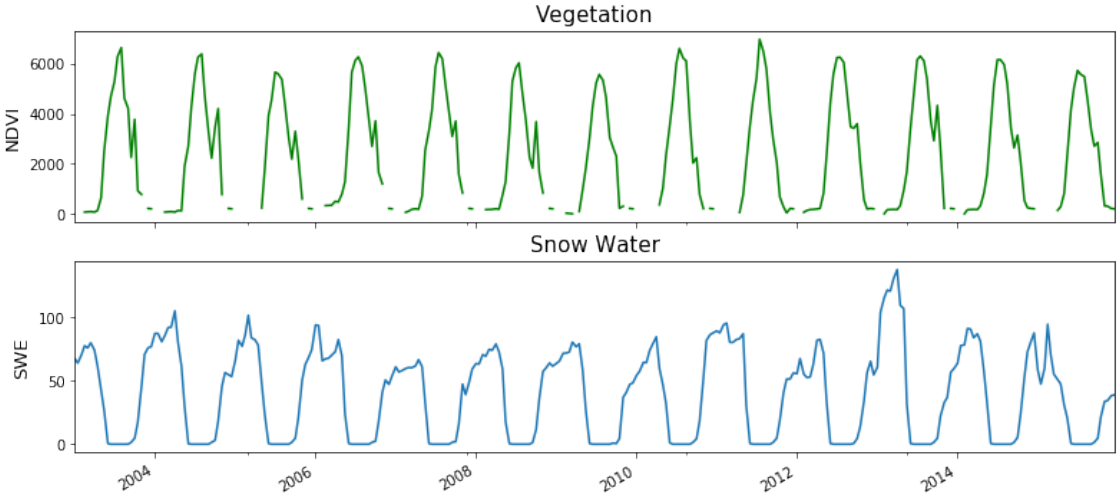



Figure 3.5: The NDVI and SWE time series for a pixel in Inuvik, Canada (69.25°N)

methods were considered (Gerber et al., 2016; Kandasamy et al., 2013), but were eventually abandoned due to their complexity and the limited scope of this thesis. Thus, a quick insight into the specific properties of each dataset is given as to motivate the chozen pre-processing procedure. As stated before, these data gaps often extend into three dimensions. Thus, both spatial and temporal approaches are possible when performing gap-filling. To aid in the understanding of how these approaches are different, one can imagine the data cube in Figure 3.4 as a large cuboid potato.

The spatial gap-filling approaches take the data potato and slice it up along the time dimension to produce one potato chip for every time step. Each of these slices is now a two dimensional map of the variable around the globe at one specific time. As an example, Figure A.6 in appendix A shows three such slices with the data daps in white. The data gaps are areas on a map, and to fill these one can utilize information from neighbouring pixels around the gap.

The temporal data-filling approaches take the data potato and push it through a french fry cutter to cut up the spatial dimensions. This produces a large number of french fries, which are the time series. Every pixel has one time serie per variable, some examples are presented in Figure A.4 in Appendix A. The data gaps are now periods in time and to fill them one can utilize the data points before and after the gap.

### 3.3.1 NDVI

NDVI is the response variable, thus the preprocessing of this variable has to be done with care. It is already in the desired resolution so only gap-filling is required. First, a small exploration of the data is performed.

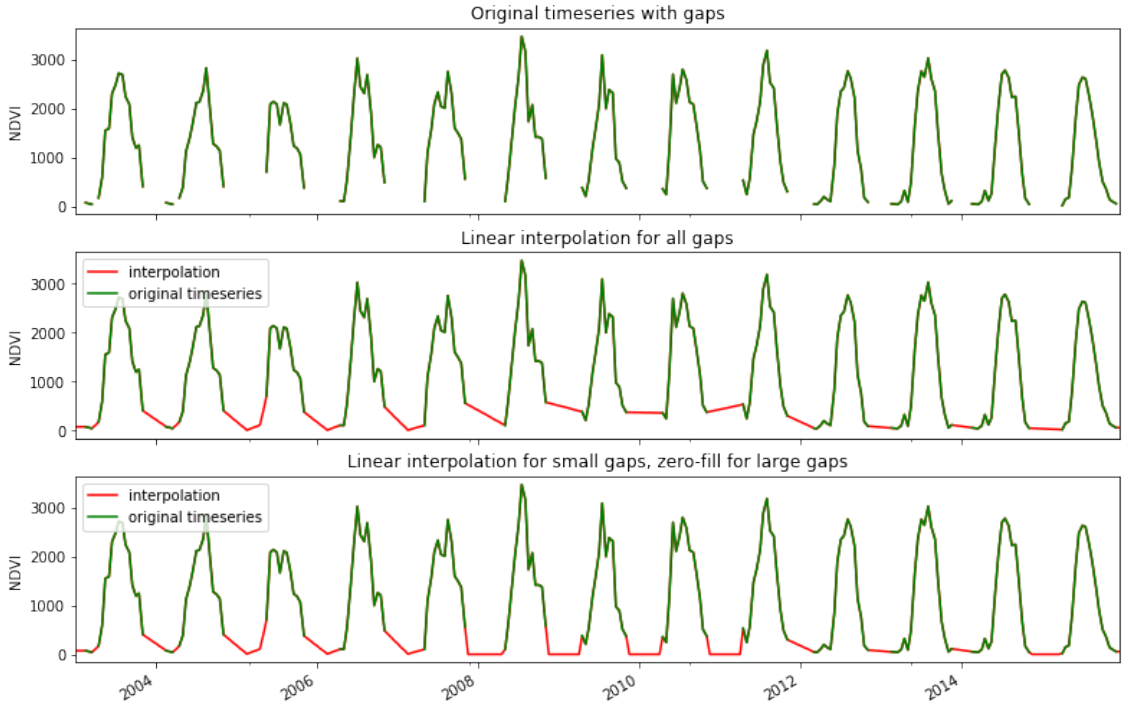



Figure 3.6: A comparison of gap-filling procedures for NDVI time series

In the raw NDVI data, 76.0% of the pixels has no data (see Figure 3.3). This is a permanent water or ice surface, barren soil or land that cannot be monitored via remote sensing. 19.7% of the locations has full time series without missing values. 4.3% of the locations have time series with some (but not all) missing values. These time series are the ones that need gap-filling. As observed from Figure 3.3, these data gaps are only located in the Northern regions. They are also very seasonal and occur during the Northern winters. The gaps are strongly related to snow cover, which prevents remote sensing of NDVI. This is observed in Figure 3.5, which shows time series from a pixel in Inuvik, Canada (69.25° N). This NDVI time series has 19.9% of it's values missing, which seems like alot. However, due to the seasonal nature of these gaps, they are short (in time) and easily fillable (in time). Thus, the NDVI time series with less than 50% missing values are gap-filled and included in the analysis.

Due to the spatial isolation of the regions that need to be gap-filled and their seasonal nature, spatial gap-filling is not very usefull for the NDVI dataset. Thus, a temporal gap-filling procedure is used.

First, a classification is made for seperating small and large gaps in time. A small temporal gap has no more than six consecutive data points missing (no more than 3 months). The number six is carefully chosen through visual exploration of different possibilities. A large gap is every gap that has more than six consecutive data points missing (more than 3 months). A small temporal gap can be filled easily through linear interpolation, which is preferable. However, when using linear interpolation

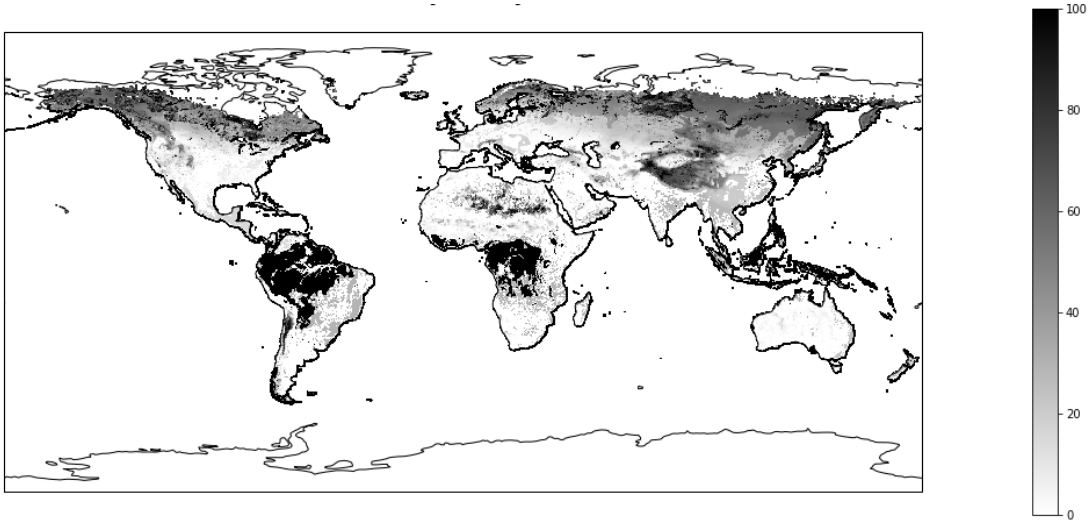



Figure 3.7: Near-surface soil moisture: Percentage of missing values

for larger gaps, weird artifacts end up in the resulting time series. An interesting alternative approach for filling these large gaps is by estimating the seasonal cycle of the variable and then using this as an estimate to fill the gaps. This is however not possible with NDVI, because the gaps are always located in the same season and it is impossible to calculate the seasonality for that time of year. It is chosen to fill the larger gaps with the value zero, which consistenly gives the best results (see Figure 3.6). This is further motivated by the fact that the few data points that are present in these winter months are close to zero (verified for time series from a number of different locations). As a final argument: all time steps for which no NDVI values were available in the original dataset are dropped before training the models. As such, the gap-filling procedure has a very limited impact on the final analysis.

### 3.3.2 Near-surface Soil Moisture

The dataset of near-surface soil moisture (ESA CCI combined) is already in the desired resolution, both in space and time. Thus, only gap-filling is required. However, these gaps of missing values are extensive. Large seasonal gaps are present at higher latitudes in the Northern Hemisphere. Furthermore, large areas of tropical rainforest contain no data over the entire 13 year time span. This is apparent from Figure 3.7, where the percentage of missing values is represented for every pixel. The coverage of rainforest is shown in Figures B.3 and B.2 in Appendix B. For additional insight into the raw soil moisture data and the extent of the data gaps, some additional figures are presented in Appendix A and an animation of the evolution of the gaps over time is also available online[3].

[3] www.linkedin.com/feed/update/urn:li:activity:6390884647424917504

Now that the extent and temporal behaviour of the data gaps is identified, a suitable gap-filling strategy is to be formulated. As a side note: the areas that have no data for the entire time span cannot be filled and are further ignored. The areas with more limited data gaps can be filled using different methods. One possibility is to exploit the spatial information around these gaps, by taking the values of neighboring pixels as an estimate of the values in the gap. Another possibility is to use the temporal information around these gaps, by taking the values of the time before and after the gaps as estimates for the missing period. Different combinations of these options were tried and tested, the final stategy is decided to have alternating interpolations in time and space.

Firstly, the small temporal gaps (up to 3 months) are filled through linear interpolation in time. Secondly, the edges of the remaining gaps space are filled in space with the average of the neighboring pixels at that time. Thirdly, the intermediate temporal gaps (up to 6 months) are filled through linear interpolation in time. Finally, the remaining missing values are all part of a big gap, for which no nearest neighbor or interpolation in time is sensible. Thus, these values are filled with the calculated seasonal cycle for that pixel. The procedure is explained in detail below.

The first step is a linear interpolation in time. When scanning through the individual time series of different pixels, it becomes apparent that a lot of small temporal gaps exist, which can easily be filled through linear interpolation (Figures A.4 and A.5 in Appendix A). The following question imposes itself: what is the maximum gap length in a time series, that should be filled through linear interpolation? In order to reliably estimate that maximum gap length, a time series of soil moisture that has no missing values is selected. Then, artificial gaps of a specific length are introduced at random places into this time series. These gaps are filled through linear interpolation and compared to the original data. This procedure is repeated for different values of gap length, and the maximum length of a small gap is chozen as the highest artificial gap length that did not result in artifacts. Figures A.7 and A.8 in Appendix A show this procedure. As a result, small temporal gaps are defined as gaps with no more than six consecutive data points missing. These short gaps are filled quite accurately through linear interpolation in time. For larger gaps, other techniques are used.

The second step is an interpolation in space. Different two dimensional spatial interpolation strategies are available, pre-implemented in a python package named scipy. These algorithms are shown in Figure A.11 in Appendix A. The nearest neighbor and linear interpolation are the most sensible ones, but they both have an important drawback. The linear interpolation algorithm cannot cope with the presence of oceans in a sensible way. The oceans are large fields of missing data, so the agorithm fills these areas through linear interpolation as well. This on its own is not a problem. However,

when a gap in the data is present in a coastal area, this becomes a problem. As an example, imagine a data gap in a part of Cuba. When this gap is filled, it makes sense to use data from other parts of Cuba, since it is part of the same Island. However, to the linear interpolation algorithm this data gap is just part of the much larger gap that is the ocean around Cuba. Thus, to fill this large gap, data from the American Mainland is also used. This problem arises for all oceans, which would mean that data from the entire other side of an ocean is used for interpolating coastal data gaps. In a dataset of this scale, using far-away data for interpolation is of limited use.

An alternative would be to fill these gaps through nearest neighbor interpolation. This makes more sense, because only the closest data is used. Nearest neighbor interpolation is quite simple: for every pixel with missing data, it just finds the nearest pixel that has data and copies the data point from that pixel. To calculate proximity, the Euclidean distance between the centers of the pixels is used, with the assumption of all the pixels being true squares of identical size. This is a fine option. However, when applying this algorithm, it fills every gap present, no matter how large. This makes little sense, it is only desirable to fill the outer edges of the gaps. This is again due to the scale of the analysis: a gap can be 20° wide, and the information in pixels around the gap is just a bad estimate for the central pixels in the gap. Thus, the center of large gaps might be filled more accurately in the next interpolation steps. Now that it is established why the given spatial interpolation algorithms are insufficient, an alternative is proposed.

This algorithm is similar to the nearest neighbor, but it only fills the pixels at the edges of the gap and considers multiple nearest neighbors for a single pixel. As stated before, spatial interpolation works on a slice of the data cube that contains a single time step. In such a map, all the pixels on the edge of a gap are identified. If a pixel has a missing value, but at least one of the neighbors does have a value, it means that the pixel is part of a gap, but it’s close to the edge. As a neighborhood, the standard Moore neighborhood is used, so the pixels have eight neighbors. When a pixel is identified as being at the edge of a gap, it is filled with the average of it’s neighbors, ignoring those that do not have values. By repeating this procedure multiple times, small gaps are filled completely while only the edges of large gaps are filled. These repetitions are called generations, an example of five generations is shown in Figure A.12 in Appendix A. This gap-filling algorithm is now performed three times on the entire datacube. These three generations are presented for one timeslice in Figure 3.8.

After this spatial interpolation, another temporal interpolation was performed. The procedure was similar to the one performed during the first step. The main difference is that now all temporal gaps that are no longer than twelve time steps are filled

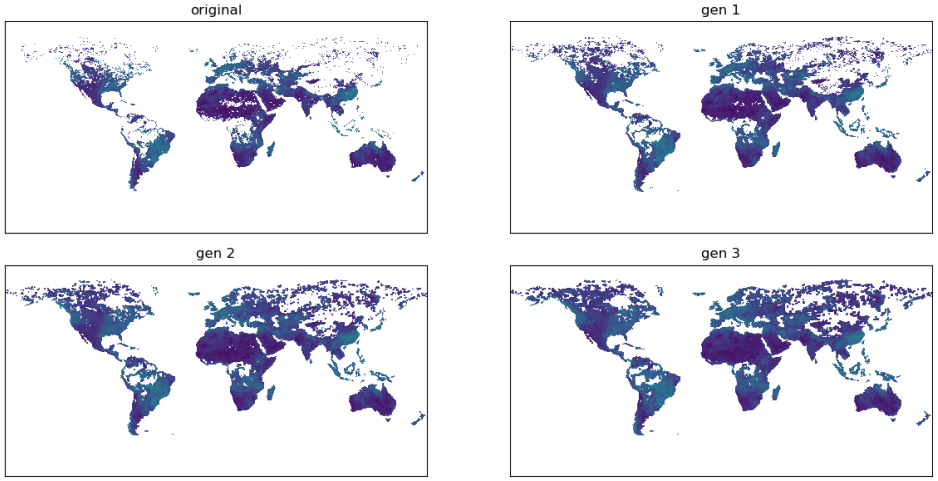



Figure 3.8: Near-surface Soil Moisture: spatial interpolation results

through linear interpolation. This corresponds to all remaining gaps up to half a year in length.

To provide a final estimate for the remaining large gaps, the seasonal cycle is calculated. This is performed on the raw data, which is divided into individual time series. Every time series has a time span of 13 years, every year has 12 months and every month has two time steps. The resulting 312 time steps can be cut into year-long strips of 24 data points each. Now, the data points in each strip are numbered one to 24. The number 1 corresponds to the first of january, the number 2 to the $15^{th}$ of january and the number 24 to the $15^{th}$ of december. The data points of these strips are now grouped together according to their number. If these groups are averaged out and ordened from one to 24, they represent the "average" year within this 13 year period. This is the basis for estimating the seasonal cycle of a variable (a mathematical description is given in Section 4.1, formula 4.4). Gaps in the seasonal cycle are filled via linear interpolation. Now, if the previously interpolated time series still had any gaps after the previous steps, the seasonal cycle is thus introduced as an estimate.

### 3.3.3 Irrigation

The irrigation dataset has no missing values, the native spatial resolution is 0.5° x 0.5° and the temporal resolution is monthly. To obtain the desired resolution, both spatial and temporal upscaling are needed. The spatial interpolation is performed first, because it is the most memory intensive step. Afterwards, the temporal upscaling is performed via simple linear interpolation.

Different pre-implemented versions are available for spatial upscaling (in two dimensions) in the python package scipy. However, these upscaling methods are all insufficient, as they don't handle the missing values of water bodies well. When these upscaling methods encounter a pixel of water (missing value) adjacent to a pixel of land (with a value), the interpolation result becomes a missing value. By using these methods, one inevitably loses information along the coast lines of continents and along water bodies. Thus, another algorithm is proposed. Figure 3.9 serves as a visual guide to accompany the explanation of this algorithm.

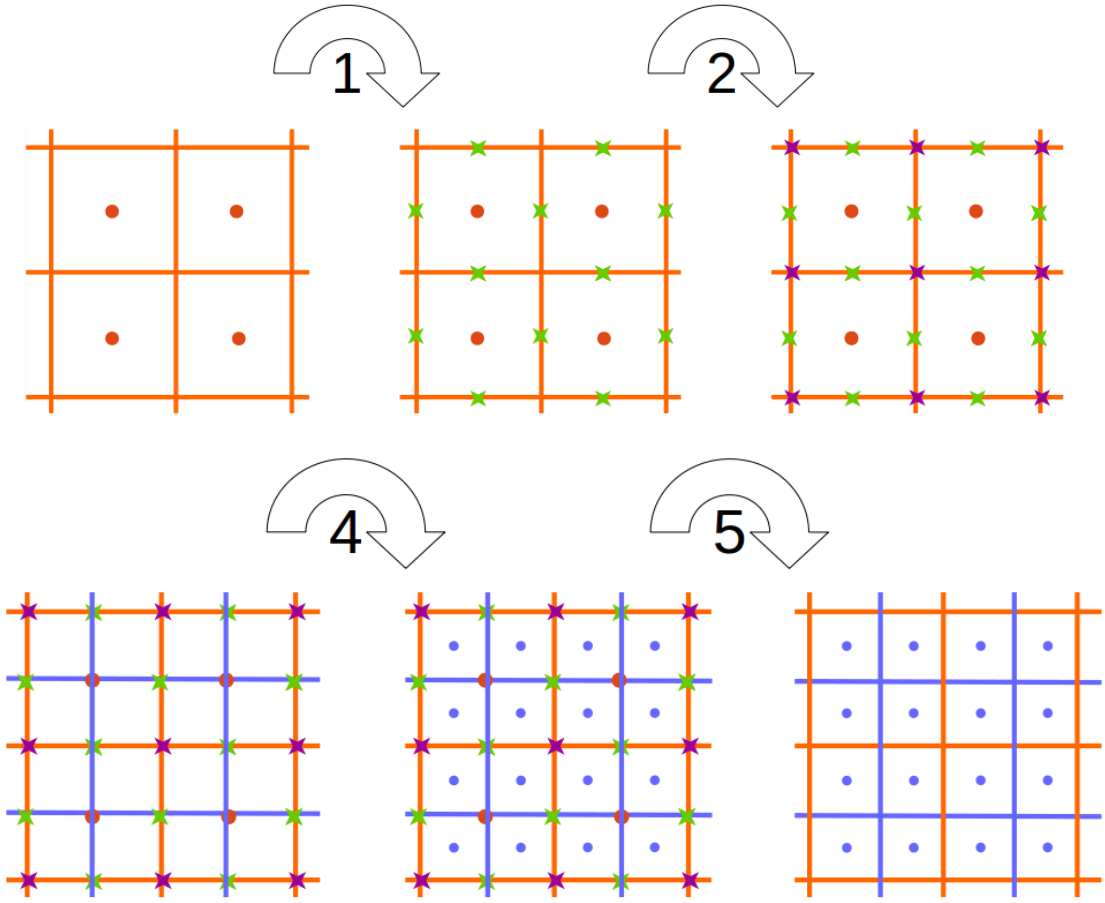



Figure 3.9: Spatial upscaling algorithm. Orange dots = original data, blue dots = new data points, green and purple stars = temporary data points

The initial situation is described in the top left of the figure. This data is arranged in a 0.5° x 0.5° resolution grid as depicted with the orange lines. Every pixel is referenced through its central coordinate, represented by the orange dots. The desired 0.25° x 0.25° resolution is depicted in the bottom right of the figure. The first step in the algorithm consists of calculating the green stars. They are all located between two orange dots and are calculated as the "value average" of their two neighboring dots. Value averaging ignores missing values and simply returns the average of the avail-

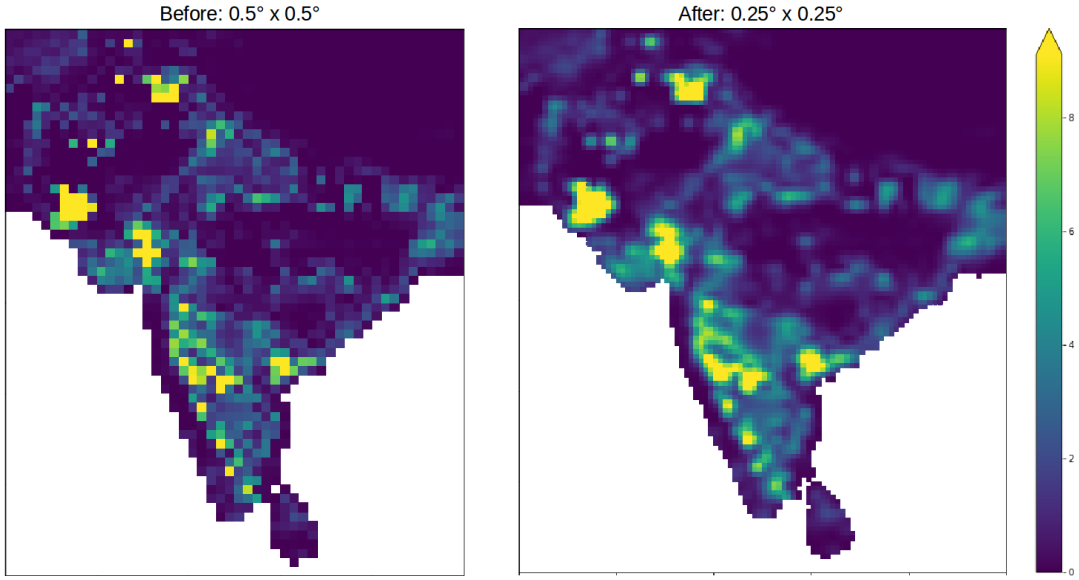



Figure 3.10: Irrigation: India before and after spatial upscaling

able values. Only if all values are missing will the value-average also be a missing value.

The second step consists of calculating the purple stars. These are each located in the middle of four orange dots and are calculated as the value average of these four dots. Now that the orange and green stars are calculated, one draws orthogonal lines through all the orange dots and green stars. The combination of old and new lines forms the new grid of double resolution. As the fourth step, a new central point can be marked with a blue dot for every pixel in this new grid. Every blue dot is surrounded by four data points: one orange dot, two green stars and a purple star. The value of each blue dot is calculated as the value average of these four data points. The final step consists of removing all purple stars, green stars and orange dots. Notice that the original data points are removed when doubling the spatial resolution. Figure 3.10 shows the result for the irrigation data set in India. Figure A.15 in Appendix A shows the result in Australia. After the spatial upsampling, every time series is upsampled from monthly to bi-weekly resolution via simple linear interpolation.

### 3.3.4 Vapour Pressure Deficit

The Vapour Pressure Deficit dataset has a native spatial resolution of 1° x 1°, with daily time steps. Thus, upscaling is needed in space and downscaling is needed in time. Data gaps are also present, most of them arising from the inherent swath width of the satellite. Firstly, the temporal downscaling will be performed, followed by the gap-filling. Finally, the spatial upscaling is performed.

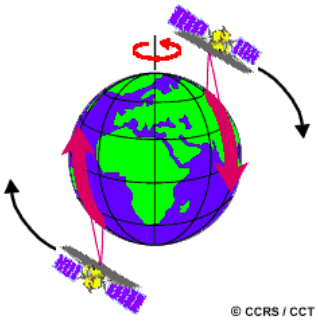



Figure 3.11: Satellite orbit, Adapted from NRCAN[4]

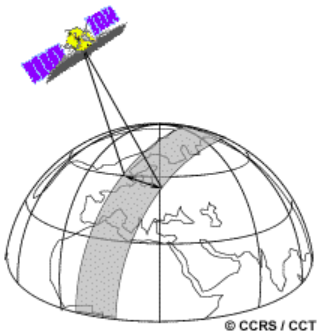



Figure 3.12: Swath Width, Adapted from NRCAN[4]

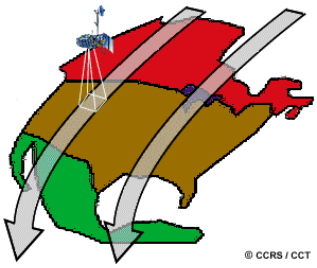



Figure 3.13: Coverage, Adapted from NRCAN[4]

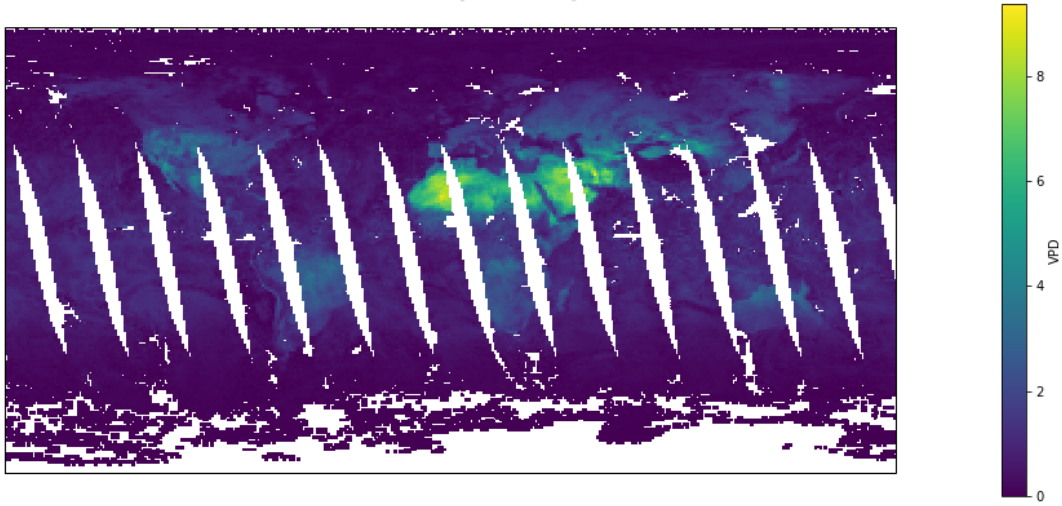



Figure 3.14: A timeslice of the raw VPD data cube (without masking of the oceans)

The data gaps mostly originate from the way these measurements are made, on a daily basis. The satellite continuously encircles the globe from the North to South Pole and then to the North pole again on the other side (Figure3.11). This orbital movement, combined with the natural rotation of the earth ensures that the satellite covers a different strip of the earth with every rotation. This width of these strips is called the swath width (Figure3.12). The combination of all these different strips determines the coverage of a satellite (Figure3.13). Because global measurements are required on a daily basis, the satellite should cover the entire globe from east to west to east (longitudinal) within 24 hours. Due to this requirement and the inherent swath width of every orbit, parts of the globe are not covered. This is observed from a timeslice of the raw VPD datacube, where strips of similar width are missing between different passages of the satellite (Figure 3.14).

These gaps are not a real problem, since temporal downscaling is required to go from the native daily resolution to bi-weekly. This means that roughly 15 days of measurements are averaged together to produce one data-point every 2 weeks. Fortunately,

[4]Natural Resources Canada, Retrieved July 23, 2018 from `http://www.nrcan.gc.ca/node/9283`

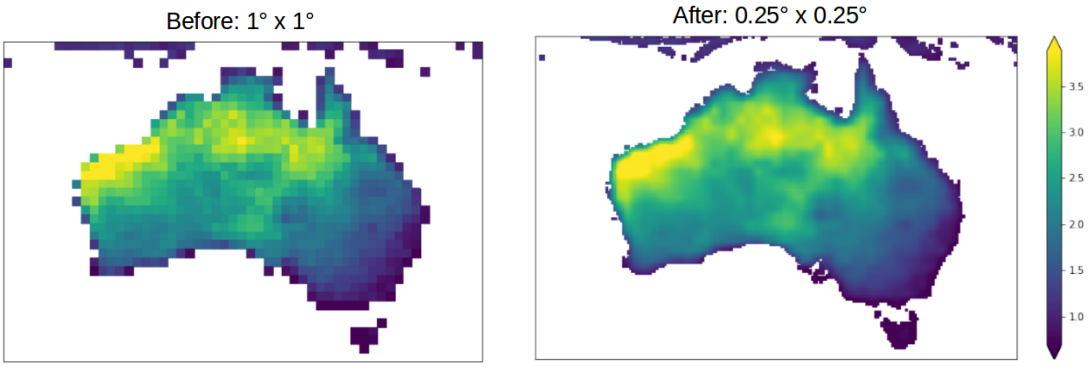



Figure 3.15: VPD: Australia before and after spatial upscaling

these gaps move spatially from one day to the next. This movement is observed in Figure A.13 in Appendix A. Because of this, few areas remain with missing data after the downscaling and those data gaps are very short. The longest temporal gap of all pixels is just five data points long. Thus, these gaps are easily filled through linear interpolation in time.

The final step to be performed is the spatial interpolation. The same algorithm showed in Figure 3.9 is performed twice to obtain the desired spatial resolution. The first application doubles the resolution from the native 1° x 1° to 0.5° x 0.5°. The second application again doubles the resolution to the desired 0.25° x 0.25° resolution. Figure 3.15 shows the results of the spatial upscaling for Australia. Figure A.14 in Appendix A shows the results for Africa.

CHAPTER 4

# METHODS

In this chapter, the general methods are explained. As a first step, the calculation of the NDVI anomalies from the raw NDVI data is explained. The theoretical framework that allows for the causal inference is Granger causality, which is explained together with the machine learning models that are used. To tackle a global problem of this size, a simplification is needed. This ultimately allows one to perform multiple local analyses in parallel. A time series decomposition is performed, and some high-level features are constructed to aid in this local analysis. In the next chapter, the results of these local analyses are combined to obtain global results.

## 4.1 NDVI anomalies

In studies that perform a Granger causality analysis on vegetation time series, the use of seasonal anomalies is commonplace (Kaufmann et al., 2003; Wang et al., 2006; Jiang et al., 2015; Notaro et al., 2006). A range of different decomposition techniques exist to extract these NDVI anomalies from the raw NDVI time series. Only decomposition techniques based on the additive STL model are considered. This model decomposes the raw time series into three additive basic elements. The first element is a longterm trend $T$, which can sometimes change over time. The second element is a seasonal cycle $S$, which captures the cyclic change of a variable within a yearly frequency. The third element is the anomalies (or residuals) $A$, that arise from substracting the trend and seasonal cycle from the raw data. The anomalies are thus defined as the variations over time that are not included in the trend or the seasonal cycle: $y_t^A = y_t - y_t^T - y_t^S$.

A popular decomposition technique is based on the Loess smoother (Cleveland et al., 1990). In this technique, the trend is not assumed to be linear and is allowed to change over time. The rate of this change is chozen via a window of time over which the trend is calculated. A small window creates a rapidly varying trend while a large window creates a more smooth trend. This is especially beneficial for climate-vegetation time series with a long timespan, where the long-term rate of change is

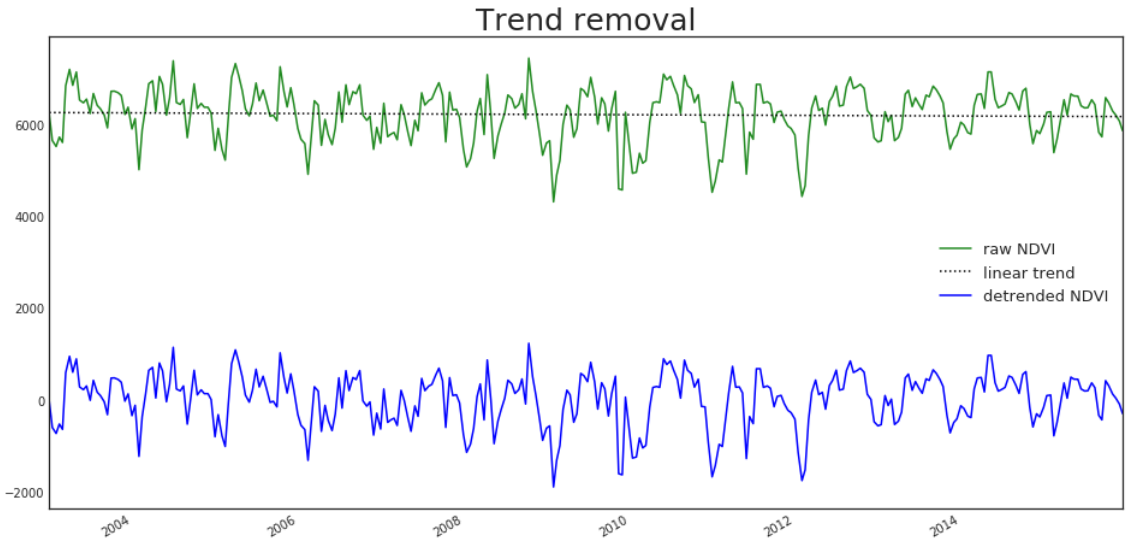



Figure 4.1: NDVI: calculation of detrended NDVI from raw NDVI

often not constant. A similar technique is the BFAST technique proposed in Verbesselt et al. (2010). In this technique, the trend is assumed to be a piecewise linear function seperated by breakpoints. This is useful for remote sensing data, where sudden persistent changes can arise from an update to a satellite.

Both decomposition techniques are considered too complex, because of the hyperparameters that need to be estimated and the difficulties that arise from missing data. Due to the global scale of the data, an immense variability exists between the different time series. Therefore, a simpler and more robust technique is sought to ensure proper decomposition of all time series. The used technique is adapted from Papagiannopoulou et al. (2017b). For the estimation of the trend $y_t^T$, a simple linear regression model in function of time is used.

$$y_t^T = a * t + b \tag{4.1}$$

The parameterisation of the slope $a$ and intercept $b$ happens with the standard method of minimizing the residual sum of squares.

$$\min_{a,b} \sum_{P+1}^{N} (y_t - y_t^T)^2 \implies a, b \tag{4.2}$$

This procedure is demonstrated in Figure 4.1 for a sample NDVI time series. The Theil–Sen estimator was also tried, as it can be computed efficiently, and is insensitive to outliers (Sen, 1968). This method was however abandonded to keep the research comparable to Papagiannopoulou et al. (2017b). Performing a detrending removes some important non-stationary effects from the data. This ensures that the mean of

the probability distribution does not change over time (with the assumption that the trend is truly linear and its parameters are constant).

The calculated trend is substracted from the raw NDVI to produce the detrended NDVI: $y_t^{dT} = y_t - y_t^T$. From this detrended NDVI, the seasonal cycle is estimated. The method used in Papagiannopoulou et al. (2017b) is based on the assumption that the seasonal cycle is annual and constant over time. From this, one can simply use the monthly expectation as an estimate for the seasonal cycle. When indexing the detrended data in a time series via a one-based indexing system (for the first point $i = 1$ and the final points $i = N$), the formulation is pretty straightforward. The number of the month is denoted a $m$, the monthly expectation $y_m^S$ is calculated as follows:

$$y_m^S = \sum_{i=1}^{N} \frac{y_i^{dT} * \delta_i}{n_{yrs}} \quad \text{with } \delta_i = \begin{cases} 1, & i \mod 12 = m \\ 0, & Else \end{cases} \tag{4.3}$$

The numbers of years is denoted as $n_{yrs}$. The inclusion of dummy variable $\delta_i$ is merely for the sake of notation, the implementation does not require it. This definition only works for time series with a monthly resolution. In this study however, we can do better by calculating the seasonal cycle on a bi-weekly basis. The notation is similar but the months are all devided in halves and index $sm$ now denotes the number of the semi-month, ranging from one to 24. The odd numbers are the first halves of the months containing days one to fifteen. The even numbers are the second halves containing the remaining days.

$$y_{sm}^S = \sum_{i=1}^{N} \frac{y_i^{dT} * \delta_i}{n_{yrs}} \quad with\ \delta_i = \begin{cases} 1, & i \mod 24 = sm \\ 0, & Else \end{cases} \tag{4.4}$$

This procedure is shown in the top part of Figure 4.2. When the resulting seasonal cycles are examined, they appear less smooth than those reported in (Papagiannopoulou et al., 2017b). This is normal since every $y_m^S$ in that studie was calculated with a lot more data. Due to the monthly resolution every datapoint is an average over 30 days, compared to the 15 days in this work. On top of that, the time series in Papagiannopoulou et al. (2017b) contain 30 years of data, while the time series in this work only contain 13 years. This makes the seasonality calculation less stable, since less years are averaged together. If a large identical anomaly occurs for three years in the same semi-month, its effect will be visible in the seasonal cycle. In a 30 year time series however, this will have less influence. In order to protect the seasonal cycle from strong anomalies, a smoother is applied. This is based on the knowledge that NDVI seldom changes drastically in a 15 day window, due to

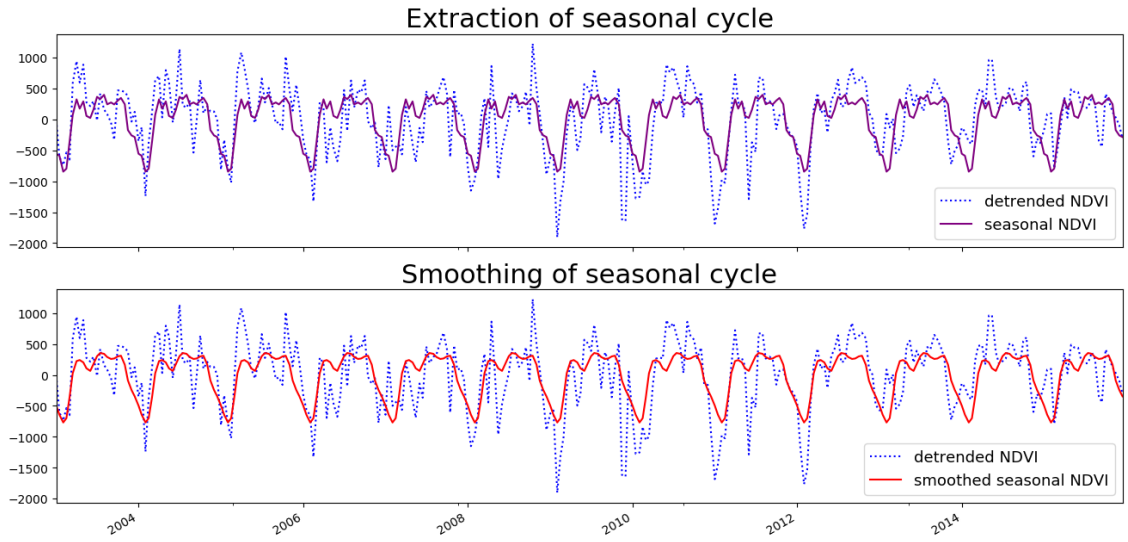



Figure 4.2: [top] NDVI: extracting seasonal cycle from detrended NDVI [bottom] NDVI: smoothing of the seasonal cycle

its lagged response and resilience to perturbations (Seddon et al., 2016). Different smoothers were tested for a variety of pixels (figure A.16 in Appendix A). The different smoothers are judged based on their force of smoothing and the possible introduction of artefacts. The chozen smoother (SM 2 in Figure A.16) is fairly basic, it just replaces the seasonal cycle $y^S_{sm}$ with the weighted average of itself and its two neighbours in time (the first and final semi-month are also neighbours).

$$y^S_{sm\_smooth} = w * y^S_{sm-1} + y^S_{sm} + w * y^S_{sm+1} \tag{4.5}$$

The weights $w$ for the neighbours are 0.5. It may seem counterintuitive to include a weighted version of the next semi-month in time. If this is not done however, the entire monthly cycle shifts slightly backwards in time because the temporal averaging is then one-directional.

Finally, the anomalies are calculated as the residual part of the raw data that was not yet integrated in the seasonal cycle or the trend.

$$y^A_t = y_t - y^T_t - y^S_{sm\_smooth} = y^{dT}_t - y^S_{sm\_smooth} \tag{4.6}$$

These anomalies are the response variabel in the analysis. There are three good reasons to prefer the NDVI anomalies over the raw NDVI as the response variable. Firstly, the anomaly construction results in a more stationary response variable, which is desirable for a Granger causality analysis (see Section 4.2). The second reason has to do with the high autocorrelation of the raw NDVI. There is a high correlation of raw NDVI with lagged versions of itself. By removing the trend and seasonal cycle, the

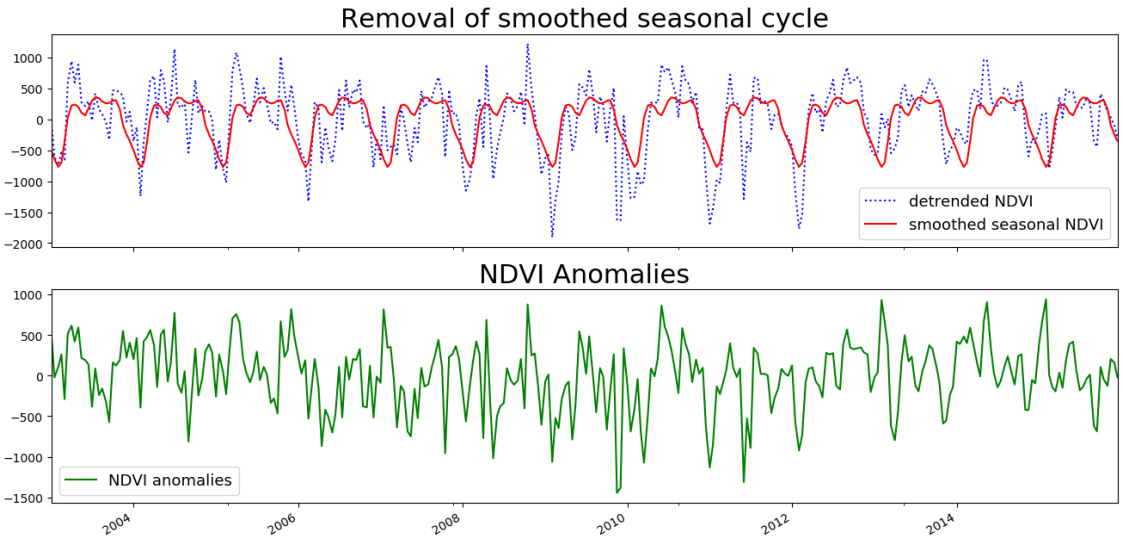



Figure 4.3: Calculation of the NDVI anomalies

autocorrelation of the anomalies is lower than the raw NDVI. This effect is demonstrated in Figures 4.4 and 4.5. These figures show the autocorrelation of the raw NDVI and the NDVI anomalies for two different measures of correlation: the Pearson correlation and the distance correlation. The Pearson correlation is the classic measure for correlation, it only accounts for possible linear correlation and ranges between -1 and 1. A higher absolute value corresponds to a higher linear autocorrelation. The distance correlation is a more recent measure of correlation that also accounts for non-linear correlation and ranges between 0 and 1 (Székely et al., 2007). A higher distance correlation corresponds to a higher correlation between variables, either linear or non-linear. From the Figures, it can be observed that the autocorrelation of the NDVI anomalies is much lower for both correlation measures and is only apparent in small lags. The level of autocorrelation that is present in the raw NDVI is undesirable for the Granger causality analysis. Thirdly, it is not very interesting to look at the raw data as a predictor since the trend and seasonal cycle often make up most of it. When predicting this raw data, one is effectively predicting mostly the seasonal cycle and trend, since they are easier to detect for most models. This is not interesting, it does not lead to new scientific insights. The existence and behaviour of both seasonal cycles and trends has been researched for decades already. The anomalies contain the causal information that we are searching for: the response of vegetation to climate perturbations.

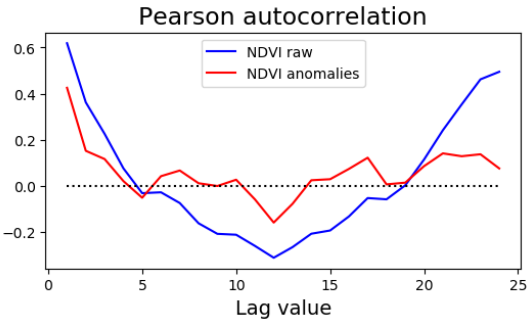



Figure 4.4: Autocorrelation in function of lag for raw NDVI and NDVI anomalies, based on Pearson correlation

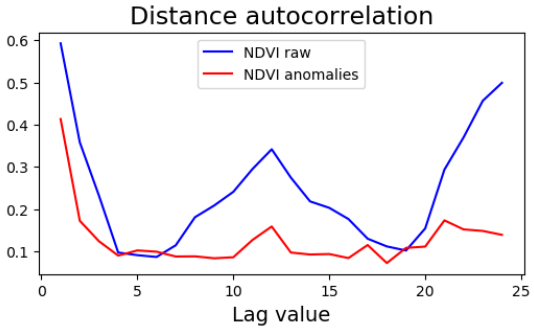



Figure 4.5: Autocorrelation in function of lag for raw NDVI and NDVI anomalies, based on distance correlation

## 4.2 Granger causality

Causality is the connection between two variables, the cause and the effect, such that the occurence or state of the cause influences the occurence or state of the effect. Although many people have an intuitive understanding of this concept, it is often incomplete or plainly incorrect. A common logic fallacy is that correlation equals causality. When two variables are highly correlated, it is often assumed that they have a causal relation. Correlation is however not a sufficient condition for causality, as a great number of examples can prove[1].

The pipeline for performing a causality analysis described in Papagiannopoulou et al. (2017b) is based on Granger causality. This theoretical framework allows for identifying causal relations between time series and was first introduced in the field of econometrics (Granger, 1969). Ever since then, it has seen used in a variety of research fields and several extensions have been proposed. The causality relationship between two variables is defined through two main principles: Firstly, the cause must happen prior to the effect. Secondly, the cause must have some unique information about the future of the effect. These general conditions for causality are then transformed into an operational definition for time series (Granger, 1980).

Suppose that one wishes to investigate the causal influence of a variable $x$ on a variable $y$. In this research, the NDVI anomalies is variable $y$. The variable $x$ is the variable that is researched for its causal influence on $y$, which depends on the research question (see Section 1.1). To investigate the causal influence of $x$ on $y$, the time series $\mathbf{x} = [x_1, x_2, ..., x_N]$ and $\mathbf{y} = [y_1, y_2, ..., y_N]$ are available. To do so with Granger causality, one wishes to prove that unique information is present in the history of $x$ about the future value of $y$. It is important to include all the other variables $z$ that may have an influence on either one of these variables. This ensures

[1] http://www.tylervigen.com/spurious-correlations

more protection against false conclusions. For these additional variables, time series $\mathbf{z} = [z_1, z_2, ..., z_N]$ are available. The simplication of a single variable $z$ is made for the sake of notation. To be fully correct, the variable $x$ that is researched for its casual influence can also be a group of variables (f.e. the water group). In that case the group is handled as a whole and the procedure is no different.

The causal influence of $x$ on $y$ is quantified through predictive modelling of time series **y**. More specifically, a datapoint $y_t$ in **y** at a time $t$ is modelled using previous data from **y** itself, **x** and **z**. Only data from within a confined window of preceding timesteps $t-p$ to $t-1$ are used, as shown in Figure 4.6. The size of the moving window is determined by parameter P, called the maximum lag. To isolate the unique information that is present in the history of **x**, two different models are used. The first model is the full model, which includes the history of all variables:

$$\hat{y_t}^{full} = f(y_{t-p}, x_{t-p}, z_{t-p}) \; \forall \; p \in 1, \ldots, P \tag{4.7}$$

The second model is the reduced model[2], which includes all variables, except $x$. Depending on the research question, either a single variable or a group of variables are omitted from the reduced model.

$$\hat{y_t}^{red} = f(y_{t-p}, z_{t-p}) \; \forall \; p \in 1, \ldots, P \tag{4.8}$$

The variable $p$ is called the lag in these models, with $P$ the maximum lag. After training the models, the prediction accurary of both models is compared on data that was not used for training (out-of-sample data) (Granger, 1980). If the prediction accuracy of the reduced model is less then that of the full model, one can conclude that unique information is contained in the history of **x** about the future of **y**. In that case, it is stated that $X$ "Granger causes" $Y$. This is reffered to as the partial Granger causality of a variable (group). This partial Granger causality is not true causality, but a notion of causality based on a number of assumptions (Granger, 1980).

A third model is also included in the analysis. This is the baseline model, which includes only past values of the NDVI anomalies to predict the current NDVI anomaly value $y_t$. The difference in prediction accuracy between the full and baseline model gives a measure for the total Granger causality of all included climate drivers on vegetation anomalies. This is relevant to identify regions with poor Granger causality. This is also why the strong autocorrelation present in the raw NDVIA is problematic. A high autocorrelation in the response variable results in a very strong baseline model. This makes it very hard for the extended and full models to improve prediction accu-

[2]This naming convention differs from Papagiannopoulou et al. (2017b), where it is named baseline model.

racy over the baseline, which makes the whole Granger causality quantification more difficult (Papagiannopoulou et al., 2017b).

$$\hat{y_t}^{base} = f(y_{t-p}) \ \forall \ p \in 1, \ldots, P \tag{4.9}$$

Prediction accuracy is quantified through the coefficient of determination $R^2$, which is defined as one minus the ratio of the residual sum of squares and the total sum of squares.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{P+1}^{N}(y_t - \hat{y}_t)^2}{\sum_{P+1}^{N}(y_t - \bar{y}_t)^2} \tag{4.10}$$

From the available data, a part is needed for determining the model parameters. This proces is called training the model and the used data is referred to as the training data. Preferably, the $R^2$ should be calculated on test data that was not used during training. Doing otherwise often results in an over-estimation of the model performance. A common practice is to devide the time series in two parts, with the first parts being used for training and the second parts for testing (James et al., 2013). This is only a viable stategy if an abundance of data is available. With only 312 datapoints per time serie, this is certainly not the case (more details in Section 4.3). Thus, 5-fold random cross-validation is used to make better use of the available data. This procedure randomly divides the datapoints into five equal parts. First, one part is excluded for testing, and the other four parts are used for training. Predictions are made for the testing set. Then, a different part is excluded for testing and predictions, with the four other parts used for training. This is repeated untill each of the five parts is used for testing and predictions are made for the entire time series. Now, the $R^2$ can be calculated as shown in Equation 4.10. It is chozen to keep the division of time steps over these 5 folds identical between different models, to reduce unnecessary sources of variance in the $R^2$ score.

Apart from random cross-validation, blocked cross-validation is also a valid option, which is often preferred for time series. In that case, the data is not randomly assigned into five batches. Instead, it is chopped up into five chronologically coherent parts along the time dimension. Blocked cross-validation is often preferable, even in a causal inference setting (Roberts et al., 2017). Due to dependence structures in the temporal data, the assumption of independent training and testing data becomes violated for random cross-validation, which randomly shuffles the data. This method was also tested, the results and discussion are provide in Section 5.4.4). Random 5-fold cross-validation is adopted for its better performance. This is defended with the fact that we are not interested in the pure $R^2$ values, but are only interested in the difference in performance between the full and reduced model. Opposed to the

prediction setting, accounting for future out-of-sample data is also of no interest in this study, effectively making one main reason for blocked cross-validation invalid (Roberts et al., 2017).

As stated before, it is important to include all the other variables $Z$ that may have an influence on either $X$ or $Y$ (Granger, 1980). This necessity is related to the assumption of causal sufficiency, which states that all hidden common causes or confounding variables are included in the data. To illustrate the impact of a confounding variable on the analysis, consider the following example. A researcher wishes to investigate causal factors related to the number of shark attacks in his/her home country. For some reason, the researcher wishes to investigate the causal influence that ice cream sales on the beaches have on the number of shark attacks. Only a time series **y** of shark attacks (red in Figure 4.6) and a time series **x** of the number of ice cream sales on beaches (blue in Figure 4.6) are available. When conducting this causality analysis, it is observed that the predictive performance of the full model including **x** is stronger than that of the reduced model excluding **x**. This leads one to conclude that the number of ice cream sales are somehow a definite causal factor in the number of shark attacks. This ridiculous conclusion is produced due to the omission of an obvious hidden common cause from the analysis. This common cause is the number of people present at the beach, which influences both the number of shark attacks and the sales of ice cream. A more correct analysis would include a time series **z** of the number of people present at the beach (yellow in Figure 4.6) in both models. In this new research, both models would have similar predictive performance and the conclusion is different. Even though the number of icecream sales on the beach is still correlated with the number of shark attacks, it is no longer identified as a causal factor.

The assumption of causal sufficiency is almost always violated in climate studies. Due to the complex nature of these global natural phenomena, it is practically impossible to include all possible confounding variables and hidden common causes. This implies that the causality conclusions from this type of research don't necessarily reflect reality. The conclusions are preferably examined by domain experts and confirmed via knowledge of physical mechanisms responsible for the causal link.

Some other assumptions are almost inherent to machine learning modelling. One such assumption is that the considered data is of sufficient quality to serve as a solid basis for the analysis. As stated in Faghmous and Kumar (2014): "Any data-driven discovery is inexorably linked to the quality of the data, their source, and sampling bias". Another important assumption is that the used model succeeds in modelling the underlying relations between the data well. Historically, mostly linear models are used in the analysis of vegetation-climate interactions. However, these interactions

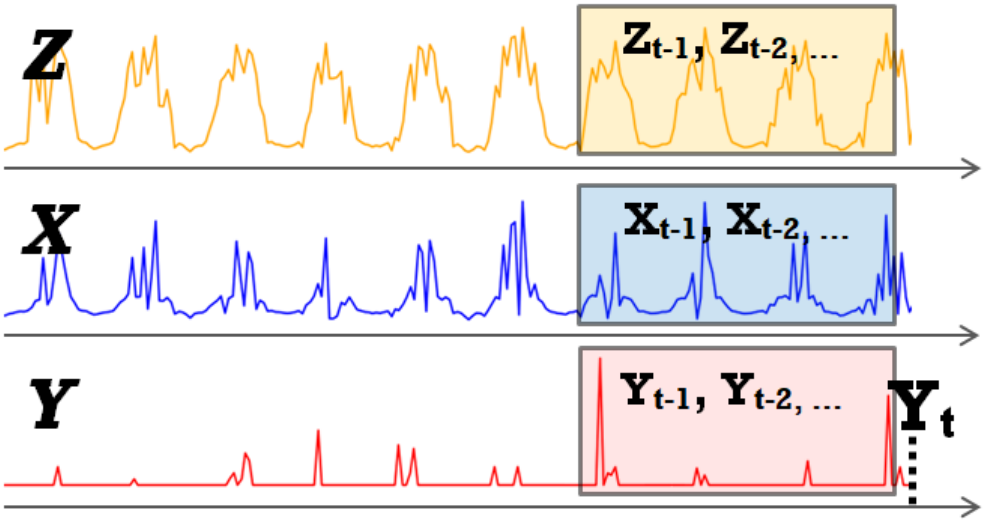



Figure 4.6: time series **x**, **y** and **z** and the window for prediction of $y_t$

have often been reported as non-linear in nature (Heimann and Reichstein, 2008; Bonan, 2011; Zhang et al., 2014; Papagiannopoulou et al., 2017b). Both linear ridge regression models and non-linear random forest models are used in this analysis, to preclude prejudice to either one. The model with the best performance will be chozen.

### 4.2.1 Linear GC modelling

The traditional Granger causality analyses in climate sciences are performed with linear models, exceptionally supplemented with a quadratic term (Kaufmann et al., 2003; Wang et al., 2006; Jiang et al., 2015; Notaro et al., 2006). The used models are multi-variate vector autoregressive models (VAR). They strive to predict $x$ and $y$ simultaneously in function of themselves and each other, without making the artificial subdivision into predictor and response variables. A multi-variate linear VAR model of order $p$ includes $p$ lags of both variables and is represented in matrix notation by the following equation.

$$\begin{bmatrix} y_t \\ x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{03} \end{bmatrix} + \sum_{p=1}^{P} \begin{bmatrix} \beta_{11p} & \beta_{12p} & \beta_{13p} \\ \beta_{21p} & \beta_{22p} & \beta_{23p} \\ \beta_{31p} & \beta_{32p} & \beta_{33p} \end{bmatrix} \begin{bmatrix} y_{t-p} \\ x_{t-p} \\ z_t - p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \tag{4.11}$$

The $\beta$ parameters are those estimated during model training, while the $\epsilon$ parameters represent white noise terms. In this study, the interest lies solely in the influence of climate drivers $x_t$ on vegetation anomalies $y_t$, thus only the first equation of this system is of interest. The standard notation of this first equation is equivalent to the

linear full model equation.

$$y_t = \beta_{01} + \sum_{p=1}^{P} \left( \beta_{11p} y_{t-p} + \beta_{12p} x_{t-p} + \beta_{13p} Z_{t-p} \right) + \epsilon_1 \tag{4.12}$$

When setting parameter $\beta_{12p}$ to zero, the formula for the linear reduced model is obtained.

$$y_t = \beta_{01} + \sum_{p=1}^{P} \left( \beta_{11p} y_{t-p} + \beta_{13p} Z_{t-p} \right) + \epsilon_1 \tag{4.13}$$

The basic model that is described by the equations above is the multivariate linear regression model. This model is widely used in statistics and machine learning. It is parameterized (= trained) through minimizing of the residual sum of squares (RSS $= \sum_{p+1}^{N} (y_t - \hat{y}_t)^2$). This process ensures that the $\beta$ parameters are chozen to best fit the training data. In this thesis, eleven basic regressor variables are available. However, each different lag time adds eleven more regressor variables and a large number of high-level features are also constructed (see Section 4.5). If a model has more regressor variables than available data points (312 for a single time series), one is working in a high-dimensional setting. This means that the classic parameterization method of minimizing the RSS no longer provides a single solution and is rendered obsolete. Even in a close to high-dimensional setting, problems arise from the minimal RSS method. When the number of datapoints is close to the number of parameters, simple RSS minimalization often leads to overfitting. The model is too flexible for the available dataset, it succeeds at fitting *all* the training data. This means that the model is also fitted to random variations in the data, which are meaningless for prediction. An overfitted model performs very bad on out-of-sample data, and is of little use for prediction (James et al., 2013).

A solution to this problem is regularization: the addition of a penalty term to the residual sum of squares. This penalty term is added to punish high absolute parameter values, if they don't lead to a significant drop in RSS. It is constructed by grouping all the $\beta$ parameters into a single vector $\boldsymbol{\beta} = [\beta_{01}, \beta_{11p}, \ldots, \beta_{13p}]$ and calculating the norm (= the sum of the absolute values) of this vector. If the penalty term is quadratic, it is an L2-norm and the resulting model is a ridge regression model (McDonald, 2009).

$$\min_{\boldsymbol{\beta}} \sum_{P+1}^{N} (y_t - \hat{y}_t)^2 + \lambda ||\boldsymbol{\beta}||^2 \tag{4.14}$$

This $\lambda \geq 0$ is a tuning parameter, which is determined before minimizing the above mention term. A higher $\lambda$ term is a stronger penalty, which punishes more parameters and sets all the weakest predictors $\beta$ parameters close to zero. Only the stronger predictors will have $\beta$ parameters significantly different from zero. A feature scaling of the regressors is performed before training. Otherwise, the model would be biased towards variables of a higher magnitude. Ridge regression will produce a different set of $\beta$ parameter estimates for each value of $\lambda$ and not all sets give good predictive performance. Selecting a good value for the tuning parameter is critical, which is further discussed in Section 4.4.

Within the framework of Papagiannopoulou et al. (2017b), the existence of Granger causality is identified by comparing the prediction accuracy of the full and reduced model on out-of-sample data. For linear VAR models, it is also possible to test these results for their statistical significance (Wang et al., 2006; Jiang et al., 2015). A standard formulation of the null hypothesis is that variable $X$ does not Granger cause $Y$. This corresponds to all of the $\beta_{12p}$ parameters in the full model being 0. An F-statistic can then be constructed to test this hypothesis (Wang et al., 2006). However, if you a large number of variables and lags is used, the partial F-test can lose power. Alternatives are available, but are less straight-forward (Ducasse, 2017). To ensure the correctness of these statistical tests an important set of assumptions should be fulfilled. The time series should be stationary and data should be transformed to eliminate strong autocorrelation. Variables and observational techniques should be independent from each other and the errors $\epsilon$ are assumed to be normally distributed. These assumptions are typically violated for vegetation-climate data on a global scale. Furthermore, these statistical tests typically do not extend well to non-linear models (Papagiannopoulou et al., 2017b). For these reasons, they are not considered for the main analysis in this work.

### 4.2.2 Non-linear GC modelling

The use of non-linear models in Granger causality analyses about vegetation-climate interactions is rare. However, they have better predictive performance compared to linear models in similar research (Papagiannopoulou et al., 2017b) and are thus also used in this work. The model of choice is a random forest model, due to its flexibility and succesful performance in a (near) high-dimensional setting (Breiman, 2001).

A random forest model is a combination of individual decision trees. Decision trees can be applied to both classification and regression problems. In this work, the random forest is constructed from regression trees, since it is a regression problem. A tree is built from a series of splits that divide the variable space into distinct regions.

These regions are chosen that way because the training data shows similar outcomes inside the region and considerable difference between regions. For the ease of explanation, Figures 4.7 and 4.8 are adopted from James et al. (2013). An example decision

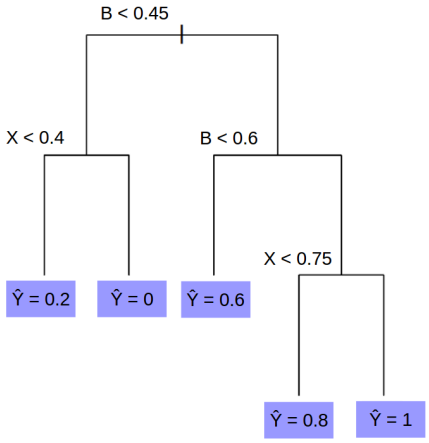



Figure 4.7: Regression tree, adopted from James et al. (2013) and altered

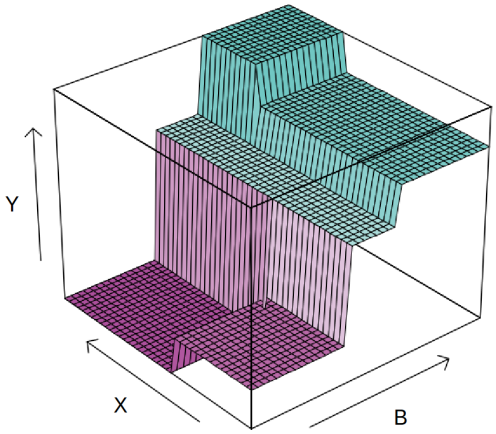



Figure 4.8: Predicted score for the variable space, adopted from James et al. (2013) and altered

tree is shown for a two dimensional regression problem. Imagine the problem in question is to estimate the quality of a new brand of coffee ($Y$) based on the quality of the beans ($B$) and the strength of the extraction ($X$). All the variables are continuous and stay between 0 and 1.

Figure 4.7 shows a regression tree fitted to the available data. The five blue boxes at the bottom are the final regions, they share the same prediction for the response variable ($\hat{Y}$). This prediction is shown in Figure 4.8. The predicted response for a region is usually the average of the observed response values for the training data in that region. The final regions are obtained through a series of splitting rules, starting at the top of the tree. The first split is based on the quality of the beans. Coffee from beans with a lower quality ($B < 0.45$) branches to the left and coffee from higher quality beans branches to the right. The low beans quality branch is then split again according to the strength of extraction. The higher beans quality branch ($B \geq 0.45$) is also split, according to bean quality and strength of extraction. In this example all splits are binary and result in two branches. Non-binary splits of three or more branches are also possible. For example, the first split ($B < 0.45$) and the second split on the right ($B < 0.6$) are equivalent to a single 3-way split.

Decision trees incorporate interaction between variables in a very intuitive way. In this example, the strength of extraction has a different effect on the predicted outcome based on the quality of the beans. For low quality beans ($B < 0.4$), a strong extraction releases more foul tasting components from the low quality coffee powder resulting in worse tasting coffee. For medium quality beans ($0.45 \geq B < 0.6$), the strength

of extraction has no influence on the quality of the coffee. For high quality beans ($B \geq 0.6$), a strong extraction releases more tasty chemicals from the beans, resulting in even better coffee.

The splits in regression trees are constructed based on minimizing the residual sum of squares (RSS). A very simple way of constructing a tree is top-down, with greedy binary splitting. At the start of the algorithm, a large variety of splits are possible. The RSS is calculated for each candidate split using the average within a region as the new prediction within that region. The split that results in the lowest RSS is applied. This procedure is repeated for each new branch untill a stop criterium is reached. This can be related to a minimum number of observations required to make a new split, a maximum number of splits or a required decrease in RSS to justify a new split.

A strong point of decision trees is that they don't assume a certain relation between the predictor and response. They are not limited to linear, quadratic or exponential relationships. This flexibility however comes at a price. A single decision tree is very non-robust. A small change in the data can cause a large change in the final constructed tree. This is undesirable, since the available data can be very noisy. Random forests present a solution to this problem by constructing a large number ($B$) of individual regression trees (this makes it an ensemble method).

The random forest algorithm involves some steps to decorrelate the trees, since constructing a lot of identical trees is of little use. A first step involves the use of a resampling method on the data. Cross-validation is a popular example of resampling methods, which allow for more efficient use of the data. The resampling method used in the random forest algorithm is bootstrap aggregating, or bagging for short. This consists of generating a large number of "new" samples from the original one. If the dataset contains $N$ datapoints, a new dataset is constructed by picking $N$ random elements from the set, with replacement. Each datapoint has the same chance of being picked at any time, even when it was already picked before. Some datapoints from the original dataset will appear multiple times in a bootstrap sample, while others don't appear. A bootstrap sample will only contain around two-thirds of the original data on average (James et al., 2013). By constructing a bootstrapped data set for every tree, each tree only has access to a limited part of the information. Each tree is grown deep with few limits to the number of splits.

The use of bootstrapped samples will still lead to a large number of similar trees. If there is a very strong predictor in the data, most trees will use this predictor in their first split. Hence, most of the trees are still very similar. Random forest further decorrelates the trees by forcing each split to only consider a small subset of the variabels. This subset of variables is picked at random each time (without replacement) and

usually contains $\sqrt{p}$ variables with $p$ the number of variables. Even though it may seem counterintuitive, it is beneficial because it reduces variance. When a strong variable is present, it is not considered in many of the splits, giving a better chance to moderately strong predictors to also play an important part in the tree. By averaging the predictions from all trees, a more robust model is obtained that performs better on unseen test data.

Another strong point of the random forest algorithm is the scale-invariance, meaning that it is invariant under scaling and some other transformations of the variables. Furthermore, the algorithm performs well in a high-dimensional setting due to its robustness to the inclusion of irrelevant features (Hastie et al., 2001). The implementation of the scikit-learn package in python is used.

## 4.3 Simplifying the problem

A first step in tackling the problem at hand is simplifying it. This involves moving away from the global setting that was necessary during data pre-processing. The Granger causality analysis will be performed locally, for every pixel individually. The individual results from these local analyses are aggregated and mapped to enable global conclusions.

For each pixel, its twelve corresponding time series are extracted from the datacubes. The Granger causality analysis is now performed for every pixel seperately, a simplified example is given in Figure 4.9. The models get a number of features (time series) to predict the NDVI anomalies time series. These features are not the climate variables as such, but rather the higher level features constructed from them. This process is explained in Section 4.5. Baseline, reduced and full models are constructed based on both random forest and ridge regression. These models only get the high-level features coming from the variables they are allowed to use (see Equations 4.7, 4.8 and 4.9). As such, random forests and ridge regression can be compared on their prediction accuracy and the results from the best of both models are used to draw the further conclusions.

Because of the way these time series are used, any knowledge of neighbourhood between pixels is not exploited. This may not be the best use of this abundance of data, but it allows for highly parallelized calculations and an immense speed-up. Furthermore, the local analyses allow for a very natural translation of the Granger causality theory, which is based on time series. The model training and predictions are performed using the Tier-2 infrastructure of the VSC centrum (Vlaams Supercomputer Centrum).

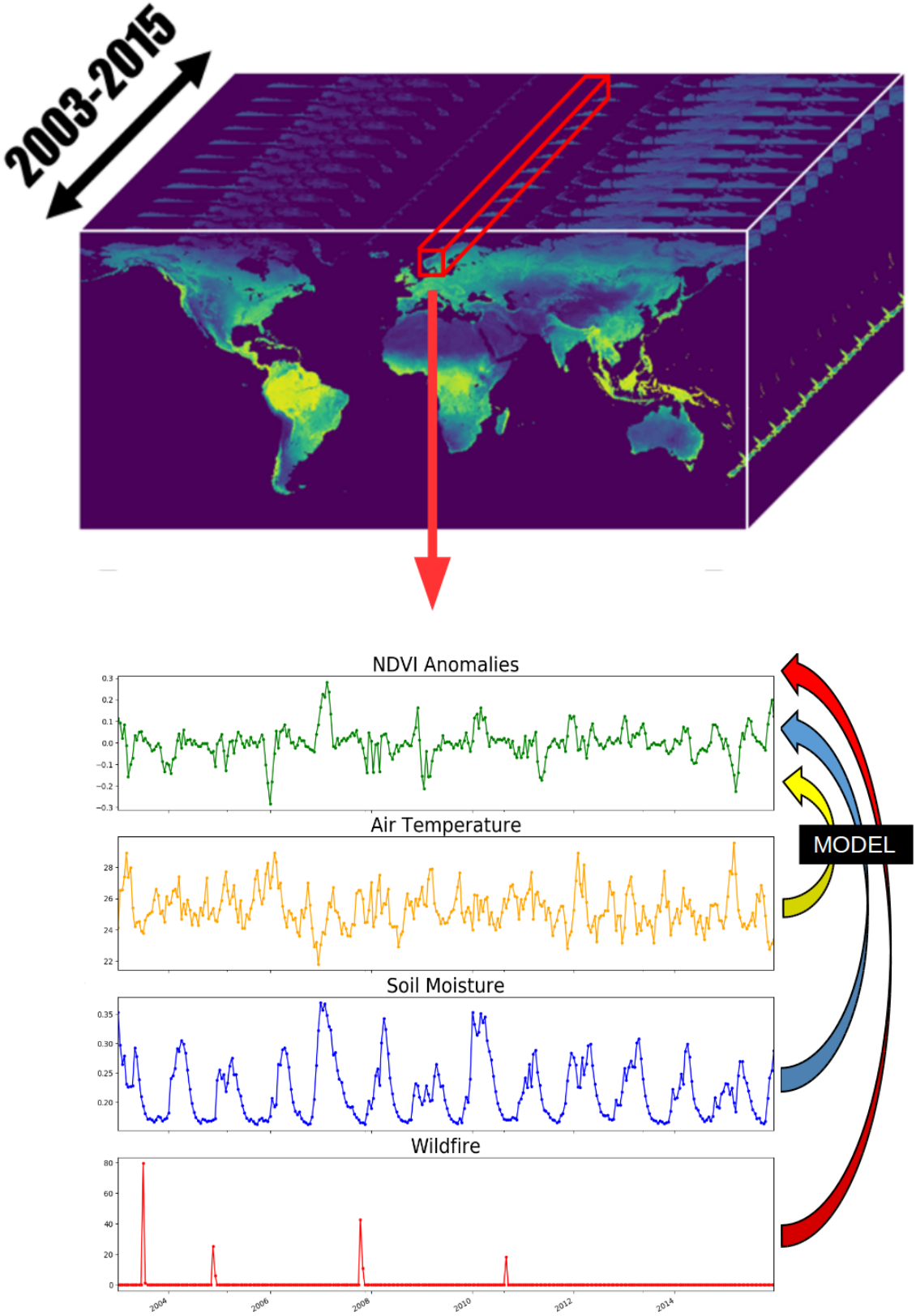



Figure 4.9: Simplified example of the procedure for one pixel

## 4.4 Model parameterization

Since every model operates independently, the hyperparameters of each model can be tuned seperately. This is very useful for ridge regression, since the optimal choice of the tuning parameter $\beta$ will vary between different pixels and models. If many strong predictors exist that are not cross-correlated, the optimal $\beta$ is low. For the ridge regression models, the optimal value of $\beta$ is picked from a set of 20 candidate values between $10^1$ and $10^{15}$. The optimal value is selected from these candidates via generalized leave-one-out cross-validation (LOOCV), which only considers the training data. The explanation of LOOCV is considered out of the scope for this thesis[3]. It is sufficient to understand that it is just another form of cross-validation, which is extremely fast for linear models (Seber and Lee, 2012). The LOOCV Ridge Regression implementation of the scikit-learn package in python is used. The training data was already selected through 5-fold cross-validation, and now the $\lambda$ parameterization is done via LOOCV within this training data. This is called a nested cross-validation scheme. This is needed for tuning hyperparameters such as $\lambda$, as they have to be set *before* the training of the model.

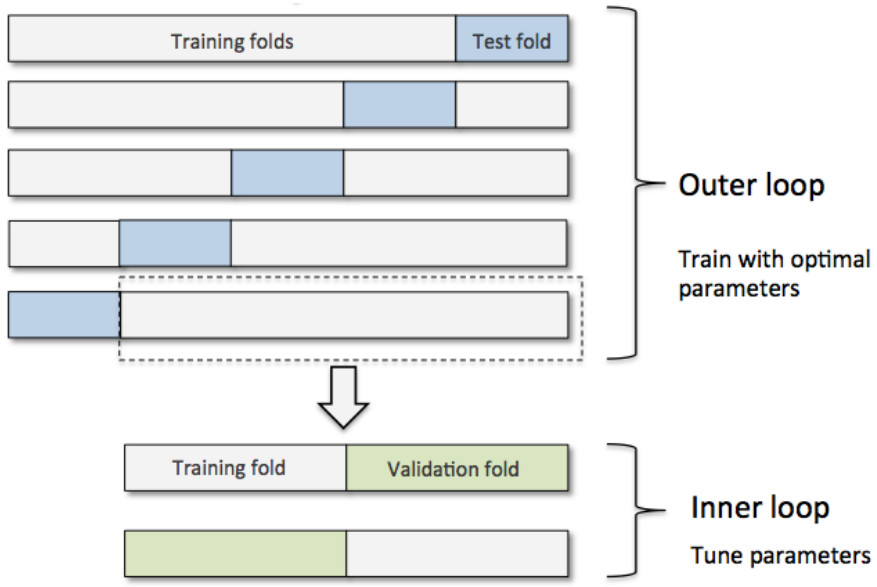



Figure 4.10: An example of a nested cross-validation scheme, Sebastian Raschka, Adopted from Machine Learning FAQ[4]

Some hyperparameters exist for the random forest models as well. They relate to the number of trees, the depth of the trees and the amount of features offered at each split. The number of trees ($B$) is a peculiar hyperparameter. Most hyperparameters have an optimal range and both smaller and larger values outside of this range give bad model predictions. For the number of trees, this is not the case. The prediction

[3] A full explanation is available in section 5.1.2 of James et al. (2013).
[4] Retrieved on July 25 2018 from `https://sebastianraschka.com/faq/docs/evaluate-a-model.html`

accuracy increases with the number of trees, but there is no optimal range after which the prediction accuracy decreases. Instead, the prediction accuracy stagnates after a certain number of trees. There is however another cost related to overestimating $B$: the time needed for model training. This increases almost linearly with the number of trees, as can be seen in Figure 4.11. From this figure, the number of trees is chozen as 150. The trees in a random forest model are typically allowed to grow deep without restriction, this is also chozen here (James et al., 2013). The amount of features offered at each split is set to the default parameter, as tuning of this parameter typically does not increase prediction accuracy by much. This default parameter is $\sqrt{p}$ with p the number of features available to the model (James et al., 2013).

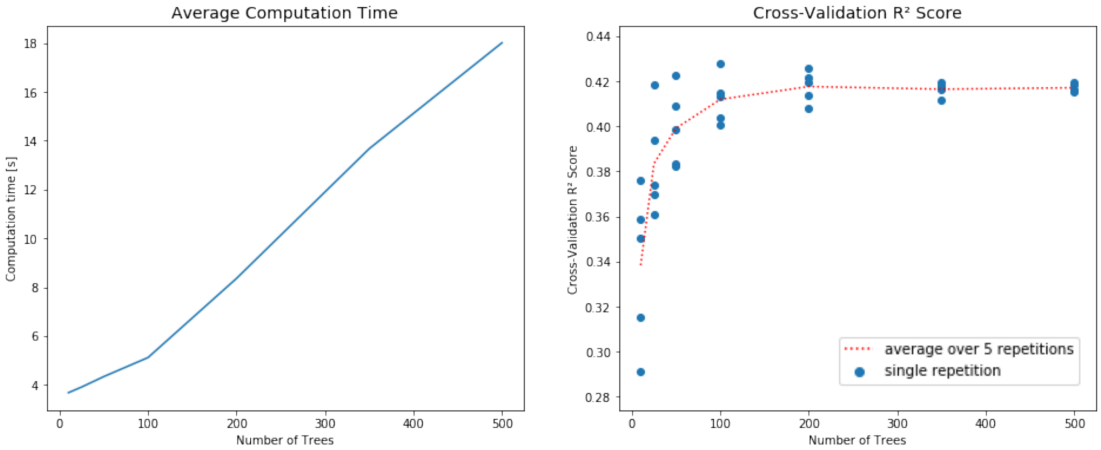



Figure 4.11: [left] Computation time needed for training and predicting of a single random forest model in function of $B$ [right] Prediction accuracy of the same random forest model in function of $B$

## 4.5 Model features

Each model gets a collection of time series from the predictor variables for training and prediction. Rather than simply using the raw variables, a large number of high-level features are constructed from them as proposed in Papagiannopoulou et al. (2017b). The construction of extreme events is not included in this work, due to the small available time span of the data and the limited scope of this thesis. Variables that have no data or only zeros are removed for that pixel.

### 4.5.1 Time series decomposition

The relevant time series that remain are put through the times series decomposition procedure described at the beginning of Section 4.1. The seasonal cycle is not smoothed this time because some variables can change rapidly in magnitude. The

result is a trend $y_t^T$ (Eq. 4.1) and a non-smoothed seasonal cycle $y_t^{SM}$ (Eq. 4.4) for each variable. The trend is then substracted from the raw time series to produce the detrended time serie: $y_t^{dT} = y_t - y_t^T$. The anomalies are produced by substracting both the trend and seasonal cycle from the raw time series: $y_t^A = y_t - y_t^T - y_t^{SM}$. This allows for extracting useful information from the raw data and it lowers the cross-correlation between predictors.

For every variable, the raw, trend, detrended, seasonal and anomaly time series are available as possible features. The trend time series are monotonous linear functions. They all offer the same information to the random forest and the ridge regression models, hence they are all functionally equivalent. Those time series are dicarded and replaced by one linear function with a positive slope and one linear function with a negative slope, to offer the same information. The raw time series contains mostly the same information as the detrended time series, but it is less stationary. For this reason, the raw time series are also discarded. Thus, from every time series we have a detrended time series $y_t^{dT}$, a seasonal time series $y_t^{SM}$ and an anomalies time series $y_t^A$. This results in a maximum of 33 predictor time series per pixel.

### 4.5.2 Lagged variables

Vegetation responds to most perturbations with a certain delay in time, called a lag. This encourages the inclusion of lagged variables in the analysis, which is done in most vegetation-climate studies of this scale (see chapter 2). The formulation of a lagged variable $x$ at lag $p$ is very straightforward: $x_t^p = x_{t-p}$ with $p$ ranging from 1 to $P$. Lagged versions of the 33 time series components are constructed, as well as lagged versions of the NDVI anomalies: $y_t^p = y_{t-p}$

What is the maximum lag $P$ that has to be incorporated in the analysis? How far back in time do climate variables have a *unique* influence on vegetation, in such a way that they cause anomalies? To truly identify the optimal maximum lag, one should perform another Granger causality analysis. In such an analysis, the different lagged versions of the same variable are all considered to be different variables. The reduced model contains all lagged variables up to a small lag value $P-1$, and the full model contains all lagged variables up to $P$. If the full model outperforms the reduced model, the lags up to P should be included in the model. Different possible values are tested and the best performing value is chozen for the maximum lag $P$. This procedure requires multiple repetitions of a time-intensive calculation and is thus considered to be out of scope for this thesis. It was reported in Papagiannopoulou et al. (2017b) that including lagged variables of more than six months lag no longer improved model predictions. Thus, the maximum lag adopted in this study is set to twelve (two lags per month).

A maximum of 441 features are obtained per pixel (33 original + 34 $*$ 12 lagged versions) for the full model.

### 4.5.3 Cumulative variables

Vegetation is resilient against short disruptions, and often prolonged climate anomalies are needed to induce a strong response on vegetation. Thus, vegetation anomalies don't always reflect the influence of a single lagged variable, but rather the cumulative impact of that variable during the preceding months. It is concluded in Ji and Peters (2003) that cumulative moisture variables up to three months are more succesful in predicting the vegetation response to drought than the single month variables. A cumulative variable of lag $k$ is constructed as the sum of all lags smaller or equal to $k$. For the NDVI anomalies, the present timestep is not included since it is the response variable $y_t^k = \sum_{p=1}^{k} y_{t-p}$. For other variables, the cumulative variable includes the present timestep: $x_t^k = \sum_{p=0}^{k} x_{t-p}$. The maximal lag $K$ is set to 12, which is the same as for ordinary lagged variables. However, not all cumulative variables with $k$ ranging from 1 to 12 are included. When $k$ becomes higher, the difference between consecutive cumulatives $x_t^k$ and $x_t^{k+1}$ becomes smaller. To prevent the abundance of highly correlated features, the cumulative variables with $k$ = 6, 8, 10 and 11 are not constructed. In total 8 cumulative variables are constructed for the 33 time series components ($k$ = 1,2,3,4,5,7,9,12), together with 7 cumulative variables for the NDVI anomalies (the cumulative variable at $k$ = 1 is identical to the lagged variable at $p$ = 1 and is not included). In the end, a maximum of 712 features are available to the full model per pixel.

CHAPTER 5

# RESULTS AND DISCUSSION

The general methods have been explained and the three research questions can now be tackled. Is water the most important climate driver for vegetation anomalies globally? What is the impact of the new variables on the analysis? What is the influence of a higher spatial and temporal resolution on the Granger Causality Framework, compared to previous work of Papagiannopoulou et al. (2017a)? First, some general results are presented for the entire analysis. Then, the results for each research question are graphically presented and discussed.

## 5.1 General results

Before further analyses, it is important to determine which machine learning model should be used. To do so, the predictive performance ($R^2$) of random forest and ridge regression models are presented in Figure 5.1. This comparison is based on the full models, as they contain all possible features constructed from all variables (see Section 4.2), which is the most complex setting for prediction. The difference in prediction accuracy between ridge regression and random forest models is also shown. For the ridge regression model, the distribution of the used $\lambda$ terms falls nicely within the given range of possible values (see Figure B.1 in appendix B). This is an indicator for correct hyperparameter tuning.

From the figures on prediction accuracy, it is clear that the random forest models outperform the ridge regression models on a global scale. This was also reported in Papagiannopoulou et al. (2017b) and all further analyses are thus performed with the random forest models. The random forest model performs well in mostly the same regions as reported in Papagiannopoulou et al. (2017b).

The results for the baseline model are presented in Figure 5.2. The difference in $R^2$ score between full and baseline model is used as a measure for the total Granger causality (see Figure 5.3). In theory, only positive Granger causality is possible: the full model has all the same features available as the baseline model, combined with many additional features. Even if all additional features are useless for prediction, the

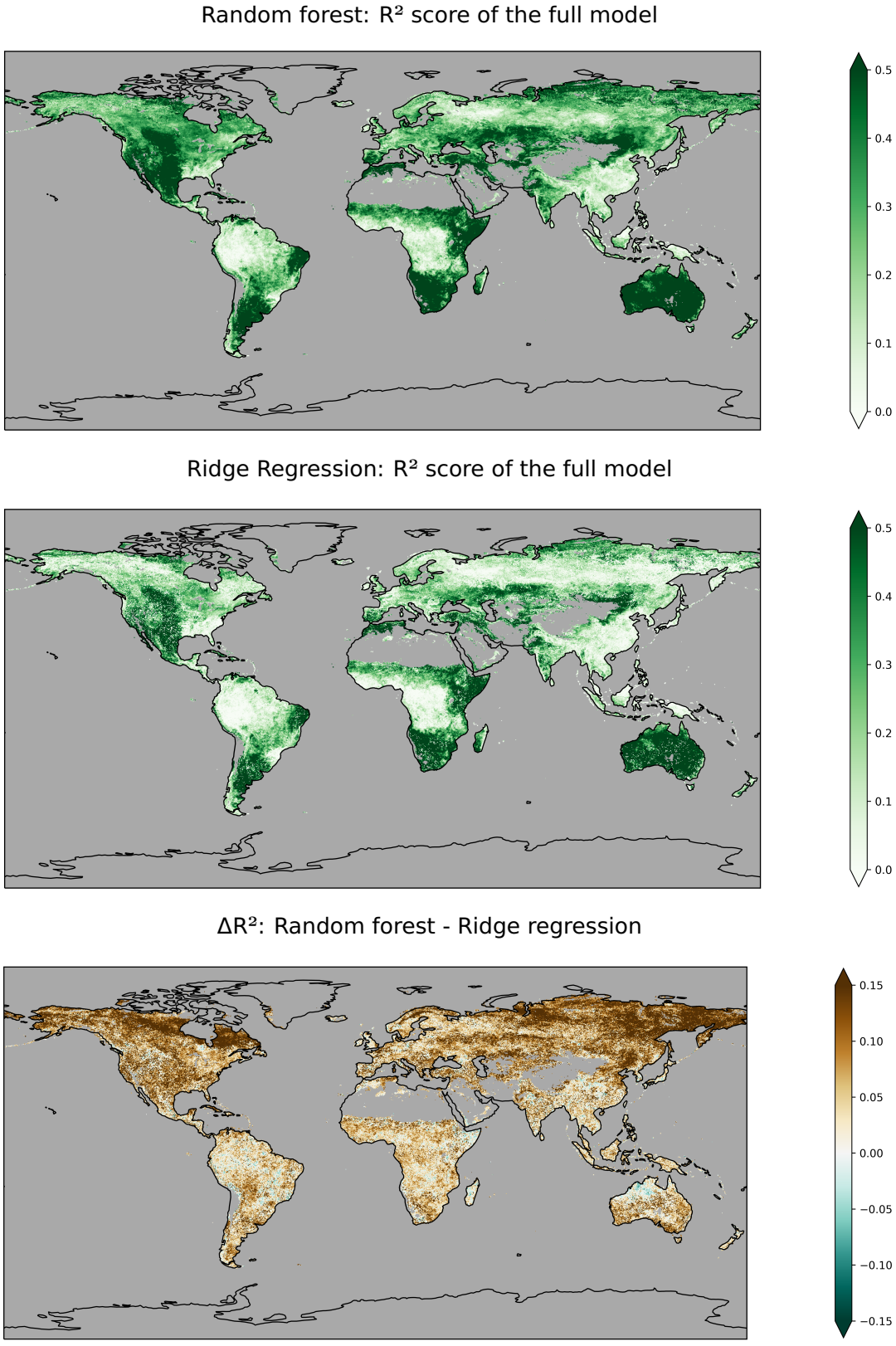



Figure 5.1: Predictive performance of Ridge regression vs Random forest in terms of $R^2$ score [Top] Random forest full model $R^2$ score [Middle] Ridge regression full model $R^2$ score [Bottom] Difference in $R^2$ score of the full models: random forest - ridge regression

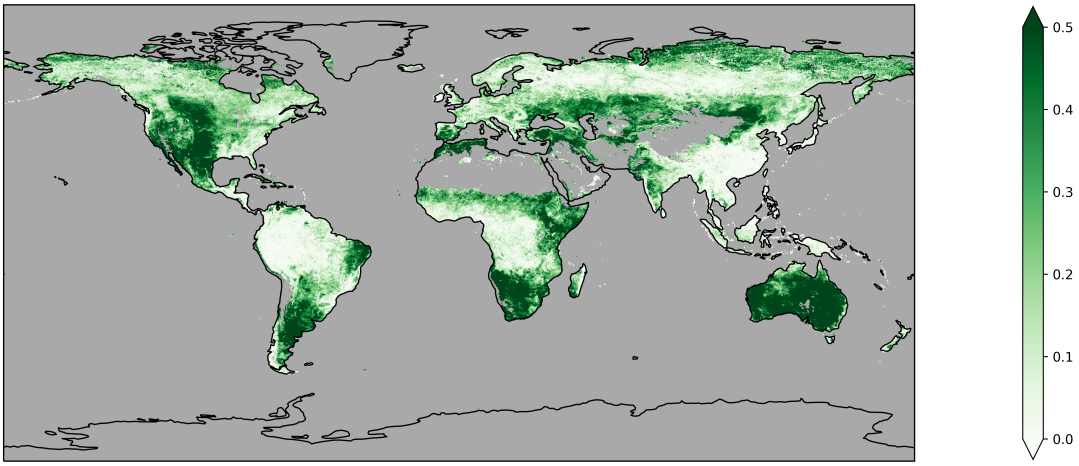



Figure 5.2: Predictive performance of Random forest baseline model: $R^2$ score

full model should still score equally well as the baseline model. In reality however, this is not always true. The difference in $R^2$ score between the full and baseline model can be negative. This happens when the climate features in the full model have very low predictive power for that pixel. A high number of useless variables are present and it becomes harder for the full model to find the few useful features that are present. The full model has to work a in high-dimensional setting and some overfitting is almost inevitable, while the baseline model is used in a low-dimensional setting. As such, the performance can drop below that of the baseline model (for more details see Section 5.3.2).

The full model outperforms the baseline model in most regions of the world, which validates the Granger causality analysis as a whole. The total Granger causality is similar to that reported in Papagiannopoulou et al. (2017a) for most of the world, with higher values in much of the Northern latitudes of North America and Asia. This total Granger causality serves as a reference for further Figures on partial Granger causality. This partial Granger causality only describes unique information present in that variable (group), it does not account for the shared information between variables. This shared information is part of the total Granger causality, so the sum of all partial Granger causalities is lower then the total Granger causality.

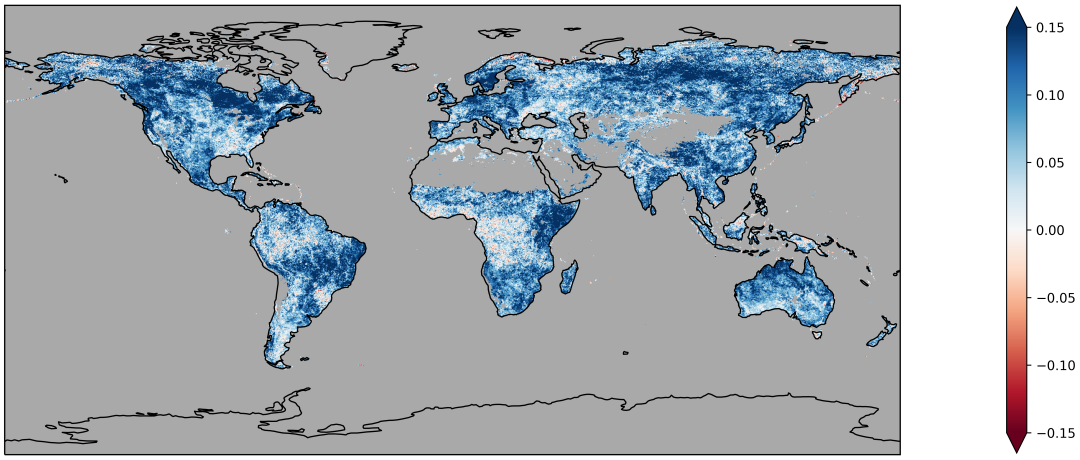



Figure 5.3: The difference in $R^2$ score between full and baseline model, used as a measure for the total Granger causality

## 5.2 Impact of the water availability group

In the work of Papagiannopoulou et al. (2017a), the partial Granger causality of three main climate driver groups on vegetation anomalies is researched. Temperature, radiation and water availability are ranked based on their partial Granger causality. The main conclusion is that water availability is the most important climate driver that Granger causes NDVI anomalies worldwide. It is reported that 61% of the global vegetated area is primarily controlled by water availability. Temperature and radiation are the primary climatic controls in 23% and 15% of the global vegetated area, respectively. The same analysis from Papagiannopoulou et al. (2017a) is now repeated with new data in an attempt to validate or nuance those findings.

The first step in ranking the different climate driver groups is calculating the partial Granger causality of each group in the vegetation anomalies. A reduced model is constructed for the water availability that uses all features except those related to precipitation, soil moisture and snow water. The $R^2$ score of the reduced model is then substracted from the $R^2$ score of the full model. This is the partial Granger causality of the water group, as shown in Figure 5.4. This procedure is repeated for the temperature and radiation group, to quantify their respective partial Granger causality. In regions where the considered group is useless for predicting the NDVI anomalies, the reduced model will perform slightly better then the full model and the reported Granger causality is negative.

It is clear that the water group provides the strongest Granger causality at a global scale. Regions with tropical or monsoonal rainforest vegetation (based on the Köppen–Geiger climate classification) coincide with regions of low Granger causality, as

can be observed in Figure B.2 in Appendix B. This is as expected, since these climates are characterized by abundant precipitation throughout the year. Tropical rainforest climate has no dry season and all months have an average precipitation value of at least 60 mm. Monsoonal rainforest climate does have a driest month with rainfall less than 60 mm. This month still contains more than $1/25^{th}$ of the total annual precipitation, by definition (Kottek et al., 2006). When using the global land cover map described in Chen et al. (2015), it becomes clear that many of the areas that have low Granger causality for water availability are forested areas or cultivated land (see Figure B.3 in Appendix B). This is related to the fact that trees generally have access to deeper layer soil moisture. The root-zone soil moisture is included in this model by means of the GLEAM dataset (see Section 3.1). Forests grow mostly in non-arid regions, as arid land forests and wooded lands only count for 6% of the total forest area (Malagnoux, 2007). This ultimately results in most of the global forests not being limitated by water availability, which is in accordance with the results of Boisvenue and Running (2006) and Seddon et al. (2016). As a side-note, it should be mentioned that it is hard to get a complete image of the actual soil-moisture content in dense forests through remote sensing. This is especially true for rain-forests due to the obscuring effect of the constant cloud cover, as could be observed from Figure 3.7. Thus, the impact of the water group may be underestimated in these regions.

Radiation shows partial Granger causality mainly in South America, Southeast Asia and the Philippines. This corresponds to the forested areas that are not water-driven. These same areas have been reported in Boisvenue and Running (2006) for the potential limitation of available sunlight to net primary production. The temperature group only shows limited partial Granger causality, mostly in Europe and near the East Coast of North America. The impact of temperature is significantly lower than was reported in Papagiannopoulou et al. (2017b).

The three main variable groups can be ranked per pixel based on their Granger causality. For every variable group, its respective place in the ranking is determined per pixel and plotted in Figure 5.5. If a group shows no total Granger causality in that pixel, no ranking is shown. The conclusions from the partial Granger causality Figures are reinforced here, with the water availability showing the best overal rankings. The radiation group ranks highest in most of the forested areas, where the water group has poor Granger causality. The temperature group mostly shows highest rankings in Europe and near the East Coast of North america. This high ranking in Europe may be linked to the increased length of heatwaves over Western Europe as reported by Della-Marta et al. (2007). During the extreme heatwave of 2003 (Russo et al., 2015), the strong reduction in vegetation net primary production has been reported as driven by temperature rather then rainfall deficit for Western Europe (Ciais et al., 2005).

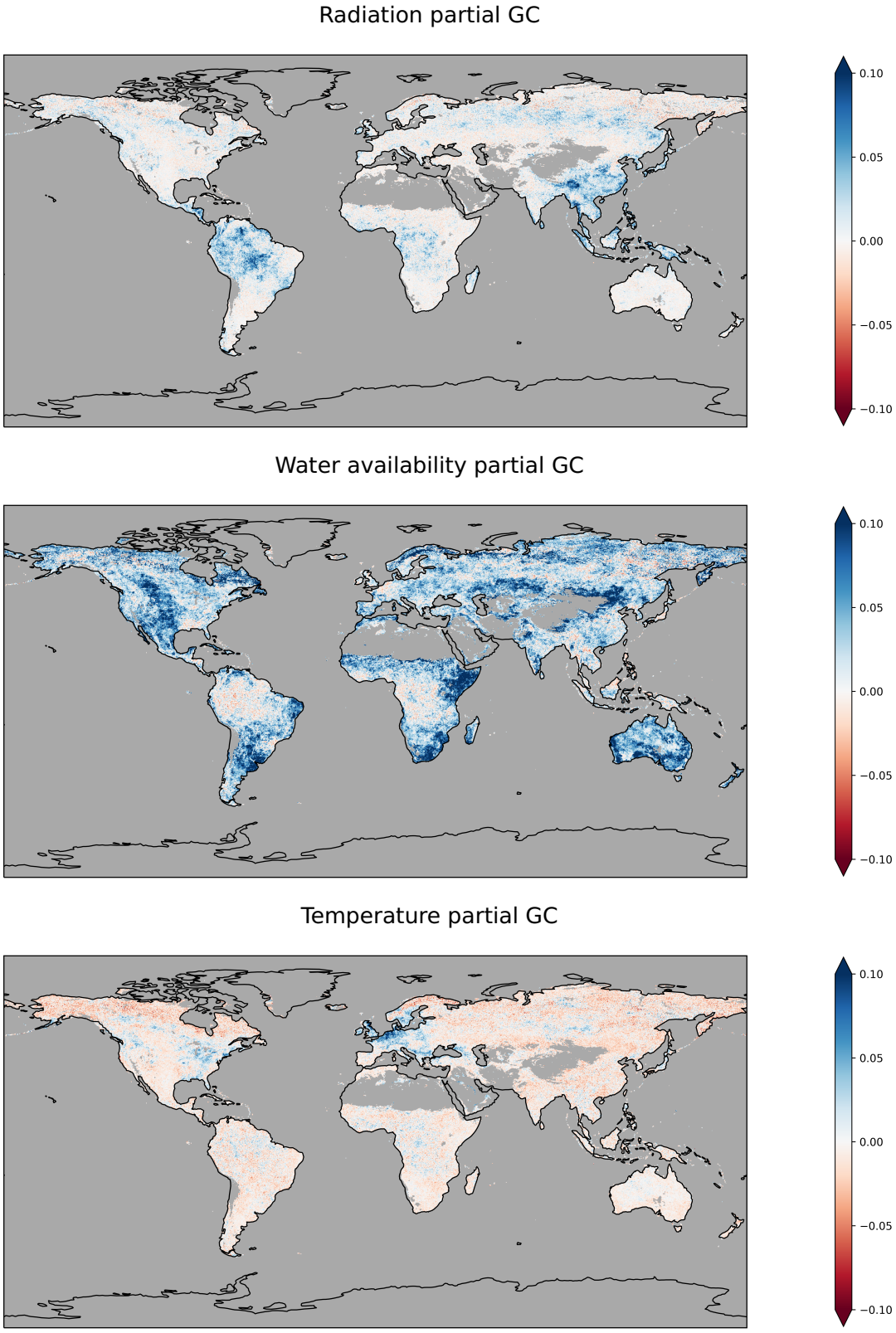
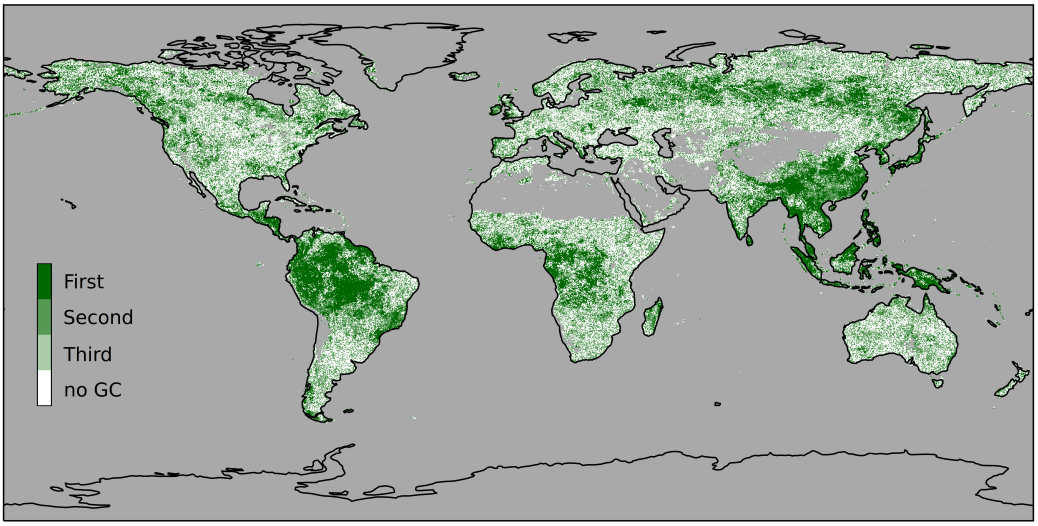
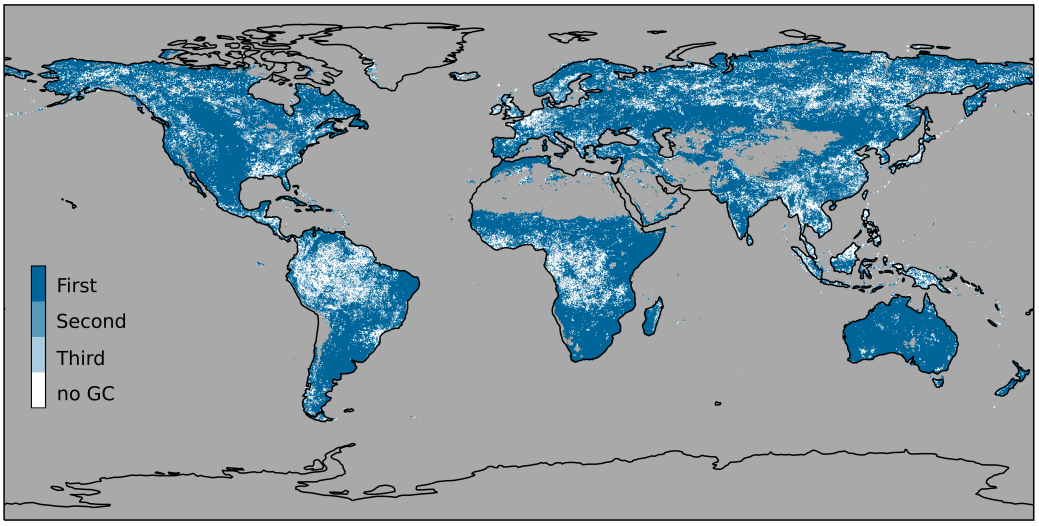



Figure 5.4: partial Granger causality (GC) of the three main groups, expressed as the difference in $R^2$ score between the full and reduced models

Radiation rank




Water availability rank




Temperature rank

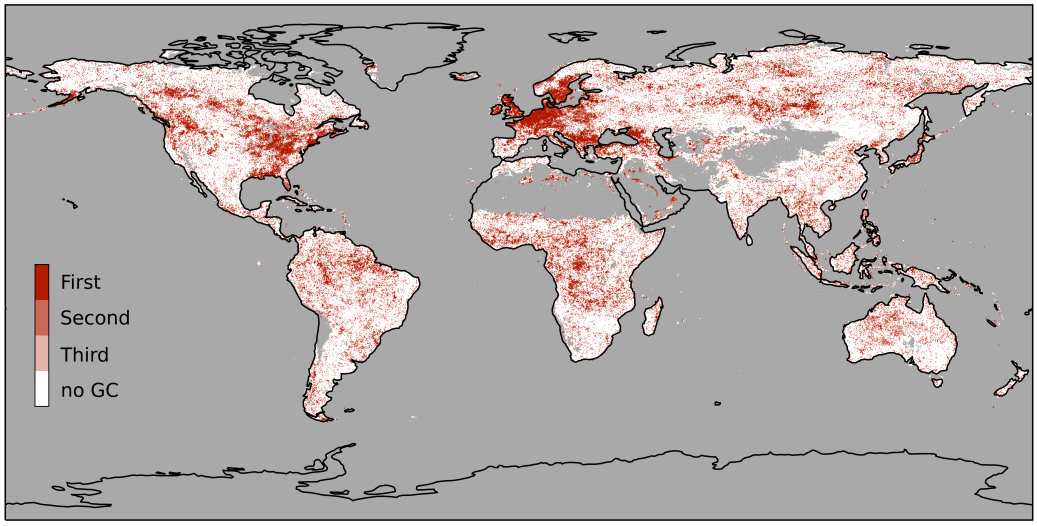



Figure 5.5: Ranking of the three main group in terms of Granger causality. The "no GC" label is used if that group shows no Granger causality for that pixel

A global map of the group with the highest Granger causality per pixel is shown in Figure 5.6. This Figure is similar to the reported results in Seddon et al. (2016) (see Figure A.3 in Appendix A) with the cloudiness replaced by radiation, as both are linked (Nemani et al., 2003). A clear resemblance exists to the global land cover map reported in Chen et al. (2015), suggesting a strong causal link between the dominant climate driver and the vegetation type. In Figure 5.6, the water availability group is the dominant group in terms of partial Granger causality, with 64% of the vegetated land being mainly water-driven. The radiation and temperature group are the dominant group in 18% and 8% of the vegetated land. No Granger causality is measured in 10% of the vegetated land. If the pixels without Granger causality are not taken into account, the relative percentages for water, radiation and temperature are 71, 20 and 9%, respectively.

In the construction of Figure 5.6, the only requirement for the inclusion of a pixel is that the total Granger causality is larger than zero. Even if the full model outperforms the baseline with only the slightest margin, the results are included and they have the same weight over the final outcome as regions with very strong Granger causality. An attempt is made to isolate the results that are based on stronger Granger causality. A minimum total Granger causality is required, and the resulting new global ranking of the three groups is shown in Figure B.4 in Appendix B. If a pixel does not meet this requirement, it is not shown. The aforementioned relative percentages of 71, 20 and 9% for water, radiation and temperature stay relatively stable over a wide range of cut-offs. The water availability percentage rises to 73%, the radiation percentage drops to 18% and the temperature percentage stays the same. The global division of the dominant climate drivers holds up, even if only strong Granger causality is taken into account.

In the global ranking, the temperature group shows the lowest control over vegetation anomalies, which is not in accordance with the results of Papagiannopoulou et al. (2017a). The primary control percentages reported were 61, 15 and 23% for water, radiation and temperature, respectively. Large areas in the Northern latitudes are no longer reported to be controlled by temperature in this thesis. Most of those regions are now reported to be controlled by radiation instead. This shift in control can be linked to an actual decreased impact of temperature in these Northern latitudes. However, it is also possible that this shift in control is not linked to reality, as Granger causality does not equal true causality. A range of factors can influence the conclusions: missing data, new data, hidden common causes and confounding variables can all influence the results. In further Sections, some of these possibilities are tested for their possible influence on the analysis. It turns out that part of the high Granger causality for temperature in previous work is related to the fact that the VPD variable

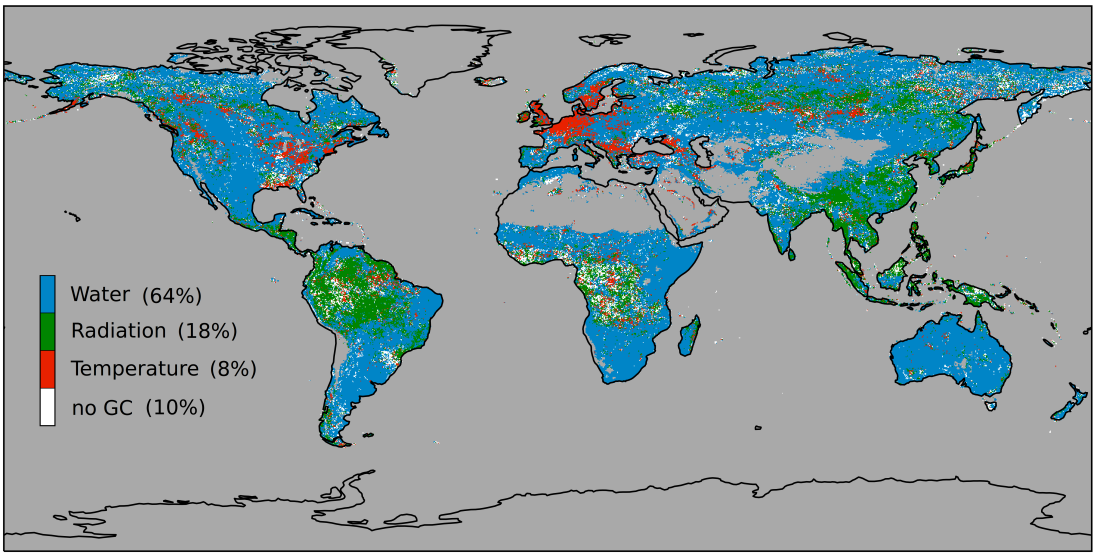



Figure 5.6: Group with the highest Granger causality for every pixel. The "no GC" label is given for pixels that have no total Granger causality

was not included in that database (see Section 5.3.3). Another factor that has an influence on the low reported temperature Granger causality in this study is the high resolution (see Section 5.4.3).

From Figures 5.5 and 5.6, the main conclusion from Papagiannopoulou et al. (2017a) is confirmed: global vegetation anomalies are primary controlled by the water availability. The areas previously shown to be water-driven remain primarily water-controlled in this study. For completeness, it must be mentioned that one factor can cause an artificial preference in the random forest models towards the water group in my thesis. This factor is the unbalanced number of variables, as every variable is represented by a single dataset (except for soil moisture). The water group has 5 raw datasets, that incorporate both remote sensing and other data sources (see Section 3.1). The temperature group has only 3 raw datasets, that are often highly correlated. The radiation group only contains data from a single source. This can cause problems, because a single dataset can be inaccurate for a certain region or period in time. Using only one data source per variable is related to the limited scope of this thesis, but it may be a source of bias. The unbalanced variable groups also present another possible source of bias. Due to the nature of random forest models, only a small subset of variables are available for each split (a maximum of 27 features in this setting). On average, almost half of the variables in these subsets will be related to the water availibility. Thus, if every variable has only one relevant lag-time, the water group simply has a better chance of being incorporated in any given random split.

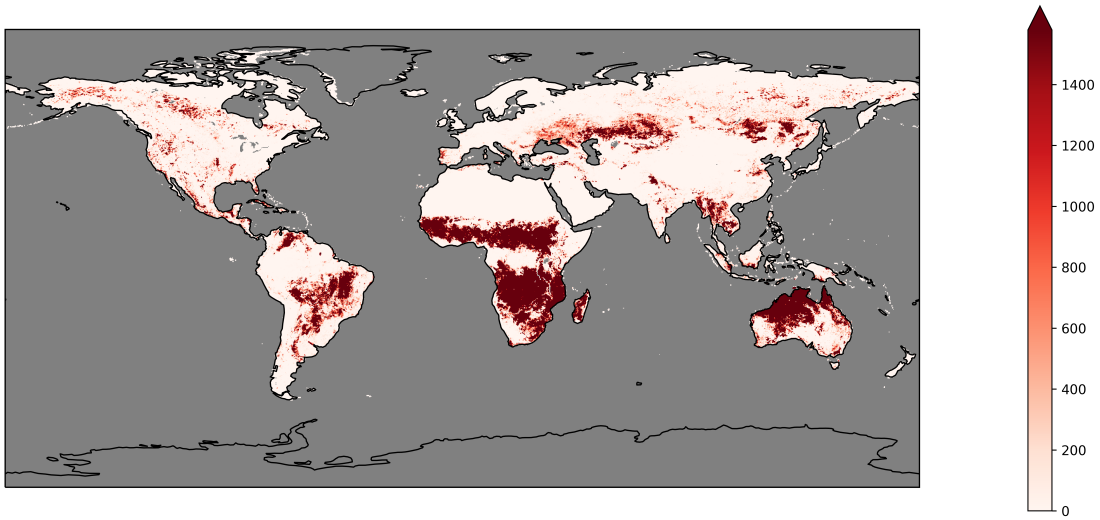



Figure 5.7: Cumulative burned area over time

## 5.3 Impact of the new variables

In this dataset, three variables are included that were not present in the research of Papagiannopoulou et al. (2017b): wildfire, VPD and irrigation. Since they are new to the analysis, it might be interesting to test these variables for their respective Granger causality, to see if they are useful for further analyses. This is done by constructing a seperate reduced model for each variable and calculating the decrease in predictive performance compared to the full model.

### 5.3.1 Wildfire

Wildfire is a climate variable that does not belong to any of the three main groups (radiation, temperature and water availability). To be fully correct, it is a variable group of its own. If the wildfire data on burned area is summed up over time per pixel, the most heavily impacted regions become clearly visible (see Figure 5.7). In the most intensely burned regions, more than 50% of the vegetated area burns on average every year (see Figure B.5 in Appendix B). It is a reasonable assumption that the amount of burned area will show strong Granger causality in these heavily affected regions.

This assumption does not hold however, as it can be seen from Figure 5.8. Most regions show only very limited Granger causality. The strongest Granger causality is present in the Northern latitudes and South America. However, even in these regions, the Granger causality is low compared to that of the three main groups. If the wildfire variable is incorporated in the ranking procedure from Figure 5.5, it is never the first or even second Granger causality factor (Figure not included).

The areas with strongest Granger causality are not really associated to the areas with the highest cumulative burned area over time. This is a counterintuitive conclusion, as much research suggests a strong impact of large-scale wildfires on vegetation (Lu and He, 2014; Puig-Gironès et al., 2017; Wakeling et al., 2012; Bond et al., 2005). A partial explanation for this phenomenon might be related to the vegetation type. The areas where a large percentage of the vegetated area burns on an annual basis are mostly grasslands and shrublands (see Figure B.3 in Appendix B). It is known that these vegetation types have many different adaptations to enable a quick recovery after a large wildfire (Mares, 2017; Lu and He, 2014; Wakeling et al., 2012). Furthermore, an important part of the impact that frequent wildfires have in these areas is that they hold back forest vegetation (Wakeling et al., 2012; Bond et al., 2005). This influence is not really detectable with this setup, as it does not investigate causal influences on the average raw NDVI over time. The only visible correlation that the areas with stronger Granger causality have to the GFED wildfire database is in another variable, that was not included in this study: the average fuel consumption. The Northern areas with stronger Granger causality also show a high average fuel consumption per square meter (Figure B.5 in Appendix B). This high fuel consumption is related to the vegetation type, as forests have more available biomass for the fire to burn. It might be that the burned area alone is just a bad predictor for how harsh wildfires impact vegetation. The burned area is a consequence of vegetation destruction, rather than a causal factor.

Another possible reason for the absence of strong Granger causality in Africa is the large amounts of smoke coming from the fires. These would be detectable through remote sensing and might interfere with the measurements for other variables used by the models. If this were to happen in such a way that the smoke disturbance were detectable in that data to the random forest models, the occurence of wildfires would no longer be unique information. This is however considered unlikely due to the advanced state of remote sensing measurements and processing of the data before publishing.

### 5.3.2 Irrigation

Irrigation is not a climate variable, but it is included in the analysis as it could be a hidden common cause or confounding variable. For example, it might be so that positive anomalies in the soil moisture data coincide with the application of irrigation. The partial Granger causality is shown in Figure 5.8. No areas with clear partial Granger causality are visible. There are some variations visible, but these variations are scattered and they don't prove anything. Due to the random nature of the random

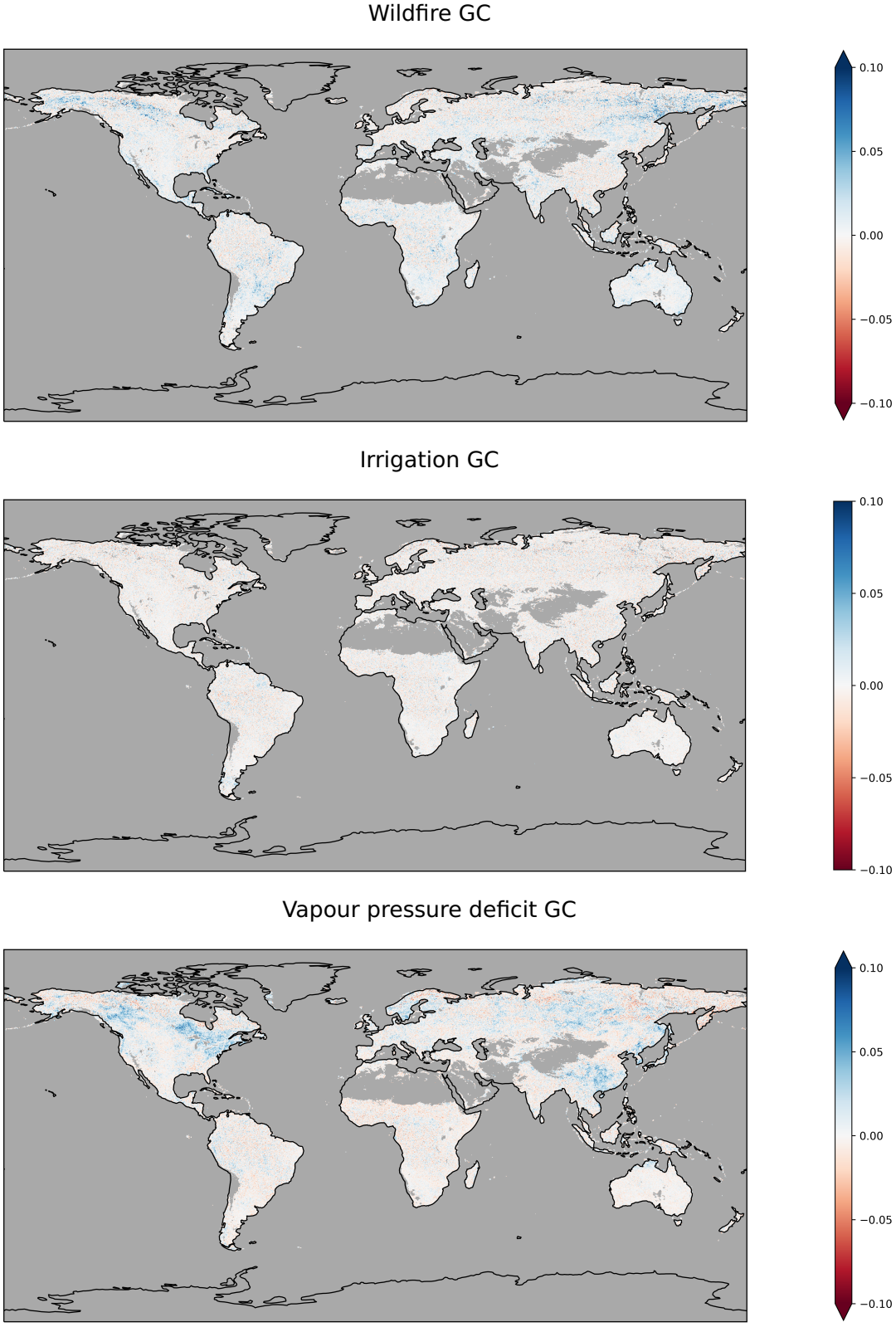



Figure 5.8: partial Granger causality for wildfire, irrigation and VPD

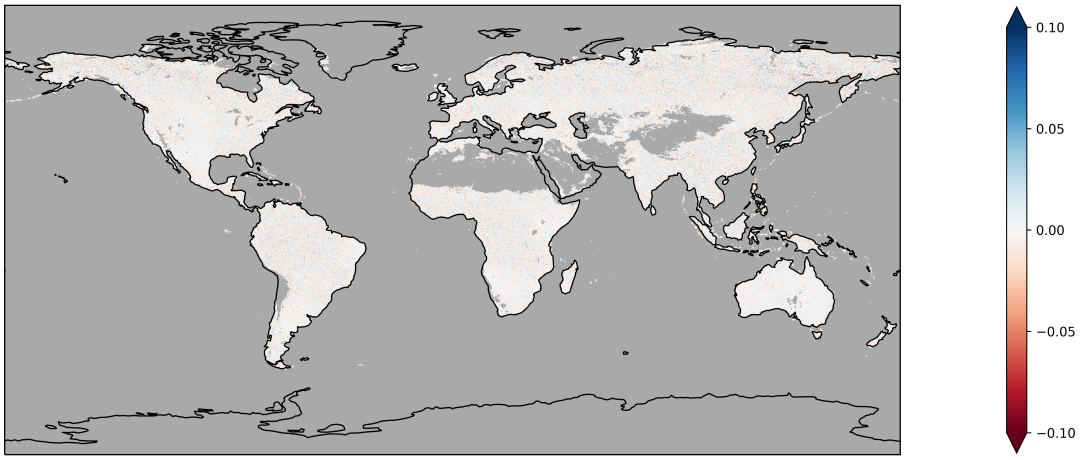



Figure 5.9: Difference in $R^2$ score between two repetitons of full model (no GC). This is used as a baseline to compare low partial Granger causality to.

forest models, such variations are also visible in the difference in $R^2$ score between two repetitions with the same full model 5.9.

In order to test for low partial Granger causality in scattered pixels, a different approach is adopted. The frequency distribution of the Granger causality values from all pixels is plotted in Figure 5.10 for irrigation. If Granger causality is present, the distribution should lean towards more positive values in theory. This is not really the case, and the mean of the distribution is only $2 * 10^{-4}$. One might conclude from this that no significant Granger causality is present, and that the addition of the irrigation variable is the same as adding a randomly created meaningless variable to the model. This is however not the case. A very large number of features are offered to the full random forest model (max. 712) and only few will have strong predictive power over the NDVI anomalies. If one were to add a useless variable, its time series components and all high-level features, it becomes harder for the model to find the stronger predictors amongst the now larger pile of weak predictors. The cost of adding a useless variable to the features is not zero. This cost is simulated by performing a partial Granger causality analysis for a random variable. The performance of the full model with the random variable (and its high-level features) is compared to the performance of the full model without the random variable. This difference in $R^2$ score is used as a measure for what the partial Granger causality of a truly useless variable would be. The frequency distribution for this difference in $R^2$ score is shown in Figure 5.10 and it is clearly negative on average.

The fact that these distributions are so spread out is also related to the random nature of the models. The difference between the $R^2$ score of two repetitions of the same full model is also shown in Figure 5.10 (labeled no GC), as a reference for the variability

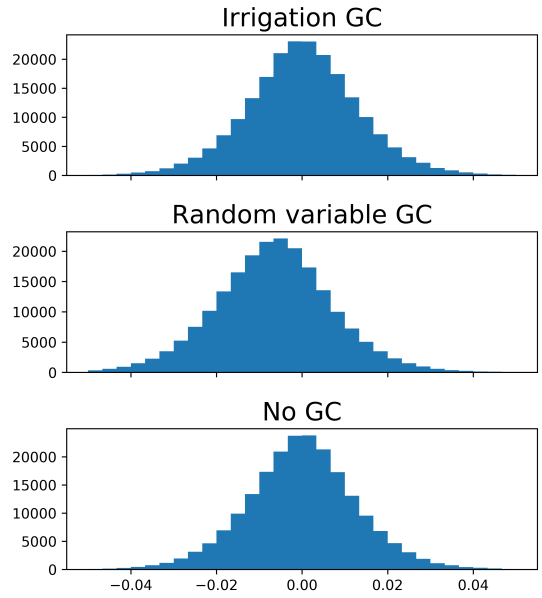



Figure 5.10: Frequency distribution of the Granger causality (GC) values: [Top] GC for irrigation [center] GC for a random variable [Bottom] Difference between two repetitons of full model (no GC)

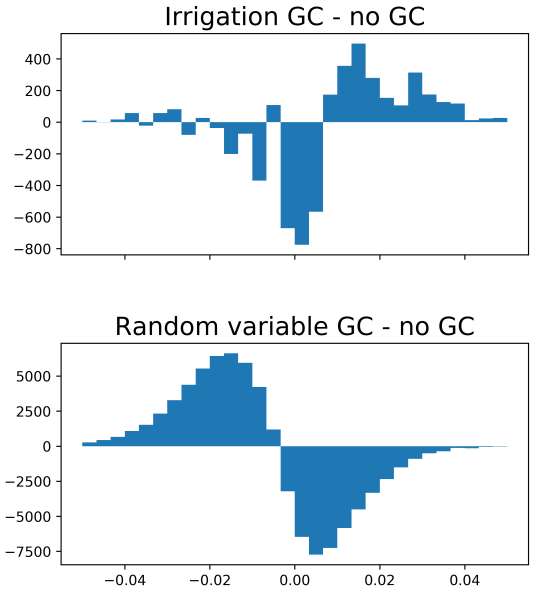



Figure 5.11: Difference between two frequency distributions: [Top] Irrigation Granger causality - no Granger causality [Bottom] Random variable Granger causality - No Granger causlaity

that affects all results. It is now possible to assess whether the distribution of irrigation Granger causality is more positive than the distribution of the difference between two repetitions in the full model (no GC). The difference between these two distributions is calculated and shown in Figure 5.11. For reference, the difference between the distributions of a random variable GC and no GC is also calculated and shown. These Figures show that more positive Granger causality values are present in the irrigation distribution than in the no GC distribution, while quite the opposite is true for a random variable. The Mann–Whitney U test can be used to test for the statisical significance of this conclusion (Mann and Whitney, 1947). The null hypothesis of this test is that no variable is stochastically larger then the other and the one-sided alternative hypothesis is that the irrigation partial Granger causality is stochastically larger then the difference in $R^2$ score between two repetitions of the same full model (no GC). The resulting p-value of $2.5 * 10^{-8}$ suggests that the alternative hypothesis is correct and that partial Granger causality exists for irrigation. The combined evidence of Figure 5.11 and the Mann–Whitney U test proves that irrigation does have partial Granger causality over NDVI anomalies. However, this partial Granger causality is limited and not really noticable on a global scale.

This leads one to wonder why irrigation shows so little Granger causality, when we know it is applied to succesfully prevent vegetation withering? This is likely related to the nature of partial Granger causality, which only accounts for the unique information

that a variable (group) holds over the response variable. In this case, the near-surface soil moisture is incorporated in both the reduced model and full model. Part of the information present in the irrigation data is likely also present in the near-surface soil moisture data. As such, the reported Granger causality for irrigation may be lower then the actual causality. Another factor that may play a role is the spatial resolution of the study. A timeseries for NDVI represents the average NDVI anomalies over 775 km² (near the equator). The application of irrigation may be too localized to induce a noticable response in the average anomalies over 775 km², for many pixels. The data quality may also have an impact on the outcome, as it is derived from modelling efforts on a monthly resolution, instead of direct remote sensing at a bi-weekly resolution as for most other sources (see Section 3.1).

Another possible explanation for the low Granger causality is related to the anthropogenic nature of irrigation. Irrigation is applied to promote plant growth and prevent drought stress. This makes it fundamentally different from the climate variables. Large application of irrigation is often an induced response to periods of lower water availability, with the specific goal to prevent vegetation anomalies. If irrigation is applied in that manner over the entire period for a pixel, periods of drought will produce less negative NDVI anomalies, compared to a neighbouring pixel without irrigation. It would not be possible to trace this causality back to irrigation, since there is no period without irrigation to compare to. This is a weakness originating from the simplification procedure described in Section 4.3. By isolating every pixel, all spatial information is lost and it becomes impossible to quantify causality arising from spatial differences.

Even though the irrigation variable might not show much partial Granger causality, it is still advised to include the variable in future analyses. It may contain important shared information with variables from the water group, thus preventing an inflation of the partial Granger causality of that group.

### 5.3.3 Vapour pressure deficit

The VPD is a measure for air drought that is included in the water group. The partial Granger causality of the VPD is shown in Figure 5.8. The observable partial Granger causality suggests that new information is present in the VPD data that was not detectable in the other (water) variables. Mainly forests in Northern latitudes and South-East Asia show Granger causality for the VPD. If the resulting NDVI anomalies are mainly negative and they are related to positive VPD anomalies, VPD is a measure for the impact of drought on Northern Hemisphere forest vegetation. This remains speculation as the sign of the explained anomalies is not known.

The inclusion of VPD as a water variable may be part of the reason for the low Granger causality of the temperature group. As mentioned in Section 3.1, the VPD is calculated as the difference between the saturated vapour pressure and the actual vapour pressure. The actual vapour pressure is the mass of air moisture per volume and is measured via remote sensing. The saturated vapour pressure is a function of the air temperature ($°C$) and is calculated as $P_v^{sat} = 611 * exp\left[\frac{19.65 * T_a}{T_a + 273}\right]$. It is likely that some of the previously reported Granger causality for the temperature group (Papagiannopoulou et al., 2017a) is no longer unique information for that group, due to the addition of the VPD variable. This would result in a lower partial Granger causality for temperature.

To test this hypothesis, the analysis of Section 5.2 is repeated. In this scenario however, the VPD variable is removed from all models, as if it were not used in this study. The temperature ranking in the models without VPD is higher in areas of North America and continental Asia, which are areas with a high Granger causality for the VPD (Figure not included). The map of the first ranking variable group is also remade. The new ranking without VPD is shown in Figure 5.12, the old correct ranking in Figure 5.6. It is clear that the dominance of the water group has dropped in favor of both the temperature and radiation group. The absolute percentages have gone from 64, 18 and 8% to 58, 19 and 13% for water, radiation and temperature. Some of the information about the NDVI anomalies held in the temperature group is thus not unique, and also available in the VPD time series. This is logical, the VPD is strongly linked to temperature by definition and it was calculated (partly) from the temperature data. It is thus likely that the VPD was a confounding variable in the results of (Papagiannopoulou et al., 2017b), as the variable was not included there. This would account for part of the reason why the temperature partial Granger causality was reported higher in those previous papers.

It could be argued that VPD represents the combined influence of the water group and temperature group rather than just the water group, as it is calculated from the vapour pressure (a measure for air water content) and temperature. This could be tested by replacing the VPD with a another water variable such as air moisture content or actual vapour pressure. This is considered out of scope for this work and could be investigated in future analyses.

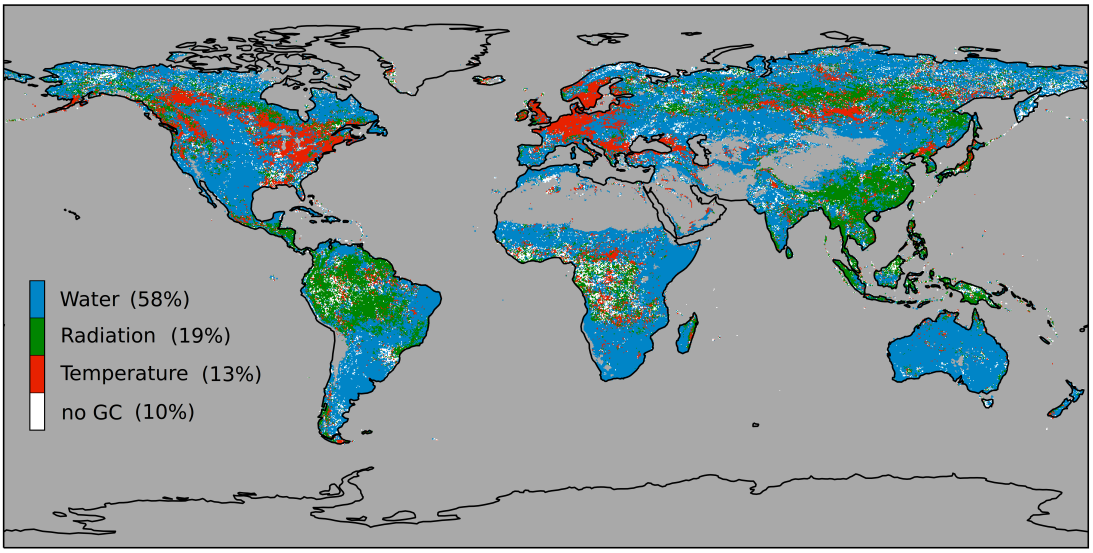



Figure 5.12: Ranking of the three main group in terms of Granger causality when the VPD variable is removed from all models. The "no GC" label is given to pixels with no total Granger causality

## 5.4 Impact of a higher spatial and temporal resolution

The data used in this study has a higher spatial and temporal resolution than the data from Papagiannopoulou et al. (2017b). The effect of the individual temporal and spatial upscaling steps is discussed. The effect of a combined upscaling is also investigated. The different resolutions may allow for the use of different techniques, such as blocked cross-validation and the exclusion of lag zero from the models. Both options are evaluated on their possibility and usefulness for future studies.

### 5.4.1 Impact of higher spatial resolution

The spatial resolution has gone from 1° x 1° in the work of Papagiannopoulou et al. (2017b) to 0.25° x 0.25° in this thesis. This means that what used to be a single pixel is now divided into sixteen different pixels. This makes it easier to pick up on local patterns, which could improve prediction accuracy. The increase in spatial resolution comes at a cost, as the datacubes become much larger. Older personal computers can show memory problems for simple mathematical operations on these datacubes. Due to the extreme parallelization described in Section 4.3, the associated increase in computation time is limited. If the problem were to be tackled in a more complex way (e.g. per region instead of per pixel), such extreme parallelization is not possible. A random forest model can become very slow in a regional setting (Papagiannopoulou

et al., 2018). It is investigated whether the higher resolution results in better performance of the models.

As a way to test for the effect of this higher resolution, the basic Granger causality analysis is performed again with a lower spatial resolution dataset. All available data is downsampled to a spatial resolution of 1° x 1° (the temporal resolution is still bi-weekly). Each new pixel contains sixteen old pixels. If less then eigth of these were included in the original analysis, the pixel is not used. The difference in $R^2$ score between the high and low resolution models is calculated for the full and baseline model : 0.25° x 0.25° - 1° x 1°. The results are presented in Figure 5.13. These values corresponds to the effect of moving from a lower to a higher spatial resolution: if the difference is positive, the higher resolution results in better predictive performance. It can be concluded that the baseline model becomes a bit worse on average, but regions with improvements in $R^2$ scores are also visible. The predictive performance of the full model drops more on average, which results in a lower total Granger causality, as is shown in the bottom of the Figure. The relation between the difference in total Grange causality (bottom of Figure) and the difference in $R^2$ score of the baseline and full models (top and middle of figure) is as follows:

$$\begin{aligned} \Delta GC_{1,2} &= GC_1 - GC_2 = \left(R^2_{full1} - R^2_{base1}\right) - \left(R^2_{full2} - R^2_{base2}\right) \\ &= \left(R^2_{full1} - R^2_{full2}\right) - \left(R^2_{base1} - R^2_{base2}\right) = \Delta R_{1,2full} - \Delta R_{1,2base} \end{aligned} \tag{5.1}$$

This lower predictive performance is a bit counterintuitive: one would expect that the availability of finer resolution data would make both models stronger, as more detailed anomalies can be explained. This effect is probably present in some regions, but another effect plays a more important role. The 1° x 1° data is an average-pooled representation of the higher resolution data, which removes some of the noise that was present. By lowering the random variance in the data, it becomes easier for the models to detect actual patterns of prediction (Costanza and Maxwell, 1994). This effect seems to dominate over the effect of the additional information and moving to a higher spatial resolution makes both the predictive performance and the total Granger causality lower.

### 5.4.2 Impact of higher temporal resolution

The temporal resolution has increased from monthly to bi-weekly between this thesis and the work of Papagiannopoulou et al. (2017a). This (combined with the spatial increase in resolution) comes at direct cost. The common timespan of the used variables only goes back to 2003 in such a high resolution. This means that only 13 years

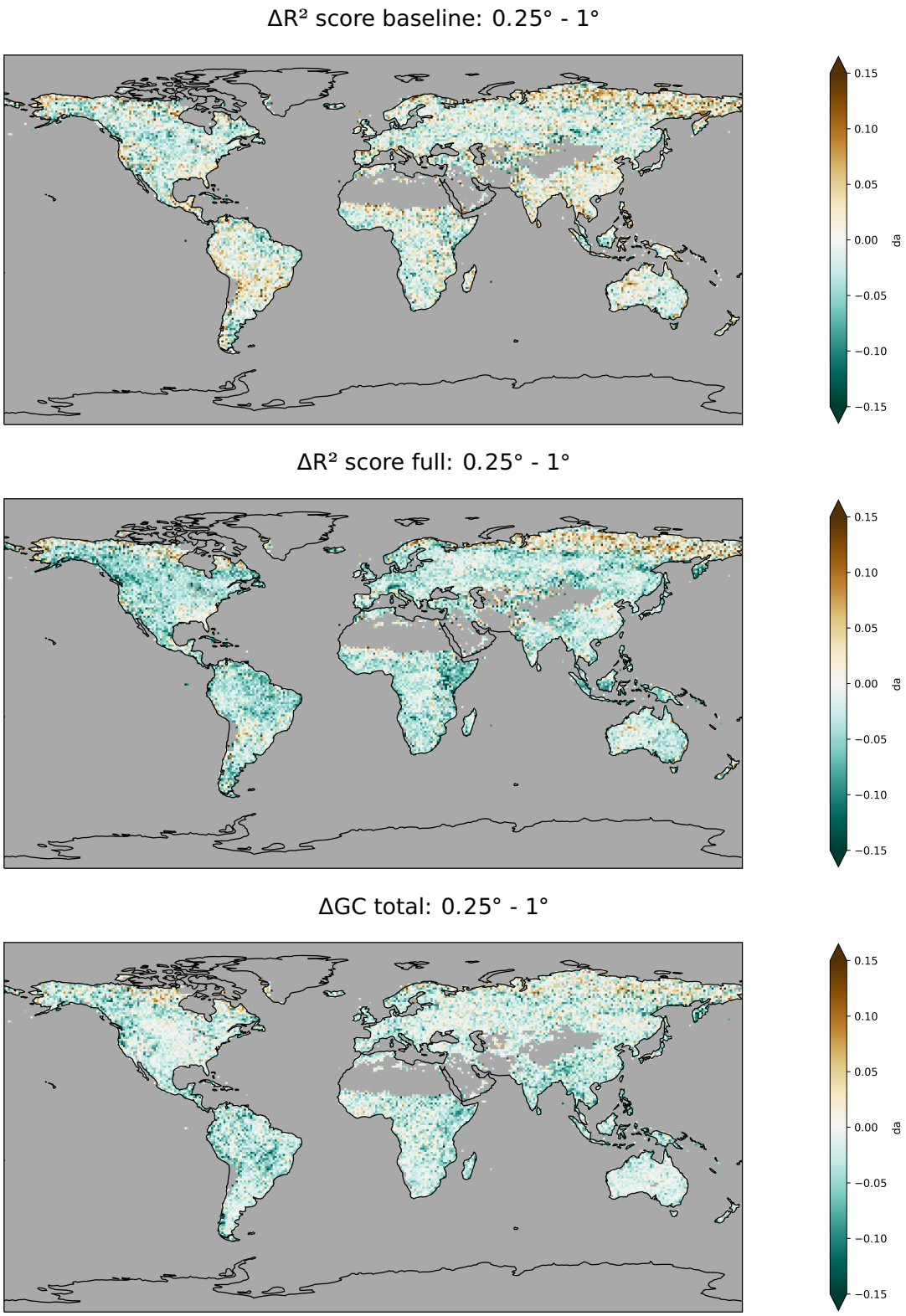



Figure 5.13: Effect of a higher spatial resolution: difference between 0.25° x 0.25° bi-weekly and 1° x 1° bi-weekly resolutions [Top] Difference in $R^2$ score between baseline models [Middle] Difference in $R^2$ score between full models [Bottom] Difference in total Granger causality (GC)

of data are available for this study, compared to 30 years for Papagiannopoulou et al. (2017b). As a consequence, it becomes harder to identify the "normal" state of NDVI when strong anomalies are present over multiple years. Long term NDVI anomalies may also be partly removed in the extraction of the seasonal cycle, which is not desirable. In this study, no climate extremes were defined. If this were the case, the short timespan would make it harder to define these extremes, as they are infrequent by definition. By doubling the temporal resolution, the amount of features also roughly doubles for every time series.

The higher temporal resolution likely has a strong impact on the predictive performance of the random forest models. Due to the shorter time period between lags, the autocorrelation of the NDVI anomalies at lag 1 becomes higher. This makes the predictive performance of all models stronger. However, if the predictive information in the anomalies at lag one is related to the influence of climate variables at earlier lags, the associated increase in $R^2$ will not be that high for the full model as this information was allready available at a low resolution. The full model gets additional features for the climate variables at lag one. This does not necessarily result in a large increase in predictive performance for the full model, as this information was already partly available to the monthly model in the form of the lag zero climate features. This effect is absent from baseline models which don't have lag zero features. All the information at lag one is new to the baseline model, as it was not used for prediction in a monthly resolution. The increase in temporal resolution might actually result in a lower total reported Granger causality. The effect of the higher resolution is investigated on the performance of the models and the reported Granger causality.

This effect is investigated in a similar manner as for the spatial resolution. The basic analysis is repeated with a monthly datacube, that was downsampled in time (the spatial resolution is still 0.25° x 0.25°). The resulting difference in prediction accuracy is shown for the baseline and full models in Figure 5.14. The baseline model shows a very strong increase in predictive performance, on a near global scale. The full model also shows an increased predictive performance in most of the world, but not as much as the baseline. The combined effect on the total Granger causality is also visible in Figure 5.14. Due to the lower increase in performance of the full model compared to the baseline, the total Granger causality decreases for most of the earth. This is as anticipated.

### 5.4.3 Combined impact

The upscaling that happened between the work of Papagiannopoulou et al. (2017b) and my work is a combined upscaling in both time and space. An isolated upscaling

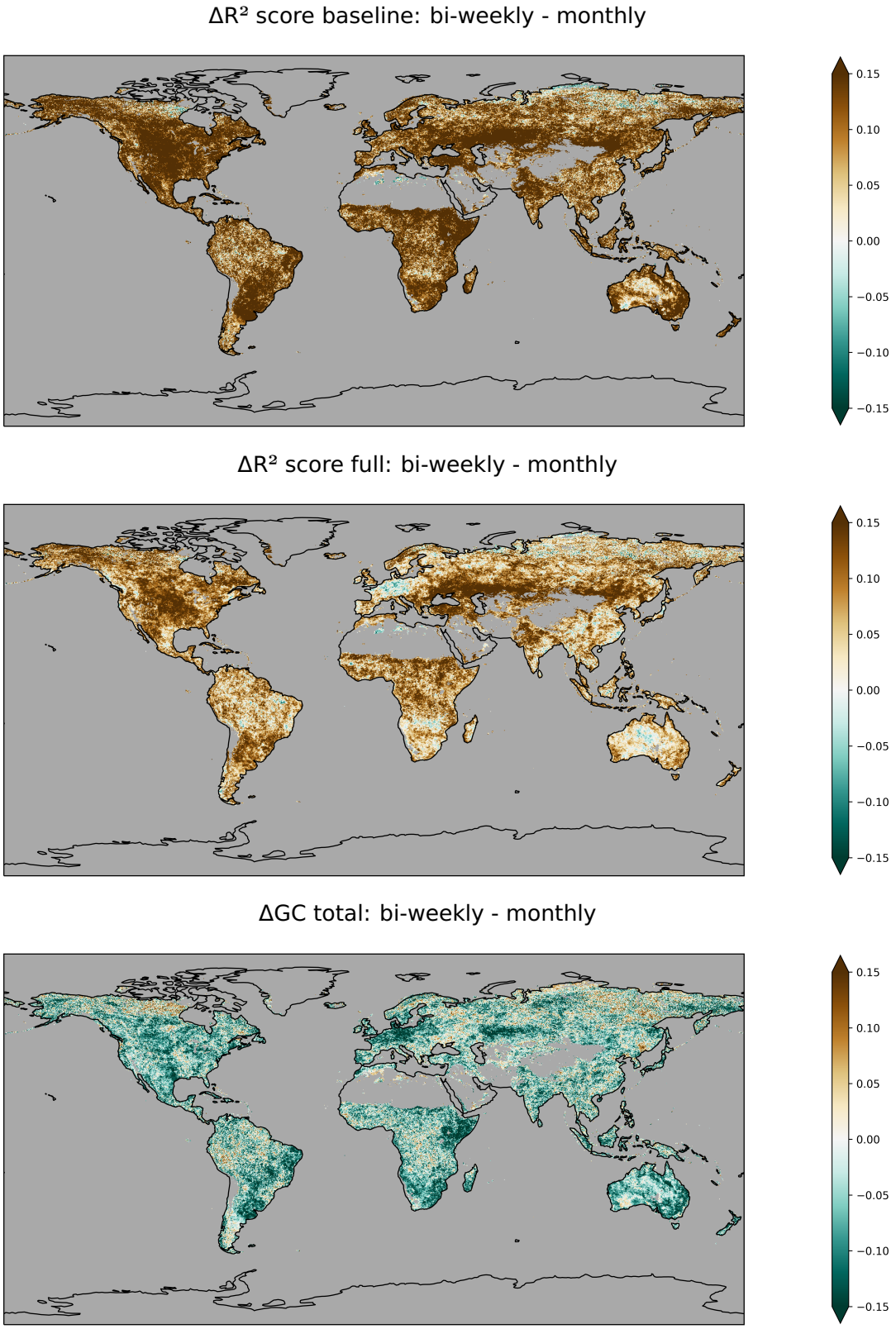



Figure 5.14: Effect of a higher temporal resolution: difference between 0.25° x 0.25° bi-weekly and 0.25° x .25° monthly resolutions [Top] Difference in $R^2$ score between baseline models [Middle] Difference in $R^2$ score between full models [Bottom] Difference in total Granger causality (GC)

in either spatial or temporal resolution results in a near global drop in total Granger causality. The drop from the spatial upscaling is related to a general decrease in the predictive performance of individual models while the drop related to the temporal upscaling resulted from a general increase in model performance. This different impact makes it harder to estimate what the effect is of both upscaling steps combined on the total Granger causality. One could intuitively assume that the effect of the combined upscaling equals the combined effects of each individual upscaling. If this were the case, the combined higher resolution would result in a much lower total Granger causality.

Another basic analysis is performed to investigate the effect of the combined upscaling. The used dataset is downscaled both spatially and temporally, to the original 1° x 1° monthly resolution of Papagiannopoulou et al. (2017b). The combined effect on the baseline model, full model and total Granger causality is shown in Figure 5.15.

It seems that the effect of a combined upscaling is indeed similar to the combined effects of the individual upscaling steps. The baseline model becomes much stronger, while the full model only partly improves. This results in a much lower total Granger causality. For this reason, the previously reported results should be checked on their consistency across the different resolutions. To this end, the main analysis is repeated with the 1° x 1° monthly dataset. The partial Granger causality of the three main groups and the three new variables is shown in Figures 5.16 and 5.17. The conclusions for both previous research questions are mostly the same in a low resolution setting. This further validates the results that were previously shown, as they are consistent between different resolution models. The predictive performance of irrigation did not visibly increase, suggesting that its low partial Granger causality was not related to the higher resolution. The global ranking of the three main variable groups is also repeated. The resulting percentages are shown in Figure 5.18.

The global ranking between the three groups is now a little more balanced. The water group is still the dominant factor in 62% of the world, which is 2% lower then before. The temperature group ranks first in 11% of the pixels, which is a 3% increase compared to the reported result in a higher resolution. Part of the lower reported importance of the temperature group in the aforementioned conclusions (compared to previous work) is partly related to the higher resolution that was used. The reported global ranking in a low resolution setting stays pretty constant over different cut-offs for the minimum required Granger causality (see Figure B.6 in Appendix B). The fact that 55% of the pixels have a total Granger causality above 0.15 again shows the large increase in total Granger causality (In the higher resolution, this was only 13%).

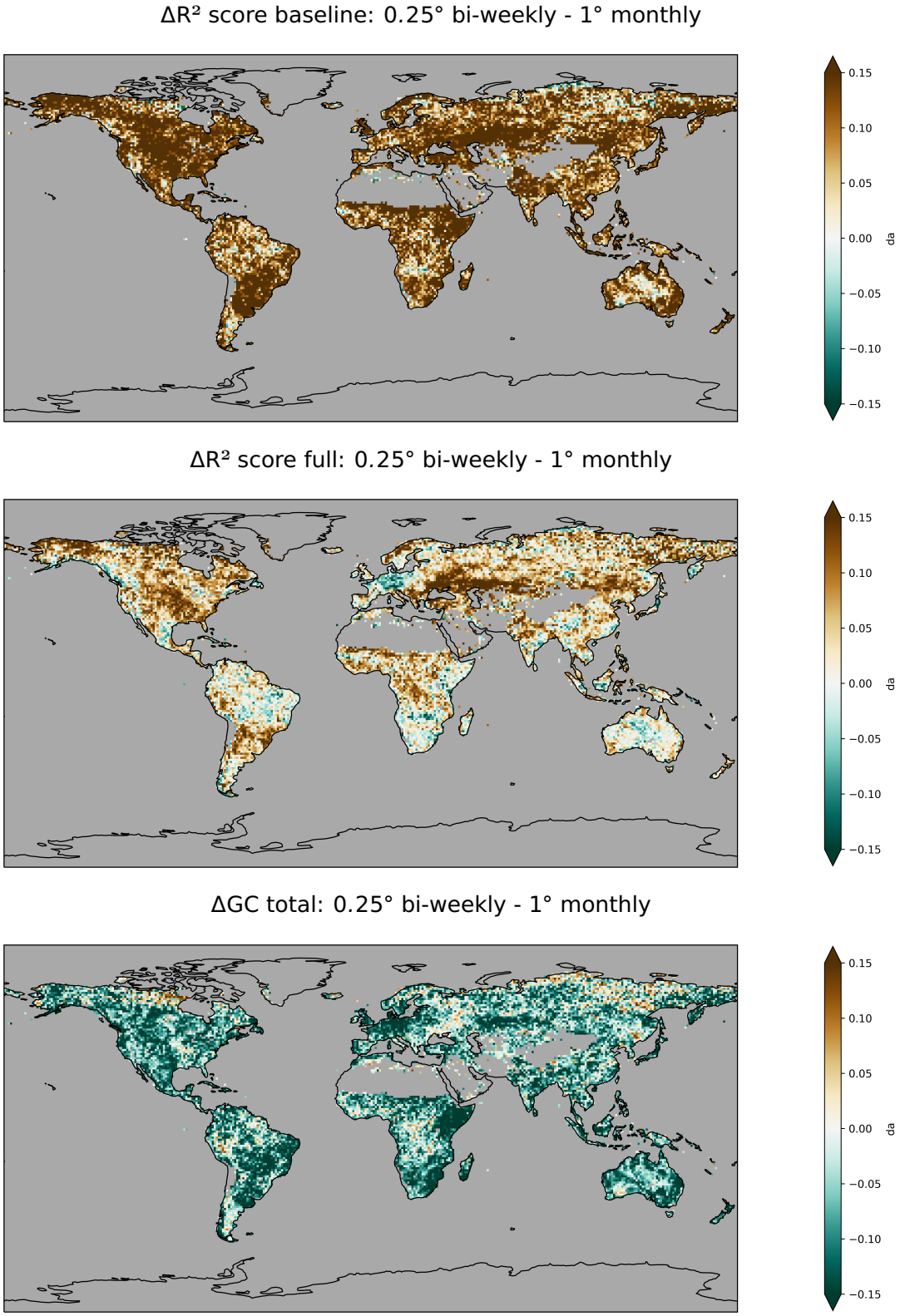



Figure 5.15: Effect of a higher combined resolution: difference between 0.25° x 0.25° bi-weekly and 1° x 1° monthly resolutions [Top] Difference in $R^2$ score between baseline models [Middle] Difference in $R^2$ score between full models [Bottom] Difference in total Granger causality (GC)

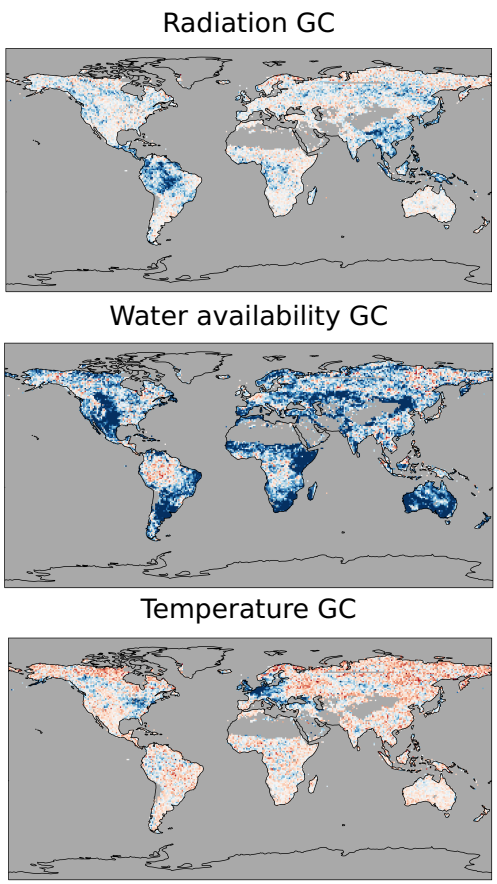



Figure 5.16: partial Granger causality (GC) of the three main groups in a 1° x 1° monthly resolution

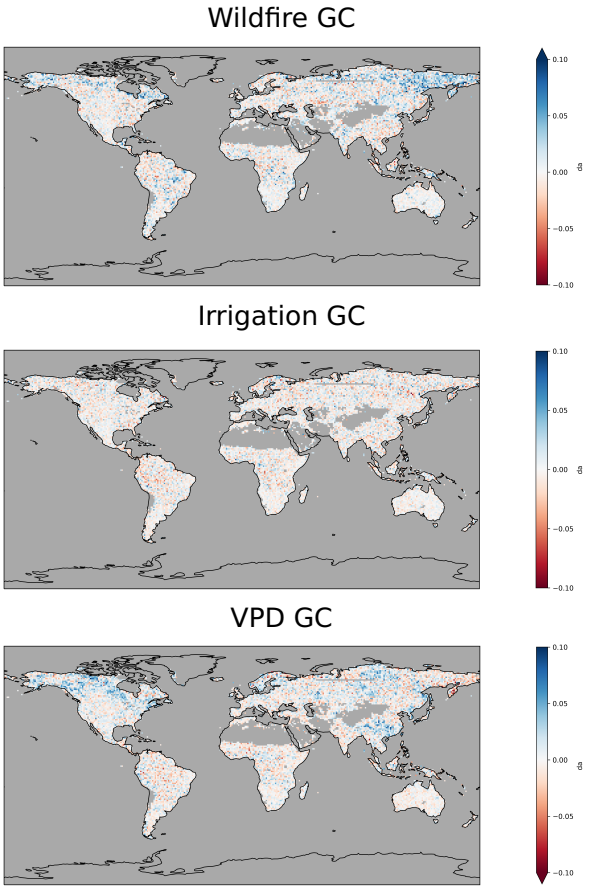



Figure 5.17: partial Granger causality of the new variables in a 1° x 1° monthly resolution

Due to the strong similarity in reported results with those of a higher resolution and the higher total Granger causality, the 1° x 1° monthly resolution is deemed superior for this framework. Additional benefits from adopting this resolution are much smaller datasets, faster calculations and a much longer available timespan of data at a lower resolution (1981 to 2015). The number of features roughly halves for all models, which is especially valuable for the full model, which performs in a high-dimensional setting. In further analyses, the term "high resolution" is used for the 0.25° x 0.25° bi-weekly resolution and the term "low resolution" for the 1° x 1° monthly resolution.

### 5.4.4 Blocked cross-validation

As stated in Section 4.2, blocked cross-validation is an alternative cross-validation scheme for time series. The data is not shuffled, instead it is split in 5 chronological blocks. Blocked cross-validation is often preferable, even in a causal inference setting (Roberts et al., 2017). However, blocked cross-validation generally comes at a loss of predictive power, especially over a short timespan. Imagine that a given time series starts with one very dry year, followed by 12 "normal" years of limited droughts. In

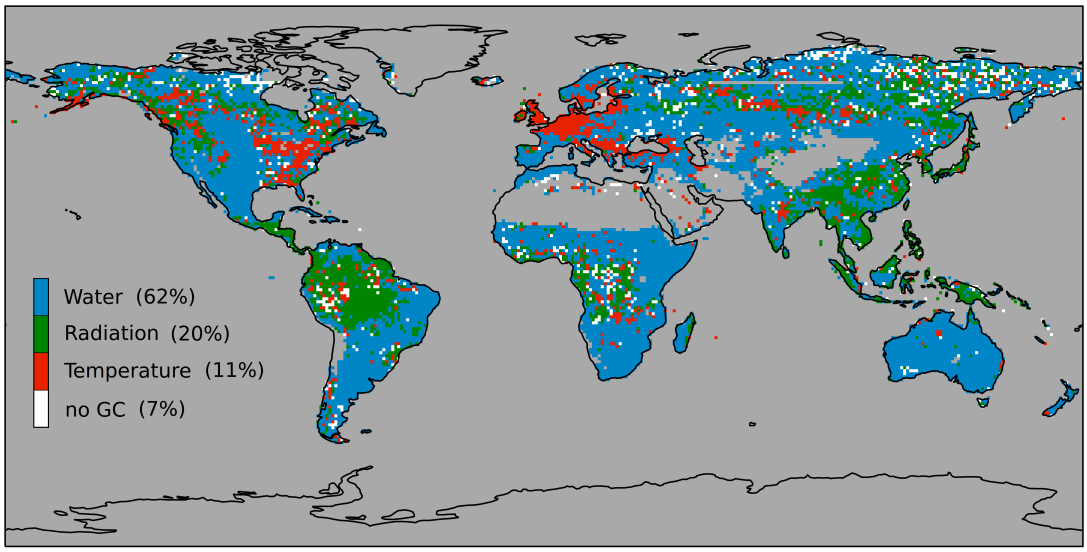



Figure 5.18: Group with the highest Granger causality for every pixel in a 1° x 1° monthly resolution. Pixels with no total Granger causality are labeled as "no GC".

that case, the dry year is completely put into the first cross-validation fold. When this fold is used for testing and the other four folds are used for training, the model does not encounter any data from very dry conditions and is not trained to know what happens to the NDVI in a very dry year. For that fold, the model will not perform well in the testing phase, resulting in low $R^2$ values. If a random cross-validation scheme is used instead, the information from the dry year will be spread more evenly over the five folds. The model can encounter some of these very dry datapoints during training and is able to better predict the anomalies for datapoints in the test set from that very dry year.

Blocked cross-validation was tested for the models at high resolution. The predictive performance of the full model becomes much lower, while the performance of the baseline model becomes a little lower (Figures not included). As a result, the total Granger causality becomes negative in nearly half of the world (see Figure 5.19). This is clear evidence that the Granger causality analysis falls apart when a blocked cross-validation scheme is introduced at a high resolution.

Blocked cross-validation was also tested for the models at low resolution. Due to the large total Granger causality that is present with random cross-validation at a low resolution, a blocked cross-validation might work in this setting. The $R^2$ score for both the baseline and full model are shown in Figure 5.20. The $R^2$ score is low in many regions for both the full and baseline model, but the full model outperforms the baseline in most of the world. The total Granger causality is shown to be strong and positive in most pixels. To investigate the potential for the future use of blocked cross-validation, the ranking procedure of the three main groups is repeated (see Figure

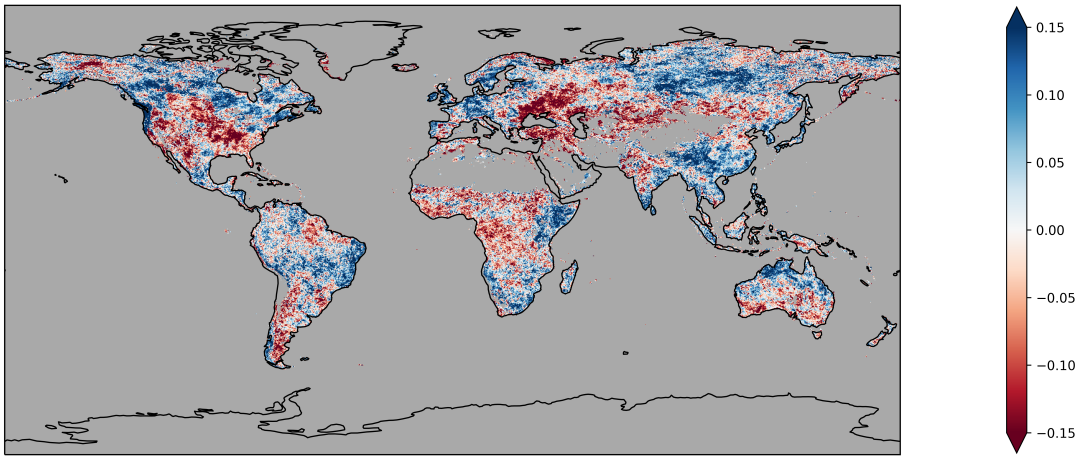



Figure 5.19: Total Granger causality with blocked cross-validation in a 0.25° x 0.25° bi-weekly (high) resolution

5.21). The amount of water-driven area is 49%, lower then what was reported from random cross-validation. This corresponds to a relative percentage of 62%, which is very close to the reported percentage in Papagiannopoulou et al. (2017a). The water and radiation groups show slightly higher relative percentages, compared to random cross-validation results. From this, it seems that blocked cross-validation produces reasonable results. As such, it is a newly available option for future research in a low resolution setting.

The main advantage of blocked cross-validation in this framework is that it gives a more true error estimates. The reported $R^2$ values from random cross-validation can be too optimistic (low), due to temporal dependencies that exist in the data. Blocked cross-validation does not scramble these dependence structures, which results in more accurate $R^2$ values in theory. However, this may not be that important in this setting, because the baseline model is allready used to control the reported $R^2$ values. Furthermore, as stated in Roberts et al. (2017): "If blocking structures follow environmental gradients, blocking may hold out entire portions of the predictor space, introducing extrapolation between cross-validation folds." No extrapolation is desired between folds in this study, as it leads to bad predictions for less common climate conditions (as illustrated in the aforementioned dry year example). The introduction of blocked cross-validation results in 21% of the pixels having no Granger causality at all (see Figure 5.21), which is a cost for the analysis. As such, no strong recommendation is given in whether to use random cross-validation or blocked cross-validation. Each researcher should evaluate personally the pros and cons of blocked cross-validation and deside whether the inclusion is useful for the analysis.

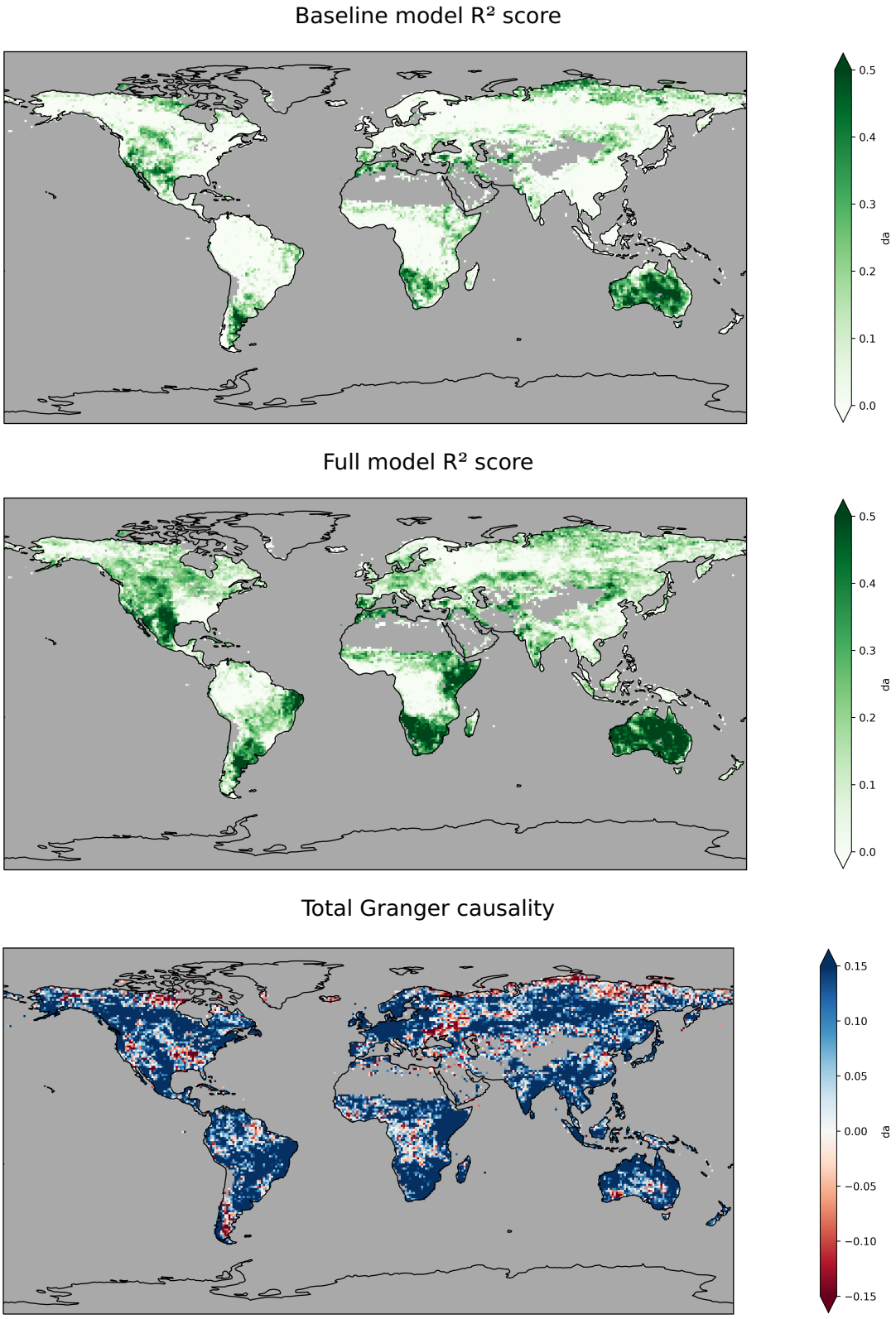



Figure 5.20: $R^2$ score of the baseline and full model and total Granger causality for blocked cross-validation in a 1° x 1° monthly (low) resolution

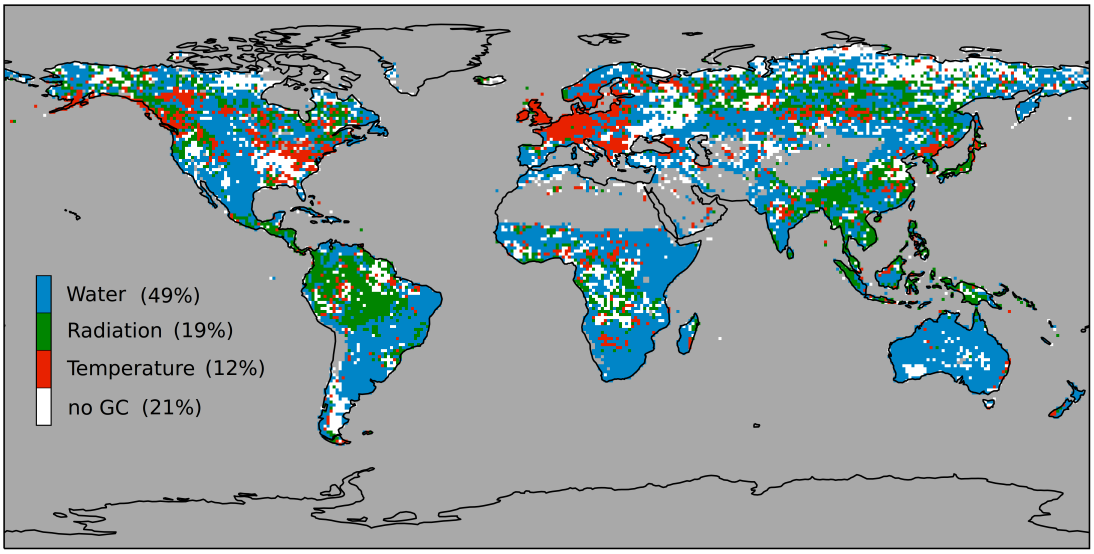



Figure 5.21: Group with the highest Granger causality for every pixel, based on blocked cross-validation in a low resolution setting. Pixels without total Granger causality are labeled as "no GC".

### 5.4.5 Exclusion of lag zero

The used framework (described in Section 4.2) is very useful for tackling the described research questions, but it is not completely in line with the definition of Granger causality. One of the main principles in Granger causality is that the cause must precede the effect. This is translated for the used framework as follows: the influence from climate variables should happen at least one timestep before the NDVI anomalies that they cause. In this framework, to predict an NDVI anomaly, climate data from the same time step is used. This data is used as features by the full and reduced models. This actually violates that principle, since both cause and effect are allowed to come from the same timestep. This approach was necessary in previous studies due to the lower temporal resolution. If all climate impacts from within the same 30 day window as the anomaly are excluded, short-lived influences on vegetation are lost. The impact of radiation is almost completely gone after that 30 day window, thus lag zero of the climate variables had to be included (Papagiannopoulou et al., 2017a). Another practical problem is that by removing lag zero, the reduced and full models lose predictive features and become weaker. The predictive power of the baseline model stays the same, as it never had a lag zero variable.

Because of the available bi-weekly data, it may now be possible to remove lag zero from the models in the high resolution setting without losing too much predictive power. The lag zero time series components are removed and all cumulative features are reconstructed as $x_t^k = \sum_{\mathbf{p=1}}^{k} x_{t-p}$. The performance for the full model turns out

to be somewhat lower with lag zero removed (Figure not included). This drop in performance is limited, however it is mostly visible in regions where the original model with lag zero included was allready showing low predictive performance. For these regions, the new full model can only barely outperform the baseline. The total Granger causality is now lower, and is even totally lost in some of those regions, as can be seen in Figure 5.22.

The effect on the Granger causality analysis for the three main groups is also investigated. The Granger causality for radiation has almost disappeared (see bottom of Figure 5.22) while the Granger causality for the temperature and water group stayed mostly the same (Figures not included). This shows that the impact of radiation on vegetation is relatively fast and short-lived. This may be due to the less destructive nature of anomalies in radiation compared to anomalies in temperature and water availability. It is also possible that those NDVI anomalies are mainly positive, related to the impact of abundant radiation on tropical rainforest vegetation, which has been reported as energy-limited on average (Williams et al., 2012; Mallick et al., 2016). The fact that temperature and water availability keep most of their Granger causality for the models with lag zero removed shows the more latent response of vegetation to those anomalies. Similar findings were reported in Papagiannopoulou et al. (2017b), where the impact of the water group shows clear Granger causality up to 3 months later. The ranking procedure for the three main groups is also repeated and the resulting map of the first rank reveals similar conclusions (see Figure 5.22). The radiation group only has first ranking in 8% of the pixels. The combination of these results shows that the higher temporal resolution is not enough to allow for the removal of lag zero from the models in the high resolution setting. This would lead to an underestimation of the global importance of radiation as a climate driver (Nemani et al., 2003; Wu et al., 2015; Seddon et al., 2016).

The removal of lag zero from the low resolution setting is not considered, as this would have an even larger impact on the full and reduced models. It is shown in Papagiannopoulou et al. (2017a) that the influence of radiation largely disappears after lag zero on a monthly resolution. In the end, the removal of lag zero is not recommended for both high and low resolutions.

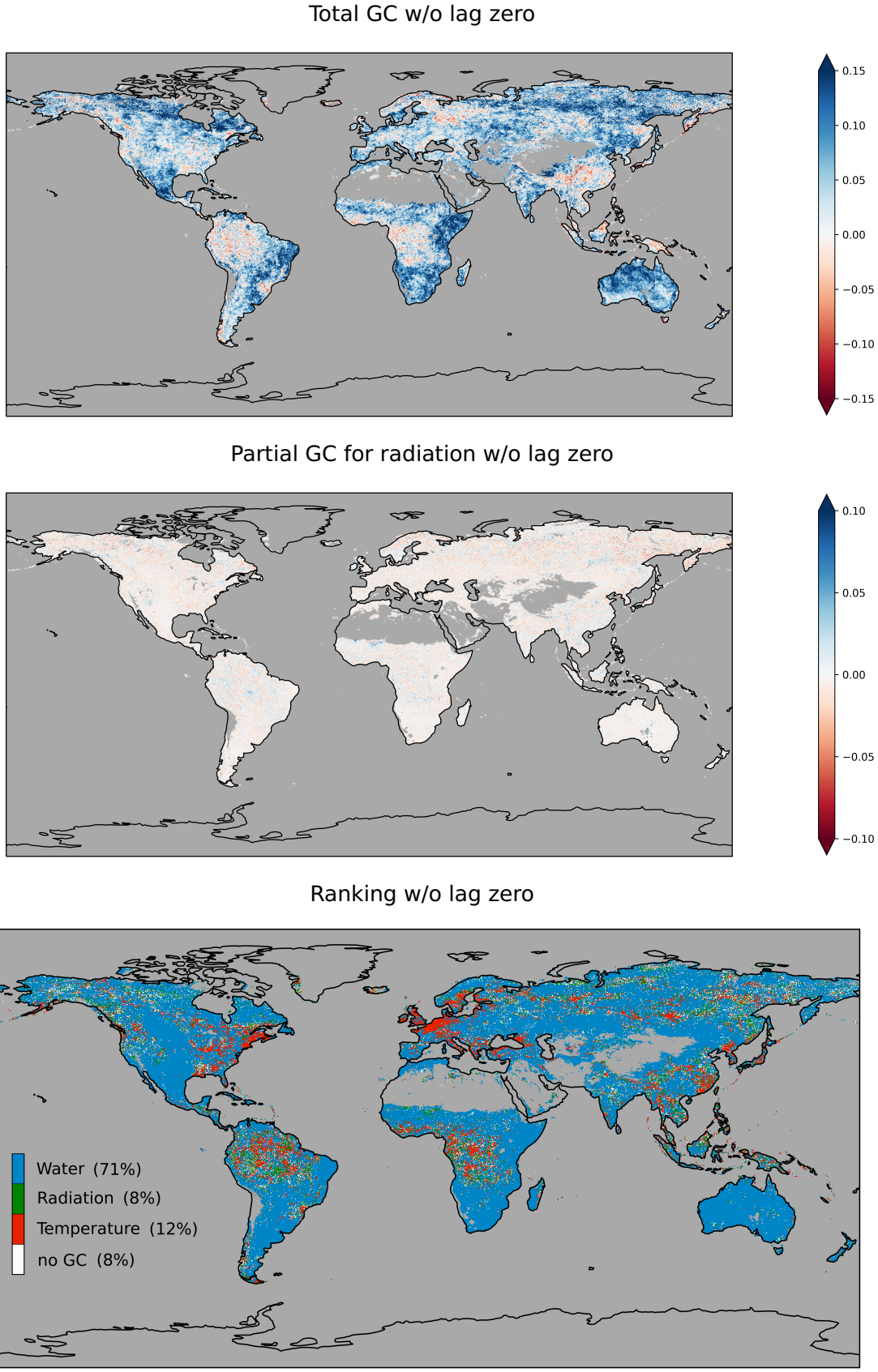



Figure 5.22: Models without lag zero in a high resolution setting (0.25° x 0.25° bi-weekly). [Top] Total Granger causality (GC) without lag zero [Middle] partial Radiation Granger causality without lag zero [Bottom] Group with the highest Granger causality for every pixel without lag zero

CHAPTER 6

# CONCLUSION

One of the main conclusions is that water availability is indeed the most important climate driver influencing global vegetation anomalies, as was reported in Papagiannopoulou et al. (2017a). The impact of water availability is strongest for vegetation types other than forests and cultivated land. The global impact of the radiation group is higher than previously reported and mostly visible in areas with tropical rainforest vegetation. The temperature group shows the strongest importance in Europe and North America but has only limited Granger causality in other regions.

The relative importance of the temperature and radiation group was reported to be the other way around in the research of Papagiannopoulou et al. (2017b). The new relative ranking between the two groups was consistent over a wide range of experimental set-ups in this thesis. Only removing lag zero switches this ranking, but this method is considered too biased for interpretation. The low partial Granger causality for temperature partly originates from the inclusion of VPD in this dataset.

The new variable VPD holds unique Granger causality for the water group in Northern forested areas. It is of use to disentangle the causal impact of temperature and water availability related to their respective influence on plant transpiration. The new wildfire variable shows some Granger causality in Northern regions with forest vegetation. It does not show strong Granger causality in regions with a large annual burned percentage of vegetation in Africa or Australia. This may be partly due to the inability of the models to incorporate spatial information. The burned area alone may be a bad predictor of the wildfire impact on vegetation and a higher Granger causality may be observable by including the fuel consumption as a variable. The new irrigation variable shows only a very small amount of Granger causality, which is negligible on a global scale. This may be related to the local effect of irrigation and possibly to the data quality. Irrigation may serve to control the reported partial Granger causality of the water group. As such, the inclusion of these three new variables is advised for future studies.

The higher resolution of the new data leads to a much lower reported total Granger causality. This originates from the combined effect of the temporal and spatial up-

scaling. The aforementioned conclusions are mostly invariant to the used resolution. A low resolution results in faster calculations, and remote sensing data is available for a much longer time span at a lower resolution (1981 to 2015). From these collective arguments, the 1° x 1° monthly resolution setting is deemed superior for this framework. One theoretical advantage of the setting with bi-weekly resolution is that it might allow for the removal of lag zero from the models. This proved possible in reality, but it leads to an underestimation of the casual impact of the radiation group and is therefore not advised. Blocked cross-validation was tested and proved possible only in the low resolution setting. No strong recommendation is given in whether to use random or blocked cross-validation, as this choice has implications for the reported results and their meaning.

The used framework has its flaws, as is the case with any framework in global vegetation-climate research. Causal information in the spatial dimension is not used and Granger causality itself only reflects unique information in variables. The importance of the water availability group might be overestimated due to the larger number of variables included in this group. The full model works in a high-dimensional setting that may contain a lot of less useful predictors for some pixels, while the baseline model works in a low-dimensional setting. The large difference in number of available features to both models can be a source of bias towards lower reported total Granger causality. As such, it might be interesting to decrease the amount of predictors in the full and reduced models by applying some dimensionality reduction technique such as principal component analysis or a non-linear alternative (Abdi and Williams, 2010). In the end, the reported Granger causality is not a measure of 'real' causality, but only of pseudo-causality. The presented results should be considered together with the previous research of Papagiannopoulou et al. and with other research based on different methods (see Chapter 2 for some examples).

## 6.1 Further research

There are plenty of opportunities for future research in this setting. An interesting approach is the further inclusion of new variables. One such variable is the fuel consumption of wildfires, which could be used complementary to the data on burned area. Other possible new variables are wind speed and turbulence, as both are relevant for the rate of vegetation transpiration on a global scale (Bonan et al., 2014, 2018). This data would likely have to origin from modelling efforts and weather station data (Fick and Hijmans, 2017), as remote sensing wind characteristics is almost exclusively performed over the ocean (Martin, 2014; Lillesand et al., 2014). The influence of VPD on the analysis could be compared to the that of the absolute air moisture content.

Another possibility would be to reverse the analysis and investigate the causal impact of global vegetation on climate drivers.

When investigating the impact of new variables, the adoption of a statistical test should be investigated. The null hypothesis is that this variable holds no new information about the state of the NDVI anomalies. A null distribution could be based on the frequency distribution of the difference between two repetitions in the full model. Different statistical tests should be tested for their use in this setting. The Granger causality distributions are not normally distributed (Figures and tests not included), so statistical tests that rely heavily on that assumption should be avoided. Possibilities include non-parametric rank tests, such as the Mann–Whitney U test (Mann and Whitney, 1947).

An interesting challenge is to utilize the spatial information present in the data. Defining regions of similar vegetation types and training a multi-task learning model per region seems like a succesful stategy (Papagiannopoulou et al., 2018). This enables one to detect and predict vegetation anomalies in space as well, which is useful for localized impacts such as land use change, deforestation and irrigation. Time series would no longer be forced to have equal amounts of negative and positive anomalies over time. This approach is also much better use of the immense data pool available, as it does not force turning to a high-dimensional setting. It might be interesting to isolate the drivers for positive and negative anomalies. By doing so, the positive and negative effects of climate drivers can be isolated and the difference between water-limitation and water fertilization would become visible. The seasonal evolution in this balance could also be investigated. To summarize: there are plenty more research questions to be answered and the findings from this thesis and (Papagiannopoulou et al., 2018) can be used as a starting point for future research.

# BIBLIOGRAPHY


Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Ackerly, D. D., Cornwell, W. K., Weiss, S. B., Flint, L. E., and Flint, A. L. (2015). A geographic mosaic of climate change impacts on terrestrial vegetation: which areas are most at risk? *PloS one*, 10(6):e0130629.

Andela, N., Morton, D., Giglio, L., Chen, Y., Van Der Werf, G., Kasibhatla, P., DeFries, R., Collatz, G., Hantson, S., Kloster, S., et al. (2017). A human-driven decline in global burned area. *Science*, 356(6345):1356–1362.

Bastos, A., Gouveia, C., Trigo, R., and Running, S. W. (2014). Analysing the spatio-temporal impacts of the 2003 and 2010 extreme heatwaves on plant productivity in europe. *Biogeosciences*, 11(13):3421.

Beck, H. E., Van Dijk, A. I., Levizzani, V., Schellekens, J., Gonzalez Miralles, D., Martens, B., and De Roo, A. (2017). Mswep: 3-hourly 0.25 global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21(1):589–615.

Boisvenue, C. and Running, S. W. (2006). Impacts of climate change on natural forest productivity–evidence since the middle of the 20th century. *Global Change Biology*, 12(5):862–882.

Bonan, G., Williams, M., Fisher, R., and Oleson, K. (2014). Modeling stomatal conductance in the earth system: linking leaf water-use efficiency and water transport along the soil–plant–atmosphere continuum. *Geoscientific Model Development*, 7(5):2193–2222.

Bonan, G. B. (2011). Forests and global change. In *Forest Hydrology and Biogeochemistry*, pages 711–725. Springer.

Bonan, G. B., Levis, S., Sitch, S., Vertenstein, M., and Oleson, K. W. (2003). A dynamic global vegetation model for use with climate models: concepts and description of simulated vegetation dynamics. *Global Change Biology*, 9(11):1543–1566.

Bonan, G. B., Patton, E. G., Harman, I. N., Oleson, K. W., Finnigan, J. J., Lu, Y., and Burakowski, E. A. (2018). Modeling canopy-induced turbulence in the earth system: a unified parameterization of turbulent exchange within plant canopies and the roughness sublayer (clm-ml v0). *Geoscientific Model Development*, 11(4):1467–1496.

Bond, W. J., Woodward, F. I., and Midgley, G. F. (2005). The global distribution of ecosystems in a world without fire. *New phytologist*, 165(2):525–538.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al. (2015). Global land cover mapping at 30 m resolution: A pok-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:7–27.

Chen, T., De Jeu, R., Liu, Y., Van der Werf, G., and Dolman, A. (2014). Using satellite based soil moisture to quantify the water driven variability in ndvi: A case study over mainland australia. *Remote Sensing of Environment*, 140:330–338.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., et al. (2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437(7058):529.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition. *Journal of Official Statistics*, 6(1):3–73.

Costanza, R. and Maxwell, T. (1994). Resolution and predictability: an approach to the scaling problem. *Landscape Ecology*, 9(1):47–57.

Della-Marta, P. M., Haylock, M. R., Luterbacher, J., and Wanner, H. (2007). Doubled length of western european summer heat waves since 1880. *Journal of Geophysical Research: Atmospheres*, 112(D15).

Devaraju, N., Bala, G., Caldeira, K., and Nemani, R. (2016). A model based investigation of the relative importance of co2-fertilization, climate warming, nitrogen deposition and land use change on the global terrestrial carbon uptake in the historical period. *Climate dynamics*, 47(1-2):173–190.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., et al. (2017). Esa cci soil moisture for improved earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, 203:185–215.

Ducasse, S. (2017). Granger causality: Definition, running the test. *Online source, http://www.statisticshowto.com/granger-causality/*.

Faghmous, J. H. and Kumar, V. (2014). A big data guide to understanding climate change: The case for theory-guided data science. *Big data*, 2(3):155–163.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Foley, J. A., Levis, S., Prentice, I. C., Pollard, D., and Thompson, S. L. (1998). Coupling dynamic models of climate and vegetation. *Global change biology*, 4(5):561–579.

Gerber, F., Furrer, R., Schaepman-Strub, G., de Jong, R., and Schaepman, M. E. (2016). Predicting missing values in spatio-temporal satellite data. *arXiv preprint arXiv:1605.01038*.

Giglio, L., Randerson, J. T., and van der Werf, G. R. (2013). Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (gfed4). *Journal of Geophysical Research: Biogeosciences*, 118(1):317–328.

Glenn, E. P., Huete, A. R., Nagler, P. L., and Nelson, S. G. (2008). Relationship between remotely-sensed vegetation indices, canopy attributes and plant physiological processes: What vegetation indices can and cannot tell us about the landscape. *Sensors*, 8(4):2136–2160.

Gouveia, C., Bastos, A., Trigo, R., and DaCamara, C. (2012). Drought impacts on vegetation in the pre-and post-fire events over iberian peninsula. *Natural Hazards and Earth System Sciences*, 12(10):3123–3137.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. econometrica, vol 37, pp 424-438, granger cw.

Granger, C. W. (1980). Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*, chapter Random Forest. Springer-Verlag New York Inc.

Heimann, M. and Reichstein, M. (2008). Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature*, 451(7176):289.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Ji, L. and Peters, A. J. (2003). Assessing vegetation response to drought in the northern great plains using vegetation and drought indices. *Remote Sensing of Environment*, 87(1):85–98.

Jiang, B., Liang, S., and Yuan, W. (2015). Observational evidence for impacts of vegetation change on local surface climate over northern china using the granger causality test. *Journal of Geophysical Research: Biogeosciences*, 120(1):1–12.

Kandasamy, S., Baret, F., Verger, A., Neveux, P., and Weiss, M. (2013). A comparison of methods for smoothing and gap filling time series of remote sensing observations–application to modis lai products. *Biogeosciences*, 10(6):4055–4071.

Kaufmann, R., Zhou, L., Myneni, R., Tucker, C., Slayback, D., Shabanov, N., and Pinzon, J. (2003). The effect of vegetation on surface temperature: A statistical analysis of ndvi and climate data. *Geophysical Research Letters*, 30(22).

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. (2006). World map of the köppen-geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3):259–263.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19(1).

Li, Z. and Kafatos, M. (2000). Interannual variability of vegetation in the united states and its relation to el nino/southern oscillation. *Remote Sensing of Environment*, 71(3):239–247.

Lillesand, T., Kiefer, R. W., and Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

Liu, Z., Notaro, M., Kutzbach, J., and Liu, N. (2006). Assessing global vegetation–climate feedbacks from observations. *Journal of Climate*, 19(5):787–814.

Loeb, N. G., Manalo-Smith, N., Su, W., Shankar, M., and Thomas, S. (2016). Ceres top-of-atmosphere earth radiation budget climate data record: Accounting for in-orbit changes in instrument calibration. *Remote Sensing*, 8(3):182.

Lu, B. and He, Y. (2014). Analyzing a north american prairie wildfire using remote sensing imagery. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pages 832–835. IEEE.

Luo, Y., Gerten, D., Le Maire, G., Parton, W. J., Weng, E., Zhou, X., Keough, C., Beier, C., Ciais, P., Cramer, W., et al. (2008). Modeled interactive effects of precipitation, temperature, and [co2] on ecosystem carbon and water dynamics in different climatic zones. *Global Change Biology*, 14(9):1986–1999.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., and Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, 51:47–60.

Malagnoux, M. (2007). Arid land forests of the world: global environmental perspectives. In *International Conference on Afforestation and Sustainable Forests as a Means to Combat Desertification, Jerusalem, Israel*, pages 16–19.

Mallick, K., Trebs, I., Boegh, E., Giustarini, L., Schlerf, M., Drewry, D. T., Hoffmann, L., Von Randow, C., Kruijt, B., Araùjo, A., et al. (2016). Canopy-scale biophysical controls of transpiration and evaporation in the amazon basin. *Hydrology and Earth System Sciences*, 20(10):4237–4264.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Mares, M. A. (2017). *Encyclopedia of deserts*. University of Oklahoma Press.

Martens, B., Gonzalez Miralles, D., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck, H. E., Dorigo, W., and Verhoest, N. (2017). Gleam v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, 10(5):1903–1925.

Martin, S. (2014). *An introduction to ocean remote sensing*. Cambridge University Press.

McAdam, S. A. and Brodribb, T. J. (2015). The evolution of mechanisms driving the stomatal response to vapour pressure deficit. *Plant Physiology*, pages pp–114.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100.

Metsämäki, S., Pulliainen, J., Salminen, M., Luojus, K., Wiesmann, A., Solberg, R., Böttcher, K., Hiltunen, M., and Ripper, E. (2015). Introduction to globsnow snow extent products with considerations for accuracy assessment. *Remote Sensing of Environment*, 156:96–108.

Mueller, R. C., Scudder, C. M., Porter, M. E., Talbot Trotter III, R., Gehring, C. A., and Whitham, T. G. (2005). Differential tree mortality in response to severe drought: evidence for long-term vegetation shifts. *Journal of Ecology*, 93(6):1085–1093.

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., and Döll, P. (2014). Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrology and Earth System Sciences*, 18(9):3511–3538.

Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., Myneni, R. B., and Running, S. W. (2003). Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *science*, 300(5625):1560–1563.

Nicolai-Shaw, N., Zscheischler, J., Hirschi, M., Gudmundsson, L., and Seneviratne, S. I. (2017). A drought event composite analysis using satellite remote-sensing based soil moisture. *Remote Sensing of Environment*, 203:216–225.

Notaro, M., Liu, Z., and Williams, J. W. (2006). Observed vegetation–climate feedbacks in the united states. *Journal of Climate*, 19(5):763–786.

Papagiannopoulou, C., Miralles, D., Depoorter, M., Verhoest, N. E., Dorigo, W., and Waegeman, W. (2016). Discovering relationships in climate-vegetation dynamics using satellite data. In *Proceedings of AALTD 2016: Second ECML/PKDD International Workshop on Advanced Analytics and Learning on Temporal Data*, page 46.

Papagiannopoulou, C., Miralles, D., Dorigo, W., Verhoest, N., Depoorter, M., and Waegeman, W. (2017a). Vegetation anomalies caused by antecedent precipitation in most of the world. *Environmental Research Letters*, 12(7):074016.

Papagiannopoulou, C., Miralles, D. G., Decubber, S., Demuzere, M., Verhoest, N. E., Dorigo, W. A., and Waegeman, W. (2017b). A non-linear granger-causality framework to investigate climate–vegetation dynamics. *Geoscientific Model Development*, 10(5):1945–1960.

Papagiannopoulou, C., Miralles, D. G., Demuzere, M., Verhoest, N. E., and Waegeman, W. (2018). Global hydro-climatic biomes identified via multi-task learning.

Pinzon, J. E. and Tucker, C. J. (2014). A non-stationary 1981–2012 avhrr ndvi3g time series. *Remote Sensing*, 6(8):6929–6960.

Puig-Gironès, R., Brotons, L., and Pons, P. (2017). Aridity influences the recovery of vegetation and shrubland birds after wildfire. *PloS one*, 12(3):e0173599.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Russo, S., Sillmann, J., and Fischer, E. M. (2015). Top ten european heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, 10(12):124003.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 329. John Wiley & Sons.

Seddon, A. W., Macias-Fauria, M., Long, P. R., Benz, D., and Willis, K. J. (2016). Sensitivity of global terrestrial ecosystems to climate variability. *Nature*, 531(7593):229.

Sen, P. K. (1968). Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, 63(324):1379–1389.

Solomon, S., Plattner, G.-K., Knutti, R., and Friedlingstein, P. (2009). Irreversible climate change due to carbon dioxide emissions. *Proceedings of the national academy of sciences*, 106(6):1704–1709.

Stohlgren, T. J., Chase, T. N., Pielke, R. A., Kittel, T. G., Baron, J., et al. (1998). Evidence that local land use practices influence regional climate, vegetation, and stream flow patterns in adjacent natural areas. *Global Change Biology*, 4(5):495–504.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

Tepley, A., Veblen, T., Perry, G., and Anderson-Teixeira, K. (2015). Vulnerability and resilience of temperate forest landscapes to broad-scale deforestation in response to changing fire regimes and altered post-fire vegetation dynamics. In *AGU Fall Meeting Abstracts*.

Tian, B. (2016a). Atmospheric infrared sounder/advance microwave sounding unit (airs/amsu) obs4mips v2 air temperature description.

Tian, B. (2016b). Atmospheric infrared sounder/advance microwave sounding unit (airs/amsu) obs4mips v2 specific humidity description.

Verbesselt, J., Hyndman, R., Zeileis, A., and Culvenor, D. (2010). Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment*, 114(12):2970–2980.

Wakeling, J. L., Cramer, M. D., and Bond, W. J. (2012). The savanna-grassland 'treeline': why don't savanna trees occur in upland grasslands? *Journal of Ecology*, 100(2):381–391.

Wang, G., Yu, M., Pal, J. S., Mei, R., Bonan, G. B., Levis, S., and Thornton, P. E. (2016). On the development of a coupled regional climate–vegetation model rcm–clm–cn–dv and its validation in tropical africa. *Climate dynamics*, 46(1-2):515–539.

Wang, W., Anderson, B. T., Phillips, N., Kaufmann, R. K., Potter, C., and Myneni, R. B. (2006). Feedbacks of vegetation on summertime climate variability over the north american grasslands. part i: Statistical analysis. *Earth Interactions*, 10(17):1–27.

Williams, A. P., Allen, C. D., Macalady, A. K., Griffin, D., Woodhouse, C. A., Meko, D. M., Swetnam, T. W., Rauscher, S. A., Seager, R., Grissino-Mayer, H. D., et al. (2013).

Temperature as a potent driver of regional forest drought stress and tree mortality. *Nature Climate Change*, 3(3):292.

Williams, C. A., Reichstein, M., Buchmann, N., Baldocchi, D., Beer, C., Schwalm, C., Wohlfahrt, G., Hasler, N., Bernhofer, C., Foken, T., et al. (2012). Climate and vegetation controls on the surface water balance: Synthesis of evapotranspiration measured across a global network of flux towers. *Water Resources Research*, 48(6).

Wu, D., Zhao, X., Liang, S., Zhou, T., Huang, K., Tang, B., and Zhao, W. (2015). Time-lag effects of global vegetation responses to climate change. *Global change biology*, 21(9):3520–3531.

Zhang, W., Jansson, C., Miller, P., Smith, B., and Samuelsson, P. (2014). Biogeophysical feedbacks enhance the arctic terrestrial carbon sink in regional earth system dynamics. *Biogeosciences*, 11(19):5503–5519.

APPENDIX A

# ADDITIONAL FIGURES: DATA AND METHODS

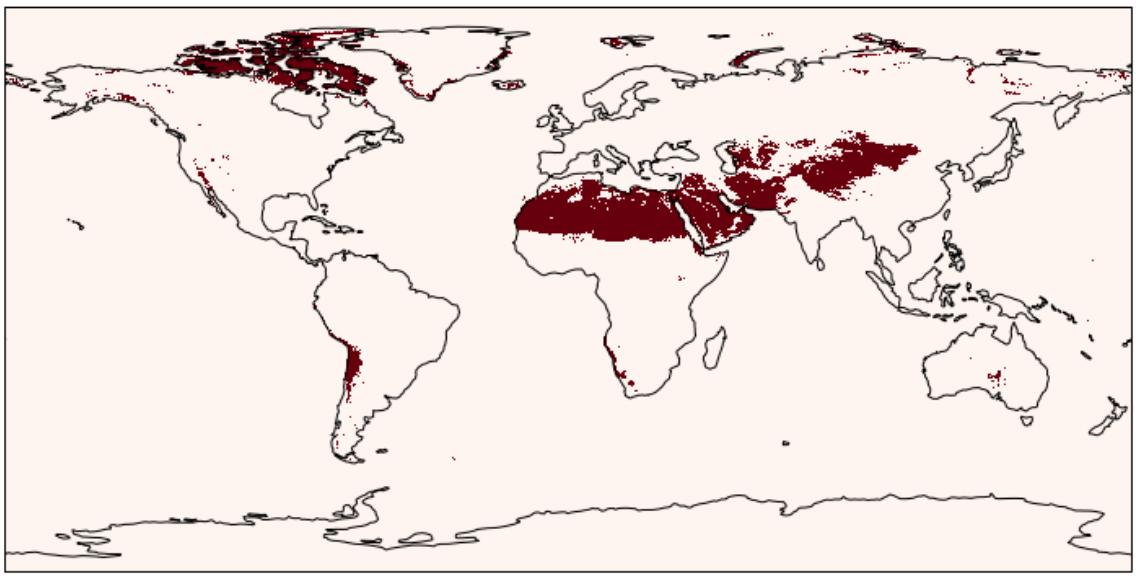


Figure A.1: Areas with total NDVI below 400000 that are exluded from the analysis

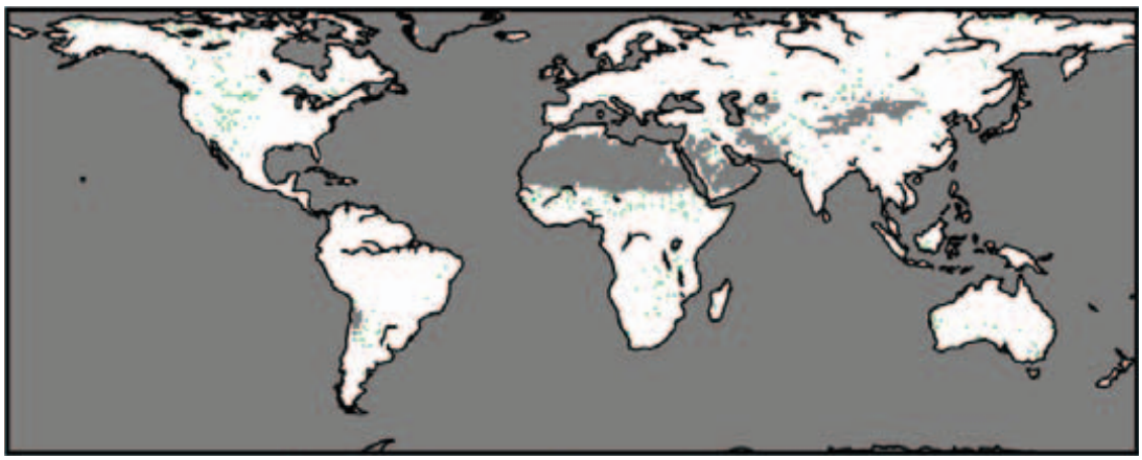


Figure A.2: Result from Papagiannopoulou et al. (2017a), the grey areas shown have no data

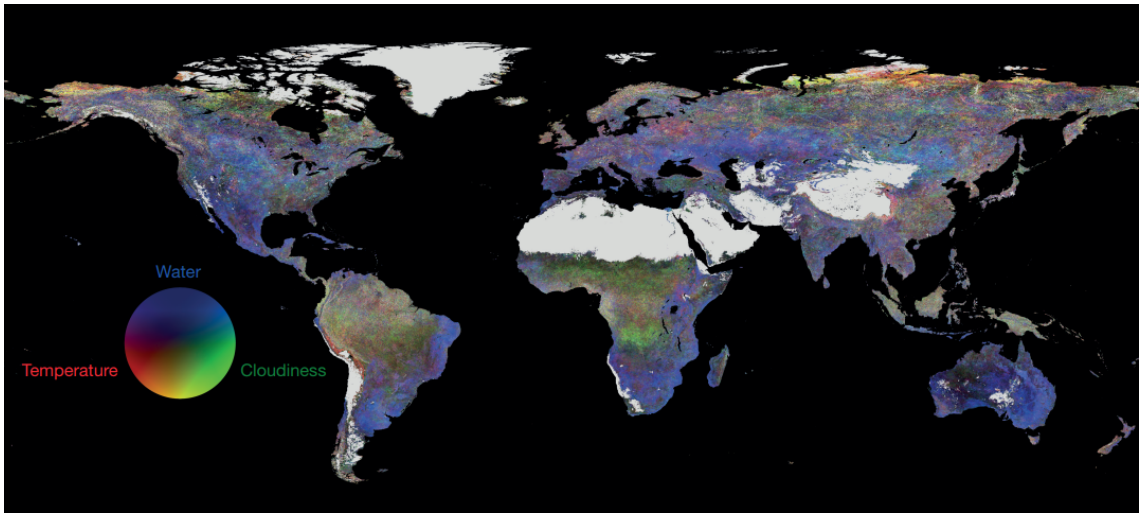



Figure A.3: Result from Seddon et al. (2016) that shows the contribution of three climate variables to the vegetation sensitivity index. White areas are excluded from the analysis

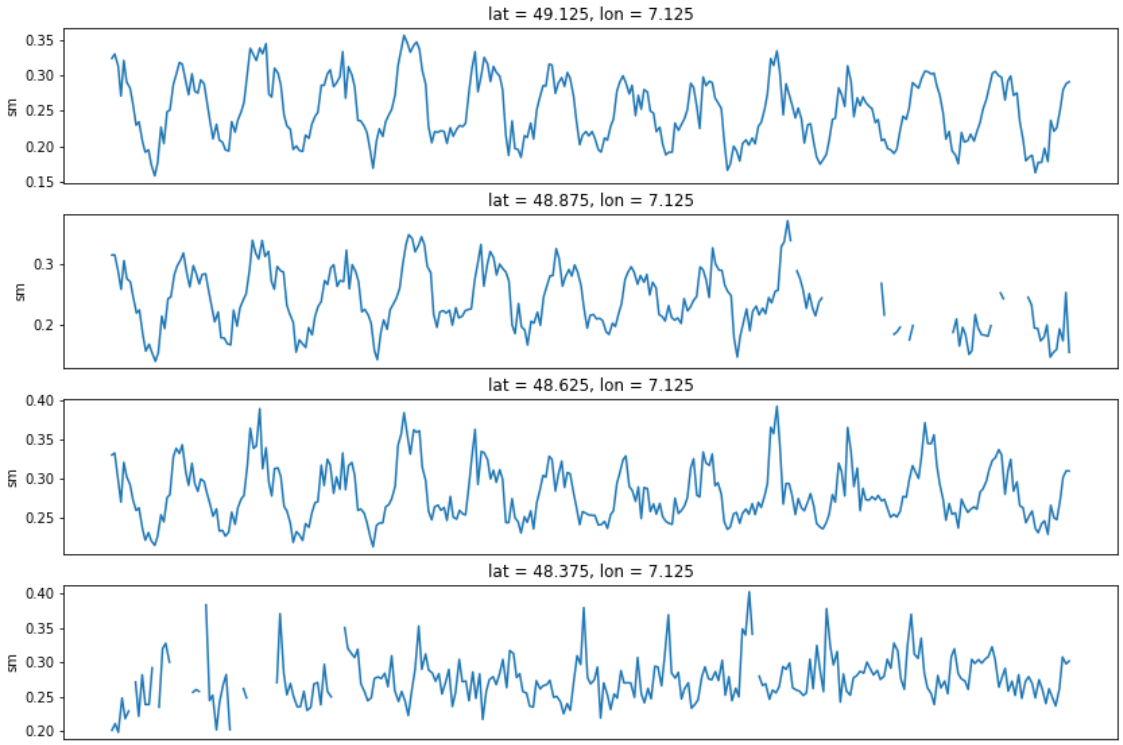



Figure A.4: Near-surface soil moisture: different time series from a vertical slice of European mainland (Longitude = 7.125)

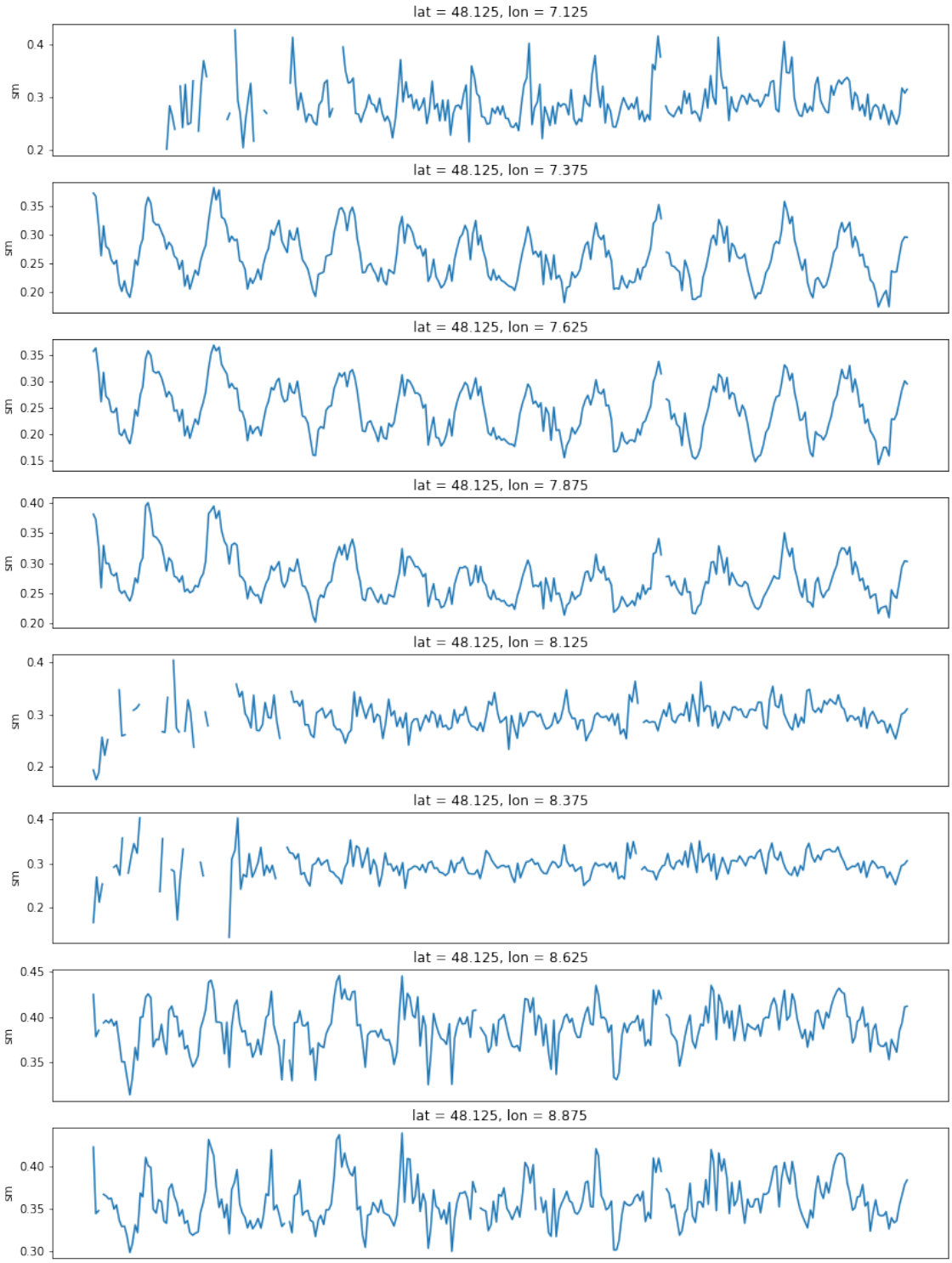



Figure A.5: Near-surface soil moisture: Different time series from a horizontal slice of European mainland at (Latitude = 48.125)

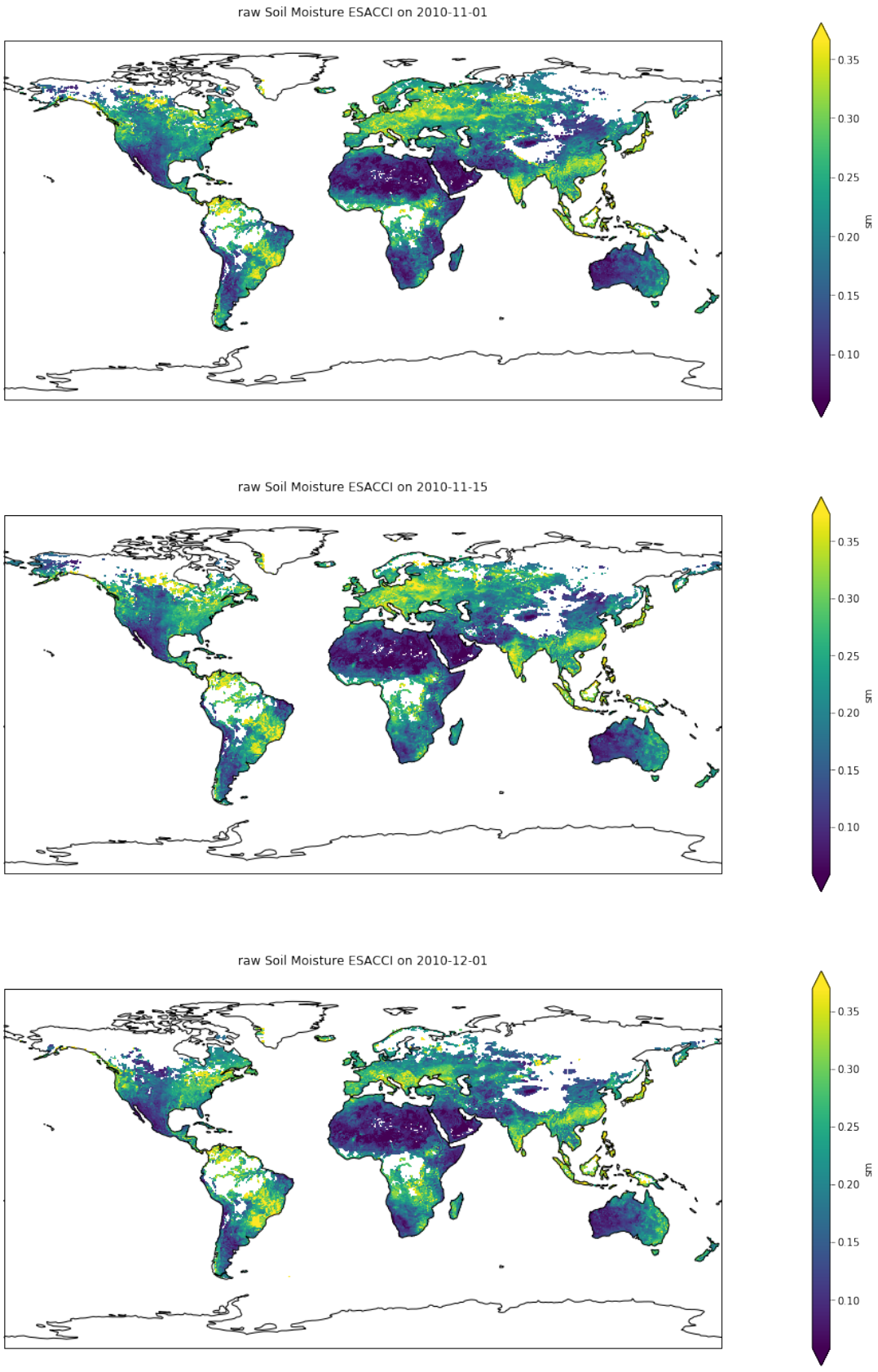



Figure A.6: Near-surface soil moisture: 3 consecutive timesteps from the raw data

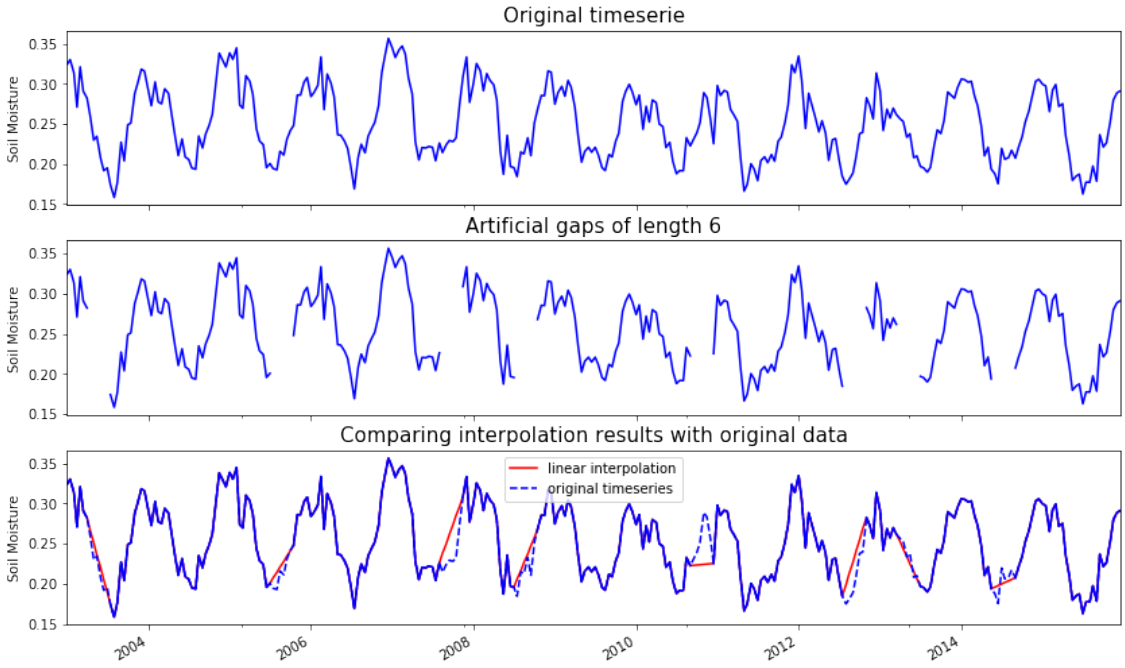



Figure A.7: soil moisture time series with artificial gaps of length 6, filled with linear interpolation

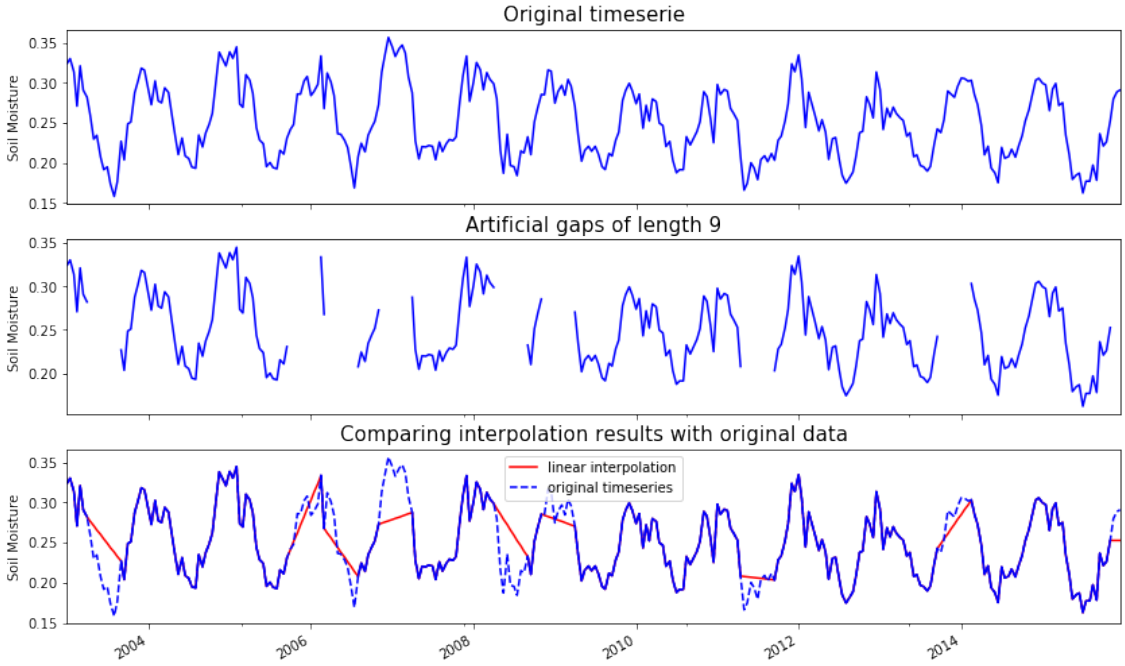



Figure A.8: soil moisture time series with artificial gaps of length 9, filled with linear interpolation

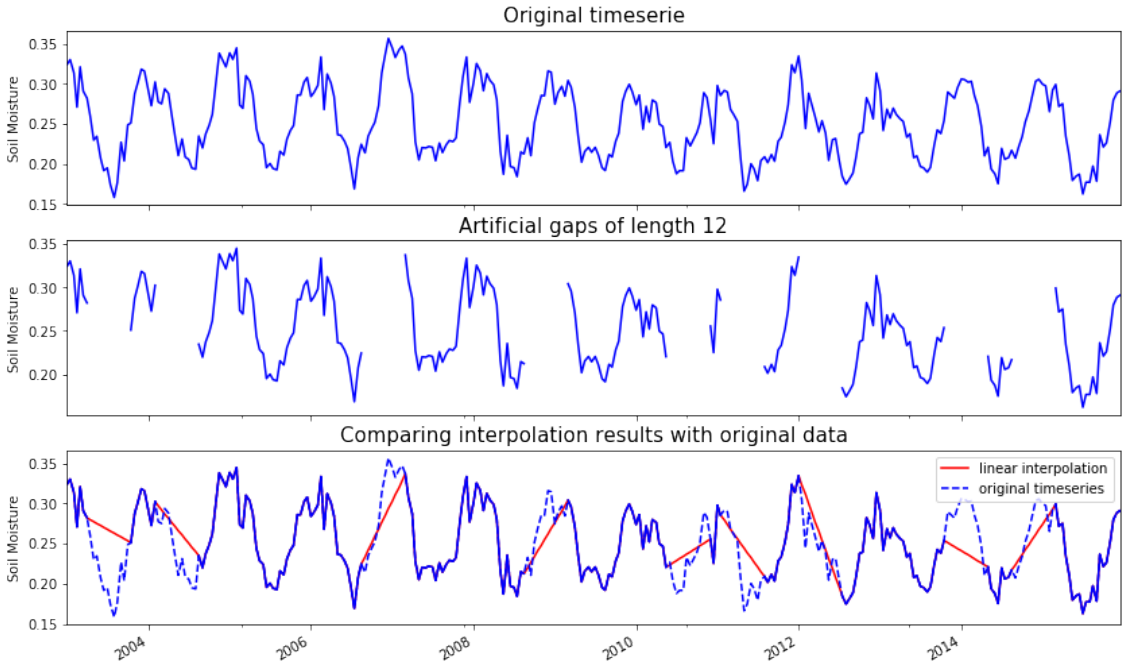



Figure A.9: soil moisture time series with artificial gaps of length 12, filled with linear interpolation

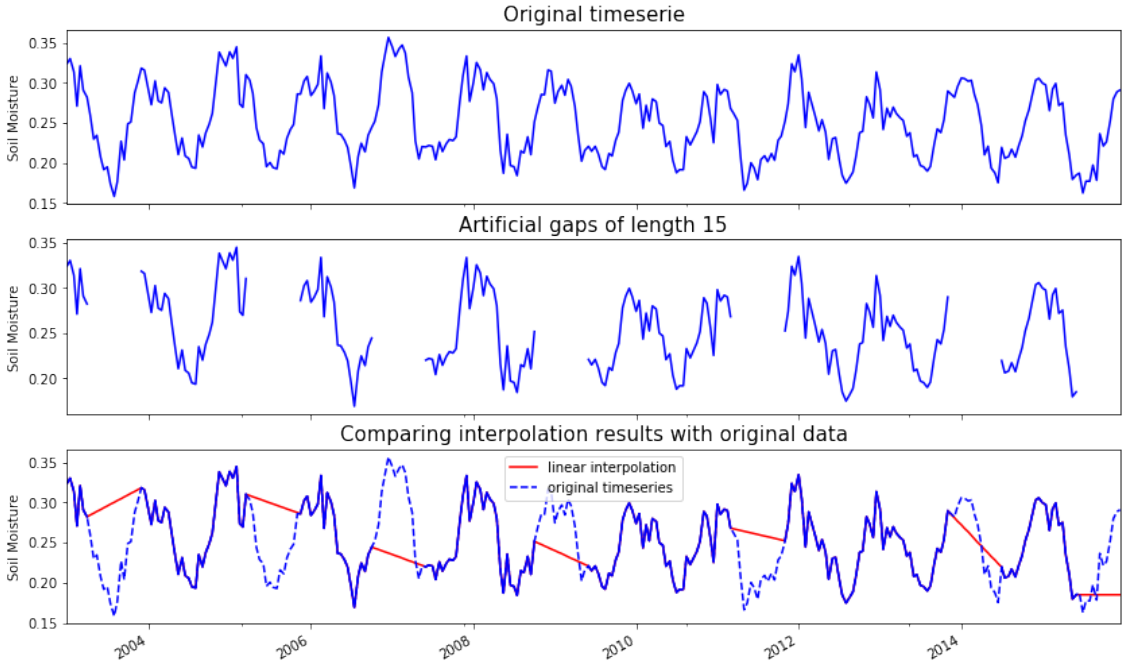



Figure A.10: soil moisture time series with artificial gaps of length 15, filled with linear interpolation

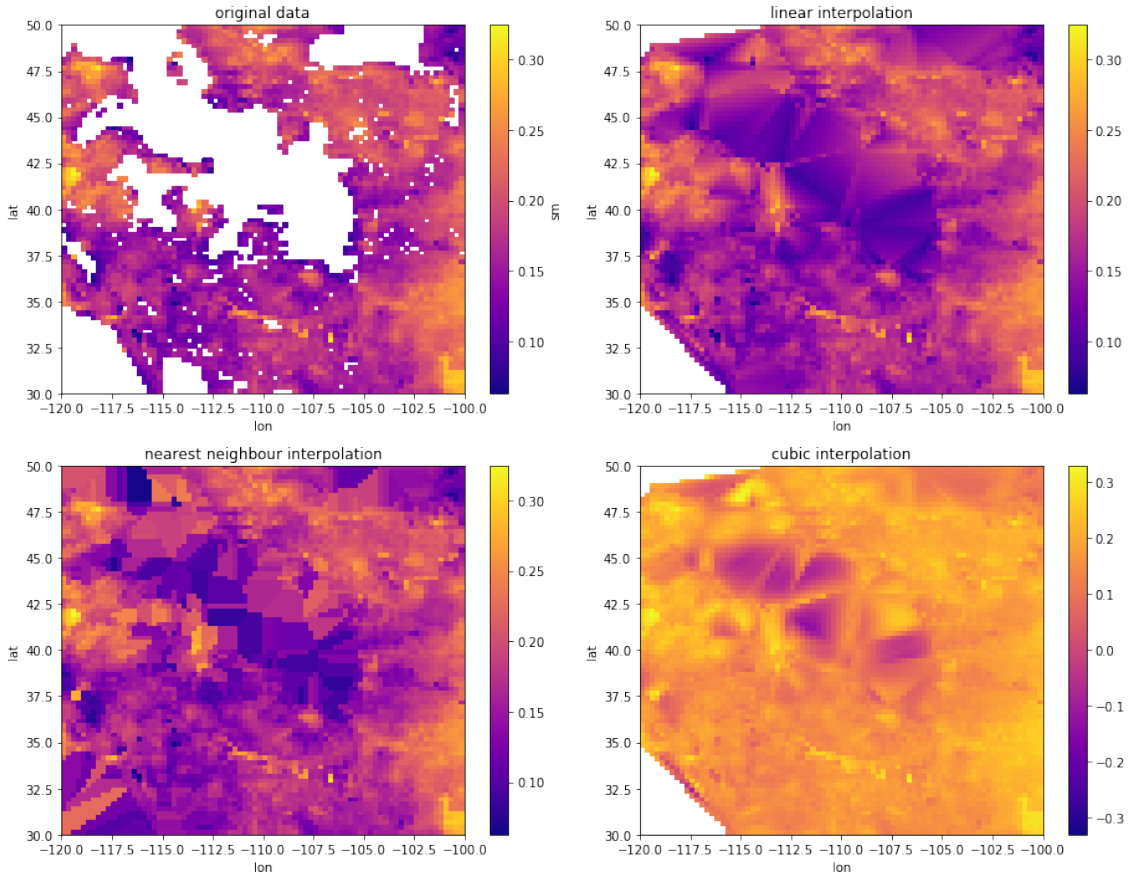



Figure A.11: Different pre-implemented interpolation algorithms. *Upper left: original data, Upper right: linear interpolation, Lower left: nearest neighbor, Lower right: cubic interpolation*

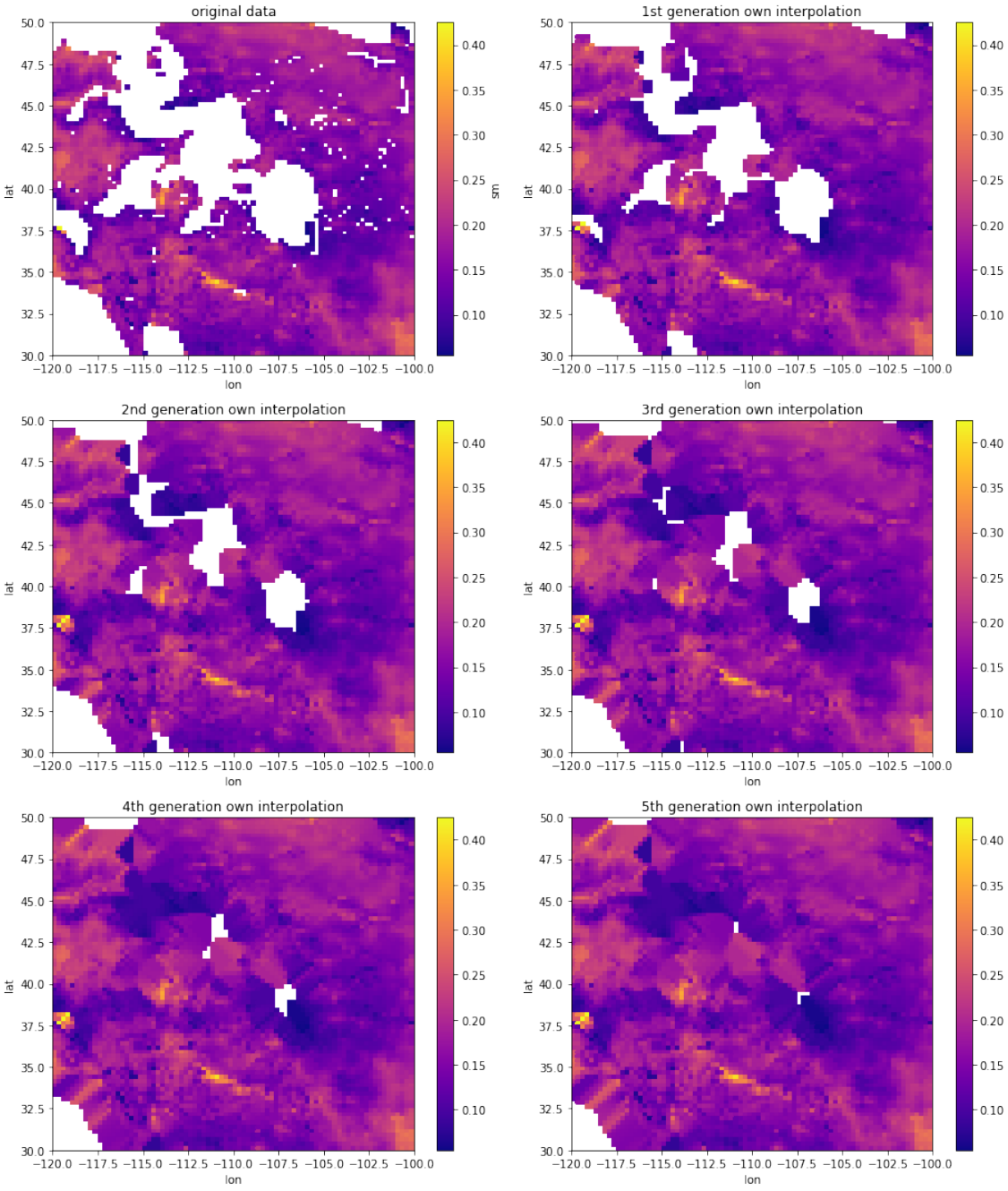



Figure A.12: Own interpolation algorithm: 5 generations. *Upper left: Original data*

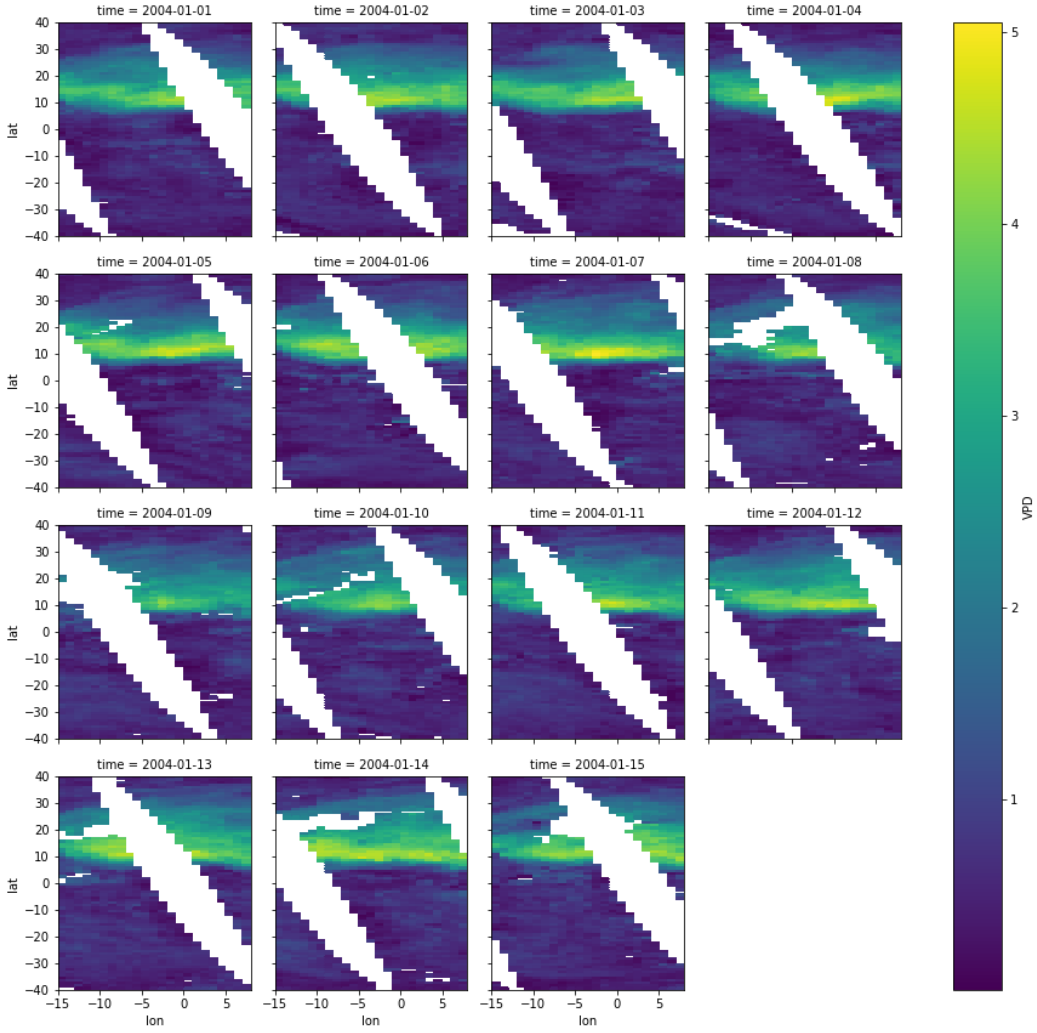



Figure A.13: VPD: Movement of the data gaps over 15 days

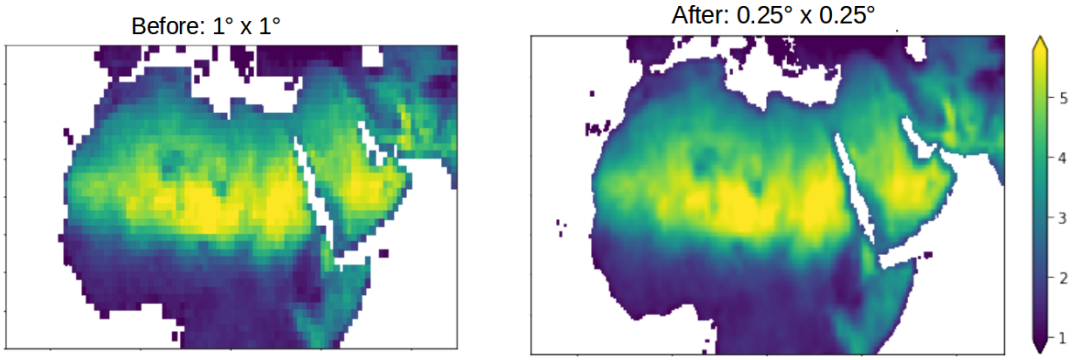



Figure A.14: VPD: Africa before and after spatial upscaling

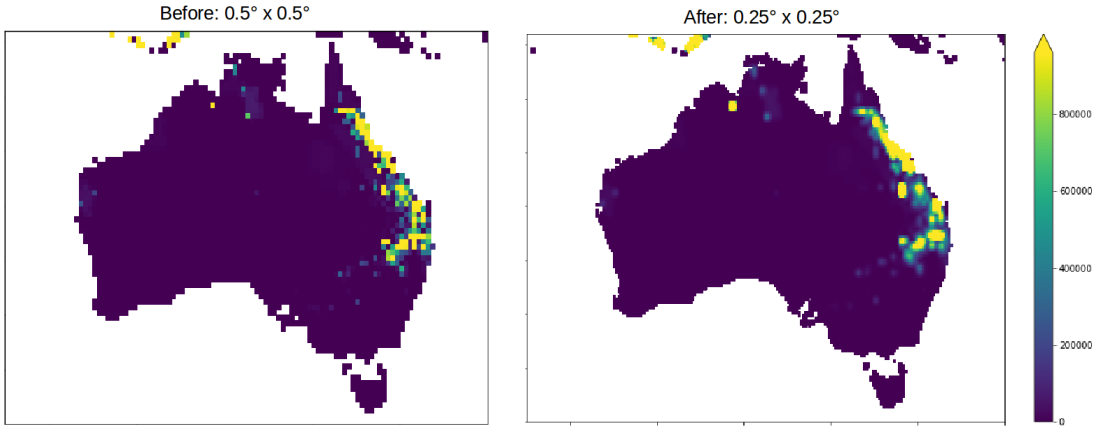



Figure A.15: Irrigation: Australia before and after spatial upscaling

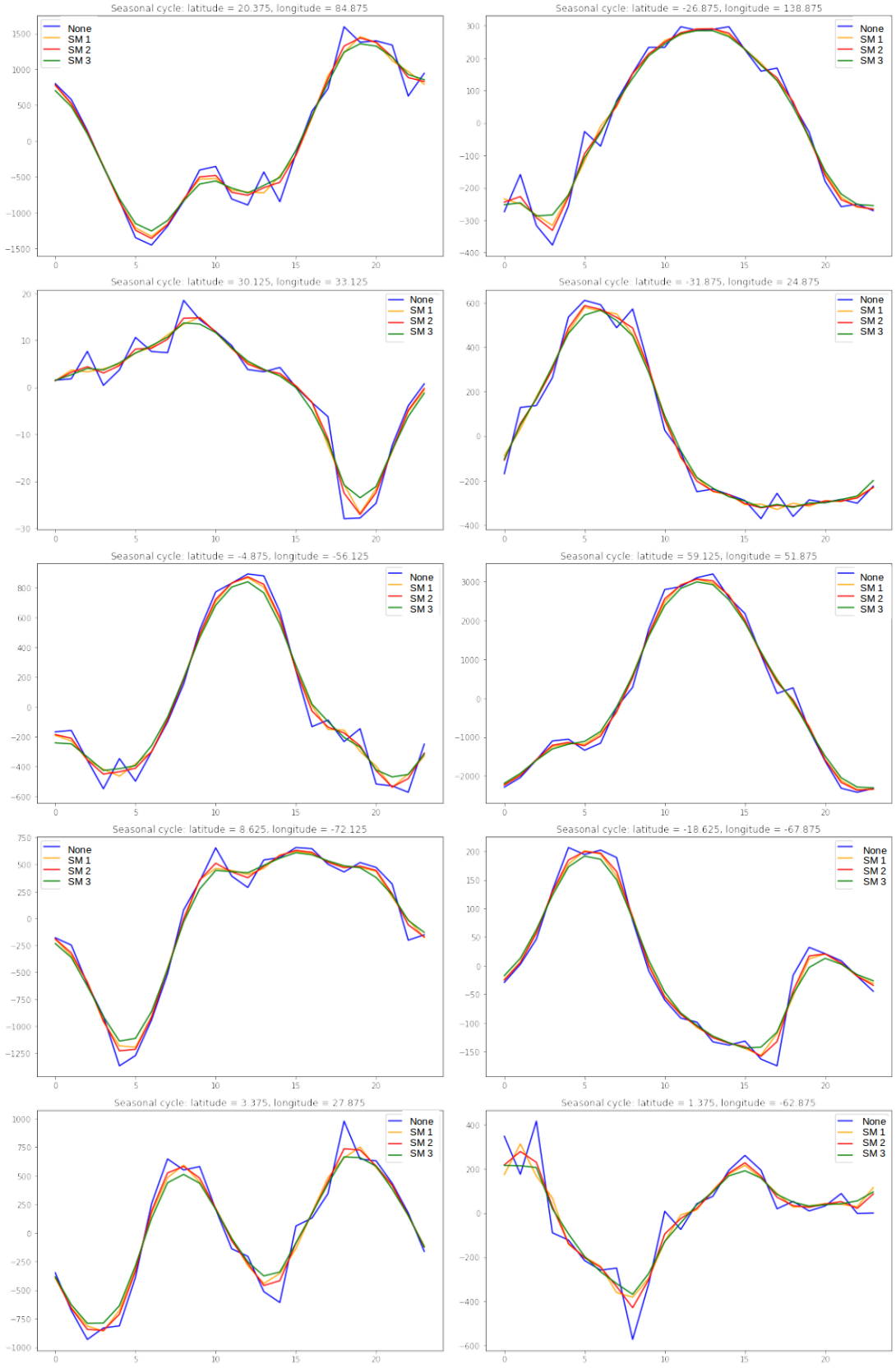



Figure A.16: Different smoothers for the NDVI seasonal cycle, for 10 random pixels. the used smoother is SM 2

APPENDIX B

# ADDITIONAL FIGURES: RESULTS

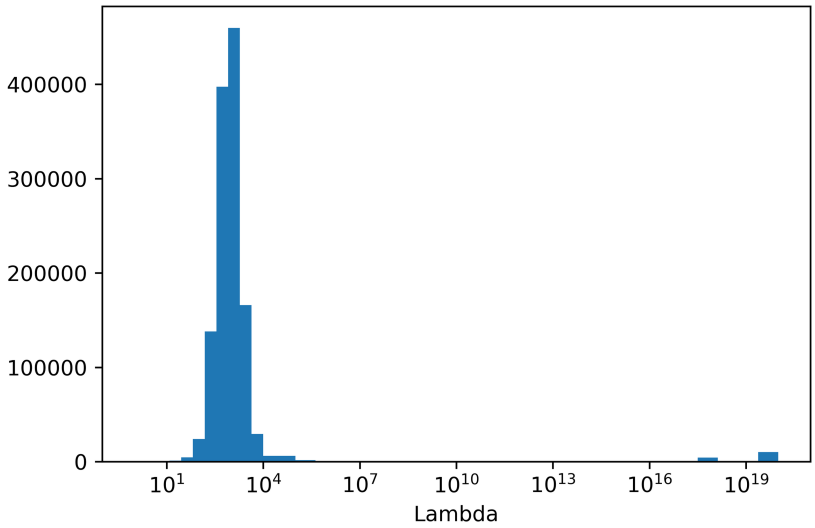



Figure B.1: Histogram of $\lambda$ terms for the ridge regression full model.

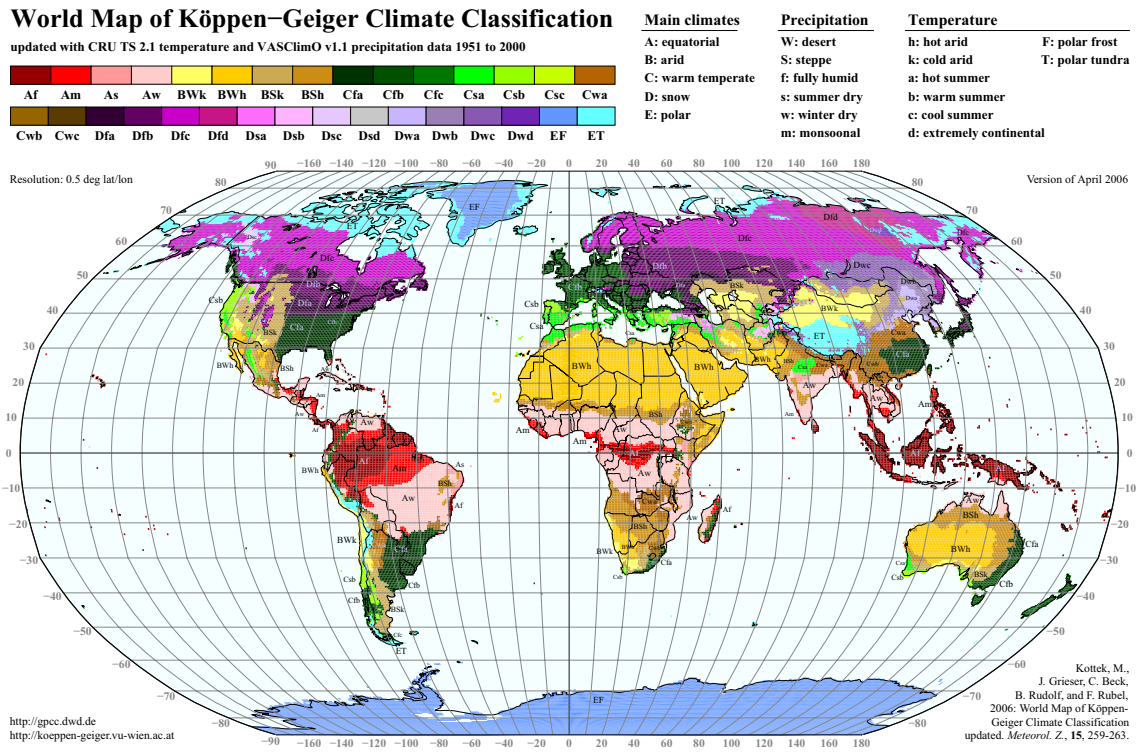



Figure B.2: Global classification of climate types, adopted from Kottek et al. (2006). Af and Am indicate tropical and monsoonal rainforest respectively

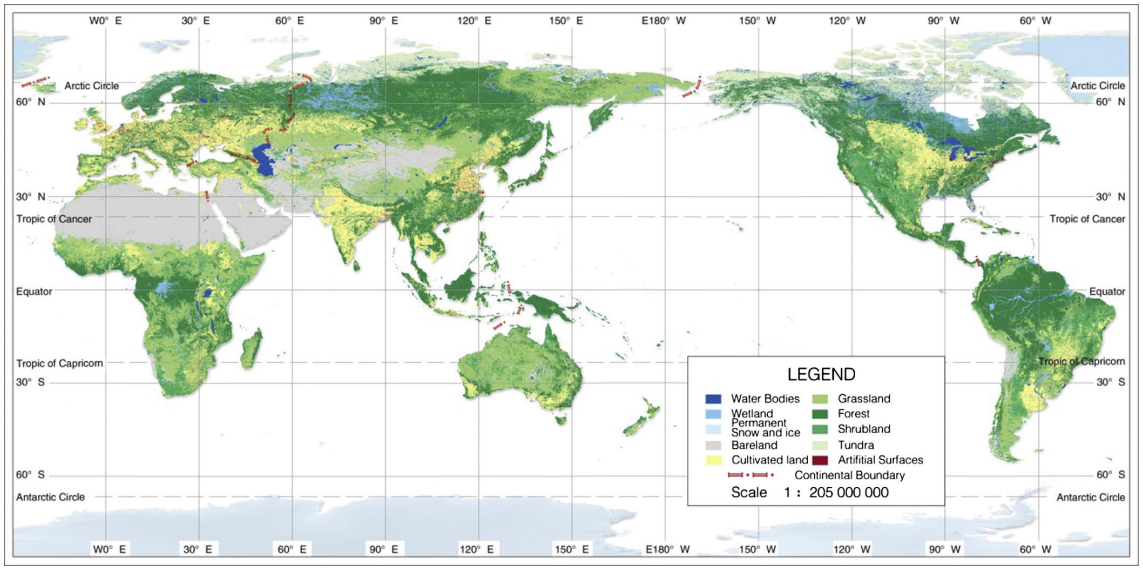
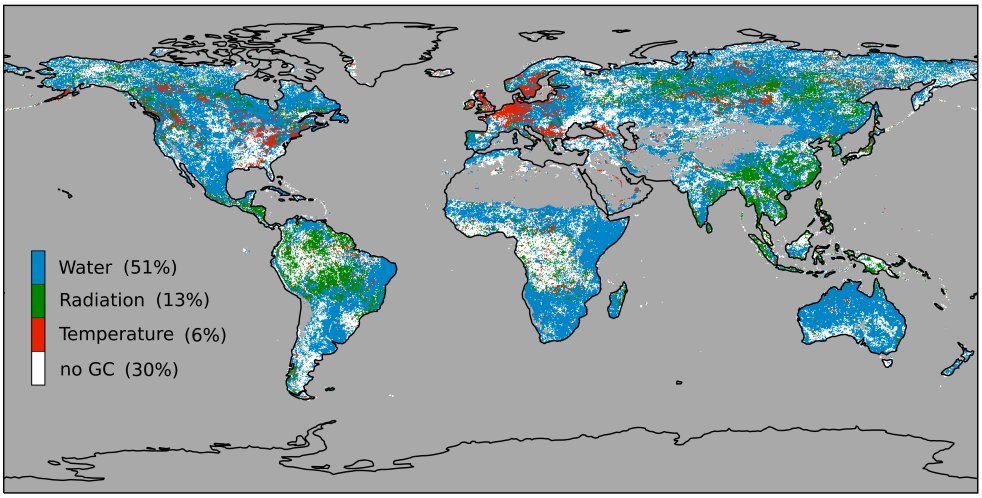
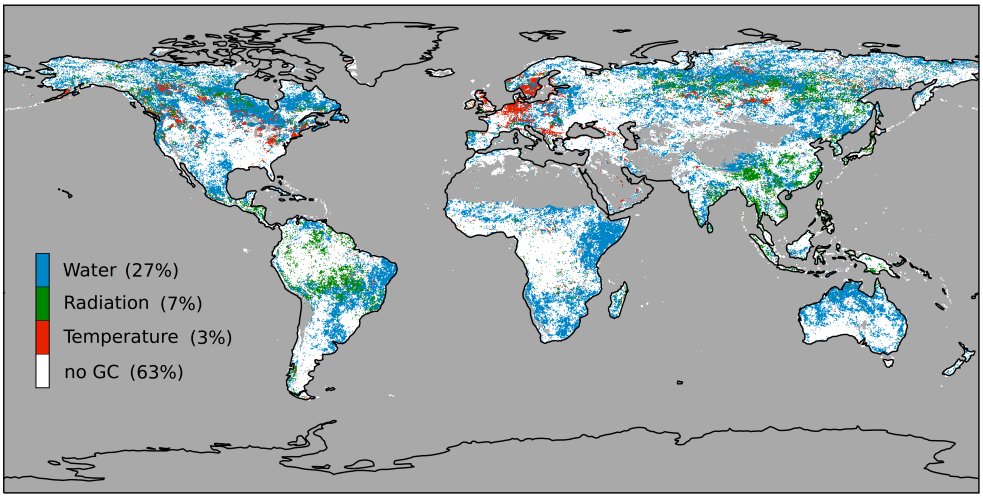



Figure B.3: Global vegetation zones mapped through remote sensing as described in Chen et al. (2015)

Total GC ≥ 0.05




Total GC ≥ 0.10




Total GC ≥ 0.15

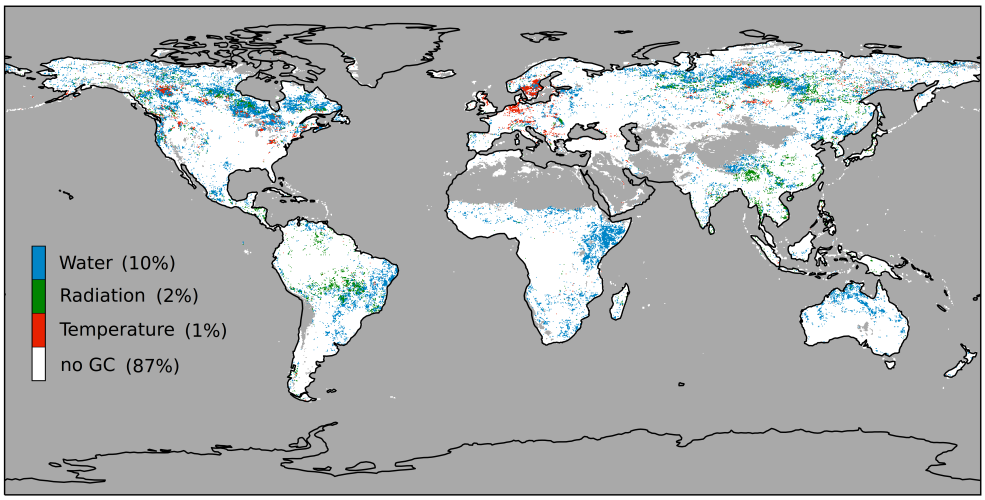



Figure B.4: Group with the highest Granger causality (GC) for every pixel. Different cut-offs are set for the minimum total Granger causality, the "no GC" label is used for pixels below that cut-off

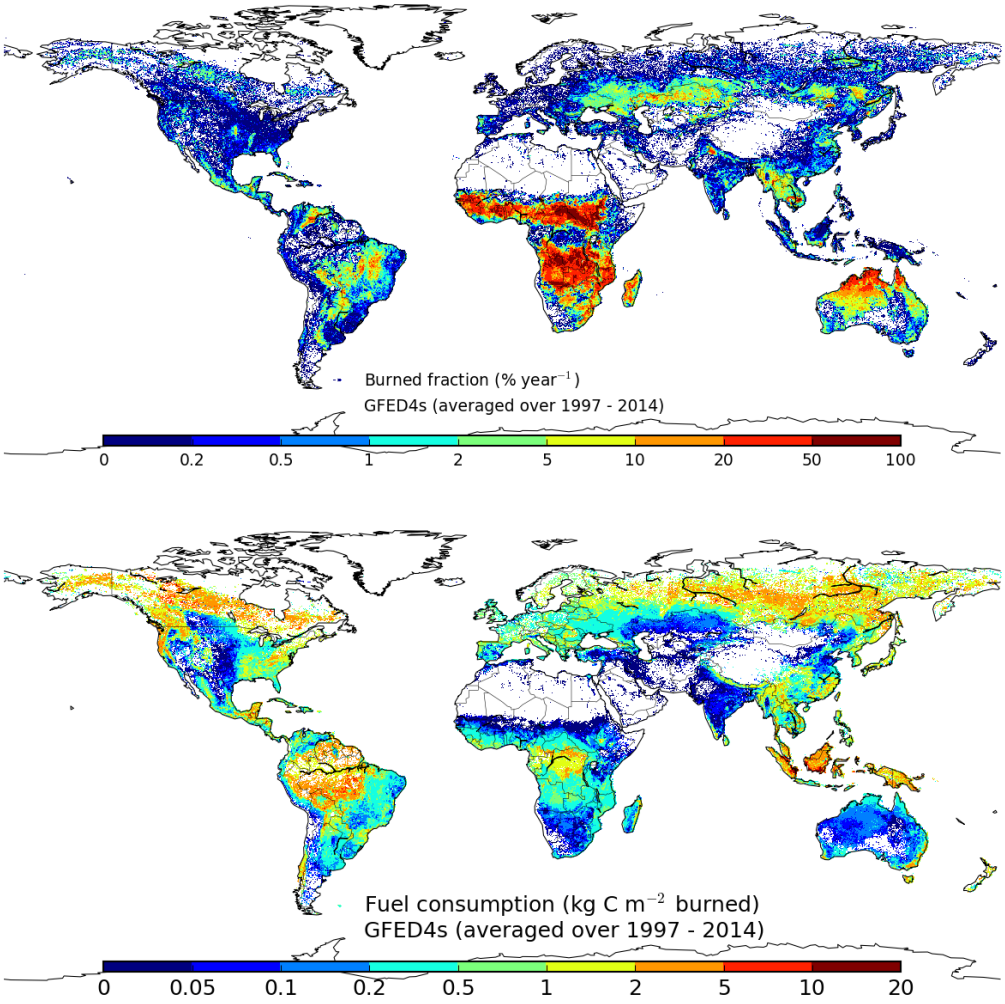
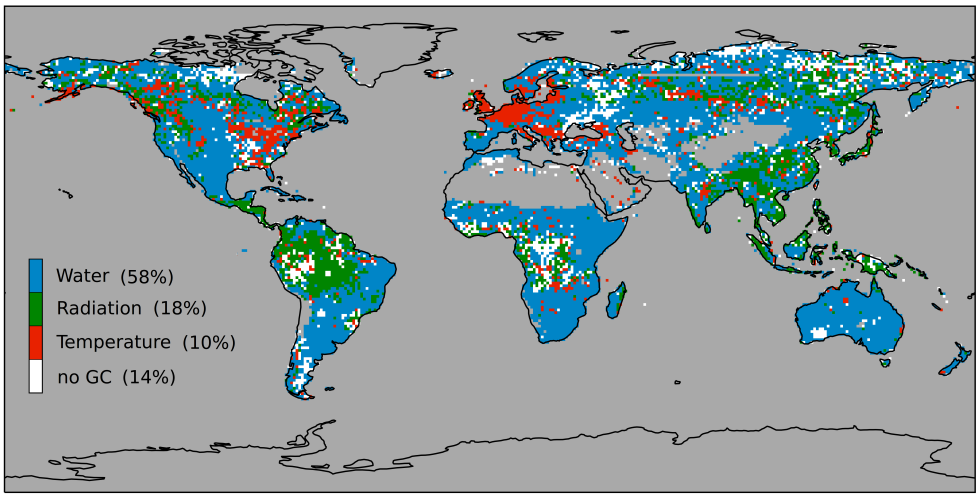
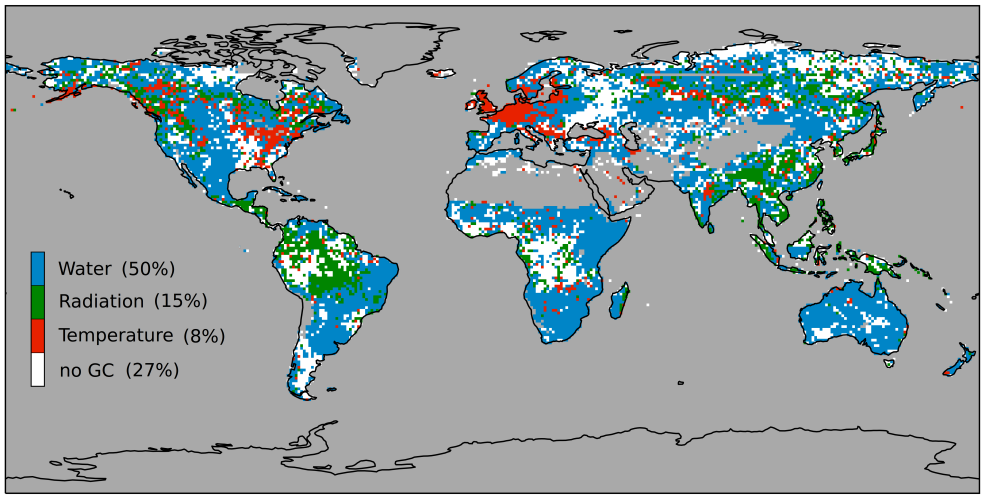



Figure B.5: Average annual burned fraction of vegetation and fuel consuption, Adopted from GFED4, Retrieved August 16, 2018 from `https://www.globalfiredata.org/figures.html`

Total GC ≥ 0.05




Total GC ≥ 0.10




Total GC ≥ 0.15

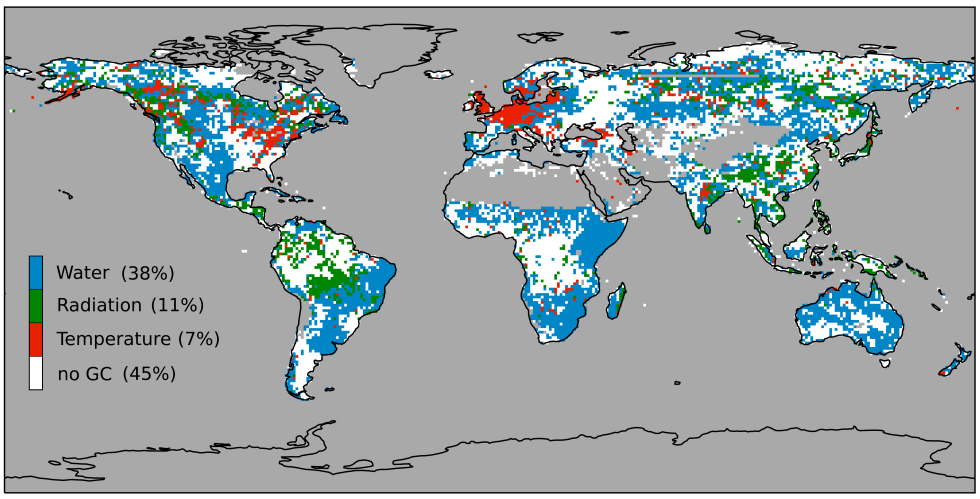



Figure B.6: Group with the highest Granger causality (GC) for every pixel in a 1° x 1° monthly resolution. Different cut-offs are set for the minimum total Granger causality, the "no GC" label is used for pixels below that cut-off