Department of Data Analysis and Mathematical modelling

A data-guided insight into global climate–vegetation dynamics

Christina Papagiannopoulou

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (Ph.D.) of Applied Biological Sciences

Academic year 2017-2018

Supervisors:	Prof. dr. Willem Waegeman Department of Data Analysis and Mathematical Modelling Ghent University, Belgium
	Prof. dr. Diego Miralles Department of Environment Ghent University, Belgium
Examination committee:	Prof. dr. ir. Geert Haesaert (Chairman)
	Prof. dr. ir. Pieter De Frenne Dr. Miguel Mahecha Prof. dr. ir. Francis Wyffels Prof. dr. ir. Jan Verwaeren Dr. Matthias Demuzere
Dean:	Prof. dr. ir. Marc Van Meirvenne
Rector:	Prof. dr. ir. Rik Van de Walle

M.Sc. Christina Papagiannopoulou

A data-guided insight into global climate–vegetation dynamics

Thesis submitted in fulfilment of the requirements for the degree of

Doctor (Ph.D.) of Applied Biological Sciences

Academic year 2017-2018

Dutch translation of the title:
Een data-gedreven inzicht in globale klimaat–vegetatie interacties,

Please refer to this work as follows:
C. Papagiannopoulou (**2018**). A data-guided insight into global climate–vegetation dynamics, PhD Thesis, Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisors Prof. Willem Waegeman and Prof. Diego Miralles for their support, patience and guidance all these years. Without them this dissertation would not have been possible. I would also like to thank Matthias Demuzere and Stijn Decubber for the time that we spent on our long brainstormings. Thanks also to Prof. Niko Verhoest, Mathieu Depoorter and Wouter Dorigo for the cooperation and fruitful discussions. This thesis is really a result of a team effort.

I would also like to express my gratitude to the members of my Ph.D. thesis committee, Prof. Francis Wyffels, Prof. Jan Verwaeren, Prof. Pieter De Frenne, Dr. Miguel Mahecha, Dr. Matthias Demuzere, Prof. Geert Heasaert for the time and effort they dedicated to reading my thesis and providing valuable feedback. During these years I had the opportunity to tutor four master students. I would also like to thank them for our cooperation; thanks goes to Anouk, Stijn, Xingxi, Ruben. A sincere thank you also goes to Prof. Grigorios Tsoumakas, who was my advisor during my master thesis, for encouraging me to pursue graduate studies and for his advice about my plans.

I have been truly lucky to become colleague and friend of many talented people in the department. I would like to thank Bram, Chaim, Kamal, Demir, Ann, Joris, Stijn, Mengzi, Wouter, Gang, Hilde, Marc, Michael, Juan Pablo, Prof. De Baets and the rest of the department for the friendly atmosphere. Thanks to Michiel and Marlies for the nice time we shared during conferences abroad. Tinne for the delicious desserts. Steffie for her support and help during this stressful last period of my Ph.D. Timpe and Ruth for their support for all the administration issues and even more. Jan for helping me with the technical stuff. Jim for our discussions about neural networks. Peter for his positive thinking. Special thanks to Jose for his friendship (not for the spicy candies!) during these years. David for our nice scientific (or not) discussions. Andreia and Niels for hosting and organizing cool events. Aisling for her nice tips and support. Bac for being my office buddy the last year. Raul and Laura for their positive attitude. Thank you all for making my stay in Belgium unforgettable. I would also like to thank our visitors Gisele, Shunyun, Wenwen, Yuanyuan, Giacomo, Giulio, Nubia, Tiago, Pamela, Han, Hasan, Hanzel for the great moments that we have shared.

Thanks to the Greek community, my stay in Ghent felt even more like home. I would like to thank Grigoris, Vasilis M., Anouk, Stavros, Eleni, Foivos, Andreas, Maria, Dimitris, Vasilis Ch. and Takis for this. Many thanks go to my good friends from Greece, Eirini and Niki. Even though we only meet once or twice per year, I know that we think and care for each other a lot.

I would like to thank from the bottom of my heart my family back in Greece; my parents, Prodromos and Fani, my sister Eirini and my brother-in-law Panagiotis. This dissertation was made possible due to your infinite and unconditional love, support and encouragement. Last but not least, I would also like to warmly thank Giannis for being in my life. His support, patience and care during the writing of this dissertation and beyond are immeasurable. Thank you for everything you have brought to my life.

Thank you.

Christina
Ghent, Fall 2018

Contents

Summary

Climate research contributes to the direction of understanding the complex climate system. Research questions are commonly related to either climate projection or climate change attribution. Climate projection aims at forecasting future states of the variables in the various ecosystems, typically over the next decades. On the other hand, climatic attribution studies focus on identifying and quantifying causal relationships between climate variables and natural or anthropogenic factors (e.g., fires, deforestation). Standard modelling approaches in the field of climate science involve the use of mechanistic climate models. Climate models consist of sets of equations and derivations that mathematically represent climate systems. This kind of models rely on prior knowledge and physical laws and they do not directly take observational data into account (i.e., data coming from satellites and/or *in situ* measurements). On the contrary, due to the ever increasing amount of observational data, data-driven models become more and more popular in the field of climate science. Data-driven models are not based on conceptual information nor predefined hypothesis; they are applied directly to the data. Their main goal is to explore the data and discover (or confirm) knowledge related to the climate system.

In this thesis, the relationship between climate and vegetation is investigated by using data-driven approaches. Specifically, methods coming from the fields of machine learning and data mining are introduced in order to model complex relationships between climatic variables, such as temperature, precipitation, radiation, and vegetation. Our work focuses on understanding climate–vegetation interactions due to the crucial role of vegetation, which characterizes the different ecosystems. Therefore, by investigating vegetation, one can measure the response of a given ecosystem to the climate variability. Hence, a better understanding of the relationship of climate–vegetation dynamics can lead to a better understanding of the effect of the projected climate change on the different ecosystems. To this end, methods that are able to model cause-effect relationships between variables can be applied. The use of these methods is investigated throughout this dissertation. Given this general view, the following research objectives are outlined:

1. Climate–vegetation interactions are characterized by complexity and thus, they are highly non-linear. Because of that, a first research objective is the development of a causality framework that takes into account these non-linearities, by extending existing methods and by incorporating machine learning algorithms into the new framework.

2. A second research objective concerns the physical interpretation of the analysis obtained by the proposed framework. This analysis allows for investigating

(i) the effect of climate on global vegetation, (ii) the most important climatic drivers for each region, (iii) the role of extremes in different ecosystems and (iv) the lagged effect of the climate on vegetation.

3. In the third research objective, we aim for developing an approach that detects coherent regions with similar climate–vegetation dynamics. For the first two research objectives, the analysis is performed for each location separately, without considering any spatial interaction between the different locations. In the third research objective, the spatial interactions between the different regions are investigated.

4. For the last objective, we analyze the effect of climate on 'browning events', i.e., periods of anomalously low vegetation greenness. In particular, we aim to detect regions where vegetation response is sensitive to climate extremes.

Chapters 3 to 8 tackle these four research objectives. After a general introduction about basic concepts and approaches in machine learning (Chapter 2), we start Chapter 3 by discussing the data sets used in this study. Since we apply a data-driven approach to investigate climate–vegetation dynamics, the construction of the database is of great importance. The database consists of the most important climatic variables. A variable that measures vegetation greenness is also included. Specifically, in this chapter, we provide the data resources and we describe the preprocessing steps followed before their use. In addition, we describe the construction of extreme indices and other features from the raw data, which encode complex patterns, forming a more expressive representation. We also perform an exploratory pre-analysis on the created database. Subsequently, in Chapter 4, we introduce our novel non-linear Granger-causality framework, which is applied on the climate–vegetation database. This framework extends traditional linear Granger-causality approaches by using more complex (non-linear) machine learning algorithms. Our results indicate that this approach is able to detect non-linear relationships between climate and vegetation that are much less visible with other simpler (and linear) approaches.

An application of the proposed framework is presented in Chapter 5. In this application, we focus on the importance of each climatic variable separately in order to find the most important climatic factor with respect to vegetation for each region. We find that in most of the global vegetated surface, water availability is the most important driver for vegetation. Our results also reveal a prolonged effect of water-related variables on vegetation. Meanwhile, the impacts of temperature and radiation are shown to be more immediate, indicating a higher resilience of vegetation to these factors. Concerning the impact of hydro-climatic extreme events (e.g., extremes in precipitation and temperature), even though they are infrequent by definition, we find that they do have an impact on vegetation variability during the study period, particularly in water-limited ecosystems.

In Chapter 6, we explore the spatial coherence of the response of vegetation to

climate. Our goal is to detect regions with similar climate–vegetation dynamics. To this end, we apply a multi-task learning technique by considering the different locations as different tasks. This approach models the global spatio-temporal data set in a multi-task learning setting without taking into account any prior knowledge about the similarity between the different tasks. Therefore, the spatial structure is learned in a purely data-driven way. We also combine this technique with a clustering algorithm in order to form regions where vegetation responds to climate in a similar way. Experimental results using our global observation-based data sets indicate that our method is able to identify regions of coherent climate-vegetation interactions, which agree well with the expectations derived from traditional global land cover maps. These regions, called 'hydro-climatic biomes', can be used in other applications, such as the exploration of the anomalous behaviour of specific ecosystems in response to climate extremes.

This last potential application is addressed in Chapter 7. We begin Chapter 7 by discussing the different definitions of browning events, which may constitute anomalous behaviour in response to climate extreme events. These definitions are applied directly to the vegetation data streams. As there are various definitions of this kind in the literature, we discuss their possible limitations and we propose some alternatives. In addition, we extend the non-linear Granger-causality framework, introduced in Chapter 4, in order to investigate the response of vegetation extremes to climate. The main conclusions of this chapter mostly include the benefits/limitations of the various modelling settings. The physical interpretation of the results will be the subject of future investigations. In the same direction, Chapter 8 elaborates on the same problem of understanding vegetation extremes with the use of time series classification algorithms. Specifically, we examine the potential of time series classification algorithms to automatically extract informative patterns from the time series.

Finally, in Chapter 9, we summarize the general conclusions and present some ideas for future research.

Nederlandse samenvatting

Klimaatonderzoek draagt bij tot het begrijpen van het complexe klimaatsysteem. Onderzoeksvragen in die context gaan meestal over klimaatprojecties en de toeschrijving van klimaatverandering. Klimaatprojecties streven er naar om, meestal in de volgende decennia, de waarden van variabelen van verschillende ecosystemen te voorspellen. Anderzijds concentreren studies voor attributie zich op de identificatie en de kwantificatie van de betrekkingen tussen klimaatvariabelen en antropogene factoren (zoals bosbranden en ontbossing). Traditionele benaderingen voor het modelleren maken gebruik van mechanistische klimaatmodellen. Deze modellen bestaan uit een verzameling van vergelijkingen en afleidingen die klimaatsystemen wiskundig voorstellen. Ze steunen op voorkennis en natuurkundige wetten, maar ze houden slechts indirect rekening met waargenomen data, zoals de data van satellieten en *in situ* metingen. Data-gebaseerde modellen gebruiken daarentegen waarnemingen, zonder zich te baseren op conceptuele informatie of vooraf gedefinieerde hypothesen. Door de steeds toenemende hoeveelheid verzamelde waarnemingen, worden data-gedreven modellen steeds populairder in klimaatwetenschap.

Dit werk concentreert zich op het begrijpen van de wisselwerkingen tussen klimaat en vegetatie, vanwege de belangrijke rol die vegetatie speelt in het karakteriseren van verschillende ecosystemen. Door het bestuderen van vegetatie kunnen we de reactie van een bepaald ecosysteem meten wanneer het klimaat verandert. Dus, een beter begrip van het verband tussen klimaat en vegetatie kan leiden tot een beter van het effect van de voorspelde klimaatverandering op verschillende ecosystemen. Hiertoe kunnen we modellen gebruiken die de oorzaak-gevolg wisselwerkingen tussen variabelen modelleren. In dit doctoraat bestuderen we het verband tussen klimaat en vegetatie door het gebruik van machine learning methoden. In het bijzonder introduceren we methoden om de ingewikkelde verbanden tussen vegetatie en klimaatvariabelen, zoals temperatuur, neerslag en straling te modelleren. Meer specifiek bestuderen we de volgende onderzoeksdoelen:

1. De wisselwerkingen tussen klimaat en vegetatie zijn ingewikkeld, dus ze zijn ze niet-lineair. Ons eerste doel is de ontwikkeling van een wetenschappelijk kader voor causaliteit, rekening houdend met niet-liniearitеiten, door het uitbreiden van bestaande modellen en door het gebruik van machine learning algoritmes.

2. Een tweede onderzoeksdoel gaat over de fysieke interpretatie van de analyse die we in het voorgestelde kader hebben bekomen. Met de hulp van deze analyse kunnen we de volgende punten onderzoeken: (a) het effect van klimaat op de globale vegetatie, (b) de belangrijkste klimaatfactoren voor elke regio, (c) de rol van de extreme waarden in verschillende ecosystemen, en (d) het

vertraagde effect van het klimaat op de vegetatie.

3. In het derde onderzoeksdoel streven we er naar om een aanpak te ontwikkelen die coherent de interactie tussen klimaat en vegetatie kan opsporen. Voor de eerste twee onderzoeksdoelen wordt de analyse afzonderlijk voor elke locatie uitgevoerd, zonder de ruimtelijke wisselwerkingen tussen de verschillende locaties te beschouwen. In dit onderzoeksdoel worden de ruimtelijke wisselwerkingen tussen de verschillende locaties onderzocht.

4. Voor het laatste doel analyseren we het effect van klimaat op 'browning' gebeurtenissen. 'Browning' gebeurtenissen zijn periodes met abnormale lage vegetatiegroenheid. In het bijzonder streven we ernaar om regio's op te sporen waar de reactie gevoelig is voor klimaatextremen.

Hoofdstukken 3 tot 8 pakken deze vier onderzoeksdoelen aan. In Hoofdstuk 2 geven we een overzicht van enkele fundamentele concepten met betrekking tot machine learning. In Hoofdstuk 3 bespreken we de gegevens die we voor ons onderzoek gebruiken. Doordat we een datagebaseerde aanpak toepassen om de dynamica tussen klimaat en vegetatie te bestuderen, is de constructie van de gegevens van groot belang. De gegevens bestaat uit de belangrijkste klimaatsvariabelen. Een variabele de groenheid van vegetatie beschrijft is ook opgenomen. Bovendien beschrijven we de constructie van extreme indices en andere kenmerken van de ruwe data. Wij voeren ook een verkennende analyse uit op de gecreëerde dataset. Vervolgens introduceren we in Hoofdstuk 4 een nieuw niet-lineair kader voor Granger-causaliteit. Dit kader breidt traditionele lineaire Granger-causaliteit uit door het gebruik van complexe (niet-lineaire) machine learning algoritmen. Onze resultaten tonen aan dat deze aanpak niet-lineaire relaties tussen klimaat en vegetatie kan opsporen. Deze relaties zijn minder zichtbaar als we traditionele lineaire modellen gebruiken.

Een toepassing van het voorgestelde kader wordt in Hoofdstuk 5 gepresenteerd. In deze toepassing besteden we afzonderlijk aandacht aan het belang van elke klimaatvariabele, zodat we voor elke regio de belangrijkste klimaatfactor voor vegetatie kunnen vinden. Wij tonen aan dat, in de meeste gebieden met normale begroeiing, de beschikbaarheid van water de belangrijkste factor voor vegetatie is. Onze resultaten tonen ook aan dat er een verlengd effect van watergerelateerde variabelen op vegetatie bestaat. De impact van temperatuur en straling daarentegen is onmiddellijker en snel uitdovend over tijd. Dus vegetatie heeft een grotere weerstand tegen deze factoren. In het geval van de impact van extreme hydrologische gebeurtenissen (extrema in neerslag en temperatuur), illustreren we dat deze gebeurtenissen een impact hebben op de veranderlijkheid van de vegetatie tijdens de onderzoeksperiode, hoewel deze gebeurtenissen zeldzaam zijn.

In Hoofdstuk 6, onderzoeken we de ruimtelijke samenhang van de reactie van vegetatie ten gevolge van klimaat. Ons doel is om regio's met vergelijkbare klimaat-vegetatie interacties op te sporen. Hiervoor passen we multi-task learning

methoden toe, door de verschillende locaties te zien als verschillende taken. Onze aanpak modelleert de ruimtelijk-temporele dataset zonder rekening te houden met enige voorafgaande kennis over de gelijkheid tussen de verschillende taken. De ruimtelijke structuur wordt dus geleerd puur op een data-gebaseerde manier. We combineren deze techniek met een clusteralgoritme, zodat we regio's kunnen vormen waar klimaat-vegetatie interacties vergelijkbaar zijn. Experimentele resultaten, die bekomen werden door het gebruik van onze wereldwijde data, tonen aan dat onze methode regio's met coherente wisselwerking tussen klimaat en vegetatie kan opsporen. Deze regio's, die 'hydro-climatic biomes' genoemd worden, kloppen met de verwachtingen die bekomen worden via traditionele globale vegetatiekaarten en klimaatzone's. Ze kunnen in andere toepassingen worden gebruikt, zoals het onderzoek van het abnormaal gedrag van bepaalde ecosystemen als reactie op klimaatextrema.

Deze toepassing wordt in Hoofdstuk 7 bediscussieerd. We bespreken de verschillende definities van 'browning' gebeurtenissen die abnormaal gedrag als reactie op klimaatextremen vertegenwoordigen. Deze definities worden direct toegepast op data van vegetatie. Doordat er verschillende definities in de literatuur bestaan, bespreken we hun mogelijke beperkingen en we stellen een aantal alternatieven voor. Verder breiden we ons niet-lineair Granger-causaliteitskader uit Hoofdstuk 4 uit om de reactie van extrema in vegetatie op het klimaat te bestuderen. De hoofdconclusie van dit hoofdstuk handelt over de voordelen en beperkingen van de modellen. De fysieke interpretatie van de resultaten is een piste voor toekomstig onderzoek. Daarnaast wordt dit hoofdstuk uitgebreid met algoritmes die automatisch informatieve patronen uit tijdreeksen halen.

Finaal vatten we in Hoofdstuk 9 enkele algemene conclusies samen en presenteren we ideeën voor toekomstig onderzoek.

List of symbols

Defined in Chapter 2:

$\mathcal{X}$	SPACE OF INPUT OBJECTS
$\boldsymbol{x}$	VECTOR OF FEATURES
d	DIMENSION OF INPUT SPACE
$\mathcal{Y}$	SPACE OF OUTPUT OBJECTS
$D = \{(\boldsymbol{x_1}, y_1), ...\}$	DATA SET OF N OBSERVATIONS
K	NUMBER OF FOLDS
N	NUMBER OF OBSERVATIONS
$D - D_k$	DATA SET MINUS THE OBSERVATIONS OF THE K-TH FOLD
$f(\boldsymbol{x}; D)$	MODEL PREDICTION FOR THE OBSERVATION $\boldsymbol{x}$
$I : \{F, T\} \rightarrow \{0, 1\}$	INDICATOR FUNCTION
$\boldsymbol{w}$	VECTOR OF PARAMETERS
ϵ	ERROR TERM
i, k, m, b, t	INDICES
λ	REGULARIZATION PARAMETER
$\|\|.\|\|$	L2-NORM OF A VECTOR
$\|\|.\|\|_1$	L1-NORM OF A VECTOR
Pr	PROBABILITY MEASURE
$\mathcal{L}$	LOSS FUNCTION
ϕ	BASIS TRANSFORMATION
d^*	NUMBER OF BASIS FUNCTIONS
P_i	DATA PARTITION i
c_i	CONSTANT RESPONSE FOR DATA PARTITION i
B	NUMBER OF DECISION TREES
$\boldsymbol{z}$	TIME SERIES $\boldsymbol{z}$
τ	CONSTANT
P	LAG TIME
μ	MEAN
σ^2	VARIANCE

Defined in Chapter 3:

t	TIME INDEX
$\boldsymbol{y}_t$	TARGET TIME SERIES
$\boldsymbol{x}_t, \boldsymbol{w}_t$	PREDICTOR TIME SERIES
$\boldsymbol{y}_t^{Tr}$	TREND
$\boldsymbol{y}_t^{S}$	SEASONAL COMPONENT
$\boldsymbol{y}_t^{R}$	RESIDUALS
k	NUMBER OF MONTHS
w_0	INTERCEPT OF LINEAR REGRESSION
w_1	SLOPE OF LINEAR REGRESSION

Defined in Chapter 4:

t	time index
$\boldsymbol{y}$	target time series
$\boldsymbol{x}, \boldsymbol{z}$	predictor time series
y_t	value at timestamp t
N	length of time series
i, p, j	indices
R^2	coefficient of determination
P	length of lag-time window
$\hat{\boldsymbol{y}}$	predicted time series
$\bar{y}$	mean value of time series $\boldsymbol{y}$
w_{ij}	parameter values
ϵ_i	error term
H_0	null hypothesis

Defined in Chapter 6:

L	number of locations
d	number of predictors
N	length of time series
i, l	indices
$\boldsymbol{X}^{(l)}$	input feature vectors of location l
$\boldsymbol{y}^{(l)}$	target time series of location l
D	spatio-temporal data set
$D^{(l)}$	data set of location l
$f^{(l)}(.)$	linear regression model for task l
$\boldsymbol{x}_i^{(l)}$	observation of task l
$\boldsymbol{w}^{(l)}$	weight vector for task l
$\mathcal{L}$	loss function
Ω	regularization term
$\boldsymbol{I}$	identity matrix
Θ	low-dimensional feature map
h	dimensionality of shared feature space
$\boldsymbol{u}^{(l)}$	high-dimensional weight vector for task l
$\boldsymbol{v}^{(l)}$	low-dimensional weight vector for task l
$\lambda^{(l)}$	regularization parameter for task l
$\boldsymbol{W}$	weight matrix with vectors $\boldsymbol{w}^{(l)}$ as columns
$\boldsymbol{V}$	weight matrix of low-dimensional space
$\boldsymbol{V_1 D V_2^T}$	SVD decomposition

List of acronyms

ARIMA	Autoregressive Integrated Moving Average
ASO	Alternative Structure Optimization
AUC	Area Under the Curve
CID	Complexity Invariant Distance
CMAP	CPC Merged Analysis of Precipitation
CMTL	Clustered Multi-Task Learning
CPC-U	Climate Prediction Center Unified analysis
CRU	Climate Research Unit
DM-test	Diebold-Mariano test
DTD	Derivative Transform Distance
DTW	Dynamic Time Warping
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA	ECMWF Re-Analysis
ESA CCI	European Space Agency's Climate Change Initiative
ESM	Earth System Models
ETCCDI	Expert Team on Climate Change Detection and Indices
fAPAR	fract. of Absorbed Photosynthetically Active Radiation
GEWEX	Global Energy and Water cycle Exchanges
GIMMS	Global Inventory Modelling and Mapping Studies
GISS	Goddard Institute of Space Studies
GLEAM	Global Land Evaporation Amsterdam Model
GPCC	Global Precipitation Climatology Centre
GPCP	Global Precipitation Climatology Project
IGBP	International Geosphere-Biosphere Program
IPCC	Intergovernmental Panel on Climate Change
ISCCP	International Satellite Cloud Climatology Project
LASSO	Least Absolute Shrinkage and Selection Operator
LST	Land Surface Temperature
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
MLOST	Merged Land-Ocean Surface Temperature
MSM	Move-Split-Merge distance
MSWEP	Multi-Source Weighted-Ensemble Precipitation
MTL	Multi-Task Learning
NASA	National Aeronautics and Space Administration

NDVI	Normalized Difference Vegetation Index
PCA	Principal Component Analysis
RF	Random Forests
RSS	Residual Sum of Squares
SRB	Surface Radiation Budget
STL	Single-Task Learning
SVD	Singular Value Decomposition
TSS	Total Sum of Squares
UDel	University of Delaware
VAR	Vector Autoregressive model
VARMA	Vector Autoregressive Moving Average
1-NN	one Nearest Neighbour

List of Figures

List of Tables

1 Introduction

'What is needed is the development of data-driven methodologies that are guided by theory to constrain search, discover more meaningful patterns, and produce more accurate models.'

Faghmous and Kumar (2014)

1.1. Assessing causality in geosciences

The field of geosciences consists of various disciplines such as biology, geology, hydrology, geophysics, ecology and aims to understand the complex and dynamic system of our planet. The term 'geosciences' is often conflated with the arguably broader 'climate science'. In this thesis we use these terms interchangeably, since both fields focus on the study of the Earth system. The importance of a better understanding of Earth's system is crucial, since there are several problems related to humanity that require solutions. Some of these problems include research questions that are also related to climate. Climate research studies contribute to the direction of understanding and addressing these problems by facing challenges which are related to either climate projection or climate change attribution. Climate projection or forecasting aims at predicting the future state of the climatic system, typically over the next decades. The goal of climatic attribution, on the other hand, is to identify and quantify cause-effect relationships between climate variables and natural or anthropogenic factors. Particularly, one can think of specific challenges, such as the effect of human greenhouse gas emissions on global temperature, the prediction of water and food availability, the detection of factors responsible for extreme events, etc. (Karpatne et al., 2017).

With the development of satellite and sensor technology, climate science has become one of the most data-rich domains. Thus, machine learning and data mining techniques are able to contribute in the direction of unraveling complex relationships in climate, by extracting useful information from the data, modelling important variables and providing tools for causal inference (Lary et al., 2016; Srivastava et al., 2017). However, the use of these techniques on geoscientific data is not straightforward, since this kind of data is commonly organized in spatio-temporal structures and characterized by high auto-correlation, non-stationarity, non-linearity, small sample size, incompleteness and uncertainty. In addition, climate data are characterized by heterogeneity due to the fact that the raw data are coming from various resources in multiple spatial and temporal resolutions and formats. These challenges become a barrier to most data science techniques, in

which assumptions such as independence or stationarity are held. On top of these challenges, the general idea behind data science techniques is not directly applicable in geoscientific applications, since phenomena such as a hurricane cannot be represented by a single observation in a data set. Therefore, collecting, preprocessing and forming a database that can be handled by data science methods is already a big challenge. However, we should stress that in climate science, there are research questions in which machine learning and data mining techniques offer the means to provide the answers. In light of this rationale, these means should be used in combination with the prior domain knowledge and the laws underlying climate processes. That way, these methods are able to produce physically interpretable results leading to significant contributions to our understanding of the climate system (Faghmous and Kumar, 2014).

1.1.1. Approaches based on climate models

The standard modelling approach in the field of climate science is based on simulation studies with mechanistic climate models (IPCC, Intergovernmental Panel on Climate Change, 2007; Moss et al., 2008). Climate models provide a huge amount of data by either simulating future climate or reconstructing past climate. This kind of models is based on conceptual representations of the global water, atmospheric and biological systems, mathematically formalized through complicated differential equations. In the simulation process, observational data are used only for initialization purposes. Since the data are produced by simulations, the outcomes are continuous, without gaps in space and time. However, they highly depend on the different parameters and initializations. In addition, the model outputs suffer from uncertainties, due to incomplete physical understanding of certain processes. To reduce the variance and the uncertainties of a single model, an ensemble approach is commonly used. This approach is based on the averaged result of multiple models, which are initialized in various conditions with different parameter values.

To illustrate the way that physically-based frameworks are used for climate attribution studies, we present the basic steps of a case study. Let us focus on disentangling the effect of anthropogenic greenhouse emissions on the occurrence of a particular climatic event, e.g., a hurricane. At a first step, the climate model runs without the use of information about the emissions. Then, the climate model runs again, considering the emissions. Finally, the difference in likelihood of occurrence of the climatic event in the two scenarios (with and without the emissions) is assessed. That way, one can draw conclusions such as, in our case, that the CO_2 emissions doubled the chance of experiencing a storm. The same framework is commonly followed to attribute climate trends. As such, conclusions which involve, e.g., that the probability of temperature increase based on a particular factor, are also possible.

1.1.2. Data-driven approaches

In contrast to concept-based models that rely on strong assumptions (i.e., prior knowledge, physical laws), data-driven models do not follow any kind of predefined hypothesis for the climate system. Specifically, the relationships between different variables are modeled by learning functions based on observational data. Thus, these statistical and machine learning models are directly applied on the input data. Note that in data-driven approaches, the underlying assumptions of every modelling approach, i.e., linear – non-linear assumptions, representation of reality based on data etc., are still valid. As already mentioned above, there are recent improvements in the field of satellite and *in situ* technology resulting in an ever increasing amount of fine resolution input data. As such, data-driven models can exploit the available data sets to answer research questions related to climate. The main challenges remain (a) the adaptation of data-driven models to the complexity of the climate data sets, as well as (b) the interpretability of their results. In this dissertation, we focus on the choice of data-driven models in climate science.

Similar research questions to the one described in Sect. 1.1.1 have been addressed by a large number of recent studies with the use of data-driven approaches. In these studies, simple linear regression models have been commonly used to model relationships between variables (Attanasio, 2012; Attanasio et al., 2012). However, recently, with the development of more complicated algorithms, other non-linear methods (such as neural networks) have been applied to climate data (Pasini et al., 2017; Attanasio and Triacca, 2011). Particularly, in this dissertation, we investigate the relationship between climate and vegetation by using machine learning methods. We introduce approaches that are able to model complex relationships between climatic variables, such as, temperature, precipitation, radiation, and vegetation. Our work is focused on understanding the climate–vegetation interactions due to the crucial role of vegetation in the different ecosystems (Bonan, 2008; Nemani et al., 2003). This is due to the fact that by investigating vegetation, one can measure the way that the different ecosystems respond to climate variability. Hence, a better understanding of climate–vegetation interactions can lead to a better understanding of the effect of climate change on the ecosystems. This valuable information can be extracted by methods that are able to model cause-effect relationships between variables. This kind of methods is used throughout this thesis.

1.1.3. Causal inference

The goal of statistical causality is to understand complex systems (i.e., climate dynamics, biological systems) by using observational data to detect causal relationships between variables. Although the definition for causality seems rather intuitive, it is a complex concept that can be mathematically formulated. However, this kind of mathematical models cannot be directly applied without first introducing some

strong assumptions. One of the most important assumptions is causal sufficiency. According to causal sufficiency, all the necessary observational data are included in the analysis (i.e., there are no hidden causes or other unobserved relevant variables). This implies that the analysis for a particular causal relation is subject to the inferred variables. Any hypothetical change in the observed variables (e.g., addition of a newly-observed variable) affects the resulting causal relationships, since other possible causes can emerge. In climate studies, it comes natural that the causal sufficiency assumption is never satisfied. For instance, climate systems are highly complex and thus there might be hidden interactions due to unobserved variables. Therefore, conclusions can be drawn based on the analysis of specific data sets. Note that expert view/knowledge on the field is necessary for the evaluation of the causal relationships that can be detected. This means that causal frameworks can confirm relations that can be physically explained, or detect new hypothetical relations, which should be further investigated. Yet, new undiscovered relations can also be identified, contributing to our knowledge about the complex climate system.

In the previous section (Sect. 1.1.2), we stress that discovering relationships between climate variables is of a great importance. Instead of just detecting correlations among climate variables, research questions also include the investigation of causal relationships between them. Arguably, the most commonly used approaches for detecting causality are: (i) Granger causality (Granger, 1969) (Granger was awarded the Nobel prize for this method), (ii) probabilistic graphical model approaches (Koller and Friedman, 2009), (iii) independence-based methods (e.g., see Runge et al. (2017)) and (iv) non-linear state-space methods (e.g., see Sugihara et al. (2012)). Each of the aforementioned approaches has its own advantages and limitations. For instance, Granger causality (Granger, 1969) is based on the predictive performance of the involved variables and is defined under the following assumptions: (a) The past and the present might cause the future but the future cannot cause the past and (b) there is no redundant information in the examined system (e.g., there is no need including both degrees in Fahrenheit and Celcius for a temperature variable). In climate studies, Granger causality has been used in the bivariate and multivariate setting (Triacca, 2005; Attanasio, 2012). For instance, Sun and Wang (1996) analyzed time series of global CO_2 emissions and global temperature anomalies, concluding that there is a positive cause-effect relation between them. The main strengths of Granger-causality approaches include the simplicity, scalability and interpretability of the method, whereas the main limitations are due to the causal sufficiency assumption and the causal loop between the involved variables. Probabilistic graphical models have also been used in detecting causal relations (Koller and Friedman, 2009). They use a graphical representation of the causes between the variables, in which the nodes represent the variables and the edges represent the relations between them. In climate studies, this kind of models has rarely been used, due to their increased complexity, espe-

cially for large scale data (Ebert-Uphoff and Deng, 2012). However, there are some climate applications in which probabilistic graphical models have been successfully used, see e.g., Cano et al. (2004); Monteleoni et al. (2011). In the category of independence-based approaches, there are methods which combine concepts from both categories described above. For instance, Runge et al. (2017) estimate the causal time series graph using conditional independence tests for time series, while Sun et al. (2015) investigate the usage of causation entropy as a measure to discover the direct parents of a given node. In climate science, these approaches have been only recently applied (Runge et al., 2014; Yi and Imme, 2014). Note that these approaches also suffer from the causal loop between the involved variables. Finally, non-linear state-space methods (Sugihara et al., 2012) assess causation by using the convergent cross-mapping method, which tests whether the historical records of one variable can reliably reconstruct the values of another variable. Examples of applications in climate science that use this kind of approaches include the studies of Van Nes et al. (2015); Ye et al. (2015). Scalability to high-dimensional data is the main issue of these approaches. Undoubtedly, there are other methods which have been used in discovery of causal relationships. Other approaches include the use of penalization methods (Lozano et al., 2009a; Shojaie and Michailidis, 2010) and other causality algorithms (Spirtes et al., 2000).

1.2. Overview of the research objectives and achievements

In this thesis, we investigate the relationship between climate and vegetation by using data-driven methods. Specifically, we combine the concept of Granger causality with feature construction techniques and machine learning algorithms to investigate climate–vegetation interactions based on spatio-temporal observational data at global scale. With the proposed framework, we are able to detect the main vegetation drivers of each region and/or to assess the importance of climatic extremes and lagged-values of climate variables for vegetation. We also explore the spatial coherence of the response of vegetation to climate and the occurrence of global vegetation extremes (browning events). Finally, in the last part of the thesis, we study the link between climate and browning events detected on observational data. In particular, we investigate the different definitions of vegetation extremes, existing in the literature, and we introduce some alternatives. Based on these definitions, we reformulate our causality framework in order to explore the main climatic factors that affect browning events in the different regions of the world. To this end, we formulate several research questions to guide our investigations, which are elaborated and motivated in subsequent chapters:

1. Is climate an important factor to determine the state of vegetation? If yes, in which regions does it play a more important role?

2. Which are the most important climatic drivers for each region?
3. Which climate extremes are important for vegetation?
4. Which is the lagged effect of the climatic variables on vegetation?
5. In which regions does vegetation respond to climate in a similar way?
6. How can we define browning events based on vegetation data?
7. In which regions does climate affect vegetation extremes the most?

These research questions are studied through the use of satellite and *in situ* data sets.

The contribution of this thesis is summarized as follows:

- We have developed a Granger-causality framework that combines the steps of data collection and pre-analysis, time series decomposition techniques, feature construction and predictive modelling approaches. This is the first time that a framework, which uses the aforementioned techniques, is applied in the context of understanding climate–vegetation interactions.
- We have also explored the use of multi-task learning modelling to exploit the spatial dependencies on our spatio-temporal climate–vegetation data sets. This modelling approach is also novel with respect to the application domain.
- We have extended the Granger-causality framework in a classification setting in order to investigate the effect of the climatic drivers on the browning events. This is also a new approach in the study of climate–vegetation interactions.
- We have performed a comparative study on the ability of various machine learning algorithms to extract useful patterns from climatic time series. This study is a new approach in the application domain.
- Finally, we have analysed and discussed the results of our analysis with the selected modelling approaches, giving new insights in the understanding of climate–vegetation dynamics.
- Note also that the proposed methodology can be applied, with some modifications if needed, in other application domains to address similar research questions.

1.3. Structure of the thesis

In Chapter 2, we provide some basic background knowledge about machine learning. Specifically, we discuss in more detail some basic concepts and commonly used methods that are also used in this thesis. We then outline some basic terminology

about time series, as climate data collected by sensors have the form of time series. We also describe various machine learning applications in geosciences.

In Chapter 3, we discuss the data sets collected for this thesis. We provide the data resources, as well as their original spatio-temporal resolution, and we describe the preprocessing steps followed before their use. We also explain the construction of extremes indices and other information from the raw data and we perform an exploratory pre-analysis on the created database.

In Chapter 4, we introduce a novel non-linear Granger-causality framework, which is able to detect non-linear relationships between climate and vegetation. We compare the results with traditional Granger-causality methodologies and we experimentally prove that complex relationships can be revealed by the use of more complex predictive modelling techniques.

An application of the framework explained in Chapter 4 is described in Chapter 5. In this application, we focus on the importance of each climatic variable separately in order to find the most important climatic factor with respect to vegetation for each region. We also investigate the importance of climatic extremes and the past variability of the climatic factors for the global vegetation.

In Chapter 6, we explore the spatial coherence of vegetation response to climate. More particularly, we apply a multi-task learning technique which is able to reveal common predictive structures among the different locations. We then describe the way of forming coherent regions based on these characteristic structures.

We begin Chapter 7 by discussing the different definitions of vegetation extremes (i.e., browning events) that exist in the literature. We then propose some alternative definitions of vegetation extremes based on vegetation data. We also extend the non-linear Granger-causality framework introduced in Chapter 4 in order to investigate the relationships between climate and vegetation extremes.

In the same direction, Chapter 8 elaborates the same problem of understanding vegetation extremes with the use of time series classification algorithms. Specifically, we examine the potential of time series classification algorithms to automatically extract informative patterns from the time series.

Finally, in Chapter 9, we summarize the modelling studies described in this thesis, as well as their conclusions, and discuss their implications for the field of climate science. We also outline the promising research avenues opened by the work contained in this thesis.

2 Machine learning background

In this chapter, we present a brief overview of machine learning to introduce the basic concepts and methods used in this thesis. Therefore, the provided explanations are in an introductory level. For a more detailed description of machine learning, the reader is directed to Hastie et al. (2001); Bishop (2007); James et al. (2014). In the next sections, some important concepts of machine learning are discussed: train-test splitting, cross-validation and over-fitting. A basic notation is also defined. Afterwards, some of the most simple, but effective, linear and non-linear machine learning methods are introduced. Then, as climate data are mostly represented as spatio-temporal data sets, basic concepts and modelling techniques for time series and spatial data are described. Finally, some examples of common challenges and machine learning applications in geosciences are presented.

2.1. Introduction

Machine learning is the field of study which focuses on algorithms that are able to learn from data. A learning algorithm can be defined according to a famous statement:

'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.'

Mitchell (1997), *Machine Learning*

This kind of algorithms has been successfully applied on various problems, such as in weather forecasting, computer vision, bioinformatics, etc. But what can one consider as task T, experience E, and performance measure P? In the application of weather forecasting, the task T can be defined as the prediction of temperature (or another climatic variable) for the next day (days) of a specific area, the experience E can be gained by using the daily temperature values of this particular location of the last 30 years and the performance measure P can be specified as the average error between the observed and the predicted temperature value. Considering an application from computer vision, such as the classification of an image to a specific category according to its theme, the task T is the application itself, the experience E is a set of images already classified in a particular category (e.g., by a human annotator) and the performance measure P can be the percentage of the correctly classified images. Therefore, it is obvious that based on the goal of each application, the definition of these three concepts varies, often making these decisions for researchers and practitioners not straightforward.

Based on the way that this experience is gained, machine learning methods are typically divided in three main categories: supervised methods, unsupervised methods and reinforcement learning approaches. In supervised methods, there are always (historical or annotated) data available to train the model on them (such as in the two applications of weather forecasting and image classification mentioned in the previous paragraph). Unsupervised methods are typically used to discover the structure of the data, to cluster the data in different groups based on their characteristics or reduce the dimensionality of the data. Thus, this kind of models is not based on labelled data sets. Finally, reinforcement learning is mainly used in robotics where agents learn to interact with their environment by using a trial and error strategy. In these systems, there is a reward function which indicates whether the moves of an agent led to a successful attempt or not. The methods used in this thesis are supervised and unsupervised algorithms.

2.2. Basic concepts

In supervised learning, algorithms build a mathematical model from the given input data. We symbolize with $\mathcal{X}$ the space of the input data objects, i.e., the space of the representation of the objects. An instance of this space is denoted with $\mathbf{x}$ and it is usually a vector of features, i.e., $\mathbf{x} = (x_1, ..., x_d)$, with d being the dimension of the input space. In addition to this space, in supervised methods there is another space, commonly symbolized as $\mathcal{Y}$, which refers to the experience used by the learning algorithm. Based on the type of the space $\mathcal{Y}$, supervised learning problems are separated into the following main categories: classification and regression problems. In classification problems, the variable to predict is a categorical variable, i.e., each object is classified to one category (or label). A special case of this type of problems is binary classification, where the target variable can take only two values $\mathcal{Y} = \{0, 1\}$, i.e., the data objects can be assigned to one out of the two categories. In regression problems the target variable takes continuous numerical values, i.e., $\mathcal{Y} = \mathbb{R}$. In practice, the instances of a data set are often denoted as pairs of an input vector $\mathbf{x}$ and the corresponding answer label or scalar y, i.e., $(\mathbf{x}, y)$. As such, a data set can be denoted as a set of pairs: $D = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_N}, y_N)\}$, with N being the number of instances (observations).

The input data is commonly split into three parts, namely training set, validation set and test set. Each part is used in different phases of the model development. The model is initially fit on the training data set. In the most common scenario, the trained model contains parameters that should be adjusted based on the data. So in this case, the validation data set is used in order to compare the performances of the models for different parameter values. That way, one can decide about the most appropriate parameter values for the model. Finally, the test data set is used to provide an unbiased evaluation of a final model fit on the training data

set (which also includes this time the validation set). An important note is that the observations in both the test data set and the validation data set should follow the same probability distribution as the training data set. The partition of a given data set to two different parts is known as the hold-out method and common proportions of the total number of instances are 70%/30% for the training and the test set, respectively.

However, when the given data set is small, the hold-out method is not feasible. The solution to this problem is K-fold cross-validation. In K-fold cross-validation, the data set is divided into K parts. In each of the K folds, one part is used for testing, while the remaining parts are used for training. Figure 2.1 illustrates a five-fold cross-validation procedure. The K-fold cross-validation error is calculated as:

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in D_k} I(f(\mathbf{x_i}; D - D_k) \neq y_i) \tag{2.1}$$

where $D - D_k$ is the data set minus the observations of the k-th fold and $f(\mathbf{x_i}; D - D_k)$ is the model prediction for the observations $\mathbf{x_i} \in D_k$. The model f is learned on $D - D_k$ and I is an indicator function, which can be substituted with any other function that measures the deviation from the true values y_i. This error is an unbiased estimate of the error of our learning algorithm when given $N\frac{K-1}{K}$ observations.

Another basic concept in machine learning is the concept of over-fitting, which is related to the ability of the model to generalise well on unseen data. A perfect performance on the training data is useless and easy, since the algorithm can only memorise all the observations in the training set with their corresponding values of the target variable. However, this kind of memorization is not considered as learning, because the model will perform poorly on unseen data. This scenario, where the error on the training data is very low, but the error on unseen data is high, is called over-fitting. On the other hand, under-fitting occurs when the

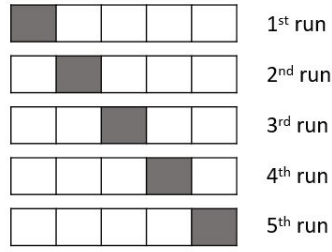


Figure 2.1: Illustration of a five-fold cross-validation procedure. The data set is split into five parts. One part serves as test data set in each iteration, while the other four are concatenated to constitute the training set. The final result is obtained by averaging the performance results of the five runs.

model is too simple to even model the training data. In this case, both the training and the testing errors are high. As the model complexity increases, the model starts to perform well on both the train and test data. However, once the model becomes too complex, the testing error increases, while the training error continues to decrease and the model is over-fitting. This phenomenon is illustrated in Fig. 2.2. In the next chapter, we discuss about a well-known term for controlling the model complexity and a commonly used way to fine-tune this term.

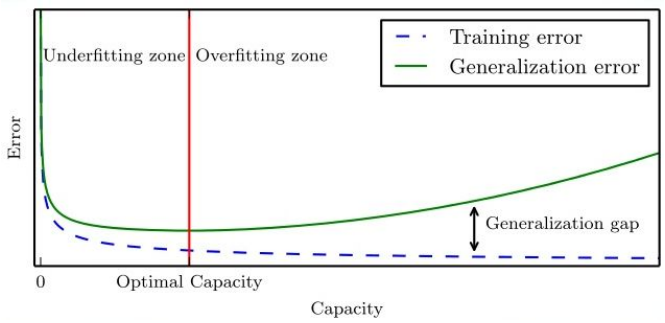


Figure 2.2: The train and the test error as a function of the model complexity (Goodfellow et al., 2016).

2.3. Linear models

2.3.1. Linear regression

We start with the simplest and the most well-known model for regression problems, the linear model. In a linear model, it is assumed that the target variable y can be expressed as a linear combination of the input features $\mathbf{x}$. By denoting with f the function that takes as input a vector $\mathbf{x}$ and gives a predicted value $f(\mathbf{x})$ (denoted also as $\hat{y}$) for the target variable, the linear model is written as:

$$f(\mathbf{x}) = w_1 x_1 + ... + w_d x_d + w_0 = \sum_{i=1}^{d} w_i x_i + w_0, \tag{2.2}$$

with the function $f : \mathcal{X} \to \mathbb{R}$, the vector $\mathbf{w} = (w_1, w_2, ..., w_d)$ being the weight parameters and w_0 being the bias term. By using a simpler notation, the above equation is represented as:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}. \tag{2.3}$$

In this case, the constant bias term can be included in the feature vector $\mathbf{x}$ as an additional constant value of 1 to introduce the model offset, so that $\mathbf{w} = (w_0, w_1, w_2, ..., w_d)$. Note that the underlying relationship between y and $\mathbf{x}$ is

expressed involving an error term ϵ by:

$$y = f(\mathbf{x}) + \epsilon = \hat{y} + \epsilon. \tag{2.4}$$

It is assumed that the errors in the regression are normally distributed.

In order to learn the function f, one should actually learn the weight vector $\mathbf{w} \in \mathbb{R}^d$. This weight vector is learned by using a training data set such as the one described in Sect. 2.2. A commonly used method to do so, is the least squares method. By using as loss function the mean squared error, one solves the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x_i})^2. \tag{2.5}$$

This problem comes with an analytical solution due to the convexity of the objective function.

2.3.2. Ridge regression

As it has been mentioned in Sect. 2.2, over-fitting is a common phenomenon in machine learning, and it is related to the complexity of the model. Specifically, when there are more model parameters than training observations, the model perfectly learns the training set, while it is not capable of making good predictions on the test data set.

A common way for controlling the model complexity is called regularization. Regularization is an extra term that is added to a loss function. This term serves as a penalty to the complexity of the function and often comes with a parameter λ, which is tuned during the training phase. The goal of regularization is to keep a balance between a small error in the training set and over-fitting. In this thesis, we use a well-known regularization method called ridge regression (Tikhonov, 1963). In ridge regression, the mean squared error is used as loss function and the L2-norm of $\mathbf{w}$ is used as regularization term, as it is known that the model complexity in least squares regression is related to the magnitude of the weights. The ridge regression optimization problem is defined as:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x_i})^2 + \lambda ||\mathbf{w}||^2. \tag{2.6}$$

The above optimization problem has two objectives; to learn a function f that well-fits the training observations and keep the function f simple to avoid over-fitting. This trade-off is controlled by the λ parameter. An explanation about the way that the different values of the parameter λ affect the learning model is coming from statistics by using the concepts of bias and variance. A low value of λ leads

to a better fit of the model to the training set, resulting in a more complex model. Thus, the bias of the model is low, since the error in the training set decreases. At the same time, the variance is increasing, as small changes in the training set lead to large changes in the parameter values. On the other hand, a high value of the parameter λ leads to a simpler model that possibly is not expressive enough to model the training data. Therefore, the opposite scenario happens in terms of bias and variance; the bias of the parameters is high, while the variance is low. And when the model is too simple, then its generalization ability is low, leading to low performance on a test data set. As such, we conclude that properly tuning the λ parameter is crucial. The tuning of the parameter is commonly done by using K-fold cross-validation or a validation set.

2.3.3. Least Absolute Shrinkage and Selection Operator (LASSO) regression

LASSO (Tibshirani, 1996) is an alternative (to ridge regression) commonly-used method which penalizes the regression coefficients in order to improve estimation. The difference between the ridge regression and the LASSO comes from the penalty term that each method uses, i.e., LASSO uses the L1-norm of $\boldsymbol{w}$ while ridge regression uses the L2-norm of $\boldsymbol{w}$. The advantage of the L1-penalty term is that it can lead to sparse estimation of regression coefficients, and therefore, LASSO is able to automatically perform variable selection. Mathematically, LASSO regression is formulated as:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x_i})^2 + \lambda ||\mathbf{w}||_1. \tag{2.7}$$

In this thesis, LASSO regression has been used only in some early experiments. However, there are some references in this well-known method throughout the next chapters.

2.3.4. Logistic regression

In the case of classification problems, the data samples are assigned to one of a fixed number of categories (classes). In Sect. 2.2, we introduced a special case of classification problems where the target variable $\mathbf{y}$ can take only two possible values. There are many classification algorithms that have been proposed in the literature for this binary classification problem. These algorithms can be extended into multi-class classification problems where more that two classes are considered. Here, we describe one of the most well-known methods of the field called logistic regression.

Despite its misleading name, logistic regression is a probabilistic approach for

classification, which returns a probability estimate to a given sample for each class. The values of the two classes in the binary classification case are typically notated as 0 (negative) and 1 (positive). That way, one can assess the correctness of the model response based on the following probability measures:

$$P_{i_1} = Pr\{y_i = 1|\mathbf{x_i}\}, \tag{2.8}$$

$$P_{i_0} = Pr\{y_i = 0|\mathbf{x_i}\}, \tag{2.9}$$

for $i = 1, ..., N$. So, one can observe that the equality $P_{i_0} + P_{i_1} = 1$ should hold for each i. Logistic regression represents the log-odds or logit function of the above probabilities as a linear model of the input feature vectors, i.e.,

$$\log\left(\frac{P_{i_1}}{1 - P_{i_1}}\right) = \mathbf{w}^T\mathbf{x_i}, \tag{2.10}$$

in which the model function corresponds to a hyperplane that separates the data samples into the different classes (Cox, 1958).

As in linear regression, a loss function is minimized also for logistic regression. The loss function in logistic regression is called logistic loss and is defined as:

$$\mathcal{L}(f(\mathbf{x}), y) = -yf(\mathbf{x}) + \ln(1 + \exp(f(\mathbf{x}))). \tag{2.11}$$

with $f(\mathbf{x})$ being the model function.

2.4. Non-linear models

2.4.1. Non-linear regression methods

In Sect. 2.3.1, we presented the linear regression model, which is a simple and well-interpreted model. However, the expressiveness of this model is rather limited, since it takes into account only linear combinations of the input variables. In machine learning, a common way of extending a linear method to a non-linear one is by using linear combinations of non-linear functions, applied on the input variables. This simple idea leads to a model formulated as:

$$f(x) = \sum_{m=1}^{d^*} w_m\phi_m(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) \tag{2.12}$$

with $\phi_m : X \to \mathbb{R}$ being non-linear functions usually called as basis functions (Waege-

man, 2009). Considering a set of d^* basis functions, the weight vector $\mathbf{w}$ is a d^*-dimensional vector. These parameters are learned based on the training observations, as in the case of linear regression. A simple way to denote the mapping of the features of a d-dimensional space to a (usually higher) d^*-dimensional space is the following:

$$\phi : X \rightarrow \mathbb{R}^{d^*} \tag{2.13}$$

This new d^*-dimensional space is also called the feature space and the individual values are called features. Thus, one can apply the linear regression model as described in Sect. 2.3.1 on this new feature representation. As such, the model becomes more expressive compared to linear regression on the original feature space, since it can capture possible non-linear relationships in the data. Moreover, the interpretability and the easy implementation of the linear regression model are retained.

In the machine learning literature there are several examples of basis functions that are used in various applications. For a simple example, one can think of the polynomial basis function, in which for a one-dimensional input variable x, the basis function ϕ_m returns a polynomial of degree m, i.e.,

$$\phi_m(x) = x^m \tag{2.14}$$

That way, in the model of Eq. 2.12, polynomials up to degree d^* will be included. Alternatively, basis functions, such as Gaussian basis functions, sigmoidal basis functions, etc., are also commonly applied. For more details about basis functions, see Scholkopf and Smola (2001).

In machine learning applications, the predictive performance of the model heavily relies on the feature representation of the data samples. If the feature space is expressive enough, the model can successfully detect the important patterns in the data, leading to a high predictive power. To this end, in the different application domains there are specialized methods, which can select well-fit basis functions. An alternative way of obtaining an expressive representation is by using the prior knowledge of the study area. Specifically, in application domains, such as in bioinformatics, relevant features can be obtained by the experts of the field. This latter approach is the one that we adopt for the experimental analysis of this thesis. Finding the appropriate basis function or creating a relevant feature representation is a task that is usually performed during a preprocessing phase.

2.4.2. Random forests

An alternative approach for non-linear regression (or classification) problems is to subdivide the space into smaller areas. This can be managed by a machine learning algorithm which is known as decision trees. Decision trees can be applied

in classification and regression problems. In this section we refer to the way that a regression tree models the data, but the extension to the classification case is straightforward, since the steps of the algorithm and the basic idea behind remain the same in both cases.

A regression tree is constructed based on a binary recursive partitioning process (Breiman et al., 1984; Hastie et al., 2001; Bishop, 2007). This process iteratively splits the initial training data set into two partitions, constructing that way the branches of a tree. The data which belong to a specific branch are further split into two partitions. Specifically, the algorithm is initialized by considering the training samples as a whole. In a next step, it tries to break up the training data set by evaluating every possible binary split based on the values of each input variable. The evaluation of the different binary splits is performed according to a specific criterion, such as the minimization of the sum of the squared deviations from the mean value of the target variable in the two partitions. This splitting and evaluating step is repeated for each partition and afterwards sub-partition of the data. This process continues until each node reaches a minimum node size (usually a parameter) and becomes a terminal node. Following these recursive steps, the algorithm splits the initial training data set into multiple partitions, forming a tree structure. In a mathematical formulation the model can be written as:

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m I(\mathbf{x} \in P_m) \tag{2.15}$$

with M being the different partitions $P_1, ..., P_M$, c_m the constant response in partition P_m and I an indicator function that returns 1 when its argument is true and 0 otherwise.

By adopting as minimization criterion the mean squared error, i.e., $\sum_{i=1}^{N}(y_i - f(\mathbf{x_i}))^2$ the best value for the c_m constants is the average of the values y_i of the target variable in region P_m, i.e.,

$$c_m = \frac{1}{|P_m|} \sum_{i=1}^{|P_m|} y_i. \tag{2.16}$$

The problem of finding the best binary partition can be seen as a minimization problem of the loss function. However, this kind of approach is computationally intractable. Thus, the most common way to do so, is to proceed with a greedy algorithm. Different implementations of the algorithm use different criteria for evaluating the splits. These criteria involve metric calculations that generally measure the homogeneity of the target variable within the data partitions. Specifically, this kind of metrics are applied to each candidate partition, resulting in a value that reflects the quality of the split. A well-known criterion is the variance reduction which is defined in one node of the tree as the total variance reduction of the target variable $\mathbf{y}$ due to the split at this node.

Decision trees have been proven effective in modelling data with input variables interacting in a non-linear way. However, they become even more popular when they are used as base models in ensemble methods, such as in Random Forests (RF). RF algorithm developed by Breiman (2001), is an ensemble learning method for classification, regression and other tasks. It works by forming a combination of multiple decision trees, where each tree contributes with a single vote to the final output, which is the most frequent class for classification and the average for regression problems, respectively. The model averaging approach that RF applies is called bootstrap aggregating (or bagging). The basic steps can be summarized in the following paragraph.

Given a training set as described in Sect. 2.2, bagging randomly selects data samples with replacement of the training set and fits trees to these constructed sub-data sets of the initial training set. This procedure is repeated several times, e.g., B times, resulting in a set of decision trees. Specifically, in each iteration $b = 1, ..., B$ the algorithm selects n training samples denoted as $(\mathbf{X}_b, \mathbf{Y}_b)$. In the sequel, a decision tree f_b is trained based on this sub-data set. Hence, in the end of this iterative procedure, the model consists of B decision trees trained on the different sub-data sets. In the prediction phase, given a test observation $\mathbf{x}'$, the model response equals the average predicted value of the individual regression trees, i.e.,

$$f(\mathbf{x}') = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x}'), \tag{2.17}$$

or to the most popular class based on majority voting among the classification trees.

Except for resampling the data with replacement, the diversity among the different trees in RF increases due to the random selection of the input variables. Each tree is trained based on this bootstrapped sample from the initial training data set and in each node the splitting is performed by using the input variables of a specific random subset. In terms of bias-variance trade-off, one should take into account the way that these terms are related with the decision trees. Decision trees tend to overfit the data, since they can capture complex relationships in them. So, they are characterized by low bias. On the other hand, they usually show a high variability when they are trained on different samples of the training data set. This means that the models have high variance. Thus, the rationale behind RF is that it combines a set of high-variance, low-bias models, resulting in a learner which has both low variance and low bias. The low variance of the RF model is managed by the output aggregation of the individual trees. As such, the variance decreases, since the RF model gives smoothed predictions, which are more likely to be near the true values.

In RF the number of trees is a parameter that is commonly tuned during training.

In most cases, when the number of trees increases, the predictions of the RF become more robust, improving the model performance. However, a large number of trees increases the computational time of RF. In practice, a large enough number of trees, e.g. 100-200, is adequate to achieve a high predictive performance. Other parameters, such as the number of features per tree or the maximum depth, usually do not have a high impact on the predictive performance of the algorithm especially in the regression setting.

To sum up about the basic characteristics of the RF algorithm, we conclude that RF is a simple method that scales well on big data sets (it constructs the trees in parallel). Moreover, it can detect non-linear complex relationships and it hardly overfits.

2.5. Time series data

In some applications such as in climate or finance, the data sets are characterized by an extra factor, 'time'. For example, in climate, the data sets usually come from sensors which produce data measurements every minute or even every second, while in financial applications, one can think of the values in the stock market or the companies which write down their sells every day. It is clear that time is a common characteristic of these data sets, since one observation is time-dependent from the previous one or the previous ones.

Data of this type are called time series data. A time series is a sequence of measurements for the same variable over different timestamps. The time interval between the different observations can vary, i.e., it can be seconds, hours, months, years, etc. In this thesis, we have collected multiple climate data sets, such as temperature data sets, precipitation data sets, etc. These data sets have the form of time series. So, the input variables, mentioned in the previous sections are in our case, time series. As such, for this section, we denote with $\mathbf{z}$ a variable which is measured as time series. As an example, we will use monthly measurements of a temperature time series in a particular location for a time period of N months. We denote the vector of the time series as $\mathbf{z} = (z_{t_1}, z_{t_2}, ..., z_{t_N})$. In the following sections we explain the basic concept of stationarity in time series and we describe simple time series forecasting methods which are used as baselines in this thesis.

2.5.1. Stationarity

A basic concept in time series analysis is stationarity. Intuitively, one can speak of a stationary time series if the following criteria are fulfilled: (1) the mean of the time series is not a function of time but it has a constant value over time, e.g., Fig.2.3a shows a time series which satisfies this condition while in Fig.2.3b

the mean value is time-dependent, (2) the variance of the time series is not a function of time, e.g., in Fig.2.3c, the variance of the time series is time-dependent, (3) the covariance between the t^{th} observation and the $(t+P)^{\text{th}}$ observation is a constant, e.g., in Fig.2.3d the covariance between the observations is not constant over time (Shumway and Stoffer, 2000).

In mathematical notation, if $\mathbf{z}$ is a time series, one can speak of a strictly stationary or strong stationary process if:

$$(z_{t_1}, z_{t_2}, ..., z_{t_P}) \tag{2.18}$$

and

$$(z_{t_1+\tau}, z_{t_2+\tau}, ..., z_{t_P+\tau}) \tag{2.19}$$

have the same distributions for all timestamps $t_1, t_2, ..., t_P$ and all the constants τ.

Except for strong stationarity there is also weak stationarity, in which the mean value of the time series is constant, while its covariance depends on the distance between the observations only and not on time. In this thesis, when we refer to stationary times series, we mean strong stationary time series. Since most of the methods in the literature for time series modelling are applied on stationary time series and because, in machine learning the data distribution should be the same for the test and the training observations, the transformation of a non-stationary time series to stationary is necessary. There are multiple ways of bringing this stationarity. Some of them are called de-trending processes, differentiation, time series decomposition methods, etc. In climate, the time series data are highly non-stationary. In the next chapter, where we describe the data set, we explain in detail the time series decomposition method that we apply.

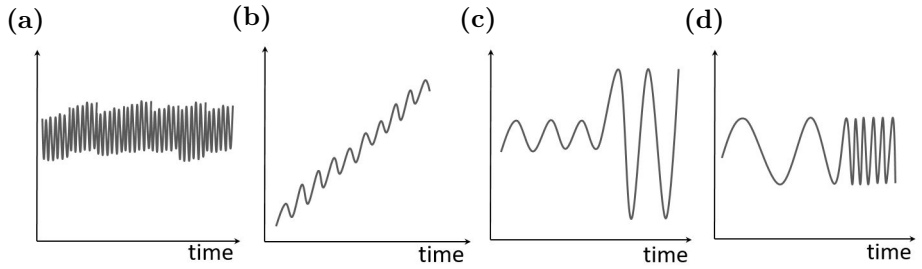


Figure 2.3: Illustration of stationary and non-stationary time series. (a) A stationary time series. (b) A non-stationary time series in which the mean value is time-dependent. (c) A non-stationary time series in which the variance is time-dependent. (d) A non-stationary time series in which the covariance is time-dependent.

2.5.2. Autocorrelation

Another basic concept in times series is the autocorrelation. Autocorrelation is a coefficient of correlation of a time series with a lagged copy of itself, i.e., it is a similarity between observations of the same times series as a function of the time lag between them. For example, the autocorrelation function for a given stationary time series $\mathbf{z_t}$ is given by:

$$Corr(\mathbf{z_t}, \mathbf{z_{t-P}}) = \frac{E[(\mathbf{z_t} - \mu)(\mathbf{z_{t-P}} - \mu)]}{\sigma^2} \tag{2.20}$$

with μ and σ^2 being the mean and the variance of the time series. The value of P indicates the time distance between the values of the two time series, i.e., the time lag. In other words, the above autocorrelation function is calculated as a correlation between the time series $\mathbf{z_t} = (z_{t_1}, z_{t_2}, ...)$ and its lagged version $\mathbf{z_{t-P}} = (z_{t_1-P}, z_{t_2-P}, ...)$. An autocorrelation for $P = 1$ is the correlation between values that are one timestamp apart.

2.5.3. Time series forecasting

Many processes produce data that are characterized by high autocorrelation. For example, in climate, one can think of the temperature measurements in which the temperature of the current month depends on the temperature of the previous month, etc. Therefore, approaches that are typically used for time series prediction use this property in order to well-model time series observations. These approaches are known as autoregressive methods and the models as autoregressive models. In an autoregressive model, the values of a time series are regressed on the corresponding lagged-values of the same time series. The value z_t in timestamp t of a time series $\mathbf{z}$ is modeled based on the value of the previous timestamp z_{t-1}, i.e.,

$$z_t = w_0 + w_1 z_{t-1} + \epsilon_t. \tag{2.21}$$

As one can observe, in an autoregressive model the lagged-value of the target variable becomes predictor (input variable) for the next value of the target variable. An error term ϵ is usually added to model the random noise in the data. In addition, in the preceding model, 1-lagged values are used each time as predictors for the forecast of the current ones. In this case the model is called a first-order autoregressive model. If one wanted to predict the temperature of the current month (z_t) by using the temperatures of the last two months (z_{t-1}, z_{t-2}), the autoregressive model would become:

$$z_t = w_0 + w_1 z_{t-1} + w_2 z_{t-2} + \epsilon_t. \tag{2.22}$$

This model is a second-order autoregressive model since the 1- and 2-lagged values are used as predictors for the forecast of the current month t. Hence, in general a P^{th} order autoregressive model is a linear regression model, which uses the values at times $t-1, t-2, ..., t-P$ as predictors for the forecast of the value at a time t. In our application, the time lag plays an important role, since future values of climatic and vegetation time series highly rely on their past values. Therefore the time window P should be carefully selected based on tuning and assessing the predictive performance of the model. In specific applications, prior knowledge about the impact of the past values is also used.

For other traditional and more advanced time series forecasting methods, the reader is referred to the analytical review papers (Gooijer and Hyndman, 2006; Gamboa, 2017).

2.5.4. Performance evaluation

Hold-out approaches are commonly used in the evaluation of time series prediction algorithms. Specifically, the model is trained on the past data and its performance is evaluated on the last block of the data observations. So in this case, one splits the initial data set in two sets, the training set, which includes the first block of the observations, and the test set, which consists of the last block of the observations, see Fig. 2.4a. This kind of evaluation process assumes that the data are stationary and follow the same distribution throughout their entire length. In climate data, where the data are highly non-stationary and scenarios, such as sensors' failure or replacement typically occur, hold-out methods lead to poor model performance.

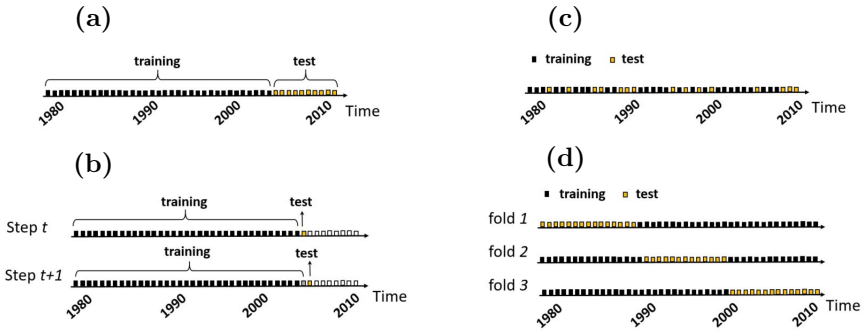


Figure 2.4: Illustration of the different evaluation procedures. (a) Train/test splitting in a hold-out approach. (b) Train/test splitting of one fold from a random three-fold cross-validation approach. (c) Two sequential steps of an online learning approach. (d) Block three-fold cross-validation.

Another approach is based on the evaluation of the learning models where a model is trained in past observations and predicts the observation value of the next timestamp. Then, the model is updated or re-trained by using the new observation as training example, see for example Fig. 2.4b. This kind of models, known as online learning models, needs many observations (thousands or millions) in order to reach a performance convergence. The process of sequential evaluation and model update is computationally intensive even for small data sets. The data set used in this thesis consists of time series of few-hundred points long.

In Sect. 2.2, we discussed about the K-fold cross-validation approach as evaluation method. In this case, randomly selected samples are assigned to the different folds. In times series analysis, this kind of evaluation approach is not commonly used. This is due to the fact that it does not come natural that future observations are used for the prediction of past observations. On top of that, the training and the test data set are not independent due to the autocorrelation between the observations. For instance, two consecutive observations are quite likely to have similar values, see Fig. 2.4c. However, since the selection of the training and the test observations is randomly performed, consecutive observations can be assigned to different training/test sets. For this reason, a more strict K-fold cross validation method has been proposed (specifically) for the time series prediction task, called the block K-fold cross-validation approach. Based on this approach, the observations are not randomly assigned to the different folds, but blocks of observations are used as folds in a cross-validation procedure. More specifically, the entire time series interval is separated in K parts in a way that consecutive observations are assigned to the same part, see Fig. 2.4d. For example, if one applies a block five-fold cross-validation in a monthly time series of 30 years, each fold will include observations from six consecutive years, i.e., the first fold will include observations from the first six years, the second one from the next six years and so on. Although this approach fits better with time series data, it cannot model well non-stationary data.

2.6. Spatial data

Except for time, space is another dimension that characterizes climate data. As mentioned in the previous chapter, climate data are coming from multiple sensors, so they might be more accurate for some regions, while they might be very noisy for others. There are several preprocessing techniques (e.g., interpolation) that are applied on the raw data. These techniques are employed in order to form gridded data sets of consistent spatial resolutions (for more details see Chapter 3).

A basic concept in spatial data sets is the one of spatial autocorrelation which is an extension of temporal autocorrelation. However, spatial autocorrelation is a bit more complicated, since time is one-dimensional, and only evolves in one

direction, ever forward. Spatial objects have (at least) two dimensions and complex shapes. Therefore, it is not always straightforward to define what is considered as adjacency between locations. Measures of spatial autocorrelation quantify at which extent two observations (values) at different spatial locations, are similar to each other. Hence, one needs two things in order to calculate this kind of correlations; observations and locations. A commonly used statistic that describes spatial autocorrelation is Moran's index (Moran, 1950), which is formulated as a correlation coefficient.

Finally, it is worth mentioning the different types of evaluation techniques in spatial data. In most studies, learning models are trained on a part of the whole spatial data set and they are evaluated on a test set which includes the rest of the observations (locations). The situation becomes more complicated if there are time series data in a spatial data set. Assuming that there are N timestamps and L different locations at a given data set; one can think of the following train/test splitting strategies: (i) all the available data of particular timestamps for all the locations are omitted for testing (Fig. 2.5a) (ii) all the available data of particular locations (for all the timestamps) are omitted (Fig. 2.5b), or (iii) finally, there is a combination of the previous strategies, in which a cross-time cross-location validation is performed, where at each time, a single time-location observation is tested, while all the corresponding time and location observations are omitted during training (Fig. 2.5c). In this thesis, we use the first approach to form our cross-validation data sets.

(a)

	t_1	t_2	t_3
l_1	train	test	train
l_2	train	test	train
l_3	train	test	train
l_4	train	test	train

(b)

	t_1	t_2	t_3
l_1	train	train	train
l_2	train	train	train
l_3	test	test	test
l_4	train	train	train

(c)

	t_1	t_2	t_3
l_1	train	omitted	train
l_2	train	omitted	train
l_3	omitted	test	omitted
l_4	train	omitted	train

Figure 2.5: Illustration of the different train/test splitting strategies in spatio-temporal data sets. The rows demonstrate the different locations and the columns the different timestamps. (a) A cross-time train/test splitting. (b) A cross-location train/test splitting. (c) A cross-time cross-location train/test splitting. Concurrent observations in time or space are excluded for both data sets (train/test).

2.7. Current applications of machine learning in geosciences

Machine learning approaches can be exploited for the effective prediction of geoscientific variables. For instance, extreme weather events such a floods and tornadoes

can be predicted by using climate data (Wang and Ding, 2015; Zhuang et al., 2016). So, forecasting and predicting future values of the Earth system (e.g., water availability) can be beneficial in deciding upfront about resource consumption. In machine learning, this problem can be modeled as a time series regression problem, in which the future values of a geoscientific variable depend on the past values of the variable itself. The most well-known approaches for this kind of problems are the autoregressive models, such as autoregressive integrated moving average (ARIMA) and vector autoregressive moving average models (VARMA), to name only a few.

However, all the aforementioned methods have been proven effective only for short-term forecasting. This is due to the fact that geoscientific variables are highly non-stationary and thus long-term predictions lead to error propagation. As a consequence, the long-term forecasts suffer from uncertainty, degrading the accuracy of the predictions. To tackle this kind of problems, researchers have exploited a recent machine learning method, namely transfer learning. The core idea of transfer learning can be used to train models for tasks which refer to present conditions, and transfer this knowledge to future tasks, which may have a small number of samples. Non-stationary geoscience data sets can be also handled by online learning algorithms. In machine learning, an online algorithm is a method that performs a model update for each training example, since the whole data set is not available at once. In other words, this kind of methods performs one update at a time, i.e., each new example is used to evaluate and update the current model. Interesting applications of online learning methods in climate data can be found in the works of Monteleoni et al. (2011) and McQuade and Monteleoni (2012). In these works, the goal is to produce robust estimates of climate variables, such as temperature, based on the outputs of climate models. Specifically, at each step the weights of the climate models' outputs are updated, taking into account the time and the space structure. Thus, at each timestep, the contribution of the climate models (which are considered as experts) at the final value of the target variable are adapted. This technique outperforms the baseline, in which an average value across the climate models is calculated at each time step without any kind of adaptation.

As mentioned in the previous sections, except for the non-stationarity, heterogeneity across space is another challenge, considering the various geoscientific variables. Machine learning techniques such as multi-task learning have been used in order to address this issue. The idea behind multi-task learning is that similar tasks are learned simultaneously, while they are able to share information between each other. Therefore, tasks with limited training examples benefit from this kind of modelling. For instance, in the work of Karpatne et al. (2014), multi-task learning modelling is applied in order to predict forest cover in Brazil. In this work, a different task is defined for each vegetation type, and based on the similarity between the different vegetation types, information between the learning tasks is shared. That way, one

model is able to learn from each homogeneous part of the data and also share information with other models trained in other homogeneous parts, improving the generalization performance. Multi-task learning approaches have been also used in prediction of climate variables. Xu et al. (2016) use a multi-task learning approach in which one model is trained based on the data of a particular station (location). The models share a common representation, based on the spatial autocorrelation between the different locations.

The large number of geoscientific variables in combination with the small sample sizes is another challenge for the machine learning community. The complex Earth system includes plenty of variables with strong dependencies between them. So, techniques, which are able to model high-dimensional data are necessary for this kind of data. Conveniently, in machine learning there are several methods that can work in these scenarios, which are known as feature selection methods. Some of them, such as the LASSO regularizer and wrapper methods, have been already applied successfully in geoscience applications (Chatterjee et al., 2012; Ma et al., 2017). In many geoscience problems, the lack of high quality ground-truth labels becomes an additional burden. In machine learning, there is a family of methods which is specialized on this kind of problems where labeled data are limited, while the unlabeled ones can be easily accessed. These approaches are known as semi-supervised methods and typically are based on the discovery of a hidden structure in the unlabeled data, which can lead to performance improvement for a given task (Zhu, 2005). Other methods, such as active learning approaches, which involve the presence of an expert to annotate and inspect the whole modelling process, have been also used in the context of geosciences (Vatsavai et al., 2005; Tuia et al., 2009). Finally, in geoscience applications one can encounter various scenarios where ground-truth labels are not available at all. In these cases, unsupervised methods, which attempt to find a structure that underlines a data set or a process, have been successfully applied. This kind of methods includes clustering techniques, dimensionality reduction approaches or breakpoint detection methods. Some applications that use the aforementioned methods are related to discovery of breakpoints on vegetation data due to fires, deforestation, etc. (Verbesselt et al., 2010b; Mithal et al., 2011), or detection of land cover classes based on climate and/or biome characteristics (Zscheischler et al., 2012).

Detecting relations between geoscientific variables is also very crucial in understanding the Earth system. For instance, the relation between the El Niño phenomenon and other extreme phenomena, such as floods or droughts, has been extensively studied (Siegert et al., 2001; Ward et al., 2014). Teleconnections are relationships that occur between variables of distant region pairs. These pairs are known as dipoles. Data-driven approaches for the discovery of such patterns typically involve graph-based representations (Steinbach et al., 2003; Runge et al., 2014). In these models, each node represents a specific location and each edge represents the correlation between the climate variables of the connected locations. Other

approaches include the representation of climate graphs as complex networks for the investigation of the climate system (Donges et al., 2009), hurricane activity, etc.

3 Database creation and variable construction

Satellite Earth observation has led to the creation of global climate data records of many important environmental and climatic variables. These come in the form of multivariate time series with different spatial and temporal resolutions. Data of this kind provide new means to further unravel the influence of climate on vegetation dynamics. In this chapter, we present the data set compiled during this PhD thesis, in the context of the SAT-EX project. We describe in detail the products that we have assembled (Sect. 3.2), the techniques that we have used to transform the data into a common spatial and temporal resolution (Sect. 3.2) and the preprossessing methodology that we have followed in order to form the final data set of our application (Sect. 3.2.1 and 3.2.2). In addition, we also present an extended exploratory analysis on the formed data set (Sect. 3.3).

This chapter is based on the content of:
Papagiannopoulou, C., Miralles, D. G., Decubber, S., Demuzere, M., Verhoest, N. E. C., Dorigo, W. A., and Waegeman, W.: A non-linear Granger-causality framework to investigate climate–vegetation dynamics, Geosci. Model Dev., 10, 1945-1960, https://doi.org/10.5194/gmd-10-1945-2017, 2017.

Decubber, S. : Spatio-temporal optimization of Granger causality methods for climate change attribution., Master thesis (tutored by C. Papagiannopoulou), Ghent University, 2017.

3.1. Introduction

Observational data in geosciences are collected by various acquisition methods such as local sensors (*in situ* measurements) or via instruments mounted on satellites (remote sensing data). These observations, in most cases, are not simultaneously available for all locations on the Earth at each timestamp. Specifically, in the case of *in situ* sensors, large regions remain unsampled due to the non-uniform distribution of sensors in space. Therefore, it is necessary to perform a preprocessing step to convert the raw data from sensors into fixed spatial grids. To this end, several techniques are used, such as simple linear interpolation methods, aggregation and reanalysis techniques (mentioned in the next paragraph). On the other hand, in the case of satellite remote sensors, continuous measurements at the same time over all locations are also unavailable, due to the orbital track of satellites around the Earth. As such, satellite data also need a number of additional preprocessing

steps, such as calibration, orbital correction, quality control, and conversion to regular grids.

Another category of observational data is based on climate model reanalysis, where model simulations are corrected by assimilating *in situ* and satellite observations. Specifically, possible values for different variables over large areas are calculated by combining data from (i) physical models and (ii) observed sensor recordings. As such, this type of observational data are adjusted into acceptable uncertainty levels and noise (due to the restrictions of physical models).

As one can observe, analysis of observational data poses several unique challenges, such as (i) the uncertainty and incompleteness of the data and (ii) the multiple spatial and temporal resolution. Therefore, this kind of data should be treated with caution in the different applications due to their special characteristics. To this end, techniques that are able to appropriately handle observational data have been recently developed both by geoscientists and machine learning experts. In this dissertation, we also develop techniques that are applied on observational data in order to investigate the relationship between climate and vegetation.

3.2. Global data sets

For this thesis, climate data sets of observational nature – mostly based on satellite and *in situ* observations – have been assembled to construct time series (see Sect. 3.2.2) that are then used to predict levels of vegetation greenness. Data sets have been selected from the current pool of satellite and *in situ* observations on the basis of meeting a series of spatio-temporal requirements: (a) expected relevance of the variable for driving vegetation dynamics, (b) multi-decadal record and global coverage available, and (c) adequate spatial and temporal resolution. The selected data sets can be classified into three different categories: water availability (including precipitation, snow water equivalent and soil moisture data sets), temperature (both for the land surface and the near-surface atmosphere), and radiation (considering different radiative fluxes independently). Rather than using a single data set for each variable, we have collected all data sets meeting the above requirements. This has led to a total of twenty-one different data sets which are listed in Table 3.1. They span the study period 1981–2010 at the global scale, and have been converted to a common monthly temporal resolution and $1^\circ \times 1^\circ$ latitude-longitude spatial resolution. To do so, we have used averages to resample original data sets found at finer native resolution, and linear interpolation to resample coarser-resolution ones. Here, we should note that other variables that include data for CO_2 emissions or other greenhouse gases, wind speed, vapour pressure deficit, fire and irrigation information, nutrient availability and land use change, are also relevant to our study. However, in this thesis we focus only on the climatic variables that may affect global vegetation dynamics.

For temperature we consider seven different products based on *in situ* and satellite data: Climate Research Unit (CRU-HR) (Harris et al., 2014), University of Delaware (UDel) (Willmott et al., 2001), NASA Goddard Institute for Space Studies (GISS) (Hansen et al., 2010), Merged Land-Ocean Surface Temperature (MLOST) (Smith et al., 2008), International Satellite Cloud Climatology Project (ISCCP) (Rossow and Duenas, 2004), and Global Land Surface Temperature Data (LST) (Coccia et al., 2015). We also included one reanalysis data set, the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim (Dee et al., 2011). In the case of precipitation, eight products have been collected. Four of them result from the merging of *in situ* data only: Climate Research Unit (CRU-HR) (Harris et al., 2014), University of Delaware (UDel) (Willmott et al., 2001), Climate Prediction Center Unified analysis (CPC-U) (Xie et al., 2007), and the Global Precipitation Climatology Centre (GPCC) (Schneider et al., 2008). The rest result from a combination of *in situ* and satellite data, and may include reanalysis: CPC Merged Analysis of Precipitation (CMAP) (Xie and Arkin, 1997), ERA-Interim (Dee et al., 2011), Global Precipitation Climatology Project (GPCP) (Adler et al., 2003), and Multi-Source Weighted-Ensemble Precipitation (MSWEP) (Beck et al., 2017). For radiation two different products have been collected (considering incoming shortwave/longwave and surface net radiation as different time series); first the NASA Global Energy and Water cycle Exchanges (GEWEX) Surface Radiation Budget (SRB) (Stackhouse et al., 2004) based on satellite data, and the second one the ERA-Interim reanalysis (Dee et al., 2011). For soil moisture we use the Global Land Evaporation Amsterdam Model (GLEAM) (Miralles et al., 2011; Martens et al., 2016), and the Climate Change Initiative (CCI) product (Liu et al., 2011a, 2012); two different soil moisture products by CCI are considered: the passive microwave data set and the combined active/passive product (Dorigo et al., 2017). Moreover, snow water equivalent data comes from the GlobSnow project (Luojus et al., 2010).

To conclude, as a proxy for the state and activity of vegetation, we use the third generation (3G) Global Inventory Modelling and Mapping Studies (GIMMS) satellite-based NDVI (Tucker et al., 2005), a commonly used long-term global record of normalized difference vegetation index (NDVI) (Beck et al., 2011). NDVI measures the vegetation greenness of each location at each timestamp, by combining information from the red and the near-infrared spectral reflectance measurements. Therefore, it reflects the vegetation state and not the plants activity as other variables, such as the solar-induced chlorophyll fluorescence (SIF). However, NDVI captures the main global vegetation patterns and trends and its data span a large period of 30 years, while most of the data sets that measure vegetation productivity include recent observations. In addition, the known limitations of NDVI, which include saturation in densely-vegetated area, sensitivity to atmospheric emissions and soil background noise, can explain some of uncertainties in the results (see next chapters). In the context of the SAT-EX project, we experimented with other

similar vegetation indices, such as the Leaf Area Index (LAI) (Demuzere et al., 2017) and the Vegetation Optical Depth (VOD) (Liu et al., 2011b), but here we preferred to present our analysis on NDVI throughout this thesis for consistency. The use of the NDVI also allows for a direct comparison of our results to previous studies. We note that this data set is used to derive the response variable in our approach (seasonal NDVI anomalies, see Sect. 3.2.1), while all other data sets are converted to predictor variables. The length of the NDVI record (1981–2010) sets the study period to an interval of 30 years.

3.2.1. Anomaly decomposition

In climate sciences, it is common that methodologies, such as Granger causality adopted in this thesis, are applied on time series of seasonal anomalies (Attanasio, 2012; Tuttle and Salvucci, 2016). The seasonal anomalies may be obtained in a two-step decomposition procedure, by first subtracting the seasonal cycle and then the long-term trend from the raw time series. Several competing decomposition methods have been proposed in the literature, including additive models, multiplicative models and more sophisticated methods based on break points (see e.g., Cleveland et al. (1990); Grieser et al. (2002); Verbesselt et al. (2010a)). In our framework, we use the following approach: in a first step, at each given pixel, the 'raw' time series of the target variable $\boldsymbol{y_t}$ and the climate predictors ($\boldsymbol{x}_t$, $\boldsymbol{z}_t$,...) are de-trended linearly based on a simple linear regression with the timestamp t as predictor variable applied to the entire study period. For the case of the target variable this can be denoted as follows:

$$y_t \approx y_t^{T_r} = w_0 + w_1 t \tag{3.1}$$

with w_0 and w_1 being the intercept and the slope of the linear regression, respectively. We obtain in this way the de-trended time series $\boldsymbol{y}_t^D = \boldsymbol{y}_t - \boldsymbol{y}_t^{T_r}$. This de-trending is needed to remove non-stationary signals in climatic time series, and allows us to draw the emphasis to the shorter-term multi-month dynamics. By de-trending one can assure that the mean of the probability distribution does not change over time; however, other moments or central moments of the probability distribution, such as the variance, might still be time-dependent. In a second step, after subtracting the trend from the raw time series, the seasonal cycle $\boldsymbol{y}_t^S$ is calculated. When the assumption is made that the seasonal cycle is annual and constant over time, one can simply estimate it as the monthly expectation. To this end, the multi-year average for each of the twelve months of the year is calculated. Finally, the anomalies $\boldsymbol{y}_t^R$ can then be computed by subtracting the corresponding monthly expectation from the de-trended time series: $\boldsymbol{y}_t^R = \boldsymbol{y}_t^D - \boldsymbol{y}_t^S$. This procedure is schematically represented in Fig. 3.1.

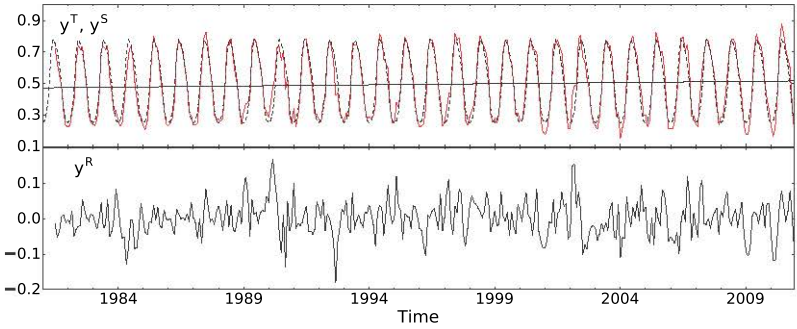


Figure 3.1: The three components of the NDVI time series decomposition of a specific pixel of the Northern hemisphere (lat: 53.5, long: 26.5). On top, the linear trend (black continuous line) and the seasonal cycle (dashed black line) fitted on the raw data (red). On the bottom the seasonal anomalies.

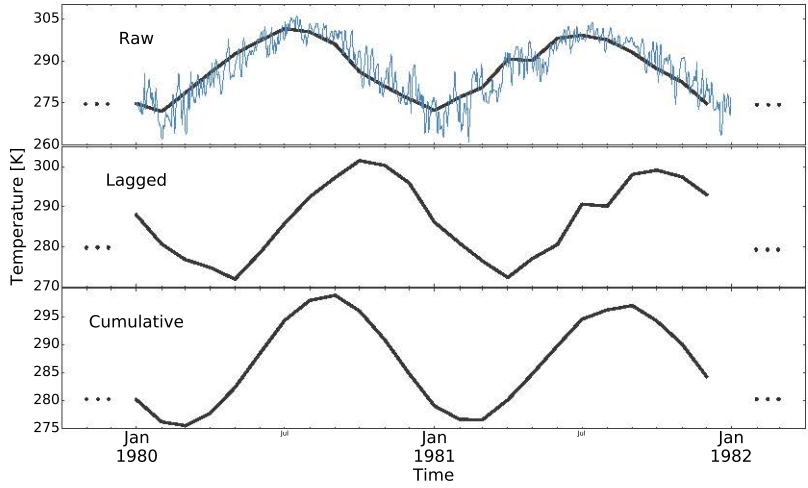


Figure 3.2: Example of lagged and cumulative variables extracted from a temperature time series. On top, part of a raw daily time series with its monthly aggregation. In the middle, the four-month lag-time monthly time series. On the bottom, the corresponding four-month cumulative variable. The pixel corresponds to a location in Kentucky US (lat: 37.5, long: -87.5).

Table 3.1: Data sets used in our experiments. Basic data set characteristics are provided, including the native spatial and temporal resolutions.

Variable	Product Name	Spatial Resolution	Temporal Resolution	Primary data source	Reference
Temperature	CRU-HR	0.5°	monthly	*in situ*	Harris et al. (2014)
	UDel	0.5°	monthly	*in situ*	Willmott et al. (2001)
	ISCCP	1°	daily	satellite	Rossow and Duenas (2004)
	ERA-Interim	0.75°	3-hourly	reanalysis	Dee et al. (2011)
	GISS	2°	monthly	*in situ*	Hansen et al. (2010)
	MLOST	5°	monthly	*in situ*	Smith et al. (2008)
	LST	0.5°	daily	satellite	Coccia et al. (2015)
Water availability	CRU-HR	0.5°	monthly	*in situ*	Harris et al. (2014)
	MSWEP	0.25°	3-hourly	satellite/*in situ*	Beck et al. (2017)
	UDel	0.5°	monthly	*in situ*	Willmott et al. (2001)
	CMAP	2.5°	monthly	satellite/*in situ*	Xie and Arkin (1997)
	CPC-U	0.25°	daily	*in situ*	Xie et al. (2007)
	GPCC	0.5°	monthly	*in situ*	Schneider et al. (2008)
	GPCP	2.5°	monthly	satellite/*in situ*	Adler et al. (2003)
	ERA-Interim	0.75°	3-hourly	reanalysis	Dee et al. (2011)
	GLEAM	0.25°	daily	satellite	Miralles et al. (2011)
	ESA CCI-PASSIVE	0.25°	daily	satellite	Dorigo et al. (2017)
	ESA CCI-COMBINED	0.25°	daily	satellite	Liu et al. (2012)
	GlobSnow	0.25°	daily	satellite	Luojus et al. (2010)
Radiation	SRB	1°	3-hourly	satellite	Stackhouse et al. (2004)
	ERA-Interim	0.75°	3-hourly	reanalysis	Dee et al. (2011)
Greenness (NDVI)	GIMMS	0.25°	monthly	satellite	Tucker et al. (2005)

3.2.2. Predictor variable construction

We do not limit our approach to considering raw and anomaly time series of the data sets in Table 3.1 as predictors, but also take into consideration different lag times, past-time cumulative values and extreme indices. These additional predictors, here referred to as 'higher-level variables', are calculated based on raw and anomaly time series. Our application can be interpreted as a way to identify patterns in climate during past-time moving windows (see Fig. 4.1 in Chapter 4) that are predictive with respect to the anomalies of vegetation time series. Therefore, by feeding predictor variables from previous timestamps to a linear (or non-linear) predictive model, one can identify sub-sequences of interest in the moving window specified for timestamp t, a technique that is similar to so-called shapelets (Ye and Keogh, 2009). In addition, vegetation dynamics may not necessarily reflect the climatic conditions from (e.g.) three months ago, but the average of the (e.g.) three antecedent months. This integrated response to antecedent environmental and climatic conditions is referred here as a 'cumulative' response. More formally, we construct a cumulative variable of k months as the sum of time series observations in the last k months:

$$\text{Cumul}[x_{t-1}, x_{t-2}, ..., x_{t-k}] = \sum_{p=1}^{k} x_{t-p} \tag{3.2}$$

Note that, unlike in the case of lagged variables, cumulative ones include always the period up to time t. Figure 3.2 illustrates an example of a four-month cumulative variable. In our analysis, we experimented with time lags covering a wide range of time-lag values and concluded that including lags of more than six months did not yield substantial predictive power.

Another type of higher-level predictor variable that can be constructed from the data sets in Table 3.1 are extreme indices. Over the last few years, several research studies have focused on defining and indexing climate extremes (Nicholls and Alexander, 2007; Zwiers et al., 2013). As an example, the Expert Team on Climate Change Detection and Indices (ETCCDI) recommends the use of a range of extreme indices related to temperature and precipitation (Zhang et al., 2011; Donat et al., 2013). Here we calculate a variety of analogous indices for the whole set of the collected climatic variables, based on both the raw data sets as well as on the seasonal anomalies (see Table 3.2). In addition, we derived lagged and cumulative predictor variables from these extreme indices to incorporate the potential impact of climatic extremes occurring (e.g.) three months ago, or during the previous (e.g.) three months, respectively. All these resulting time series appear as additional predictor variables in our framework (see Sect. 4.2.3 of Chapter 4).

Combining the different climate and environmental predictor variables described above, we obtain a database of 4,571 predictor variables per 1° pixel, covering thirty years at a monthly temporal resolution.

Table 3.2: Extreme indices considered as predictive variables. These indices are derived from the raw (daily) data and the (daily) anomalies of the data sets in Table 3.1. We also calculate the lagged and cumulative variables from these extreme indices (see Sect. 3.2).

Name	Description
STD	Standard deviation of daily values per month
DIR	Difference between max and min daily value per month
Xx	Max daily value per month
Xn	Min daily value per month
Max5day	Max over 5 consecutive days per month
Min5day	Min over 5 consecutive days per month
X99p/X95p/X90p	Number of days per month over $99^{\text{th}}/95^{\text{th}}/90^{\text{th}}$ percentile
X1p/X5p/X10p	Number of days per month under $1^{\text{th}}/5^{\text{th}}/10^{\text{th}}$ percentile
T25C[a]	Number of days per month over 25°C
T0C[a]	Number of days per month below 0°C
R10mm/R20mm[b]	Number of days per month over 10/20 mm
CHD (Consecutive High value Days)	Number of consecutive days per month over 90^{th} percentile
CLD (Consecutive Low value Days)	Number of consecutive days per month below 10^{th} percentile
CDD (Consecutive Dry Days) [b]	Number of consecutive days per month when precipitation < 1 mm
CWD (Consecutive Wet Days)[b]	Number of consecutive days per month when precipitation $\geqslant 1$ mm
Spatial Heterogeneity [c]	Difference between max and min values within 1° box

[a] Only for temperature data sets
[b] Only for precipitation data sets
[c] Only for data sets with native spatial resolution $<1^{\circ}$ lat-lon

3.3. Exploratory pre-analysis

In this section an exploratory pre-analysis is conducted for the raw climate time series and the higher-level features that serve as predictors to model vegetation in the next chapters. The target variable of our analysis is the NDVI seasonal anomalies. In fact both the raw NDVI time series as well as the seasonal anomalies are explored. In addition, autocorrelations within the vegetation time series and correlations between vegetation and climate variables are calculated and illustrated as well.

3.3.1. Correlation between climate records from different products

The database consists of multiple time series for every land pixel. These time series are records from the same climate variable. For instance, as mentioned above, there are seven time series related to temperature in our data set and five of them measure the near-surface air temperature while the rest (*GISS* and *MLOST*)

measure de-trended temperature anomalies. Intuitively, one could expect the correlations between these temperature-related time series to be very high. Figure 3.3 illustrates the Pearson correlation coefficients between the temperature-related time series for a randomly selected pixel from the data set. In general, the different temperature measurements for this pixel are highly correlated, with ISCCP being the only product that is slightly less correlated to the other products (Decubber, 2017).

Figure 3.4 shows the correlation coefficients for the pairs of temperature-related time series at global scale. The examined pairs of products from top to bottom are the following: CRU/ERA, LST/CRU and UDel/ERA. On the left part of Fig. 3.4, the correlations between the raw time series are illustrated, while on the right part the correlation between the anomalies (which are calculated based on the time series decomposition method described in Sect. 3.2.1) are presented. The global maps of the correlation coefficients between all pairs of raw temperature-related time series are very similar to the maps on the left side on Fig. 3.4 (figures omitted). As one can observe, the raw time series are highly correlated in most regions, except for the tropics. This is due to the fact that in these regions (i.e., Amazon, Congo basin), there is no clear seasonal cycle present and there are no large fluctuations in the temperature values throughout a year period. In contrast, in more temperate climates, e.g., in the North Hemisphere, there is a strong presence of seasonal cycle in the temperature time series (and in other climatic variables as well, e.g., radiation). Hence, this seasonal component is similar among the different products, since it can be easily captured by the measurements. As such, seasonal variability constitutes a strong correlated component for the time series and if it is missing the correlation between the various products becomes lower. In order to assess the correlation between the different products without the effect of the obvious seasonal component, we calculate the anomalies of the temperature-related time series. The panels on the right of Fig. 3.4 depict the correlation values of the anomalies time series at global scale. As one could expect, the time series of anomalies are not so highly correlated. This means that even though these records measure the same climate variable they may contain different information. Therefore, depending on the way that these measurements are obtained for each part of the world, they can be more accurate for different regions.

Similar to Fig. 3.3, Fig. 3.5 depicts the correlation coefficients between raw water-related time series (for the same selected pixel as before). The snow-water equivalent variable is excluded from this analysis, since no snow coverage is observed in this particular pixel. As one can observe, some water-related time series are strongly correlated, while others are only weakly or not at all correlated. A conclusion that can be drawn by this correlation analysis among the different temperature- and water-related products is that even though their records measure the same climate variables, these records are not always highly correlated. This can be explained

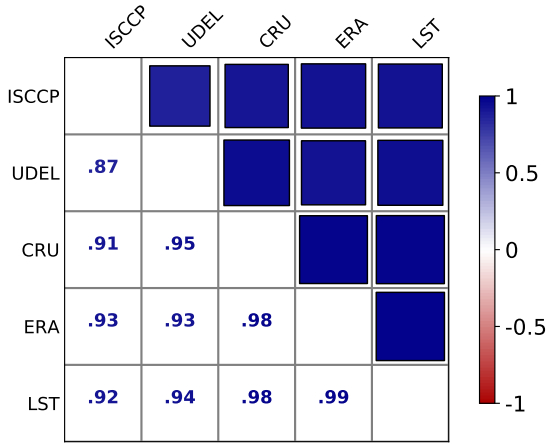


Figure 3.3: Pearson correlation coefficients of the temperature variables from a randomly selected pixel (South Africa; lat: -24.5, long: 22.5). All temperature products measure near-surface air temperature expressed in K.

by the way that these products are generated. For instance, temperature is not directly measured by the satellite equipment; instead, irradiance in different parts of the wavelength spectrum is measured, commonly by the various satellite sensors. On top of this, these sensors might be based on different technologies (National Research Council (U.S.). Committee on Earth Studies, 2000). In addition, surface temperature measurements are dependent on heterogeneities in the surface and are accurate only under cloud-free conditions. Similar factors affect precipitation measurements. Moreover, as discussed in Sect. 3.1, the final product depends on the method that is used to convert the measurements to a (e.g., temperature) data set consistent in space and time. On the other hand, *in situ* measurements are also post-processed (e.g., with interpolation techniques) in order to form the final product. Therefore, one should expect differences between the different products even if they are meant to represent the same climate variable (Hughes, 2006). Although a full technical discussion on measurement techniques and differences between satellite-based and *in situ* observations is beyond the scope of this thesis, it is clear that temperature or precipitation records coming from different observational sources do not necessarily contain equivalent information.

3.3.2. Autocorrelation of vegetation time series

The NDVI seasonal anomalies is the target time series in our analysis. The past values of this time series are used for predicting the future ones in the modelling approach that will be described in Chapter 4. Therefore, an autocorrelation analysis

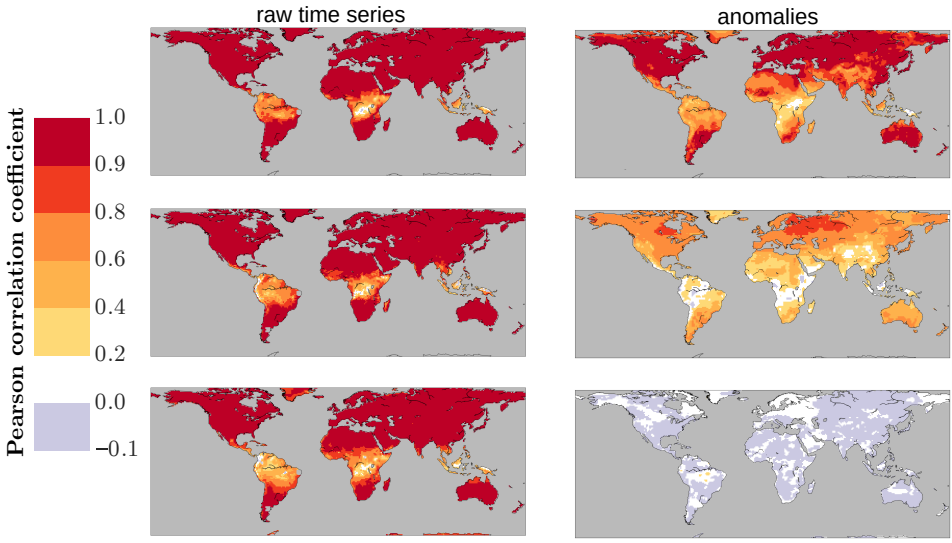


Figure 3.4: Pearson correlations between three pairs of raw temperature time series (left) and between their anomalies (right). From top to bottom: CRU/ERA, LST/CRU and UDel/ERA. Corresponding to section 3.3.1.

of this time series can reveal to what extent it is possible to predict the value of the NDVI seasonal anomalies at a next timestamp, based on the values of the previous timestamps. Figure 3.6 shows the autocorrelation value of the NDVI seasonal anomalies for every pixel, for temporal lags 1-4 months. The NDVI seasonal anomalies are positively correlated in most regions of the world for the temporal lag of one month. The highest autocorrelation values are observed in Australia, South America, North America, Central Asia, South of Africa and the Sahel region. The autocorrelation values decrease for the temporal lag of two in all regions, although the autocorrelation values remain positive in Australia. When the time lag increases (i.e., three and four) the autocorrelation of the NDVI seasonal anomalies time series decreases, having near-zero values in most of the regions.

3.3.3. Correlation between vegetation and climate data

Since vegetation needs some time to adapt to climate variability, the lagged values of the climate variables are incorporated in our analysis (see Chapter 4). In order to explore the size of the temporal window for the climate variables that should be taken into account, correlation plots between climate variables and NDVI seasonal anomalies are created (Fig. 3.7). In this figure, correlations are calculated based on different temporal lags between the contemporaneous observation of the NDVI seasonal anomalies and the past climatic time series. Four different

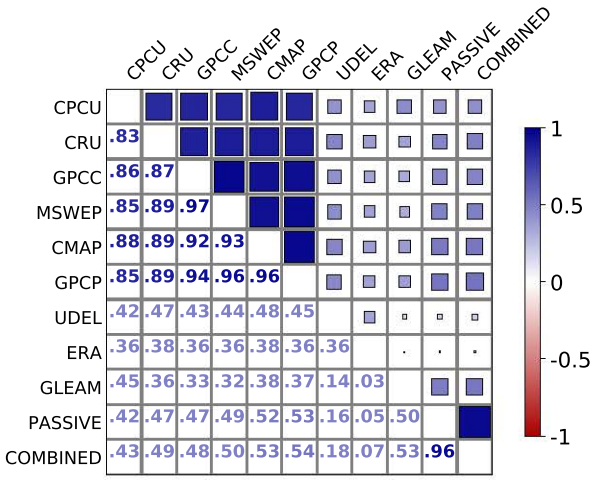


Figure 3.5: Correlation matrix of the water-related time series from a randomly selected pixel (latitude -24.5, longitude 22.5). All products measure precipitation in mm, except for *GLEAM*, *PASSIVE* and *COMBINED* which measure soil moisture. *GLOBSNOW*, which measure thickness of snow coverage, is not included in this visualization.

climate products are selected; one for each climate variable, namely, CRU for near-surface air temperature, MSWEP for precipitation, GLEAM for soil moisture and SRB for incoming shortwave radiation. In general, the climatic time series are most correlated with the NDVI seasonal anomalies when there is no lag difference between the measurements (lag 0). Correlations between current vegetation and past climate tend to fade away, when the temporal lag increases further back in the past. Specifically, for the temperature-related product, the correlation values with the NDVI seasonal anomalies are larger than 0.2 in absolute value for only a few pixels. However, correlations values become smaller than 0.1 in absolute value for temporal lags equal or larger than three months. The same conclusions can be drawn for the correlations calculated for the radiation product. On the other hand, the correlations with the precipitation and soil moisture time series are much stronger, with correlation coefficients larger than 0.2 between NDVI seasonal anomalies. Correlations with these water-related variables of even 12 months ago are still strong in some pixels in Australia. This result is in line with prior knowledge about the 'memory' of the land surface, which is longer than memory of the atmosphere (Hilker et al., 2014).

Remarkably, precipitation and soil moisture are consistently negatively correlated with vegetation seasonal anomalies obtained from the same month in pixels at higher latitudes (i.e., Europe and north of Asia). At the same time, vegetation in these pixels is positively correlated with temperature and most of them also with

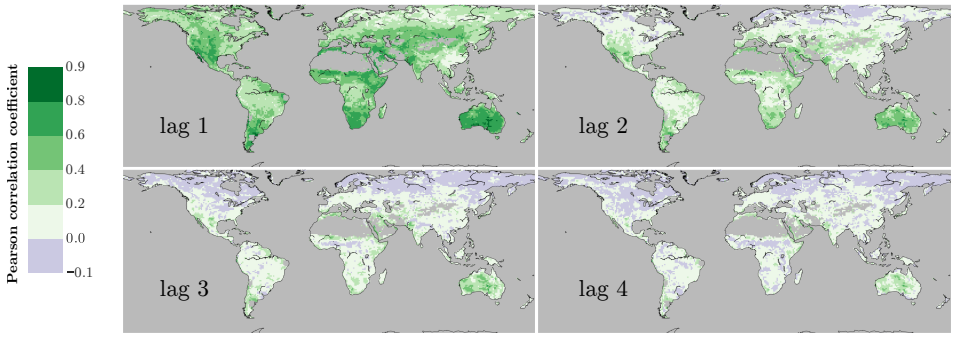


Figure 3.6: Autocorrelation between NDVI seasonal anomalies for the different temporal lags (1-4).

radiation. This result can be explained by the correlation between the climatic variables; water-related variables and temperature (or radiation) are negatively correlated. In a month with a large amount of precipitation (soil moisture is increased), the amount of radiation reaching the vegetation tends to be lower because of the increased cloud coverage. As a result, the vegetation is confronted with lower temperature or less radiation, which appears to be associated with lower NDVI seasonal anomalies (Decubber, 2017).

3.3.4. Visualization of climate data sets in two-dimensions

Principal component analysis (PCA) is a commonly used dimensionality reduction technique for data visualization and exploration. As described in the previous sections, high-level features (see Table 3.2) as well as lagged and cumulative variables have been constructed based on the raw climatic time series. The raw climatic time series have been also decomposed into the three components of seasonality, trend and anomalies. As such, one data set for each pixel has been created, with columns as many as the features extracted by the climatic time series (4,571) and with rows as many as the monthly observations of the last 30 years (360 observations in total). If one applies the PCA on a data set of a single pixel, the observations are projected from a high-dimensional space into a lower-dimensional space, while the variability between the observations is maximally retained. That way, the observation of one month, represented in a high-dimensional space, is projected to a lower-dimensional space that is based on the corresponding principal components. These principal components span the lower-dimensional subspace and consist of linear combinations of the original feature space. The weight of each feature is commonly referred to as the *loading* for a particular component, while the coordinates of the data points on each of the principal components are referred to as the *scores*. The principal components are ordered by the amount of variance that they explain in the original feature space. Here we applied the

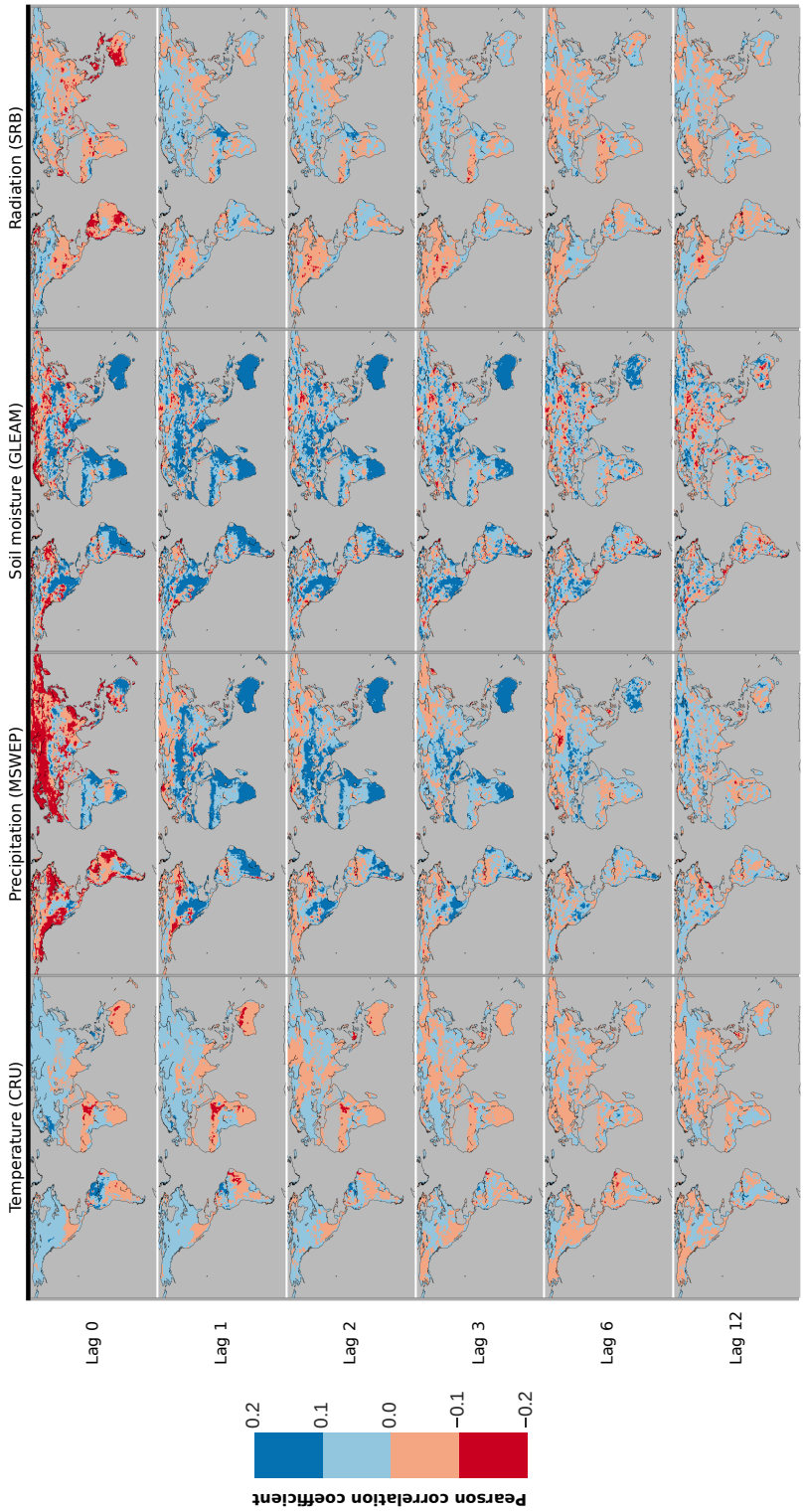


Figure 3.7: Correlation coefficients calculated between climate variables and NDVI seasonal anomalies for temporal lags of 0, 1, 2, 3, 6 and 12 months.

scikit-learn python implementation of PCA (Pedregosa et al., 2011). Figure 3.8 shows the projection of the observations of 16 randomly sampled pixels on the first two principal components. The percentage displayed on top of each subplot is the amount of variance that is explained by these two principal components.

The observations in Fig. 3.8 are colored in red and blue based on their timestamp; red observations correspond to earlier timestaps, while blue ones to more recent timestamps. There are several remarks considering the plots of Fig. 3.8: (i) in some pixels, early and recent observations seem to be scattered randomly (e.g., second plot from the left on the second row) (ii) in others, early and recent observations seem to form two separate clusters, and (iii) in the rest of the sampled pixels, there is a circular pattern in the observations (e.g., for the pixels on the third row of the plots).

Two of the plots (in dashed boxes) are further highlighted in Fig. 3.10. The data points are now visualized by their relative order based on their timestamp. The same color scheme is used as in Fig. 3.8. In Fig. 3.10a, a plot of a pixel where the observations show a main contrast along time is depicted. Two well-separated clusters are formed by roughly splitting the earliest 200 and the last 150 observations. In the plot of Fig. 3.10b, the observations corresponding to the first 12 timestamps are highlighted in a larger font for illustration purposes. As one can observe, the observations are located sequentially next to each other in the circular pattern, forming 12 clusters of observations which belong to the same month of the year. For example, observations from January form the first cluster, from February the second one, etc. In order to see the spatial distribution of the pixels, in which there is a contrast in PCA scores between early and recent observations, we use a logistic regression classifier. By using this classifier, we expect that these pixels will be distinguished from the others in which the PCA scores between early and recent observations show either a circular or a more random pattern. To this end, observations are labeled according to their timestamp, i.e., for the 200 earliest observations the '0' label is given, while for the last 160 the label '1' is assigned. The scores on the first two PCA coordinates are used as predictor variables and the classification accuracy is evaluated on a random 20% hold-out test set. As a rigid classifier with a linear decision boundary, logistic regression achieves a high accuracy when the early-recent pattern is clear, but is expected to perform poorly when the observations form more than two separate clusters or are scattered in a more random fashion. Based on this modelling approach, the highest classification accuracy is achieved in the tropical regions, see Fig. 3.9. This indicates that the pixels with a strong contrast between early and recent observations are located in these regions. The explanation for this result is coming from the fact that there is no strong climate seasonality in these regions. On the other hand, regions further away from the equator have a pronounced seasonal cycle with respect to climate. In our data sets, there are several features with strong seasonality (e.g., extreme indices), and thus this seasonal component naturally emerges in the PCA

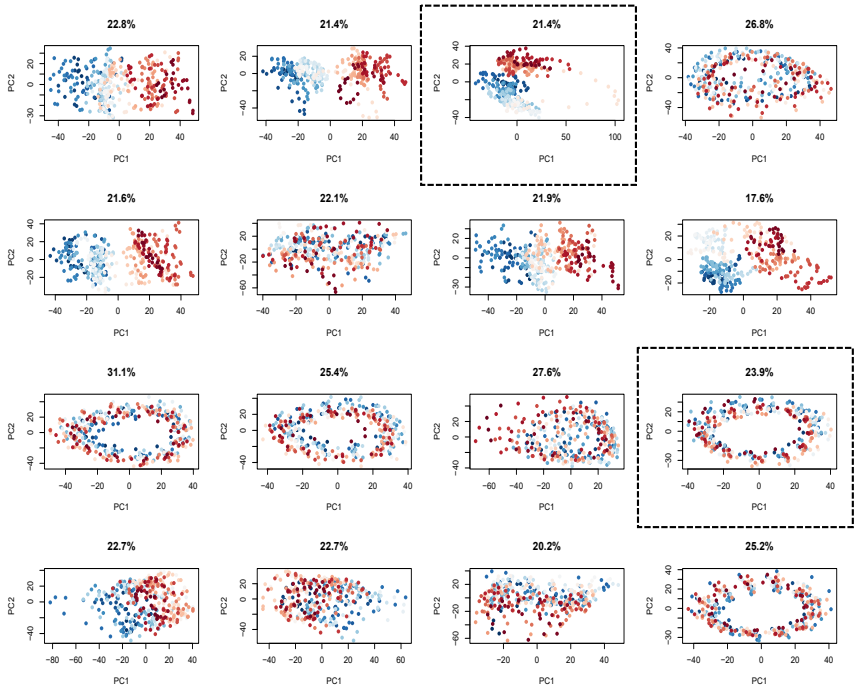


Figure 3.8: Scores of the observations on the first two PCA dimensions for 16 randomly sampled pixels. The observations are color coded from blue (early months) to red (most recent months). The plots in the dashed boxes are highlighted in Fig 3.10.

plots whenever it is present. Moreover, in our analysis, we observed that the classification performance improves, as more principal components are included as predictors. This means that in pixels with a strong seasonal cycle, the contrast between observations throughout time fades away, while this is not the case in the tropical regions.

This conclusion is also confirmed by our next experiment in which all the variables with a seasonal component are removed. In this setting, the logistic regression classifier scores a higher accuracy for most of the pixels, as expected. This result suggests that the contrast along the first two principal components between early and recent observations is much more pronounced when the seasonal variability is taken out of consideration. The contrast between early-recent observations is caused by the presence of a trend in most time series, reflected in the climatic indices calculated on the de-seasonalized data. In order to verify this result, as a final experiment, we run the PCA on the time series of the anomalies. In this case, the logistic regression classifier performs poorly in discriminating between early and recent observations, confirming that the large contrasts are mainly explained by linear trends in the time series.

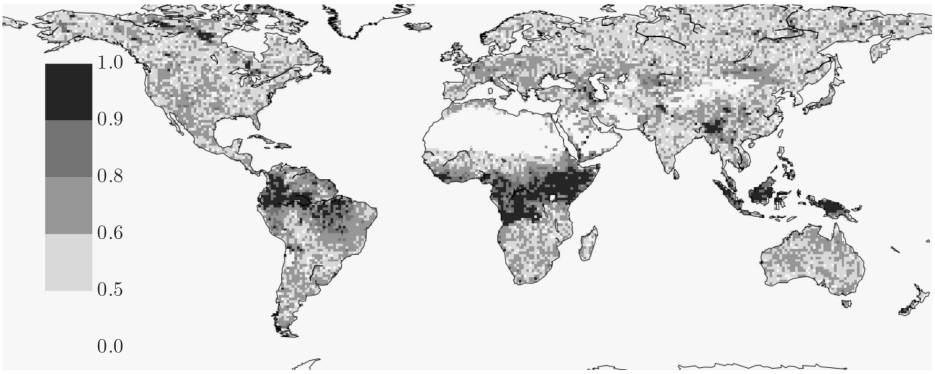


Figure 3.9: Proportion of test data correctly classified as early (first 200) or recent (last 153 months) by logistic regression, using the scores of the observations on the first two PCA dimensions as predictors.

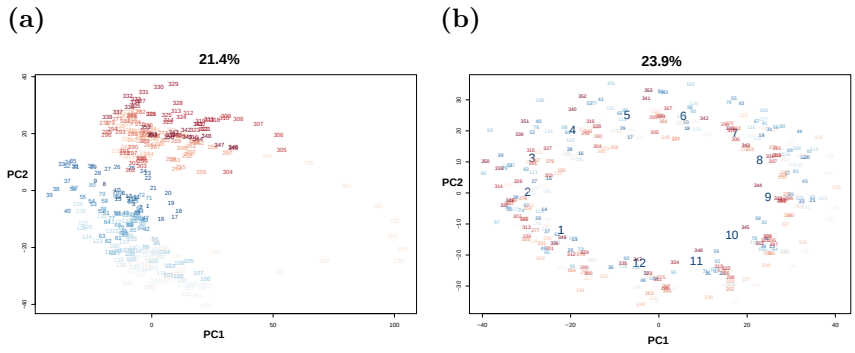


Figure 3.10: Two distinct PCA score patterns. (a) Clear contrast between early and recent months. (b) A 12-cluster circular pattern formed by the yearly observations of each month. The percentage of total variance explained by the first two principal components is shown on top of each plot.

3.4. Conclusions

In this chapter, we have presented the database used in this thesis. Specifically, the basic concepts related to data acquisition methods have been introduced and the data resources used for the database composition have been discussed. We have also described the time series decomposition technique used for the target and the predictor variables, as well as the feature construction approaches we have followed. In addition, we have presented an extended exploratory pre-analysis on the resulting database, by conducting correlation analysis between the target and the predictor variables and by visualizing the data in two-dimensions with PCA.

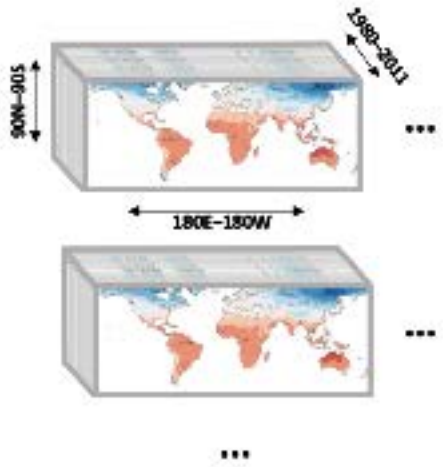


Figure 3.11: Illustration of the database. Each data cube consists of records for each 1° pixel at each timestamp from 1981-2010. Multiple data cubes correspond to multiple variables.

To sum up, our database consists of 21 climatic products (7 temperature-related products, 12 water-related products, 2 radiation-related products) as well as 1 vegetation product which span a period of 30 years (1981-2010). All the data sets have been transformed into the same spatial (1° latitude-longitude) and temporal resolution (monthly). Our database can be illustrated as a multi-data cube with dimensions the spatial coordinates and time, see Fig. 3.11. The vegetation data set serves as target variable in our analysis. We have isolated the anomalies component (by removing seasonality and possible trends) from the target variable. The same approach has been followed for the predictor variables as well. The final set of the predictor variables comprises lagged values of the raw time series and anomalies, cumulative variables, extreme indices, etc. In total, we obtain a database of 4,571 predictor variables for each 1° pixel, which covers 140 gigabytes of memory. The size of the database played an important role in the selection of the monthly temporal resolution and 1° spatial resolution.

Based on our exploratory pre-analysis, there is a strong degree of similarity between different temperature-related records and between different water-related records in the data set, in most regions of the world. However, this is mostly because of the strong seasonal component that is present in most of the raw signals. Moreover, the NDVI seasonal anomalies show fairly large autocorrelation at a temporal lag of 1 month, although the autocorrelation drops down in most pixels for larger temporal lags. Considering the correlations between the climate variables and the NDVI seasonal anomalies (Figure 3.7), the highest correlations occur between observations from the same month or one month earlier. Finally, the first two

principle components from the PCA reflected the largest source of variability in different pixels. In regions with a pronounced seasonal cycle, the seasonal pattern is responsible for the largest part of the variation. In the tropics, the largest variation between different observations occurs over time, indicating the presence of a trend in (at least) part of the features. This contrast became apparent in most other pixels as well, after any variable with a seasonal component was removed from the data set.

This database is further analysed in coming chapters in order to address the objectives of this thesis (see Chapter 1) and give answers to our specific research questions (Sect 1.2).

4 A non-linear Granger causality framework to investigate climate–vegetation dynamics

The main focus of this thesis is the study of climate–vegetation interactions. By having at our disposal the database described in the previous chapter and methods coming from the field of machine learning and data mining, we try to unravel the complex relationships between climate variability and vegetation dynamics. Commonly-used statistical methods are often too simplistic to represent complex climate-vegetation relationships due to linearity assumptions. In this chapter, we describe our core approach, which is an extension of the Granger causality analysis. Specifically, we present a novel non-linear framework consisting of several components, such as data collection from various databases, time series decomposition techniques, feature construction methods and predictive modelling by means of random forests. The first steps have been described in Chapter 3, while in this chapter the non-linear causality framework is introduced (Sect. 4.2). Experimental results on our global database indicate that, with this framework, it is possible to detect non-linear patterns that are much less visible with traditional Granger causality methods (Sect. 4.3.1 and 4.3.2). In addition, we discuss extensive experimental results that highlight the importance of considering non-linear aspects of climate–vegetation dynamics (Sect. 4.3.3).

This chapter is an edited version of:
Papagiannopoulou, C., Miralles, D. G., Decubber, S., Demuzere, M., Verhoest, N. E. C., Dorigo, W. A., and Waegeman, W.: A non-linear Granger-causality framework to investigate climate–vegetation dynamics, Geosci. Model Dev., 10, 1945-1960, https://doi.org/10.5194/gmd-10-1945-2017, 2017.

4.1. Introduction

Vegetation dynamics and the distribution of ecosystems are largely driven by the availability of light, temperature and water, thus they are mostly sensitive to climate conditions (Nemani et al., 2003; Seddon et al., 2016; Papagiannopoulou et al., 2017b). Meanwhile, vegetation also plays a crucial role in the global climate system. Plant life alters the characteristics of the atmosphere through the transfer of water vapour, exchange of carbon dioxide, partition of surface net radiation (e.g., albedo), and impacts on wind speed and direction (Nemani et al., 2003; McPherson et al., 2007; Bonan, 2008; Seddon et al., 2016; Papagiannopoulou et al., 2017b).

Because of the strong two-way relationship between terrestrial vegetation and climate variability, predictions of future climate can be improved through a better understanding of the vegetation response to past climate variability.

The current wealth of Earth observation data can be used for this purpose. Nowadays, independent sensors on different platforms collect optical, thermal, microwave, altimetry and gravimetry information, and are used to monitor vegetation, soils, oceans and atmosphere (e.g., Su et al., 2011; Lettenmaier et al., 2015; McCabe et al., 2017). The longest composite records of environmental and climatic variables already span up to 35 years, enabling the study of multi-decadal climate–biosphere interactions. Simple correlation statistics and multilinear regressions using some of these data sets have led to important steps forward in understanding the links between vegetation and climate (e.g., Nemani et al., 2003; Barichivich et al., 2014; Wu et al., 2015). However, these methods in general are insufficient when it comes to assessing causality, particularly in systems like the land–atmosphere continuum in which complex feedback mechanisms are involved. A commonly used alternative consists of Granger causality modelling (Granger, 1969). Analyses of this kind have been applied in climate attribution studies, to investigate the influence of one climatic variable on another, e.g., the Granger causal effect of CO_2 on global temperature (Triacca, 2005; Kodra et al., 2011; Attanasio, 2012), of vegetation and snow coverage on temperature (Kaufmann et al., 2003), of sea surface temperatures on the North Atlantic Oscillation (Mosedale et al., 2006), or of the El Niño Southern Oscillation on the Indian monsoons (Mokhov et al., 2011). Nonetheless, Granger causality should not be interpreted as 'real causality'; one assumes that a time series A Granger-causes a time series B if the past of A is helpful in predicting the future of B (see Sect. 2 for a more formal definition). However, the underlying statistical model that is commonly considered in such a context is a linear vector autoregressive model, which is (again), by definition, linear – see e.g., Shahin et al. (2014); Chapman et al. (2015).

In this chapter, we show new experimental evidence that advocates the need of non-linear methods to study climate–vegetation dynamics, due to the non-linear nature of these interactions (Foley et al., 1998; Zeng et al., 2002; Verbesselt et al., 2016). To this end, we have assembled a large, comprehensive database, comprising various global data sets of temperature, radiation and precipitation, originating from multiple online resources, as described in Chapter 3. We use NDVI to characterise vegetation, which is commonly used as a proxy of plant productivity (Myneni et al., 1997; Nemani et al., 2003). We followed an inclusive data collection approach, aiming to consider all available data sets with a worldwide coverage, and at least a thirty-year time span and monthly temporal resolution (see Chapter 3). Our novel non-linear Granger causality framework is used for finding climatic drivers of vegetation and consists of several steps (Sect. 4.2). In a first step, we apply time series decomposition techniques to the vegetation and the various climatic time series to isolate seasonal cycles, trends and anomalies. Subsequently,

we explore various techniques for constructing more complex features from the decomposed climatic time series, see more details in Chapter 3. In a final step, we run a Granger causality analysis on the NDVI anomalies, while replacing traditional linear vector autoregressive models by random forests. This framework allows for modelling non-linear relationships and prevents over-fitting. The results of the global application of our framework are discussed in Sect. 4.3.

4.2. Granger causality for climate studies

4.2.1. Linear Granger causality revisited

We start with a formal introduction to Granger causality for the case of two times series, denoted as $\boldsymbol{x} = [x_1, x_2, ..., x_N]$ and $\boldsymbol{y} = [y_1, y_2, ..., y_N]$, with N being the length of the time series. In this work, $\boldsymbol{y}$ alludes to the NDVI anomalies time series at a given pixel, whereas $\boldsymbol{x}$ can represent the time series of any climatic variable at that pixel (e.g., temperature, precipitation, radiation). Granger causality can be interpreted as predictive causality, for which one attempts to forecast y_t (at the specific timestamp t) given the values of $\boldsymbol{x}$ and $\boldsymbol{y}$ in previous timestamps. Granger (1969) postulated that $\boldsymbol{x}$ causes $\boldsymbol{y}$ if the autoregressive forecast of $\boldsymbol{y}$ improves when information of $\boldsymbol{x}$ is taken into account. In order to make this definition more precise, it is important to introduce a performance measure to evaluate the forecast. Below we will work with the coefficient of determination R^2, which is here defined as follows:

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=P+1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=P+1}^{N}(y_i - \bar{y})^2} \tag{4.1}$$

where $\boldsymbol{y}$ represents the observed time series, $\bar{y}$ is the mean of this time series, $\hat{\boldsymbol{y}}$ is the predicted time series obtained from a given forecasting model, and P is the length of the lag-time moving window. Therefore, the R^2 can be interpreted as the fraction of explained variance by the forecasting model, and it increases when the performance of the model increases, reaching the theoretical optimum of 1 for an error-free forecast, and being negative when the predictions are less representative of the observations than the mean of the observations. Using R^2, one can now define Granger causality in a more formal way.

Definition 1. *We say that time series $\boldsymbol{x}$ Granger-causes $\boldsymbol{y}$ if $R^2(\boldsymbol{y}, \hat{\boldsymbol{y}})$ increases when $x_{t-1}, x_{t-2}, ..., x_{t-P}$ are included in the prediction of y_t, in contrast to considering $y_{t-1}, y_{t-2}, ..., y_{t-P}$ only, where P is the lag-time moving window (Granger, 1969).*

In climate sciences, linear vector autoregressive (VAR) models are often employed to make forecasts (Stock and Watson, 2001; Triacca, 2005; Kodra et al., 2011;

Attanasio, 2012). A linear VAR model of order P boils down to the following representation:

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} w_{01} \\ w_{02} \end{bmatrix} + \sum_{p=1}^{P} \begin{bmatrix} w_{11p} & w_{12p} \\ w_{21p} & w_{22p} \end{bmatrix} \begin{bmatrix} y_{t-p} \\ x_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \tag{4.2}$$

with w_{ij} being parameters that need to be estimated and ϵ_1 and ϵ_2 referring to two white noise error terms. This model can be used to derive the predictions required to determine Granger causality. In that sense, time series $\boldsymbol{x}$ Granger-causes time series $\boldsymbol{y}$ if at least one of the parameters w_{12p} for any p significantly differs from zero. Specifically, and since we are focusing on the vegetation time series as the only target, the following two models are compared:

$$y_t = \hat{y}_t + \epsilon_1 = w_{01} + \sum_{p=1}^{P} \left(w_{11p} y_{t-p} + w_{12p} x_{t-p} \right) + \epsilon_1 \tag{4.3}$$

$$y_t = \hat{y}_t + \epsilon_1 = w_{01} + \sum_{p=1}^{P} w_{11p} y_{t-p} + \epsilon_1 \tag{4.4}$$

We will refer to model in Eq. (4.3) as the *full model* and to model in Eq. (4.4) as the *baseline model*, since the former incorporates all available information and the latter only information of $\boldsymbol{y}$. Note that the two models are nested and the baseline model should capture all the information related to the history of the target time series.

Comparing the above two models, $\boldsymbol{x}$ Granger-causes $\boldsymbol{y}$ if the full model manifests a substantially better predictive performance in terms of R^2 than the baseline model. To this end, statistical tests can be employed, for which one typically assumes that the errors in the model follow a Gaussian distribution (Maddala and Lahiri, 1992). However, our above definition differs from the perspective in research papers that develop statistical tests for Granger causality (Hacker and Hatemi-J, 2006), because we intend to move away from statistical hypothesis testing. This is because the assumptions behind such testing are typically violated when working with climate data where neither variables nor observational techniques are fully independent from each other in most cases, and errors are not normally distributed (see Sect. 4.2.4 for a further discussion).

In climate studies, the Granger causal relationship between two time series $\boldsymbol{x}$ and $\boldsymbol{y}$ has often been investigated in the bivariate setting (Elsner, 2006, 2007; Kodra et al., 2011; Attanasio, 2012; Attanasio et al., 2012). However, such an analysis might lead to incorrect conclusions, because additional (confounding) effects exerted by other climatic or environmental variables are not taken into account (Geiger et al., 2015). This problem can be mitigated by considering time series of additional variables. For example, let us assume one has observed a third variable $\boldsymbol{z}$, which

might act as a confounder in deciding whether $\boldsymbol{x}$ Granger-causes $\boldsymbol{y}$. The above definition then naturally extends as follows.

Definition 2. *We say that time series $\boldsymbol{x}$ Granger-causes $\boldsymbol{y}$, if $R^2(\boldsymbol{y}, \hat{\boldsymbol{y}})$ increases when $x_{t-1}, x_{t-2}, ..., x_{t-P}$ are included in the prediction of y_t, in contrast to considering $y_{t-1}, y_{t-2}, ..., y_{t-P}$ and $z_{t-1}, z_{t-2}, ..., z_{t-P}$ only, where P is the lag-time moving window.*

Similarly as above, we refer to the two models as full and baseline model, respectively. Therefore, in the tri-variate setting, Granger causality might be tested using the following linear VAR model:

$$\begin{bmatrix} y_t \\ x_t \\ z_t \end{bmatrix} = \begin{bmatrix} w_{01} \\ w_{02} \\ w_{03} \end{bmatrix} + \sum_{p=1}^{P} \begin{bmatrix} w_{11p} & w_{12p} & w_{13p} \\ w_{21p} & w_{22p} & w_{23p} \\ w_{31p} & w_{32p} & w_{33p} \end{bmatrix} \begin{bmatrix} y_{t-p} \\ x_{t-p} \\ z_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}, \tag{4.5}$$

where a causal relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ exists if at least one w_{12p} significantly differs from zero. As previously mentioned, the time series $\boldsymbol{z}$ might also have a causal effect on $\boldsymbol{y}$ and be correlated with $\boldsymbol{x}$. For this reason $\boldsymbol{z}$, should be included in both models (baseline and full), so that the method can cope with cross-correlations between predictors, in our case between the climatic drivers of vegetation anomalies. An extension of this definition for more than three times series is straightforward.

4.2.2. Over-fitting and out-of-sample testing

It is well known in the statistical literature that predictions made on in-sample data, that is, the same data that was used to fit the statistical model, tend to be optimistic. This process is often referred to as over-fitting, i.e., by definition, the fitting process leads to parameter values that cause the model to mimic the observed data as closely as possible (Friedman et al., 2001). In the context of Granger causality analysis, over-fitting will occur more prominently in the multivariate case, when the number of considered time series increases. The results in Sect. 4.3 are based on multivariate analysis, thus they are vulnerable to over-fitting; the situation further aggravates when switching from linear to non-linear models, because then the number of parameters typically increases to allow for a more flexible functional model form.

To prevent over-fitting, out-of-sample data should be used in evaluating the predictive performance in Granger causality studies (Gelper and Croux, 2007). The most straightforward procedure for creating out-of-sample data is to separate the time frame into two parts, a training set and a test set, which typically constitute the first and last half of the time frame. A few authors have adopted this approach for climatic attribution (Attanasio et al., 2012; Pasini et al., 2012); however, satellite

Earth observation time series are usually too short to allow for train-test splitting in that fashion. An alternative approach, which uses the available data in an efficient manner, is cross-validation. To this end, the time frame is divided in a number of short intervals, typically a few years of data, in which one interval serves as a test set, while all remaining data are used for parameter fitting. This procedure is repeated until all intervals have served once as a test set, and the prediction errors obtained in each round are aggregated, so that one global performance measure can be computed (see Chapter 2). We direct the reader to Michaelsen (1987) and Von Storch and Zwiers (2001) for further discussion.

The inclusion of a regularization term in the fitting process of over-parameterized linear models will avoid over-fitting. Typical regularizers that shrink the parameter vectors of linear models towards zero are L2-norms as in ridge regression, L1-norms as in LASSO models, or a combination of the two norms, as in elastic net (Friedman et al., 2001). Translated to VAR models, this implies that one should impose restrictions on the parameter matrix of Eq. (4.5), as done in the recent theoretical paper of Gregorova et al. (2015). In this setting, we want to identify causal relationships between a vegetation time series and various climatic time series. Hence, there is only one target variable of interest, and a simpler approach can be adopted. Denoting the vegetation time series by $\boldsymbol{y}$, one can mimic in the tri-variate setting a VAR model by means of three autoregressive ridge regression models:

$$y_t = \hat{y}_t + \epsilon_1 = w_{01} + \sum_{p=1}^{P} \left(w_{11p} y_{t-p} + w_{12p} x_{t-p} + w_{13p} z_{t-p} \right) + \epsilon_1 \quad (4.6)$$

$$x_t = \hat{x}_t + \epsilon_2 = w_{02} + \sum_{p=1}^{P} \left(w_{21p} y_{t-p} + w_{22p} x_{t-p} + w_{23p} z_{t-p} \right) + \epsilon_2 \quad (4.7)$$

$$z_t = \hat{z}_t + \epsilon_3 = w_{03} + \sum_{p=1}^{P} \left(w_{31p} y_{t-p} + w_{32p} x_{t-p} + w_{33p} z_{t-p} \right) + \epsilon_3 \quad (4.8)$$

Our goal is to detect the climate drivers of vegetation, and not the feedback of vegetation on climate (see e.g., Green et al. (2017)). Therefore, it suffices to retain Eq. (4.6) in our analysis as is stated above for the tri-variate case (Eq. 4.5). Concatenating all parameters of this model into a vector $\boldsymbol{w} = [w_{01}, w_{11p}, ..., w_{13p}]$, one fits in ridge regression the parameters by solving the following optimization problem:

$$\min_{\boldsymbol{w}} \sum_{P+1}^{N} (y_t - \hat{y}_t)^2 + \lambda ||\boldsymbol{w}||^2 \quad (4.9)$$

with λ being a regularization parameter, that is tuned using a validation set or nested cross-validation and $||\boldsymbol{w}||^2$ being a penalty term, i.e., the squared L2-norm of the coefficient vector. The sum only starts at $P + 1$ because a moving window of P lags is considered. For simplicity, we describe the above approach for the

tri-variate setting, even though the total number of variables used in our study is a lot larger (see Chapter 3); nonetheless, extensions to the multivariate setting are straightforward.

4.2.3. Non-linear Granger causality

The methodology that we develop in this thesis is closely connected to the methods explained in the previous section. However, as we hypothesize that the relationships between climate and vegetation can be non-linear (Foley et al., 1998; Zeng et al., 2002; Verbesselt et al., 2016), we replace the linear VAR-models in the Granger causality framework with non-linear machine learning models. In other fields, such as in neurosciences, kernel methods or other non-linear models have been used for the investigation of non-linear Granger causality relationships between time series (Ancona et al., 2004; Marinazzo et al., 2008). In our analysis, we stick to simple non-linear methods that are applicable to large data sets. More sophisticated approaches typically do not scale well enough in global climate–vegetation data sets. Therefore, in our approach, the machine learning algorithm we choose is random forests, due to its excellent computational scalability (Breiman, 2001). Random forests are a well-known method that has shown its merits in diverse application domains, and that has successfully been applied to Earth observations in both classification and regression problems (Dorigo et al., 2012; Rodriguez-Galiano et al., 2012; Loosvelt et al., 2012a,b). However, random forests has not been applied yet in the context of Granger causality. Briefly summarized, the random forest algorithm forms a combination of multiple decision trees, where each tree contributes with a single vote to the final output, which is the most frequent class (for classification problems) or the average (for regression problems) – see Chapter 2.

Compared to most application domains where random forests are applied, we employ the algorithm in a slightly different way, as an autoregressive non-linear method for time series forecasting. In practice, this means that we replace the full and baseline linear model of Sect. 4.2.1 by a random forest model. At each pixel, the vegetation time series is still considered as response variable, and the various climate time series serve as predictor variables (see Chapter 3 for an overview of our database). For a given value of the NDVI time series $\boldsymbol{y}$ at timestamp t, we investigate properties of the different predictor time series – i.e., temperature, radiation, etc. – by considering a moving window including a number of previous months (Fig. 4.1). In this way, the definition of Granger causality in Sect. 4.2.1 is adopted. Any climatic time series $\boldsymbol{x}$ Granger-causes vegetation time series $\boldsymbol{y}$ if the predictive performance in terms of R^2 improves when the moving window $x_{t-1}, x_{t-2}, ..., x_{t-P}$ is incorporated in the random forest, in contrast to considering $y_{t-1}, y_{t-2}, ..., y_{t-P}$ and $z_{t-1}, z_{t-2}, ..., z_{t-P}$ only. Analogous to the linear case, we will speak of a full random forest model when all variables are taken into account and of a baseline random forest model when only the moving window $y_{t-1}, y_{t-2}, ..., y_{t-P}$

of $\boldsymbol{y}$ is considered as predictor. In Fig. 4.1, this principle is extended to four time series. The baseline random forest predictions of NDVI at t_1 are based on the observations from the green moving window only, whereas the full random forest model includes the three red moving windows as well.

In our experiments, we treat each continental pixel as a separate problem, and use the scikit-learn library (Pedregosa et al., 2011) for the random forest regressor implementation, with the number of trees equal to 100 and the maximum number of predictor variables per node equal to the square root of the total number of predictor variables. Different values in these parameters do not show large impact on the results. Changes in these parameters or in the randomness of the algorithm do not cause substantial changes in the results (not shown). Model performance is assessed by means of five-fold cross-validation. The window length is fixed to twelve months because initial experimental results revealed that longer lag time windows did not lead to improvements in the predictions (results omitted). Finally, we also experimented with techniques that exploit spatial correlations to improve the predictive performance of the model (see Sect. 4.3.3).

4.2.4. Granger causal inference

Generally, the null hypothesis (H_0) of Granger causality is that the baseline model has equal prediction error as the full model. Alternatively, if the full model predicts the target variable $\boldsymbol{y}$ significantly better than the baseline model, H_0 is rejected. In some applications, inference is drawn in VAR by testing for significance of individual model parameters. Other studies have used likelihood-ratio tests, in which the full and baseline models are nested models (Mosedale et al., 2006). However, in both cases, the models are trained and evaluated on the same in-sample data. As it has been discussed above, the performance of any Granger causal model should be validated on out-of-sample data to avoid overfitting (see Sect. 4.2.2). Therefore, the null hypothesis of non-causality in the formulation stated above should be tested for by comparing out-of-sample prediction errors. To this end, statistical tests have been proposed and applied both in the econometric literature as well as in Granger causality studies in the context of climate science. This kind of tests, which compare out-of-sample prediction errors, are available for models for which parameter estimation is done through ordinary least squares or maximum likelihood estimation (Attanasio et al., 2013). Moreover, the asymptotic and finite-sample properties of a battery of tests for comparing forecasting accuracies of different models have been studied and more recently, further tests aiming specifically at nested models have been proposed (Clark and McCracken, 2001).

Unfortunately, all the tests mentioned above were designed to compare the out-of-sample prediction errors of linear parametric models (McCracken, 2007). In climate, relations between variables are non-linear and tend to become even more non-linear as the temporal resolution of the data becomes finer (Attanasio et al.,

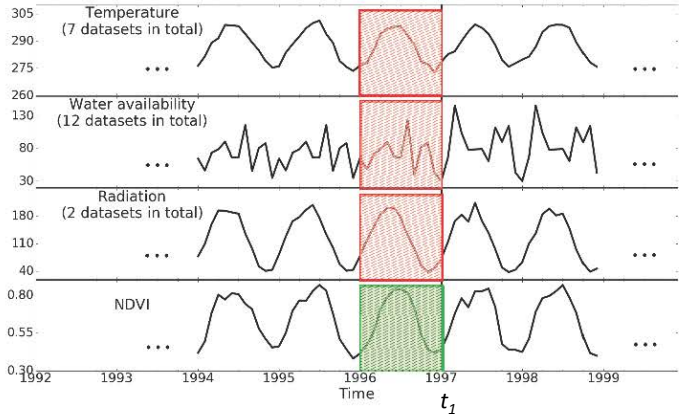


Figure 4.1: An illustrative example of the moving window approach considered in the analysis of vegetation drivers at a given timestamp t_1. NDVI takes here the role of the time series $\boldsymbol{y}$ in Eq. 4.3. In addition three climate predictor time series are shown. The baseline random forest model only considers the green moving window, whereas the full random forest model includes the red moving windows as well. The pixel corresponds to a location in North America (lat: 37.5, long: -87.5).

2013). Therefore, it would be convenient to have at our disposal a statistical test to assess the significance of any quantitative evidence of climate Granger-causing vegetation anomalies that we can find. Ideally, the test would be model-independent so that any non-linear model could be used. One well-known model-independent test to compare the accuracy of two forecasts is the Diebold-Mariano test (DM-test) (Diebold, 2015a). Although its application to Granger causality is promising, the test does not hold for nested models, because under H_0, the prediction errors from two nested models are exactly the same and perfectly correlated (McCracken, 2007). An alternative approach for comparing the predictive performance of different models is to use resampling methods such as the bootstrap or schemes such as 5×2 cross-validation (Dietterich, 1998). Methods based on the bootstrap have been used before in Granger causality studies with climate data (Diks and Mudelsee, 2000; Attanasio et al., 2013). However, these results need to be interpreted with care because, by increasing the number of bootstrap samples, the power of any paired test (such as the Wilcoxon signed rank test) to detect significant differences between the error distributions of both models (full and baseline) increases as well. For these reasons, we conclude that developing a statistical test that is able to handle non-stationary time series and non-linear models is not a trivial task. To the best of our knowledge, no such test exists in the current literature. In this paper, we focus on expressing Granger causality in a quantitative instead of a qualitative way, and stress the gained improvement with the use of a non-linear model.

4.3. Results and discussion

4.3.1. Detecting linear Granger-causal relationships

In a first experiment, we evaluate the extent to which climate variability Granger-causes the anomalies in vegetation using a standard Granger causality approach, in which only linear relationships between climate (predictors) and vegetation (target variable) are considered. To this end, ridge regression is used as a linear vector autoregressive (VAR) model in the Granger causality approach (note this ridge regression will be substituted by the non-linear random forest approach in Sect. 4.3.2). In the application of the ridge regression, we use all climatic and environmental predictor variables (Sect. 3.2.2), and adopt a nested five-fold cross-validation to properly tune the hyper-parameter λ (see Eq. 4.9). Figure 4.2a shows the predictive performance of the full ridge regression model. While the model explains more than 40% of the variability in NDVI anomalies in some regions ($R^2 > 0.4$), this is by itself not necessarily indicative of climate Granger-causing the vegetation anomalies, as it may reflect simple correlations. In order to test the latter, we compare the results of the full model to a baseline model, i.e., an autoregressive ridge regression model that only uses previous values of NDVI to predict the NDVI at time t (see Sect. 4.2.1). If climate Granger-caused the variability of NDVI at a given pixel, the full ridge regression model (Fig. 4.2a) would show an increase in the predictive power over the predictions based on the baseline ridge regression model. However, the results unequivocally show that – when only linear relationships between vegetation and climate are considered – the areas for which vegetation anomalies are Granger-caused by climate are very limited, involving mainly semiarid regions and central Europe (Fig. 4.2b).

For further comparison, we analyze the predictive performance obtained when (linear) Pearson correlation coefficients are calculated on the training data sets, selecting the highest correlation to the target variable for any of the 4,571 predictor variables at each pixel. Figure 4.2c shows that the explained variance is again rather low, and for most regions substantially lower than the R^2 of the baseline ridge regression model, here considered as the minimum to interpret this predictive power as Granger-causal. These results indicate that, despite being routinely used as a standard tool in climate–biosphere studies (see e.g. Nemani et al., 2003), univariate correlation analyses are unable to extract the nuances of the relationships between climate and vegetation dynamics.

4.3.2. Linear versus non-linear Granger causality

To analyze the effect of climate on vegetation more thoroughly, we substitute the linear ridge regression model (VAR) by the non-linear random forest model.

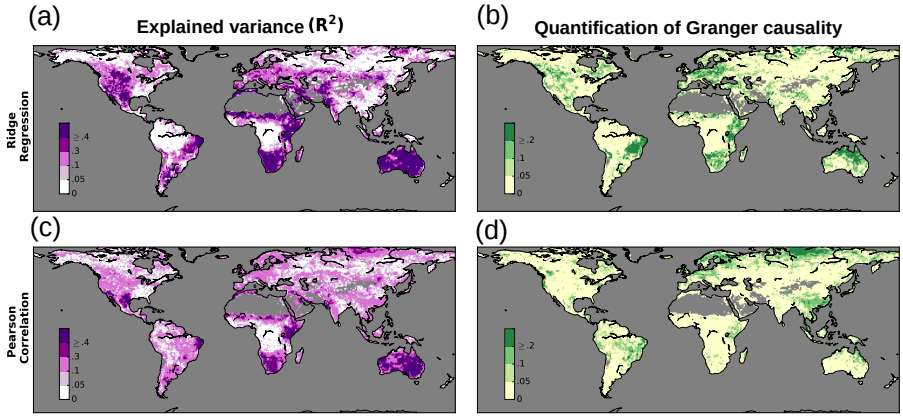


Figure 4.2: Linear Granger causality of climate on vegetation. **(a)** Explained variance (R^2) of NDVI anomalies based on a full ridge regression model in which all climatic variables are included as predictors. **(b)** Improvement in terms of R^2 by the full ridge regression model with respect to the baseline ridge regression model that uses only past values of NDVI anomalies as predictors; positive values indicate (linear) Granger causality. **(c)** A filter approach in which the variable with the highest squared Pearson correlation against the NDVI anomalies is selected. **(d)** Improvement in terms of R^2 by the filter approach with respect to the same baseline ridge regression model that uses only past values of NDVI anomalies.

Results in Fig. 4.3 highlight the differences. Compared to the results in Sect. 4.3.1, the predictive power substantially increases by considering non-linear relationships between vegetation and climate (Fig. 4.3a). This is the case for most land regions, but is especially remarkable in semiarid regions of Australia, Africa, Central and North America, which are frequently exposed to water limitations. In those regions, more that 40% of the variance of NDVI anomalies can be explained by antecedent climate variability. These results are further investigated by Papagiannopoulou et al. (2017b), who highlight the crucial role of water supply for the anomalies in vegetation greenness in these and other regions. On the other hand, the variance of NDVI explained in other areas such as the Eurasian taiga, tropical rainforests or China is again below 10%. We hypothesize two potential reasons: (a) the uncertainty in the observations used as target and predictors are typically larger in these regions (especially in tropical forests and at higher latitudes), and (b) these are regions in which vegetation anomalies are not necessarily primarily controlled by climate, but may be predominantly driven by phenological and biotic factors (Hutyra et al., 2007), occurrence of wild fires (van der Werf et al., 2010), limitations imposed by the availability of soil nutrients (Fisher et al., 2012a) or agricultural practices (Liu et al., 2015a). Nonetheless, the explained variance shown in Fig. 4.3a is again not necessarily indicative of Granger causality. As we did

in Fig. 4.2b, in order to test whether the climatic and environmental controls do, in fact, Granger-cause the vegetation anomalies, we compare the results of our full random forest model to a baseline random forest model which only uses previous values of NDVI to predict the NDVI at time t. As seen in Fig. 4.3b, in this case, the improvement over the baseline is unambiguous. One can conclude that – while not bearing into consideration all potential control variables in our analysis – climate dynamics indeed Granger-cause vegetation anomalies in most of the continental land surface, with a larger impact on subtropical regions and mid-latitudes. Moreover, a comparison between Figs. 4.2b and 4.3b unveils that these causal relationships are non-linear, as expected given the distinct resistence and resilience of different ecosystems, which is reflected by a progressive response and recovery of vegetation to these perturbations (Foley et al., 1998; Zeng et al., 2002; Verbesselt et al., 2016).

For a better understanding of the results obtained by the two models, we average the performance of each model regionally. More specifically, we use the International Geosphere-Biosphere Program (IGBP) (Loveland and Belward, 1997) land cover classification to stratify the mean and variance of R^2 for both the baseline and the full model in Fig. 4.3 per IGBP land cover class. The barplot in Fig. 4.4 shows that the full model outperforms the baseline model in all IGBP land cover classes, i.e., that Granger causality exists for all these biomes. In the parentheses, we note the number of pixels per region. The error bars indicate that the variances of the two models are analogous, i.e., they are low or high in both models in the same land cover class. For the Closed Shrublands region, one can observe the highest difference between the two models, yet only 19 pixels belong to this biome type. In savanna regions, the performance of the full model is high in comparison with other regions (see Fig. 4.3). On the other hand, the lowest performance improvement of the full model with respect to the baseline is observed for the regions of Deciduous Needleleaf Forests and Evergreen Broadleaf Forests. This shows that for these two regions, climate is not identified as a major control over vegetation dynamics (see discussion in previous paragraph about tropical and boreal regions).

4.3.3. Spatial and temporal aspects

Environmental dynamics reveal their effect on vegetation at different time scales. Since the adaptation of vegetation to environmental changes requires some time, and because soil and atmosphere have a memory, a necessary aspect to investigate is the potential lag-time response of vegetation to climate dynamics which relates to the ecosystem resistence and resilience properties. The idea of exploring lag times was introduced by several studies in the past (see e.g., Davis (1984); Braswell et al. (1997)), and it has been adopted in various studies more recently (Anderson et al., 2010; Kuzyakov and Gavrichkova, 2010; Chen et al., 2014; Rammig et al., 2014).

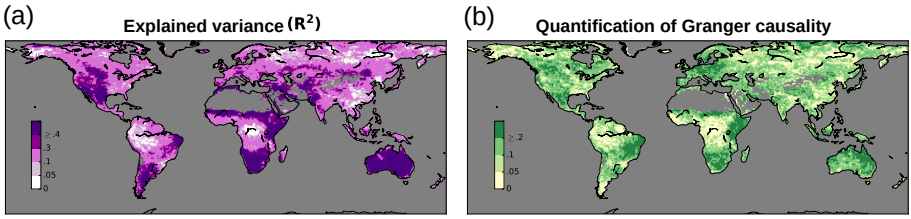


Figure 4.3: Non-linear Granger causality of climate on vegetation. **(a)** Explained variance (R^2) of NDVI anomalies based on a full random forest model in which all climatic variables are included as predictors. **(b)** Improvement in terms of R^2 by the full random forest model with respect to the baseline random forest model that uses only past values of NDVI anomalies as predictors; positive values indicate (non-linear) Granger causality.

These studies indicate that lag times depend on both the specific climatic control variable and the characteristics of the ecosystem. As explained in Chapter 3, in our analysis shown in Fig. 4.2 and 4.3, we moved beyond traditional cross-correlations, and incorporate higher-lever variables in the form of cumulative and lagged responses to extreme climate. As mentioned in Sect. 4.2.3, our experiments indicated that lags of more than six months do not add extra predictive power (not shown), even though the effect of anomalies in water availability on vegetation can extend for several months (Papagiannopoulou et al., 2017b).

To disentangle the response of vegetation to past cumulative climate anomalies and climatic extremes, Fig. 4.5a visualizes the predictive performance when cumulative variables and extreme indices are not included as predictive variables in the random forest model. As shown in Fig. 4.5b, in almost all regions of the world the predictive performance decreases substantially compared to the full random forest model approach, i.e., using the full repository of predictors (Fig. 4.3a), especially in regions such as the Sahel, the Horn of Africa or North America. In those regions 10-20% of the variability in NDVI is explained by the occurrence of prolonged anomalies and/or extremes in climate, illustrating again the non-linear responses of vegetation. For more detailed results about lagged vegetation responses for specific climate drivers and the effect of climate extremes on vegetation, the reader is referred to Chapter 5.

Because of uncertainties in the observational records used in our study to represent climate and predict vegetation dynamics, and given that ecosystems and regional climate conditions usually extend over areas that exceed the spatial resolution of these records, one may expect that the predictive performance of our models becomes more robust when including climate information from neighbouring pixels. In addition, it is quite likely that neighbouring areas have similar climatic conditions, which in their turn affect vegetation dynamics in a similar manner. We therefore also consider an extension of our framework to exploit spatial autocorrelations, inspired

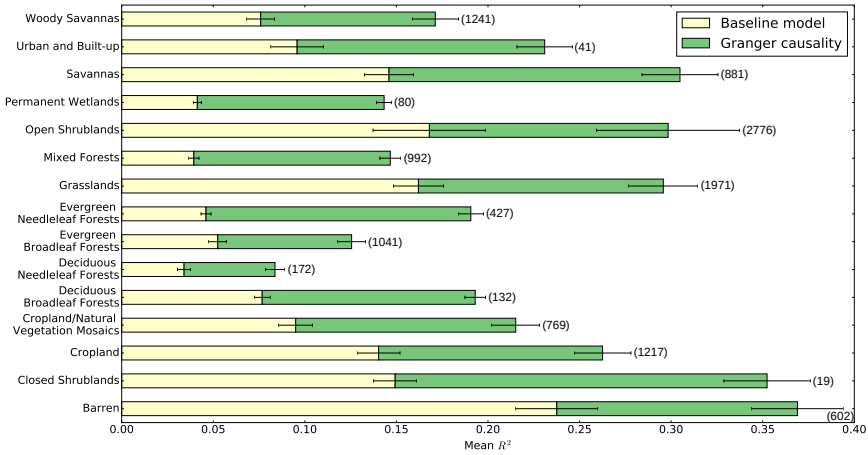


Figure 4.4: Mean R^2 and variance per IGBP land cover class for both the baseline and full random forest model. The green part indicates the improvement in performance of the full model with respect to the baseline, i.e., the quantification of Granger causality (as in Fig. 4.3b). The number of pixels per IGBP class is noted in the parentheses.

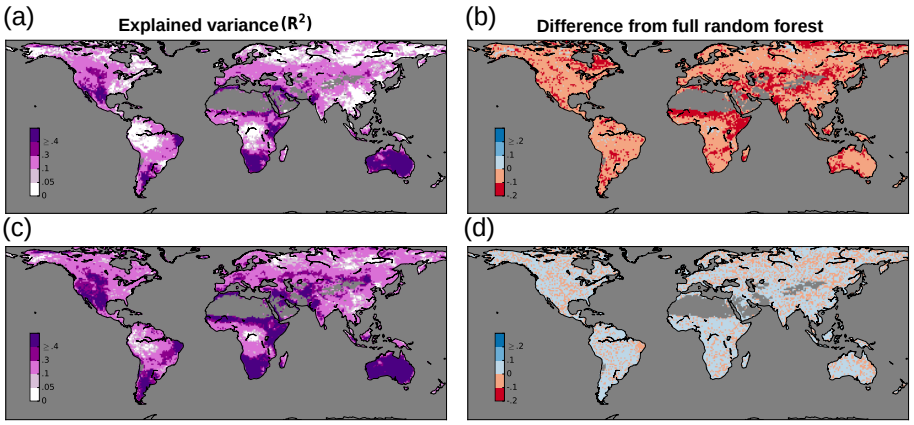


Figure 4.5: Analysis of spatio-temporal aspects of our framework. **(a)** Explained variance (R^2) of NDVI anomalies based on a full random forest model in which all climatic variables are included as predictors as in Fig. 4.3a, except for the cumulative variables and the extreme indices (see Chapter 3). **(b)** Difference in terms of R^2 between the model without cumulative and extreme predictors and the full random forest model in Fig. 4.3a. **(c)** Explained variance (R^2) of NDVI anomalies based on a full random forest model in which all climatic variables are included as predictors as in Fig. 4.3a, but including also the predictors from the eight nearest neighbours. **(d)** Difference in terms of R^2 between this full random forest model which includes spatial information from neighbouring pixels and the full random forest model in Fig. 4.3a.

by Lozano et al. (2009b), who achieved spatial smoothness via an additional penalty term that punishes dissimilarity between coefficients for spatial neighbours. In our analysis, we incorporate at a given pixel spatial autocorrelations by extending the predictor variables of our models with the predictor variables of the eight neighbouring pixels. We provide such an extension both for the full and the baseline random forest model. As such, for the full random forest model, a vector of 41,139 (4,571 $\times$ 9) predictor variables is formed for each pixel.

Figure 4.5c illustrates the performance of the full random forest model that includes the spatial information. As one can observe in Fig. 4.5d, the explained variance of NDVI anomalies remains similar to the original model that depicts the same approach without spatial autocorrelation (Fig. 4.3a). While in most areas the performance slightly increases, the explained variance never improves by more than 10%; as a result, incorporating spatial autocorrelations in our framework does not seem to further improve the quantification of Granger causality and is not considered in further applications of the framework (see Papagiannopoulou et al. (2017b)). A possible explanation for this result is that the model without the spatial information cannot be outperformed because of the large dimensionality of the feature space, which may include redundant information, in combination with the low number of observations per pixel (Fig. 4.3a). Note that in this case the number of observations per pixel remains the same as in the original model (360 observations) while the number of predictor variables is nine times larger.

4.3.4. The importance of focusing on vegetation anomalies

In Chapter 3, we advocated that Granger causality analysis should target on NDVI anomalies, as opposed to raw NDVI values. There are several fundamental reasons for this. First, by applying a decomposition, one can subtract long-term trends from the NDVI time series, making the resulting time series more stationary. This is absolutely needed, as existing Granger causality tests cannot be applied for non-stationary time series. Secondly, by subtracting the seasonal cycle from the time series, one is not only able to remove a confounding factor that may contribute predictive power without bearing causality, but is also able to remove a clear autoregressive component that can be well explained from the NDVI time series themselves. As vegetation has a strong seasonal cycle, it is not difficult to predict subsequent vegetation conditions by using the past observations of the seasonal cycle only. To corroborate this aspect, we repeat our analysis in Sect. 4.3.2, but this time considering the raw NDVI time series instead of the NDVI anomalies as the target variable. We again compare the full and the baseline random forest models.

The results are visualized in Fig. 4.6a. As it can be observed, worldwide the R^2 is close to the optimum of one. However, due to the overwhelming domination of the

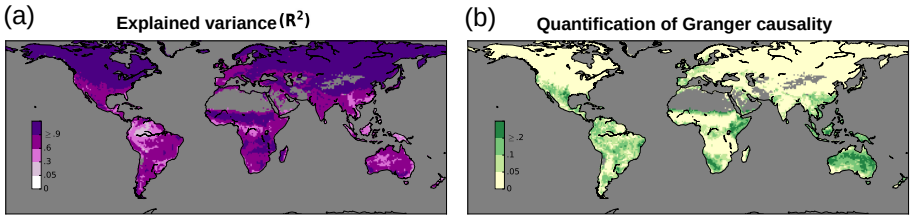


Figure 4.6: Comparison of model performance with R^2 as metric with the raw NDVI time series as target variable. **(a)** Full random forest model **(b)** Improvement in terms of R^2 of the full random forest model over the baseline random forest model.

seasonal cycle, it becomes very difficult, or even impossible, to unravel any potential Granger-causal relationships with climate time series in the Northern Hemisphere – see Fig. 4.6b. The predictability of NDVI based on the seasonal NDVI cycle itself is already so high that nothing can be gained by adding additional climatic predictor variables; see also the large amplitude of the seasonal cycle of NDVI at those latitudes compared to the NDVI anomalies, as illustrated in Chapter 3, Fig. 3.1. Therefore, a non-linear baseline autoregressive model is able to explain most of the variance in the time series. Moreover, as observed in Fig. 4.1, temperature and radiation also manifest strong seasonal cycles that often coincide with the NDVI cycle. For most regions on Earth, such a stationary seasonal cycle is less present for variables, such as precipitation. This can potentially yield wrong conclusions, such as that temperature in the Northern hemisphere is driving most NDVI variability, since the two seasonal cycles have the same pattern. However, based on the above discussion, it becomes clear that results of that kind should be treated with caution: for climate data, a Granger causality analysis should be applied after decomposing time series into seasonal anomalies.

4.4. Conclusions

In this chapter, we introduced a novel framework for studying Granger causality in climate-vegetation dynamics. Our approach consists of the combination of data fusion, feature construction and non-linear predictive modelling. The choice of random forests as a non-linear algorithm has been motivated by its excellent computational scalability with regards to extremely large data sets, but could be easily replaced by any other non-linear machine learning technique, such as neural networks or kernel methods.

Our results highlight the non-linear nature of climate–vegetation interactions and the need to move beyond the traditional application of Granger causality within a linear framework. Comparisons to linear Granger causality approaches indicate

that the random forest framework can predict 14% more variability of vegetation anomalies on average globally. The predictive power of the model is especially high in water-limited regions where a large part of the vegetation dynamics responds to the occurrence of antecedent rainfall. Moreover, our results indicate the need to consider multi-month antecedent periods to capture the effect of climate on vegetation, in particular to account for the effects of climate extremes on vegetation resilience. The reader is referred to Papagiannopoulou et al. (2017b) for a detailed analysis of the effect of different climate predictors on the variability of global vegetation using the mathematical approach described here.

5 Detecting the main vegetation drivers at global scale

Quantifying environmental controls on vegetation is critical to predict the net effect of climate change on global ecosystems and the subsequent feedback on climate. Following the non-linear Granger causality framework described in Chapter 4, we aim to uncover the main drivers of monthly vegetation variability at the global scale. Results indicate that water availability is the most dominant factor driving vegetation globally: about 61% of the vegetated surface was primarily water-limited during 1981-2010. Intra-annually, temperature controls Northern Hemisphere deciduous forests during the growing season, while antecedent precipitation largely dominates vegetation dynamics during the senescence period. The uncovered dependency of global vegetation on water availability is substantially larger compared to previous works. This is owed to the ability of the framework to (1) disentangle the co-linearities between radiation/temperature and precipitation, and (2) quantify non-linear impacts of climate on vegetation. Our results reveal a prolonged effect of precipitation anomalies in dry regions: due to the long memory of soil moisture and the cumulative, non-linear, response of vegetation, water-limited regions show sensitivity to the values of precipitation occurring three months earlier. Meanwhile, the impacts of temperature and radiation anomalies are more immediate and dissipate shortly, pointing to a higher resilience of vegetation to these anomalies. Despite being infrequent by definition, hydro-climatic extremes are responsible for up to 10% of the vegetation variability during the 1981-2010 period in certain areas, particularly in water-limited ecosystems. Our approach is a first step towards a quantitative comparison of the resistance and resilience signature of different ecosystems, and can be used to benchmark Earth system models in their representations of past vegetation sensitivity to changes in climate.

This chapter is an edited version of:
Papagiannopoulou, C., Miralles, D. G., Dorigo, W. A., Verhoest, N. E. C., Depoorter, M. and Waegeman, W.: Vegetation anomalies caused by antecedent precipitation in most of the world. Environ. Res. Lett., 12(7):074016, 2017.

5.1. Introduction

Vegetation is a key player in the climate system, constraining atmospheric conditions through a series of positive and negative feedbacks. Plants regulate water, energy and carbon cycles, through their transfer of vapour from land to atmosphere (i.e., transpiration, interception loss), effects on the surface radiation budget (e.g., albedo,

surface temperature, emission of volatile organic compounds), exchange of carbon dioxide with the atmosphere (i.e., photosynthesis, respiration), and influence on wind circulation (Bonan, 2008; Mcpherson, 2007; Teuling et al., 2017). Vegetation holds around 42% (∼28 Pg C) of the terrestrial carbon storage and assimilates about 20% of the annual anthropogenic emissions of carbon dioxide (Pan et al., 2011; Le Quere et al., 2016). This fundamental role highlights the importance of understanding the regional drivers of ecological sensitivity and the response of vegetation to climatic changes at the global scale.

Vegetation dynamics are generally driven by climate, in particular by precipitation, incoming radiation, air temperature and atmospheric humidity (Nemani et al., 2003). In addition, nutrient availability (e.g., atmospheric CO_2 concentrations, soil chemicals) and short-term natural and anthropogenic disturbances (e.g., fires, volcanic eruptions, logging, insect epidemics) can be crucial at various spatio-temporal scales (Fisher et al., 2012b; Reichstein et al., 2013; Le Quere et al., 2016; Zhu et al., 2016). Consequently, humans impact vegetation dynamics directly through land-use change or agricultural management, and indirectly through air pollution, induced changes in climate and spread of pest outbreaks (Baccini et al., 2012; Reichstein et al., 2013). In natural conditions, long-term climatological controls on vegetation dominate: this is reflected in the general distribution of continental biomes, largely based on the annual cycle of solar irradiance, mean temperature, and the intensity of dry and wet seasons (Kottek et al., 2006). However, at shorter temporal scales, the interactions between vegetation and climate become complex and species-dependent (Zimmermann et al., 2009). Some vegetation types react preferentially to specific climatic changes, with different levels of intensity, resilience and lagged response (Wu et al., 2015; De Keersmaecker et al., 2015; Seddon et al., 2016). In addition, extreme climatic events – such as droughts, heatwaves, or heavy winds and storms – may cause long-lasting impacts and even bring ecosystems to a tipping point for collapse (Anderegg et al., 2015; Ciais et al., 2005; Reichstein et al., 2013; Verbesselt et al., 2016). Ultimately, the resistance and resilience of ecosystems to these anomalies depend on both vegetation characteristics and the duration and severity of climatic events (Anderegg et al., 2015; Cole et al., 2014).

A first and necessary step to understand how vegetation will respond to future climatic changes is to quantify the sensitivity of global ecosystems to past time climate variability. Conveniently, satellites routinely collect a wealth of information about the dynamics of our biosphere, hydrosphere and atmosphere: current multi-satellite composite records of environmental and climatic variables enable the study of global vegetation–climate interactions over multi-decadal time scales. Recent studies using long-term satellite records have indicated an overall greening trend (Zhu et al., 2016) and a long-term increase in above ground biomass (Liu et al., 2015b) – particularly at high latitudes and in the tropics – that have been attributed to CO_2 fertilization, warming trends and land-use change. Dominant

ecosystem drivers at inter- and intra-annual scales have also been intensively studied, both globally (Nemani et al., 2003; Poulter et al., 2014; De Keersmaecker et al., 2015; Wu et al., 2015; Gonsamo et al., 2016; Seddon et al., 2016) and over specific regions (Zhou et al., 2014; Barichivich et al., 2014; Guan et al., 2015). A particular example of a well-studied phenomenon is the short-term response of the Amazonian forest to precipitation scarcity and radiation, which has been the subject of intense debate over the past few years (Morton et al., 2014; Saleska et al., 2016). In the context of identifying the short-term (e.g. monthly) climatic controls on global vegetation dynamics, approaches based on correlations or multilinear regressions between climate and vegetation variables have led to important steps forward in our understanding (Nemani et al., 2003; Zhao and Running, 2010; Wu et al., 2015; De Keersmaecker et al., 2015; Seddon et al., 2016). However, these approaches are not designed to infer causality directly, and are commonly subjected to artifacts emerging from auto-correlation, non-linearity and cross-correlation between climatic drivers (Papagiannopoulou et al., 2017a). As mentioned in the previous chapters, the exponential increase in the volumes of satellite, *in situ* and reanalysis records existing today, together with the consistent progress of computing science, allow for more sophisticated data-driven methods to yield robust insights into the global interactions between vegetation and climate. As such, machine learning approaches are becoming increasingly valuable to investigate complex cause-effect relationships in geosciences, as well as to evaluate the skill of climate models in representing these interactions (Faghmous and Kumar, 2014). In Chapter 4, we presented an approach which adopts the well-known Granger (1969) causality framework – originally introduced in econometrics to quantify a measure of pseudocausality in time series – and extended it to capture the non-linearity of vegetation–climate relationships. This was achieved by substituting the traditional linear autoregressive model used in Granger-causality approaches with a non-linear random forest algorithm (Breiman, 2001). This new framework has clear advantages over simpler approaches: (a) it can cope with the emerging wealth of Earth observations while preventing over-fitting, (b) it enables a robust estimation of deterministic relationships, and (c) it incorporates the non-linear nature of vegetation–climate interactions.

5.2. Materials and methods

5.2.1. Data and feature construction

For our analysis, we used the database described in Chapter 3. This database consists of several predictive features constructed from 21 data sets. These predictive features consist of monthly time series for each 1° pixel, and include: raw data time series of each data set, seasonal anomalies (after subtraction of the seasonal cycle

based on the multi-year mean for each corresponding month of the year), de-trended seasonal anomalies (after subtraction of the long-term linear trend from the seasonal anomalies), lagged variables (with monthly lags up to six months into the past), cumulative variables (corresponding to the cumulative mean over the antecedent one to six months), and extreme indices (including the maximum and minimum of a variable per month, number of days per month exceeding a given threshold, values of specific percentiles, etc.). The lagged variables, cumulative variables and extreme indices were computed based on both raw data and (de-)trended seasonal anomalies. These predictive features are used to train the non-linear Granger causality framework (see Sect. 5.2.3), targeting the variable of de-trended NDVI seasonal anomalies (Tucker et al., 2005).

5.2.2. Non-linear Granger causality framework

Given a particular target time series, one speaks of the existence of Granger causality if the prediction of this target variable improves when information from other time series is taken into account in this prediction (Granger, 1969). Here, we quantify the extent to which a variable $\boldsymbol{x}$ (i.e., a predictive feature, or a certain group of them – see Sect. 5.2.3) is Granger-causing a target variable $\boldsymbol{y}$ (i.e., the de-trended NDVI anomalies at each individual pixel) by computing the increase in the variance of $\boldsymbol{y}$ that is explained by the random forest model predictions when $\boldsymbol{x}$ is included in the set of predictive features used by the model (this set also includes past values of $\boldsymbol{y}$ to conform to the definition of Granger causality). The explained variance is then defined as $R^2 = 1 - \frac{RSS}{TSS}$, with RSS being the sum of squared errors of the predictions (relative to the true de-trended NDVI anomalies), and TSS being the sum of the squared differences between the true values and their long-term mean, as defined in Chapter 4.

5.2.3. Sequential method to evaluate the impact of specific groups of features

To explore the importance of different climatic variables for the occurrence of NDVI anomalies, all predictive features have been aggregated into one of these three groups: 'temperature' (including surface and air temperature), 'radiation' (including incoming shortwave, longwave and net) and 'water' (including precipitation, surface and root-zone soil moisture, and snow water equivalent) – see Table 3.1 of Chapter 3. Then, taking the 'water' group as example, the explained variance (R^2) of NDVI anomalies by 'water' is calculated sequentially by: (1) applying the random forest approach to predict the anomalies of NDVI based on the entire database of predictors (including 'water', 'radiation' and 'temperature' features, but also past NDVI values to conform to the definition of Granger causality); (2) applying the random forest approach to the entire database except for the 'water' group; (3)

calculating the deterioration in the predictive performance (R^2) after excluding the 'water' group. Moreover, to prevent favouring groups with a larger number of predictive features, the number of selected features in every random forest is forced to be the same for all three groups by randomly selecting the same number of features for each group of variables.

As in most statistical techniques, the assessment of causality is ultimately limited to quantifying the level of cross-covariance between predictors and target variable, thus if critical predictors are not included, the importance of the assessed variable (or group of variables) may be inflated. However, the sequential approach explained above preserves the multivariate nature of the framework, as opposed to a hypothetical case in which the contribution of a specific group of variables (e.g., 'water') is assessed in isolation. In addition, the approach goes one step beyond previous statistical analyses of global vegetation drivers by preventing the importance of a secondary driver (e.g., temperature) to be inflated due to its correlation to the primary one (e.g., water availability). Nonetheless, the resulting R^2 is still not a measure of 'real' causality but of pseudo-causality, given the unfeasibility of including all possible drivers of global vegetation. Finally, we note that since the R^2 attributed to a particular variable (or group) is quantified by subtracting it from the entire database of predictors, its causal effect (computed as R^2) will be underestimated as long as the remaining variables are strongly correlated to that one being subtracted. As such, the R^2 reported here refers to the explained variance that is 'unique' to the variable (or group), i.e., the part of the variance in the NDVI anomalies that cannot be explained by any other variable in the database.

5.3. Results and discussion

5.3.1. Detecting important vegetation drivers

More than half (61%) of the vegetated area appears primarily controlled by water availability (i.e., precipitation, soil moisture or snow dynamics) – see Fig. 5.1a. In addition, for 17% of the remaining vegetated area, water availability is the second most important limiting factor after temperature or radiation (Fig. 5.1b). Temperature and radiation are the primary climatic controls for 23% and 15% of the vegetated areas, respectively. In addition, for most of these energy-driven regions the dynamics of vegetation are largely independent from climate variability (Fig. 5.1b). That is the case for both high latitudes and tropical zones, where no climatic driver is responsible for a substantial fraction of the variability in vegetation, as suggested by the inability of the framework to explain the dynamics in NDVI anomalies (see below). Nonetheless, in boreal and temperate regions, radiation and temperature remain the two main climatic controls, respectively, and

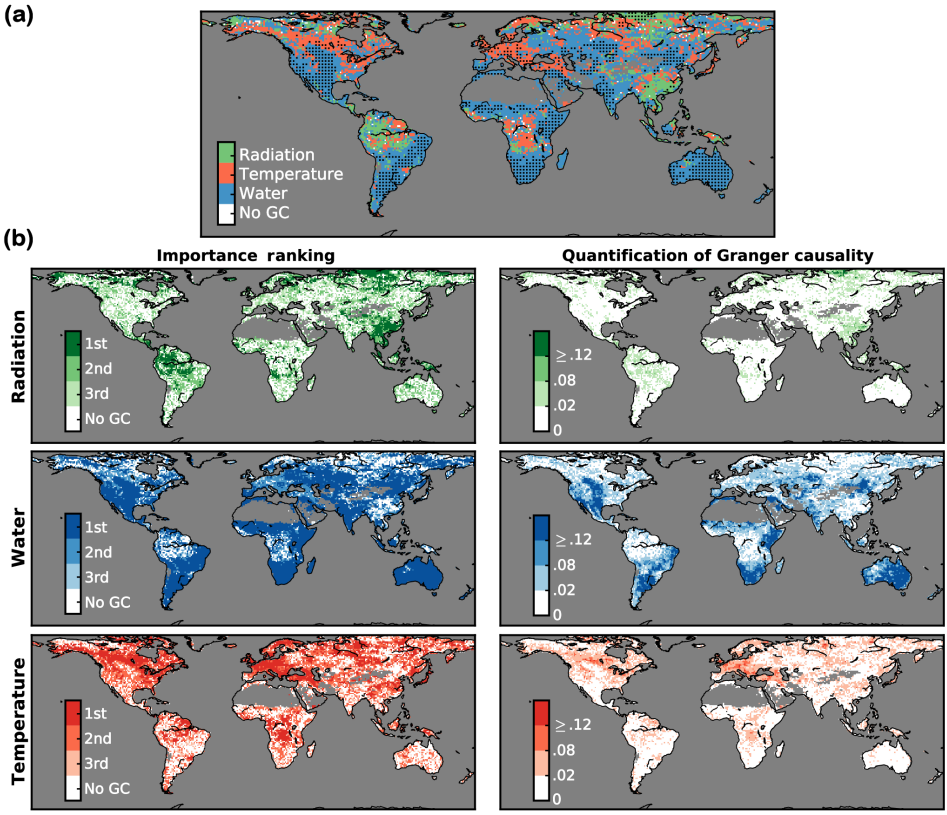


Figure 5.1: Primary climatic and environmental factors controlling vegetation dynamics. (a) Temperature, radiation and water-limited continental regions based on the non-linear Granger causality approach targeting de-trended NDVI anomalies. The net of black dots is represented at 2° resolution and indicates areas with $R^2 > 0.3$ for the full model including all variables. 'No GC' indicates no Granger causality ($R^2 \approx 0$). (b) Order of importance of each group of variables for vegetation according to the performance in terms of R^2 (left), and the corresponding R^2 (right). Grey colour indicates the regions considered as non-vegetated throughout the analysis.

most European croplands are temperature-driven (Fig. 5.1a and 5.1b). The relative importance of water availability in boreal regions such as Siberia or Alaska responds to the influence of snowmelt, which is nonetheless controlled by temperature and radiation patterns (Barichivich et al., 2014). Meanwhile, central Europe is mostly temperature-driven, while China is largely controlled by radiation. These patterns are in general agreement with those by Nemani et al. (2003); Wang et al. (2011); Wu et al. (2015); Seddon et al. (2016), bearing into consideration the different periods and seasons of focus, and the differences in methodology and data.

For the remaining vegetated land, the availability of water is the first control over ecosystem dynamics, and is particularly important in semiarid regions such as

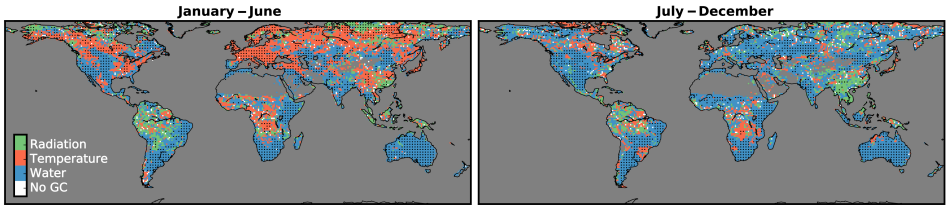


Figure 5.2: Factors controlling vegetation dynamics for two annual seasons. Temperature, radiation and water-limited regions during January-June (left) and July-December (right). The net of black dots is represented at 2° resolution and indicates areas with $R^2 > 0.3$. 'No GC' indicates no Granger causality ($R^2 \approx 0$). Grey colour indicates the regions considered as non-vegetated throughout the analysis.

eastern and central Australia, the Pampas and Caatinga region in South America, the US Great Plains, and the south and Horn of Africa. Interestingly, most of these ecoregions were recently shown to influence their own availability of water through transpiration feedbacks during dry and wet years (Miralles et al., 2016). As mentioned above, in tropical forests, none of the climatic drivers is causing a large fraction of vegetation variability. This may be explained by the subtle changes in vegetation and the ecosystems resistance to mean climate dynamics. However, this low response of tropical rainforests may also reflect aspects such as the dependency on phenological processes driven by biotic factors (Hutyra et al., 2007), occurrence of wild fires (van der Werf et al., 2008), limitations imposed by the availability of soil nutrients (Fisher et al., 2012b) and tropical deforestation (Hansen et al., 2013). In addition, it may also echo the influence of CO_2 fertilization, even though CO_2 emissions are expected to be more important for multi-decadal trends than for monthly dynamics (Liu et al., 2015b; Zhu et al., 2016). Therefore, Fig. 5.1 only partially supports the hypothesis of a radiation constraint on tropical vegetation, as defended by Nemani et al. (2003) or Seddon et al. (2016): (a) the Amazonian rainforest is affected by radiation, yet the South East Asia and Congo rainforests appear primarily driven by temperature; (b) other (non-climatic) drivers seem to dominate the dynamics in these ecosystems, as discussed above. Nevertheless, known issues of NDVI saturation in densely vegetated areas (Beck et al., 2011) may contribute to these results and should be considered.

The primary climatic controls over vegetation dynamics may shift throughout the year, both due to natural phenological cycles as well as intra-annual climate variability. In Fig. 5.2, our framework is applied to estimate the dominant factors causing vegetation variability during two distinct six-month seasons: January-June and July-December. As expected, results are markedly different in regions of ample phenological cycles, such as Northern Hemisphere mid and high latitudes. Deciduous and mixed forests in North America, Europe and China show a strong dependency on temperature during January-June, which is consistent with the expectations of the timing and length of their growing season being dependent on

temperature (see e.g., Chmielewski and Rötzer (2002); Menzel and Fabian (1999)). On the other hand, precipitation occurring during summer and autumn appears more relevant as a control of Northern Hemisphere deciduous vegetation during the senescence period (Xie et al., 2015) – see results for July-December in Fig. 5.2. Consequently, 50% of the vegetated surface appears primarily water-limited during January-June, while a larger 66% is primarily water-limited during July-December, with this seasonal dependency being mainly attributed to the phenological cycle of deciduous forests.

5.3.2. Lagged vegetation response to climate

Since vegetation, soil and atmosphere have a memory, and because some vegetation properties take time to respond to environmental changes, it is crucial to explore the latency in this response, which is ultimately related to the resistance and resilience of the ecosystems. Lag times are already considered in Figs. 5.1 and 5.2, given that our non-linear approach includes predictive features with various lags and based on several past cumulative periods (see Chapters 3 and 4). While the concept of introducing lag times in the study of these relationships is certainly not new (Davis, 1984; Braswell et al., 1997), it has become more extended in recent years (Chen et al., 2014; Wu et al., 2015; Seddon et al., 2016). The aforementioned studies suggest that the time taken by vegetation to respond to climatic and environmental anomalies, as well as its resilience, depend on both climate and ecosystem characteristics. Figure 5.3 shows that changes in water availability lead to lagged effects and longer-term impacts on vegetation than those in radiation and temperature. Semiarid ecosystems in Australia and the Americas show sensitivity to the dynamics in water availability occurring even longer than three months earlier, which partly reflects their lower resilience to drought stress (De Keersmaecker et al., 2015). In addition, Fig. 5.3 confirms that vegetation greenness typically takes several weeks to react to precipitation anomalies (Adegoke and Carleton, 2002; Seddon et al., 2016): the available water during the previous month (i.e., lag = 1) has more predictive power than during the current month (lag = 0).

On the other hand, the effect of temperature and radiation is more immediate (maximum at lag = 0), and dissipates rapidly, indicating a higher resilience of vegetation to anomalies in these variables. This is supported by the results in Fig. 5.3, which show that temperature and radiation data cannot help predict vegetation greenness in the following month, not even in energy-limited regions. These results also relate to the short memory of atmosphere compared to that of soil, implying that air temperature and radiation anomalies are less likely to prevail than those of water availability in following months (Seneviratne et al., 2006). These insights from Fig. 5.3 agree with the results by Seddon et al. (2016), but disagree with Wu et al. (2015). The latter showed a delayed response of vegetation greenness to radiation anomalies, based on multilinear regressions and a partial correlation model. However, as mentioned in Sect. 5.2.3, multilinear

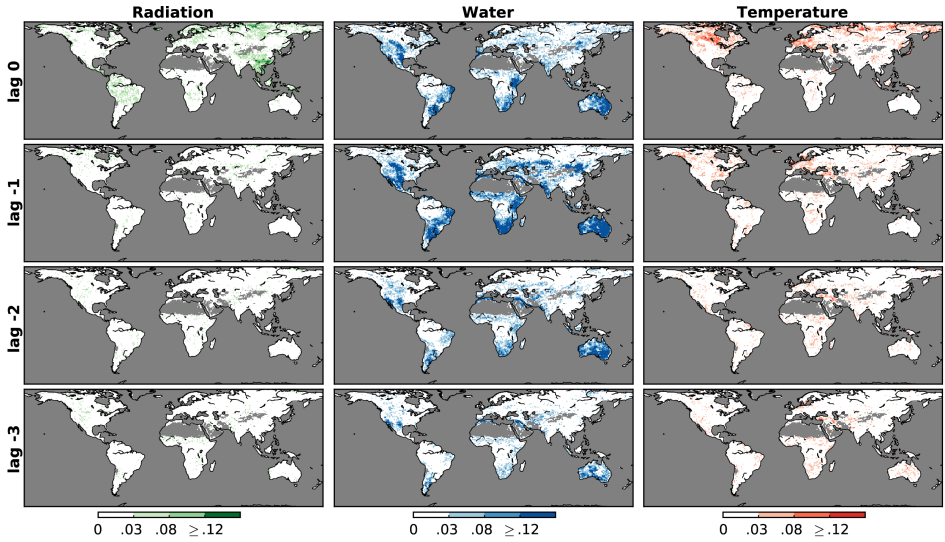


Figure 5.3: Temporal scale of the effects of hydro-climatic variables on vegetation. Influence (R^2) of each group of variables (radiation, temperature and water availability) on the NDVI anomalies considering different lag times (in months) separately.

regressions are prone to inflate the importance of temperature and radiation due to the (negative) correlations these variables hold against precipitation and soil moisture (Papagiannopoulou et al., 2017a). More complex frameworks, such as the one proposed here, allow us to disentangle and quantify the impacts of different climatic drivers independently and deterministically, which seems a necessary step to advance our understanding on climate–vegetation interactions.

5.3.3. Effect of hydro-climatic extremes in vegetation

Finally, we specifically target the net effect of hydro-climatic events – i.e., extremes in temperature, water availability and radiation – on global vegetation. Recent studies have highlighted the key role such events play for the structure and functioning of ecosystems, with their impacts depending on timing, magnitude, extent and type of event, and on the natural resistance and resilience of the ecosystem (Reichstein et al., 2013; Zscheischler et al., 2014; Sippel et al., 2016). Because our database of predictors includes climate extreme indices calculated based on the data sets in Table 3.1 (see Chapters 3 and 4), we have the means to isolate the importance of hydro-climatic extremes for global ecosystems following the sequential approach described in Sect. 5.2.3. Figure 5.4 depicts this importance in terms of R^2, i.e., the added explanatory power of these climate extremes – over the remaining predictor variables in the database – when it comes to predicting past NDVI anomalies. Hydrological extremes had an influence over the vegetation dynamics in most ecoregions on Earth during 1981-2010, being more important

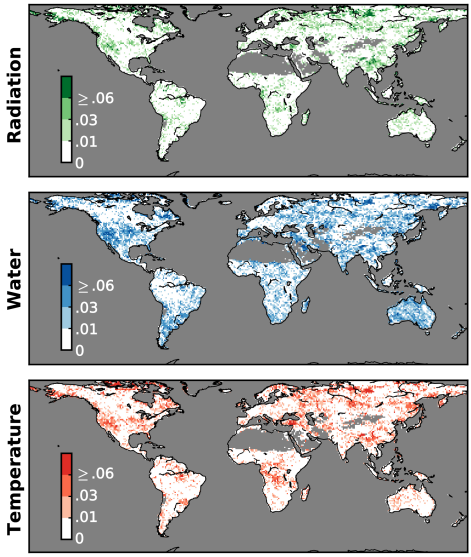


Figure 5.4: Effect of hydro-climatic extremes on vegetation. Influence (R^2) of radiation, water and temperature extremes on vegetation, calculated as their potential to predict the de-trended NDVI anomalies during the period 1981-2010.

in areas such as the US and Australia, where severe droughts occurred in recent decades. As expected, radiation and temperature extremes have an impact at higher latitudes and in the tropics; in particular, parts of boreal and tropical forests were affected by high temperature events. The apparent response of boreal forests to extremes in temperature is in line with the results by Zscheischler et al. (2014). Despite a particular type of extreme being able to explain up to 5%-10% of past vegetation variability for some regions, the importance of these events is low compared to that of the general climate dynamics. This is simply related to the fact that extremes are by definition infrequent, thus for the multi-decadal period considered here vegetation typically responds to regular environmental conditions.

5.3.4. Discussion

Despite the general agreement of our results with previous literature, an overall finding emerges from our analysis: water availability is not only the dominant control factor over vegetation in semiarid regions, but in most transitional ecoregions as well. On the contrary, Wu et al. (2015) reported that most of the vegetated land is primarily controlled by temperature, then radiation and finally water (the latter accounting only for 16% of the area where results were significant), while Nemani et al. (2003) reported 40%, 33% and 27% of the vegetated land being primarily constrained by water, temperature and radiation, respectively. Here, we estimate

a contrasting 61%, 23% and 15%, which is qualitatively more comparable to the results by Seddon et al. (2016). The latter reported water limitations in regions that were predominantly energy-limited according to Nemani et al. (2003), such as Western Europe and the American prairies. Figure 5.3 supports the hypothesis that (a) our consideration of the non-linear, lagged and cumulative impacts of water availability on vegetation, (b) the treatment of the co-linearities between temperature, radiation, and precipitation by our Granger-causality model (Sect. 5.2.2), and (c) our sequential approach to unravel the importance of these separate drivers (Sect. 5.2.3), are behind the stronger importance of water availability for vegetation dynamics revealed in our study. Nonetheless, it is important to note that differences in the accuracy of the radiation, water and temperature observations used here could affect the resulting contributions of these drivers, and may explain part of the differences with previous studies.

5.4. Conclusion

We have identified the main climatic and environmental controls on global vegetation during the satellite era following the non-linear Granger causality framework, which uses random forests as core model and is driven by a large database of global observational features (Chapter 4). Results indicate that water availability is the primary factor driving NDVI anomalies globally, with 61% of the vegetated continental surface being water limited, despite the relative importance of temperature in the Northern Hemisphere during the growing season. This overall water constraint appears more dominant than previously reported (Nemani et al., 2003; Wu et al., 2015; Gonsamo et al., 2016). In semiarid environments, water control over vegetation is reinforced by the long memory of soil moisture, which allows precipitation to affect vegetation dynamics more than three months into the future, in contrast to the more immediate and shorter-lasting impacts of radiation and temperature. We argue that this kind of non-linear interactions have not been adequately exposed by more traditional studies based on correlations and multilinear regression models.

Overall, our findings highlight a strong dependency of global vegetation on water availability, and show the imprint of hydro-climatic extremes on global vegetation during the satellite era. These results suggest that over a large part of the continents vegetation is prone to follow future trends in water availability. Critically, for most of the regions reported here as water-limited, the supply of precipitation is expected to decline following global warming (Fischer et al., 2014), and a general aggravation in hydro-climatic extremes is also expected (Seneviratne et al., 2012; Fischer et al., 2014). In the light of these projections, further studies to characterize the resistance and resilience of global vegetation to precipitation scarcity remain imperative to adequately predict the fate of these ecosystems.

6 Detecting regions with similar climate-vegetation dynamics via multi-task learning

In the previous chapters, we investigated the relationship between climate and vegetation by using a pixel-level approach. In this chapter, we explore the spatial relationship between the different locations (pixels) in order to detect locations with similar characteristics with respect to climate–vegetation dynamics. To this end, we model our spatio-temporal problem in a multi-task setting by considering the different locations as different tasks. As such, the dynamic interplay between vegetation and local climate is modelled to delineate ecoregions that share a coherent response to hydro-climate variability. Our novel framework is based on a multi-task learning approach which learns a low-dimensional representation of predictive structures (Sect. 6.2). This low-dimensional representation is combined with a clustering algorithm that yields a classification of biomes with coherent behaviour. Experimental results using our global observation-based data sets indicate that, without the need to prescribe any land cover information, our method is able to identify regions of coherent climate–vegetation interactions that agree well with the expectations derived from traditional global land cover maps (Sect. 6.3). The resulting global 'hydro-climatic biomes' can be used to analyse the anomalous behaviour of specific ecosystems in response to climate extremes and to benchmark climate–vegetation interactions in Earth system models.

This chapter is an edited version of:
Papagiannopoulou, C., Miralles, D. G., Demuzere, M., Verhoest, N. E. C., and Waegeman, W.: Global hydro-climatic biomes identified via multi-task learning, accepted in Geosci. Model Dev., https://doi.org/10.5194/gmd-2018-92, 2018.

6.1. Introduction

Approaches which aim to define regions with similar biophysical characteristics are commonly known as land cover classification schemes, and are widely used in multiple geoscientific disciplines. Land cover classifications are crucial to enable a better understanding of the spatial variability of the land surface, which can be a first and necessary step towards understanding complex spatio-temporal interactions among different environmental variables (Feddema et al., 2005). Traditional land use/land cover (change) classifications are typically based on spectral information from the land-surface coming from satellites (Loveland and Belward,

1997; Congalton et al., 2014). Amongst the most well-known and widely used are the International Geosphere-Biosphere Program DISCover Global 1km Land Cover classification (IGBP-DIS) (Loveland et al., 2000), Global Land Cover 2000 (Bartholomé and Belward, 2005) and more recently the land cover map developed within the European Space Agency's Climate Change Initiative (ESA CCI) (Poulter et al., 2015; Li et al., 2018). Similarly, climate classification schemes cluster regions with similar climate conditions and are also widely used to stratify geographical regions with different climatic expectations (Baker et al., 2009; Brugger and Rubel, 2013; Garcia et al., 2014; Herrando-Pérez et al., 2014). Here, the best known is probably the Köppen-Geiger climate classification (Köppen, 1936), which has been modified many times in recent decades (e.g. Thornthwaite, 1943; Trewartha and Horn, 1980; Feddema, 2005; Kottek et al., 2006; Peel et al., 2007). Yet to date, dynamics in these climate regimes are used as diagnostic of climate change by exploring their shifting boundaries (e.g. Diaz and Eischeid, 2007; Chen and Chen, 2013; Zhang and Yan, 2014a,b; Spinoni et al., 2015; Chan and Wu, 2015) or as a means to predict future climatic zone distributions using climate projections (e.g. Hanf et al., 2012; Gallardo et al., 2013; Mahlstein et al., 2013).

As advocated in the previous chapters, in recent years, the exponential advance in Earth observation research has made climate science one of the most data-rich scientific domains (Faghmous and Kumar, 2014). As such, data-driven methods have become popular in their use for land cover and climate classifications. For instance, Lund and Li (2009) proposed a new distance measure to define seasonal means and autocorrelations of climatic time series from weather stations, and grouped the stations using a hierarchical agglomerative clustering. Zscheischler et al. (2012) also stressed the importance of unsupervised methods for tasks, such as the classification of the land surface into zones with different climate and vegetation characteristics. Metzger et al. (2012) applied an alternative data-driven approach on climate and vegetation data that used principal component analysis (PCA) to discover informative structures in the data. In this method, the principal components of the initial climate–vegetation data set were applied as input to a clustering algorithm. Interesting results in the same direction can be attributed to Netzel and Stepinski (2016, 2017), who used distance measures of climatic variables, such as dynamic time warping, coming from time series analysis, in a data mining approach. In addition, temporal change in climate zones has been explored in the same context via clustering algorithms, such as k-means (Zhang and Yan, 2014a,b). Finally, data-driven methods have been also applied for the biome classification task, which has been commonly treated as an object recognition problem using remote sensing data. In this case, techniques coming from computer vision are frequently applied (Mekhalfi et al., 2015; Chen and Tian, 2015). Following the progress in computer science, neural networks and deep learning approaches are also becoming popular for this kind of tasks in recent years, making the whole procedure even more automated (Scott et al., 2017; Xu et al., 2017).

Previous studies rely on spectral information, supervised techniques or clustering approaches, which are applied to observations of climate variables and/or vegetation characteristics. However, these classification schemes are not based on the type of response of vegetation to climate dynamics. Recent advances in understanding vegetation response to climate variability highlight the importance of revealing the sensitivity of ecosystems to climate conditions, see Nemani et al. (2003); De Keersmaecker et al. (2015); Seddon et al. (2016); Papagiannopoulou et al. (2017b); Liu et al. (2018). Therefore, a step beyond these previous studies is a spatial characterization of the vegetation dynamics that are induced by climate variability, so that ecosystems of similar response to climate anomalies can be unveiled. This objective could be tackled by geostatistical approaches, such as geographically weighted regression (GWR) (Brunsdon et al., 1996), which assume that neighboring pixels have a similar behaviour with respect to specific variables; these methods have already been applied in studies with a regional focus (Propastin et al., 2008; Zhao et al., 2015; Georganos et al., 2017). However, here, we aim to avoid neighborhood assumptions and focus on the discovery of relationships between pixels based on the similarity in their modelled climatevegetation interaction, acknowledging that global ecosystems may experience similar interactions even if they are remotely located from each other. A previous effort towards detecting regions with similar vegetation response to climate involves the work of Ivits et al. (2014), where PCA is performed on the data matrix of drought anomalies and vegetation state, and a clustering is applied to the correlation coefficients based on the spatio-temporal patterns obtained by PCA. However, in this study, the interaction between climate and vegetation is not explicitly learned, nor the causes behind vegetation changes are inferred in a predictor–target framework.

Here, we introduce for the first time (to the best of our knowledge) a data-driven approach that aims to quantify the response of vegetation to local climate variables in a supervised setting at a global scale, and use this information to define ecoregions of consistent behaviour against hydro-climatic variability. In simple terms, our framework results in regions where vegetation responds similarly to the dynamics in temperature, soil moisture, incoming radiation, etc. The proposed framework relies on predictive modelling and clustering techniques and builds further upon recent work (described in Chapter 4) in which we investigated the global response of vegetation to local climate by applying machine learning algorithms in a Granger causality setting (Chapters 4 and 5). Since here we aim to exploit the relationships between different pixels – instead of modelling each pixel separately as in our previous studies – we propose the use of multi-task learning (MTL) methods (Caruana, 1997). These methods are commonly used for solving multiple related tasks: considering as one task the prediction of vegetation in one location and as multiple tasks the prediction of vegetation in multiple locations, we can model our problem by using an MTL approach. First, we apply an MTL approach which tries to unveil low-dimensional common predictive structures and

exploit the relationships among them. Second, we employ a clustering technique on these informative structures, which is applied on a lower-dimensional space (Sect. 6.2). This clustering technique is known as spectral clustering (Ng et al., 2002), and is one of the core assets of our framework. We refer to the emergent regions of coherent vegetation–climate behaviour as *hydro-climatic biomes* (Sect. 6.3).

6.2. Materials and methods

6.2.1. Data sets

The large database of global climate and vegetation data that will be used in the context of our framework is described in detail in Chapter 3 and is mostly based on satellite and *in situ* observations. Briefly, the database spans a 30-year period (1981-2010) at monthly temporal resolution and 1° latitude-longitude spatial resolution. The most important climatic and environmental drivers of vegetation are included in this database, namely: (i) land surface temperature, (ii) near-surface air temperature, (iii) longwave/shortwave surface radiative fluxes, (iv) precipitation, (v) snow water equivalent, and (vi) soil moisture. To characterise vegetation, we use the NDVI data set (Tucker et al., 2005). The target variable of our machine-learning framework is the de-trended seasonal NDVI anomalies, calculated as described in Chapter 3, while all other data sets are used as predictor variables. In addition, a series of 'high-level features' has been hand-crafted from the raw time series of predictors, and used as well as predictor variables. As such, our set of predictive features includes not just the raw data time series of each climate/environmental variable, but also: seasonal anomalies, de-trended seasonal anomalies, lagged variables, past cumulative variables, and extreme indices – see Chapter 3. The use of these non-linear features greatly improved causal inference and help characterise non-linear relationships between climate and vegetation dynamics in our recent work (see Chapter 4). For a further discussion about the importance of the higher-level representation adopted in our framework, we refer the reader to Sect. 6.3.1.

6.2.2. Pixel-based approach: single-task learning

In our study, we use information on climate and vegetation variables at specific time points and locations. Formally, we consider a spatio-temporal data set $D = \{(\mathbf{X}^{(1)}, \boldsymbol{y}^{(1)}), (\mathbf{X}^{(2)}, \boldsymbol{y}^{(2)}), ..., (\mathbf{X}^{(L)}, \boldsymbol{y}^{(L)})\}$, with L being the number of different locations and $(\mathbf{X}^{(l)}, \boldsymbol{y}^{(l)})$ the tuple of the predictor variables and the target variable of each location l. We denote $D^{(l)} = \{(\boldsymbol{x}_i^{(l)}, y_i^{(l)})\}_{i=1,...,N}$ the observations of a location l, while the input feature vectors are denoted as a matrix $\mathbf{X}^{(l)} = [\boldsymbol{x}_1^{(l)}, ..., \boldsymbol{x}_N^{(l)}]^T$ and the corresponding target values as $\boldsymbol{y}^{(l)} = [y_1^{(l)}, ..., y_N^{(l)}]^T$.

Specifically, $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times d}$ is the matrix of the predictor variables with d being the number of predictors, and $\boldsymbol{y}^{(l)} \in \mathbb{R}^{N}$ the response time series (i.e., NDVI seasonal de-trended anomalies), where N denotes the number of discrete timestamps, i.e., the length of the time series. In this setting, a straightforward approach is to tackle each regression problem in each location l separately, i.e., by independently training one model for each location (see Chapter 4 and Papagiannopoulou et al. (2017a)). That way, for every pixel, only the data of that particular location l is used $((\mathbf{X}^{(l)}, \boldsymbol{y}^{(l)}), l = 1, ..., L)$, not attempting to utilize the data from other regions where the target variable might have a similar response to the predictors.

We can start by defining regions of similar climate–vegetation dynamics with the most naive approach: the relationship between climate and vegetation can be caught by the weights of a regression model, i.e., the regression coefficients of the predictor variables. Specifically, if one defines a multiple linear regression model for a location l, the model for the l^{th} location is given by $f^{(l)}(\boldsymbol{x}_i^{(l)}) = \boldsymbol{w}^{(l)}\boldsymbol{x}_i^{(l)}$, with $\boldsymbol{x}_i^{(l)}$ being the input data (i.e., one observation) and $\boldsymbol{w}^{(l)}$ being the weight vector learned for particular location l, which describes the importance of each input variable for the target – see Fig. 6.1a. Even though one can assume that these weight vectors can be similar for regions in which the response of vegetation to climate is similar, the information from these other regions is not used in the prediction (i.e. each regression is applied at each individual pixel separately). This is despite the fact that these locations could be grouped (e.g., based on a similarity measure of the weight vectors) into wider regions that one may assume that share common climate–vegetation dynamics. Note also that the information captured by each weight vector $\boldsymbol{w}^{(l)}$ should be sufficient, which means that it is necessary for the models to have a good generalization performance.

6.2.3. Exploiting spatial relationships: multi-task learning

Unlike the single-task learning models that only take the data of each particular location into account, MTL models extract information of data sets with similar characteristics from other locations. As such, they can be expected to generalize better and give a higher predictive performance on unseen data. Specifically, by using the MTL approach, the generalization of the model improves if the data set of each task is expanded by observations from highly related tasks. This is crucial, especially in cases where the number of training instances per task is limited. The basic idea that underlines the MTL modelling approach is the learning of a separate model for each task and not a unique model trained on a concatenated set of observations of all tasks. Note that in our spatio-temporal data sets, each location can be seen as a different task, and neighbouring (or distant) locations with similar climate-vegetation interactions will tend to have similar (yet not identical) behaviour. In light of this observation, MTL seems to be a quite natural modelling approach to explore the interaction between climate and vegetation in different locations.

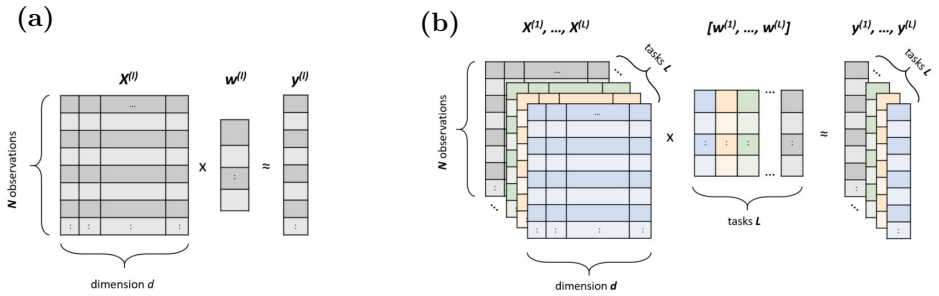


Figure 6.1: Graphical representation of the two learning approaches. (a) A single-task learning approach in which each pixel is treated separately. For each pixel l there is an input data set $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times d}$, with N being the number of observations and d being the number of predictors, and a target vector $\boldsymbol{y}^{(l)} \in \mathbb{R}^{N}$. The vector $\boldsymbol{w}^{(l)} \in \mathbb{R}^{d}$ represents the weight vector learned by the model. (b) A multi-task learning approach in which the models of L tasks are simultaneously learned. The input of the method is the data sets $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(L)}$ of all locations (i.e., all global land pixels). The corresponding target vectors are denoted with $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(L)}$. The weight matrix $[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(L)}] \in \mathbb{R}^{d \times L}$ contains the weight vectors for all tasks.

The idea of MTL is not new (Baxter, 1997; Caruana, 1997; Baxter et al., 2000), and it has been applied in many machine learning applications in medical sciences (Bi et al., 2008; Zhang et al., 2012) and computer vision (Zhang et al., 2014). It has also been used in climate science to improve the way multiple Earth System Models (ESMs) outputs are combined, by treating the locations as different tasks. In these studies, the idea is that in neighbouring locations (pixels which are close to each other), similar ESMs tend to have similar performance (Subbian and Banerjee, 2013; McQuade and Monteleoni, 2013). A recent study proposed a hierarchy of tasks, in which at a first level, tasks of each location are trained into an MTL setting, while at a second level, tasks of each variable are sharing information (Gonçalves et al., 2017). In addition, for modelling spatio-temporal data, Xu et al. (2016) introduced an MTL framework in which local models share a common representation based on the spatial autocorrelation. Although this kind of modelling is becoming more common in climate science, it has not been combined (to the best of our knowledge) with clustering approaches in the context of mapping land cover nor climate–vegetation dynamics.

In this chapter, we focus on MTL methods that can discover the relationship between different tasks (locations) and recover strong predictive structures of the vegetation response to climate. These are then used to conform hydro-climatic biomes, i.e., regions of coherent vegetation behaviour with respect to climate variability (see Sect. 6.3.4). To this end, we use the same notation as before by denoting $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times d}$ as input data matrix of the predictor variables, $\boldsymbol{y}^{(l)} \in \mathbb{R}^{N}$ as the target vector for each location l and $\boldsymbol{w}^{(l)} \in \mathbb{R}^{d}$ in which each value corresponds

to a weight. We define as $[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(L)}] \in \mathbb{R}^{d \times L}$ the weight matrix of all locations such that the $\boldsymbol{w}^{(l)}$ vector is the l^{th} column of the $[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(L)}]$ matrix - see a graphical representation of the notation in Fig. 6.1b. Given a loss function $\mathcal{L}$ (e.g., the squared error loss), the multi-task minimization problem is formulated as:

$$\min_{\boldsymbol{w}^{(1)},..,\boldsymbol{w}^{(L)}} \sum_{l=1}^{L} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{w}^{(l)} \boldsymbol{x}_i^{(l)}, y_i^{(l)}) + \Omega(\boldsymbol{w}^{(1)}, ..., \boldsymbol{w}^{(L)}) \tag{6.1}$$

where $\Omega(\boldsymbol{w}^{(1)}, ..., \boldsymbol{w}^{(L)})$ is a factor which controls the relatedness among the tasks. In our setting, we assume that there is no prior knowledge about the relationship of the tasks (locations) and we aim to apply a method that can discover these relationships.

In literature, there are many MTL methods that are trying to do two things simultaneously: learn a weight matrix $[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(L)}]$ and another matrix which captures the task relationships simultaneously (Ando and Zhang, 2005; Chen et al., 2009; Zhou et al., 2011). In real applications, there are scenarios where the tasks of an MTL problem follow a specific structure, i.e., some tasks are more related whereas some others are unrelated. In order to identify this group structure, researchers have developed various methods which have been referred to as clustered multi-task learning (CMTL) methods (Zhou et al., 2011). For instance, Xue et al. (2007) proposed a method which uses a Dirichlet process-based statistical model to identify similarities between related tasks, while Jacob et al. (2009) introduced a framework which identifies groups of tasks and performs the learning at once. In the same direction, Wang et al. (2009) used an inter-task regularization term to take into consideration tasks which have been grouped in the same cluster in a semi-supervised setting. More recently, Barzilai and Crammer (2015) suggested a method that assigns explicitly each task to a specific cluster, building a single model for each task by using linear classifiers that are combinations of some basis. An alternative approach has been proposed by Zhou et al. (2011), in which the structure of the task relatedness is unknown and is learned during the training phase. Interestingly, when case-specific conditions are fulfilled, this method is equivalent to the method by Ando and Zhang (2005), known as the Alternative Structure Optimization (ASO), which belongs to the category of MTL methods that assume the existence of a shared low-dimensional representation among the tasks. The name of the method indicates that an alternating optimization procedure is involved during the learning process, since the weight matrix and the matrix that captures the shared low-dimensional representation are learned simultaneously. Typically, in these procedures, the optimization of each part is separately performed, while the other part remains fixed. In our work, we apply the ASO method due to its simplicity and the fact that it does not need a lot of iterations to capture the information about the task relatedness that is needed. This is crucial for our

application, since the large size of the global database we use, puts severe limitations to the choice of method. Another aspect is that by learning this low-dimensional representation we can have a visual inspection of the 'most predictive common structures' for each region. In the following section we explain in detail the ASO method used in our setting.

6.2.4. Learning predictive structures from multiple tasks

The ASO algorithm proposed by Ando and Zhang (2005) learns common predictive structures from multiple related tasks that are assumed to share a low-dimensional feature space. Specifically, by applying this method, one learns one model function for each individual task and the learned weight vector is decomposed into two parts: (a) a high-dimensional space, and (b) a shared low-dimensional space based on a feature map learned during the process. This feature map is a matrix which serves as a link between a high-dimensional space and a low-dimensional space. In our case, L predictor functions $\{f^{(l)}\}_{l=1}^{L}$ are simultaneously learned by exploiting the shared feature space that underlines all tasks. This low-dimensional feature space is expressed in a simple linear form of a low-dimensional feature map $\boldsymbol{\Theta}$ across the L tasks. Mathematically, the function $f^{(l)}$ can be written as:

$$f^{(l)}(\boldsymbol{x}_i) = \boldsymbol{w}^{(l)}\boldsymbol{x}_i^{(l)} = \boldsymbol{u}^{(l)}\boldsymbol{x}_i^{(l)} + \boldsymbol{v}^{(l)}\boldsymbol{\Theta}\boldsymbol{x}_i^{(l)} \tag{6.2}$$

with $\boldsymbol{\Theta} \in \mathbb{R}^{h\times d}$ being a parameter matrix with orthonormal row vectors, i.e., $\boldsymbol{\Theta}\boldsymbol{\Theta}^T = \mathbf{I}$, where h is the dimensionality of the shared feature space, and $\boldsymbol{w}^{(l)}, \boldsymbol{u}^{(l)}$ and $\boldsymbol{v}^{(l)}$ are the weight vectors for the full feature space, the high-dimensional one (initial dimension d), and the shared low-dimensional one (based on the h parameter), respectively. As mentioned before, the ASO method is equivalent to the CMTL method (Zhou et al., 2011), under a specific condition: the parameter k, which symbolizes the number of clusters in the CMTL approach, is equal to the parameter h of the ASO method. This condition determines the number of clusters that should be used in the clustering phase of our framework, because the objective of ASO is optimized based on the value of the parameter h. We reconsider this equivalence in Sect. 6.3.3, where we discuss about the number of clusters that should be identified based on our analysis.

Formally, ASO can be formulated as the following optimization problem:

$$\min_{\{\boldsymbol{w}^{(l)},\boldsymbol{v}^{(l)}\},\boldsymbol{\Theta}\boldsymbol{\Theta}^T=\mathbf{I}} \sum_{l=1}^{L}\left(\sum_{i=1}^{N}\mathcal{L}(\boldsymbol{w}^{(l)}\boldsymbol{x}_i^{(l)}, y_i^{(l)}) + \lambda^{(l)}\left\|\boldsymbol{u}^{(l)}\right\|_2^2\right) \tag{6.3}$$

with $\left\|\boldsymbol{u}^{(l)}\right\|_2^2$ being the regularization term ($\boldsymbol{u}^{(l)} = \boldsymbol{w}^{(l)} - \boldsymbol{\Theta}^T\boldsymbol{v}^{(l)}$) that controls the task relatedness among L tasks, $(\boldsymbol{x}_i^{(l)}, y_i^{(l)})$ being the input vector and the corresponding target value of the i^{th} observation in a particular location l, and

$\lambda^{(l)}$ being a predefined parameter – see Fig. 6.2 for the graphical representation of the notation. During the learning process the weight matrix $[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(L)}]$ and the matrix $\boldsymbol{\Theta}$, which captures the shared low-dimensional representation, are learned simultaneously. The regularization term $\left\|\boldsymbol{u}^{(l)}\right\|_2^2$, based on the value of the parameter λ, penalizes the differences between the weights on the initial high-dimensional space and the weights on the low-dimensional space parameterized by $\boldsymbol{\Theta}$.

There are several ways of solving the optimization problem in Eq. (6.3) (Ando and Zhang, 2005). Our main purpose is to extract the shared feature space $\boldsymbol{\Theta}$ in order to apply a clustering on the low-dimensional feature space. In this feature space, locations with similar predictive structures will be grouped into the same broader region. For this reason, we adopt the Singular Value Decomposition (SVD)-based ASO algorithm, proposed by Ando and Zhang (2005), which achieves good performance even in the first iteration of the method. As mentioned before, this is crucial to our application, given the large number of tasks and the high-dimensional data sets. The steps of the SVD-based ASO are presented in Algorithm 1.

Algorithm 1 SVD-ASO

Input: training data $D^{(l)} = \{(\boldsymbol{x}_i^{(l)}, y_i^{(l)})\}_{i=1,...,N}$, where $l = 1, ..., L$
Parameters: h and $\boldsymbol{\lambda} = \{\lambda^{(1)}, ..., \lambda^{(L)}\}$
Output: $\boldsymbol{\Theta} \in \mathbb{R}^{h \times d}$ and $\mathbf{V} = \left[\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(L)}\right]^T \in \mathbb{R}^{L \times h}$
Initialize: $\boldsymbol{w}^{(l)} = 0, l = 1, ..., L$, and $\boldsymbol{\Theta}$ to random
repeat
  for $l = 1$ **to** L **do**
    with fixed $\boldsymbol{\Theta}$ and $\boldsymbol{v}^{(l)} = \boldsymbol{\Theta}\boldsymbol{w}^{(l)}$, solve the optimization problem of Eq. (6.3) for $\boldsymbol{u}^{(l)}$:
    $\text{argmin}_{\boldsymbol{u}^{(l)}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{u}^{(l)}\boldsymbol{x}_i^{(l)} + (\boldsymbol{v}^{(l)}\boldsymbol{\Theta})\boldsymbol{x}_i^{(l)}, y_i^{(l)}) + \lambda^{(l)} \left\|\boldsymbol{u}^{(l)}\right\|_2^2$
    $\boldsymbol{w}^{(l)} = \boldsymbol{u}^{(l)} + \boldsymbol{\Theta}^T\boldsymbol{v}^{(l)}$
  end for
  Apply an SVD decomposition on $\mathbf{W} = [\sqrt{\lambda^{(1)}}\boldsymbol{w}^{(1)}, ..., \sqrt{\lambda^{(L)}}\boldsymbol{w}^{(L)}]$:
  $\mathbf{W} = \mathbf{V_1}\mathbf{D}\mathbf{V_2}^T$ (with diagonals of $\mathbf{D}$ in descending order)
  $\boldsymbol{\Theta} = \mathbf{V_1}^T[:h, :]$ // update $\boldsymbol{\Theta}$ to the first h rows of $\mathbf{V_1}^T$
until convergence

The SVD-based ASO method can be interpreted as a dimensionality reduction technique applied to the model space (i.e., weights). It should be stressed here that this method must not be confused with PCA, which is usually employed on the data space (input space of predictors) (Metzger et al., 2012; Ivits et al., 2014). The goal of the ASO method is to detect the principal components of the parameter matrix, while PCA identifies the principal components of the input data $\mathbf{X}$. This goal can be achieved by considering the models of multiple tasks as samples of their own distribution. Therefore, these samples can only be formed by using an MTL approach, in which there is access to the models from multiple

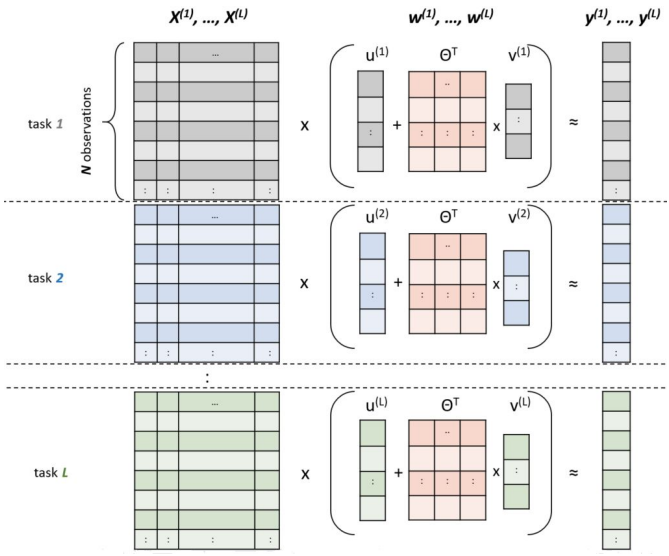


Figure 6.2: Graphical representation of the ASO method. The input of the method is the data sets $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(L)}$ of all locations. The corresponding target vectors are denoted with $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(L)}$. The weight vector $\boldsymbol{w}^{(l)} \in \mathbb{R}^d$ of the full space is decomposed in two parts; to the weight vector $\boldsymbol{u}^{(l)} \in \mathbb{R}^d$ of the high-dimensional space and the weight vector $\boldsymbol{v}^{(l)} \in \mathbb{R}^h$ of the low-dimensional one. The low-dimensional feature map $\boldsymbol{\Theta}^T \in \mathbb{R}^{d \times h}$ is common for all the tasks.

learning tasks. Moreover, in our work, we explicitly consider the climatic variables as predictors and the vegetation variable as target variable, and we learn the relationship between them in a supervised setting. As such, the regions that we define rely on the relationship between climate and vegetation in a prediction setting, and the clustering is calculated based on similarity of this relationship (i.e. the model coefficients for different locations), see Sect. 6.2.5 for more details. As such, we learn relationships between climate and vegetation in a supervised setting, whereas PCA-based methods (Metzger et al., 2012; Ivits et al., 2014) are fully unsupervised. In our study the SVD decomposition is used as part of the optimization algorithm, thus in a supervised setting. In this setting, the model weights are optimized based on a given training set. Therefore, the discovered structures are obtained during the training process.

To clarify the notation used in the ASO method, we intuitively explain the symbolization of the method in relation to our specific setting: the problem of detecting locations with similar climate–vegetation dynamics. As mentioned above (Sect. 6.2.2 and 6.2.3), the input features that constitute the $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times d}$ matrix consist of the climatic predictor variables, i.e., the extreme indices, lagged variables, etc., calculated based on raw climatic time series of a certain location l. The dimensions N and d correspond to the number of observations, i.e., the length of the time series and the number of predictor variables, respectively. The target variable for a

particular location l, which is the NDVI anomalies, is symbolized with $\boldsymbol{y}^{(l)} \in \mathbb{R}^N$. As such, an observation of a certain location l at a particular timestamp i is denoted as a pair $(\boldsymbol{x}_i^{(l)}, y_i^{(l)})$. The goal of the ASO method is to learn the weight matrix $[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, ..., \boldsymbol{w}^{(L)}]$, i.e., a single weight vector $\boldsymbol{w}^{(l)}$ for each location l. This weight vector $\boldsymbol{w}^{(l)}$ is able to capture the relationship between the predictor variables and the target, i.e., the climatic variables and the NDVI anomalies. Therefore, climatic predictors that are more important for vegetation anomalies correspond to higher absolute values in the weight vector $\boldsymbol{w}^{(l)}$. As a result, locations with similar weights are considered as regions where vegetation responds to climate in a similar way. As described in a previous paragraph of this section, the ASO method assumes that the weight vectors $\boldsymbol{w}^{(l)}$ consist of two parts the $\boldsymbol{u}^{(l)}$ and the $\boldsymbol{v}^{(l)}\boldsymbol{\Theta}$. These two parts are learned simultaneously in Algorithm 1 in an alternating fashion. The first part, i.e., the $\boldsymbol{u}^{(l)} \in \mathbb{R}^d$ belongs to the high-dimensional space, the initial one, which is equal to d. This part expresses the location-specific part of the weight vector, i.e., the deviation of each location's weight vector from the weights learned in a lower dimensional space. The second part consists of the matrix $\boldsymbol{\Theta} \in \mathbb{R}^{h \times d}$ that represents the map from the initial dimension d to the lower dimension h and the weight vector $\boldsymbol{v}^{(l)} \in \mathbb{R}^h$. The map matrix $\boldsymbol{\Theta}$ is common for all the locations (tasks) and can be learned across them due to the MTL approach. The weight vector $\boldsymbol{v}^{(l)}$ represents the projection of the initial weights to a low-dimensional space h. Intuitively, this second part of the weight decomposition expresses the coarsest and most important part of weights, since it detects the most important structures through the map matrix $\boldsymbol{\Theta}$. The matrix $\mathbf{V} = [\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(L)}]^T \in \mathbb{R}^{L \times h}$ denotes the representation of the models in the low-dimensional space h for the L locations.

Note that the ASO method can be extended into a non-linear (kernelized) version, see more details in the original paper (Ando and Zhang, 2005). Although this non-linear extension of the method fits well our setting, we preferred to stick to a simple linear approach since our preliminary results with the kernelized version showed a marginal or no improvement (depending on the region) in predictive performance. This result can be explained by the fact that kernel parameters should be carefully fine-tuned. Tuning is a computationally intensive process and cannot easily be applied in our large data set.

6.2.5. Land classification: clustering highly-predictive structures

Clustering in machine learning is the task of grouping a set of samples in such a way that those samples that belong to the same group (cluster) are more similar with respect to a specific criterion than to samples that belong to other groups. Clustering techniques are usually based on a distance (or similarity) measure that is calculated among the samples and/or group of samples. There are several clustering

approaches and an in-depth review can be found in Xu and Tian (2015).

It is known that in high-dimensional spaces, distance measures are not able to capture well differences between pairs of samples, thus clustering algorithms tend to perform better in lower dimensional spaces. In our setting, we learn the common feature map $\mathbf{\Theta} \in \mathbb{R}^{h \times d}$ and the $\mathbf{V} = [\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(L)}]^T \in \mathbb{R}^{L \times h}$ matrix, which is the representation of the models in this low-dimensional space, using the SVD-ASO method – see Sect. 6.2.4. The $\mathbf{V}$ matrix captures the information of the similar predictive structures among all the tasks, so similar tasks are closer in this low dimensional space and as a consequence, they have a similar representation (i.e., similar weights) in this matrix. That way, the clustering techniques based on distance calculations are applied on the more expressive low-dimensional space, resulting in a better performance. As it has been discussed in Chapter 4, global climate–vegetation relationships are complex and non-linear. Here, if the $\mathbf{V}$ representation is expressive enough, the clustering method can group together locations with similar models, i.e., locations in which vegetation responds to climate in a similar way. Thus, it is first necessary to evaluate the quality of the learned matrix $\mathbf{V}$. The most straightforward way to do so is by measuring the predictive performance of the MTL model in terms of, e.g., R^2. If the predictive power of the model is strong, we can conclude that the $\mathbf{V}$ matrix is able to well-capture the relationships of each task with the highly predictive structures. So, given that the $\mathbf{V}$ representation is sufficiently learned from the data, we can apply any kind of clustering algorithm on the low-dimensional representation of matrix $\mathbf{V}$. This approach is also known as spectral clustering, due to the fact that the clustering algorithm is applied on a reduced feature space, making the clustering results more robust.

In our application, we use a hierarchical agglomerative clustering approach (Ward, 1963) where the number of clusters is not predefined. In the hierarchical clustering approach, the result is usually depicted as a dendrogram in which the leaves represent the observations and the inner nodes correspond to the data clusters. The dendrogram branches are proportionally long to the value of the intergroup dissimilarity. By defining this hierarchical form of the clustering result, one can define the number of clusters by cutting down vertically (or horizontally, depending on the view) the dendrogram in a point where the dissimilarity between the clusters is high and therefore the branches are longer – see Sect. 6.3.3 for the choice of the optimum number of clusters in our analysis.

6.2.6. Experimental setup

In all the experiments, we use as predictors all the climatic data sets and the features that we have constructed from them, as well as the 12-lagged values of the target variable. A total number of 3,209 predictor variables is included, i.e., $d = 3,209$ in our setting. These variables constitute the input to our framework,

i.e., the $\mathbf{X}^{(l)}, l = 1, ..., L$ data sets. As target variable, we use the NDVI seasonal anomalies, denoted as $\boldsymbol{y}^{(l)}, l = 1, ..., L$ for each location. We examine 13,072 land pixels where each pixel constitutes a single task in our MTL setting, i.e., $L =$ 13,072. The data set of each single task consists of 360 monthly observations, i.e., $N = 360$.

For the STL modelling, evaluated for comparison, we use ridge regression for each location independently, as discussed in Chapter 2 and 4. The regularization parameter λ is tuned using a separate validation set. Note that by splitting the original data set in three parts (1) training set, (2) validation set, and (3) test set, we tune the parameters in a set of observations (validation set) that are not included in the final test set and achieve a fair evaluation of the model performance. The optimization problems of the SVD-ASO algorithm are solved by using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm.

6.3. Results and discussion

6.3.1. Importance of a higher-level representation of features

To illustrate the non-linear response of vegetation and explain our choice to use a high-level feature representation in our framework, we compare the model performance with and without the use of this high-level representation. Figure 6.3a shows the predictive performance of the ASO-MTL method when the raw variables as well as the corresponding six-lagged values are included in the model, i.e., the cumulative variables and the extreme indices are not included as predictors. Figure 6.3b visualizes the difference in predictive performance of the ASO-MTL model with and without the cumulative variables and the extreme indices as predictors. As one can observe, in regions such as Europe, North America, southern and northern parts of Asia and parts of South America, the model performance substantially decreases if these higher-level features are not used in the data representation. In these regions, more than 10% of the variability in NDVI anomalies is explained by this more complex (non-linear) representation, illustrating the non-linear nature of the relationship between climate and vegetation dynamics.

6.3.2. Single- versus multi-task learning model

In a second experiment, we compare the predictive performance of the STL model versus the MTL model. For the STL modelling, ridge regression is used. For the MTL modelling, we apply the ASO-MTL model (Ando and Zhang, 2005) described in Sect. 6.2. We use a separate validation set to tune the regularization parameter

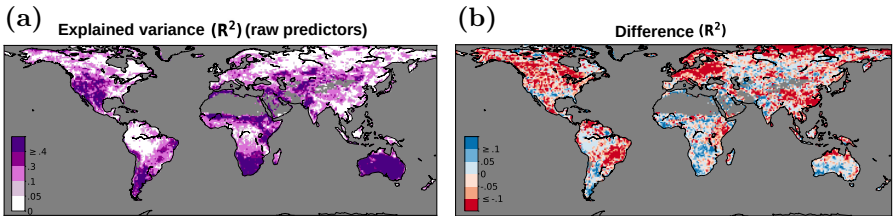


Figure 6.3: Comparison of the predictive performance in terms of R^2 of the model which does not include the cumulative variables and the extreme indices with the model which is trained with the full collection of higher-level features – Chapter 3. (a) Explained variance (R^2) of NDVI anomalies based on the raw data of the climatic variables as well as their six-lagged values (cumulative variables and the extreme indices are not included as predictors to the model). (b) Difference in terms of R^2 between the model without cumulative and extreme predictors and the model which includes all the higher-level feature representation (see Fig. 6.4 in the next section).

λ for both approaches. For the STL approach, we tune the λ parameter for each location (task) separately, while for the MTL approach we use the same λ value for all the tasks, taking into account the average performance across these tasks. For the ASO-MTL method, we have also experimented with the value of the h parameter, which is the dimensionality of the shared feature space – see Sect. 6.3.3 for more details about the influence of this parameter on the clustering results. Finally, we evaluate the performance of both approaches in terms of R^2, as in Chapter 4. Figure 6.4 depicts the result of our comparison. Figure 6.4a shows the R^2 of the ASO-MTL model while Fig. 6.4b highlights the difference in predictive performance of the MTL model in comparison with the STL model. As shown in Fig. 6.4b, in almost all regions of the world, the predictive performance increases substantially compared to the STL approach. In fact, over extensive regions (40% of the study area), more than 5% of the variability in NDVI is explained by the spatial structure of the data. In statistical terms, this implies the existence of a hidden structure between the different locations (tasks), which is informative with respect to our target variable. The dotted regions in Fig. 6.4b correspond to areas where the MTL model significantly outperforms the STL models based on the Diebold-Mariano statistical test, which compares model predictions (Diebold, 2015b). For the statistical test, we use the False Discovery Rate (Benjamini and Hochberg, 1995) method to correct the p-values at level 0.05 due to the multiple-hypothesis testing setting.

Additionally, Fig. 6.4a shows that more than 40% of the mean monthly vegetation dynamics can be explained by climate variability in some regions. In particular, in regions such as Australia, Africa and Central and North America the predictive power of the model is stronger in terms of R^2, following the same pattern and scoring similar R^2 values as the random forest approach by Papagiannopoulou et al.

(2017a) (Chapter 4). To deepen on the performance difference between the two approaches, the R^2 scores are presented as two different distributions in Fig. 6.4c. The blue histogram corresponds to the distribution of the R^2 scores of the STL approach, while the orange one corresponds to the distribution of the R^2 scores of the MTL approach. As can be observed, the distribution of the R^2 scores is shifted to the right for the MTL, meaning that values are typically greater than those derived from the STL approach. Moreover, the skew towards the left in the blue histogram, with values close to zero, is an indication of the near-zero performance of the STL models in many locations. The Wilcoxon paired statistical test (Demšar, 2006) confirms that the results of the two approaches are overall statistically different (p-value $< 10^{-9}$).

Since we are ultimately interested in investigating regions of coherent impact of climate variability on vegetation dynamics, we also evaluate the ability of the MTL model to detect Granger-causal effects of climate on vegetation. For a detailed description of the Granger causality modelling framework, see Chapter 4. This point is crucial to understand the extent to which the climatic predictors carry additional information about the dynamics in vegetation that is not contained in the past vegetation signal itself. The results of applying the Granger causality analysis using MTL modelling are shown in Figure 6.4d, which illustrates results of the full MTL model compared to the baseline MTL model. This baseline model only uses previous values of NDVI to predict monthly NDVI anomalies (Chapter 4). In this figure, it becomes clear that climate dynamics Granger-cause monthly vegetation anomalies in most regions of the world, and the ability of the MTL model to detect deterministic relationships is evidenced. This is also confirmed by the Wilcoxon paired statistical test (p-value $< 10^{-9}$). On the other hand, the ability of the STL model to detect Granger-causal relationships is rather limited compared to that of the MTL model. Figure 6.4e depicts the result of the comparison, where in almost all regions the quantification of Granger causality of the MTL approach increases substantially compared to the one of the STL approach. Analogous to Fig. 6.4c, Fig. 6.4f compares the distributions of Granger causality (i.e., the difference in predictive performance in terms of R^2 between the full and the baseline model) between the STL and MTL approach. Once again, the blue histogram corresponds to the distribution of Granger causality retrieved using the STL approach, while the orange corresponds to the results of the MTL approach. The shift to the right of the orange histogram shows the larger ability of the MTL model to reveal Granger-causality between climate and vegetation. Similar to the previous comparison, the Wilcoxon paired statistical test (Demšar, 2006) confirms that the results of the two approaches are overall statistically different (p-value $< 10^{-9}$). In summary, these findings highlight the potential of using the low-dimensional feature representation learned from the data to fulfill our final objective, which is the detection of vegetated areas holding a similar response to climate via a clustering approach.

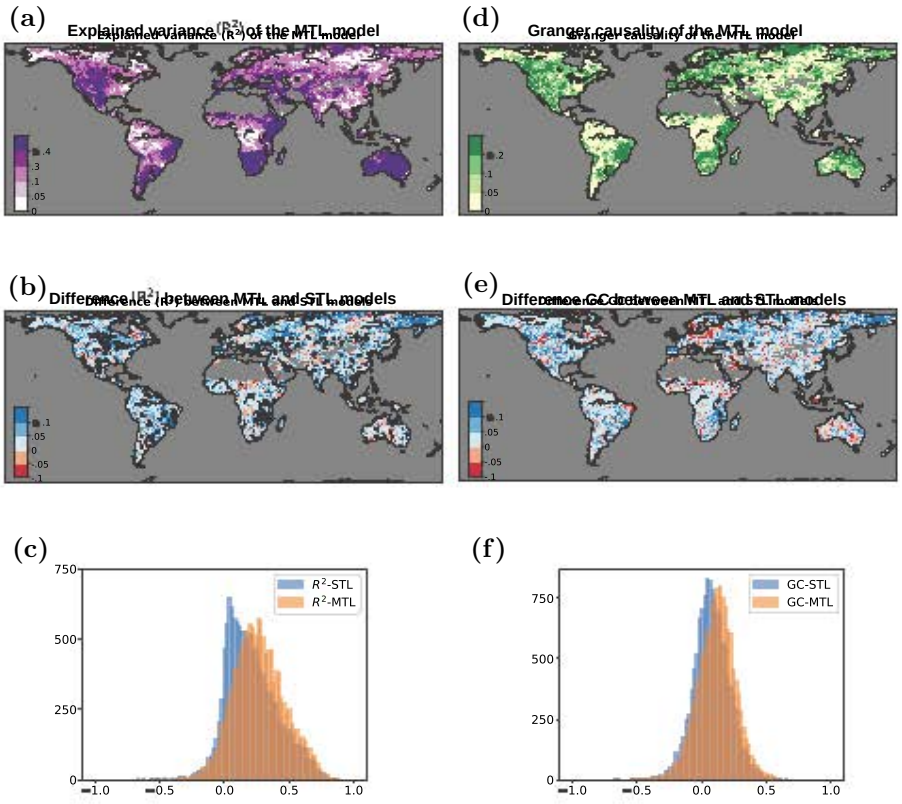


Figure 6.4: Comparison of the predictive performance between the STL and the MTL approaches. (a) Explained variance (R^2) of the NDVI monthly anomalies based on the MTL approach. (b) Difference in terms of R^2 between the MTL and the STL approaches; blue regions indicate a higher performance by the MTL. The dotted regions correspond to areas where the MTL model significantly outperforms the STL models based on the Diebold-Mariano statistical test (Diebold, 2015b). (c) Comparison of the distributions of the R^2 scores in the STL and in the MTL setting; the blue histogram corresponds to the STL, and the orange one to the MTL approach. (d) Quantification of Granger causality for the MTL approach, i.e., improvement in terms of R^2 by the full MTL model with respect to the R^2 of the baseline MTL model that uses only past values of NDVI anomalies as predictors; positive values indicate Granger causality (Chapter 4). (e) Difference in terms of Granger causality between the MTL and the STL approaches; blue regions indicate a higher performance by the MTL. (f) Comparison of the distributions of the Granger causality in the STL and in the MTL setting; the blue histogram corresponds to the STL, and the orange one to the MTL approach.

6.3.3. Appropriate number of hydro-climatic biomes

As described in Sect. 6.2.5, there are multiple approaches that can be used to define the number of classes in a clustering problem. In our framework, we define the number of clusters by using a data-driven approach. In our analysis, we choose not to use information from any predefined number of vegetation and/or climate classes existing in the literature, since the ultimate goal is to identify land classes fully independently, and only based on the observed relationship between vegetation and climate. To this end, we rely on the definition of the number of clusters on the predictive performance of the MTL model. In Sect. 6.2.3, it is stated that the ASO-MTL approach shares the objective function of the CMTL method. This only holds if the number of clusters (which is a predefined parameter in the CMTL method) is equal to the value of the parameter h in the ASO-MTL method, which is the dimensionality of the common feature space. In light of this equivalence relation, we experimented with a wide range of values for h in a validation set, aiming to select the value of h that maximises the model performance in terms of R^2. Figure 6.5 shows the median of the predictive performance (R^2) for all tasks

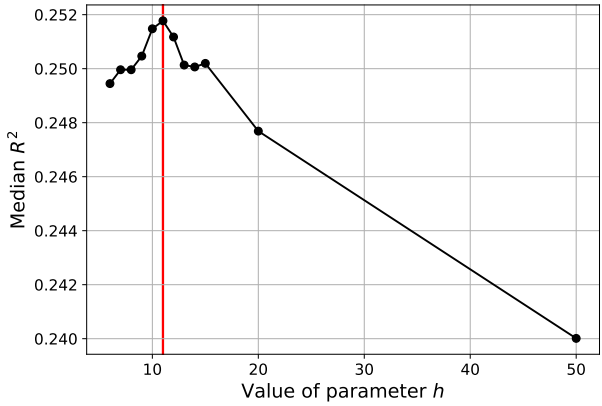


Figure 6.5: Assessing the number of biomes: Median of the predictive performance of the ASO-MTL model in terms of R^2 when the value of the h parameter varies. For $h = 11$ the model scores the maximum value of R^2. However, the differences in the predictive performance for $h = 6, [...], 15$ are marginal.

when the value of the parameter h varies. Note that for these experiments, the λ parameters remain constant in order to assess only the effect of parameter h on the model performance. As one can observe in Fig. 6.5, the maximum median value R^2 is achieved when $h = 11$. However, the differences in the predictive performance for $h = 6, .., 15$ are marginal. Therefore, we can conclude that the method gives robust results, as the strongest predictive structures are captured for the first most important components given by the SVD (see more details in

Sect 6.3.5). As such, the number of biomes we use in the clustering phase equals to 11, since the data-driven methodology with the criterion of the maximum predictive performance indicates this particular number of biomes.

The results of this hierarchical clustering (with Euclidean distance) can be visualised in a dendrogram representation. Figure 6.6b depicts the dendrogram formed by our framework, with the vertical cutting line separating the data into 11 clusters. This representation allows for a visual inspection of whether the choice of the 11 clusters is in line with the dissimilarities existing in the observations. As one can observe, our choice is reasonable, since the clusters at this point are quite dissimilar, based on the Euclidean distance metric, compared to hypothesized cutting lines either before or after this point. In other words, the branches of the dendrogram are already quite long at 11 clusters, indicating high dissimilarities between the resulting classes.

6.3.4. Hydro-climatic biomes

The final objective of this study is to uncover the regions in which vegetation responds in a analogous way to climate anomalies, here referred to as 'hydro-climatic biomes'. In the previous section, we investigated the appropriate number of such regions based on the information contained in our database. Figure 6.6a illustrates the spatial distribution of the emerging global hydro-climatic biomes. The colours depicted correspond to those of the clusters in the dendrogram of Fig. 6.6b. Further analysis of this dendrogram, in combination with the spatial distribution of the clusters in Fig. 6.6a, shows that our framework can clearly differentiate the bio-climatic behaviour of northern latitude ecosystems from those in mid- and southern latitudes. The behaviour of tropical ecoregions is unsurprisingly closer to the behaviour of sub-tropical ones, while boreal regions sharing the exposure to low temperature anomalies have a more coherent response to one another, forming the second main branch of the dendrogram. Bearing in mind the results of the Granger causality approach by Papagiannopoulou et al. (2017b), described in the previous chapter (Chapter 5), as well as the prior knowledge on climate and land use classification, we define the hydro-climatic biomes as follows: (1) Tropical, (2) Transitional water-driven, (3) Transitional energy-driven, (4) Sub-tropical energy-driven, (5) Sub-tropical water-driven, (6) Mid-latitude water-driven, (7) Mid-latitude temperature-driven, (8) Boreal temperature-driven, (9) Boreal water-driven, (10) Boreal water/temperature-driven, (11) Boreal energy-driven. This nomenclature is broadly based on latitude and main climatic drivers.

Figure 6.6c shows the main 10 climate regions of the Köppen-Geiger climate classification, which is based on precipitation and temperature, and their seasonality. On the other hand, the International Geosphere-Biosphere Program (IGBP) (Loveland and Belward, 1997) land cover classification, depicted in Fig. 6.6d, is mostly based

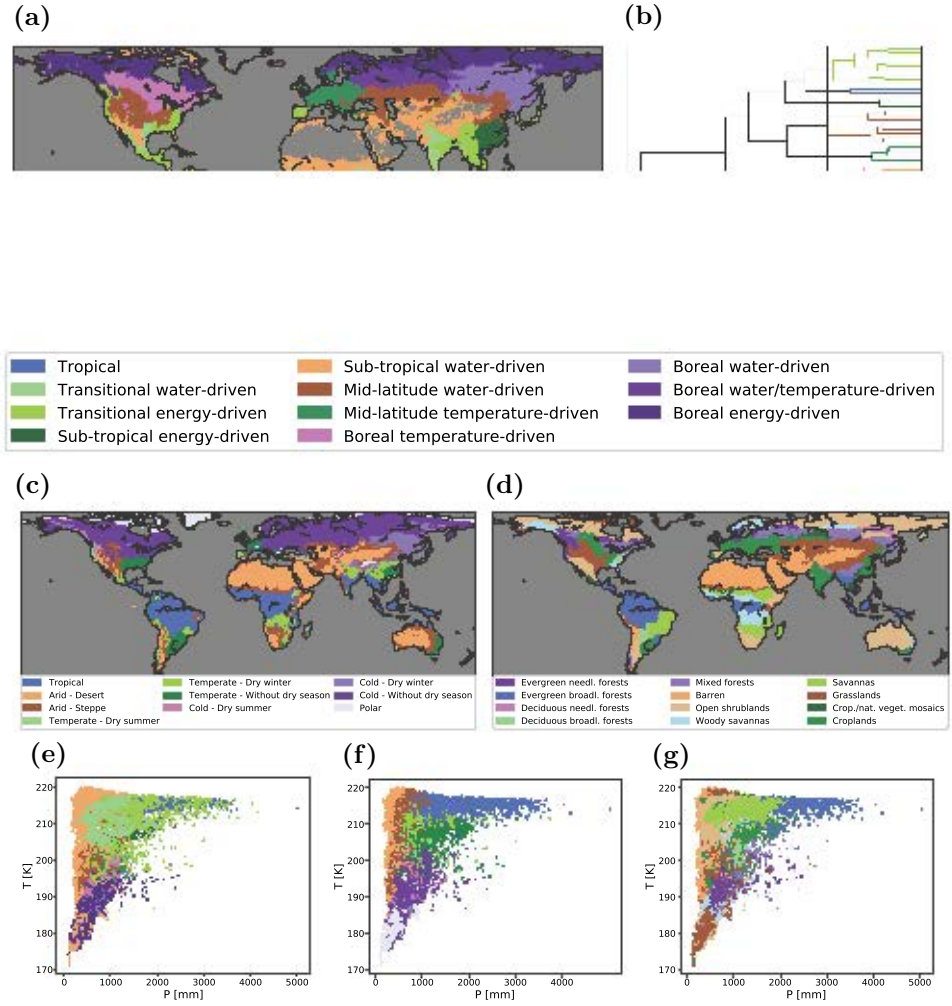


Figure 6.6: Comparison of the different land surface classification schemes. (a) Hydro-climatic biomes derived from the proposed framework. The region colours correspond to the colours of the clusters that are depicted in the dendrogram. (b) Dendrogram scheme of the clustering result derived by the hierarchical agglomerative clustering on the low-dimensional representation of our model observations. The length of the dendrogram branches is a function of the inter-cluster dissimilarities. The vertical cutting line marks the data split into 11 clusters. The denomination of the different classes is supported by the results from Papagiannopoulou et al. (2017b), described in Chapter 5. (c) Simplified Köppen-Geiger climate classification scheme. (d) IGBP land use classification scheme. (e) Climate space (i.e. mean annual temperature versus precipitation) for our hydro-climatic biomes in Fig. 6.6a. (f) Same as (e) but for the Köppen-Geiger climate classes in Fig. 6.6c. (g) Same as (e) but for IGBP in Fig. 6.6d.

on plant functional types. Without the need to prescribe any land cover or climate classification, and only relying on the spatial coherence in the vegetation response to climate anomalies, our hydro-climatic biomes in Fig. 6.6a clearly depict some

of the main characteristic patterns from these traditional classification schemes. For instance, the region of North Asia is quite coherent in terms of climate based on the 10 climate classes shown here (Fig. 6.6c), but quite diverse in terms of vegetation type (Fig. 6.6d); the hydro-climatic biomes show a clear distinction in the transition from shrublands (energy-driven) to coniferous forests (energy- and water-driven). In North America, the more energy-limited ecosystems along the coasts emerge from the water-driven regions inland, and a latitudinal behaviour is also depicted, partly reflecting the transition from croplands and grasslands into temperate and boreal forests. Patterns in the tropics clearly differentiate between rainforest and transitional savannas, and in South America the different drivers of vegetation dynamics in the Arc of Deforestation lead to a class change that is not depicted by neither the Köppen-Geiger climate classification nor the IGBP land cover classes. Finally, the patterns found for arid and warm semiarid regions (here referred to as 'sub-tropical water-driven'), and their transition towards wetter and more vegetated ecosystems, agree with the expectations based on vegetation (Fig. 6.6d) and climate (Fig. 6.6c).

The comparison to the Köppen-Geiger and IGBP maps serves only as a general evaluation or proof of concept for our hydro-climatic biomes map, since in the end such maps are based on a different rationale, and thus, there is no intent to 'outperform' these classification schemes. However, it can be observed in this comparison that the hydro-climatic biomes map in Fig. 6.6a combines information on climate and vegetation zones by illustrating regions where vegetation similarly interacts with the multi-month dynamics in climatic and environmental conditions. This conclusion is confirmed by the scatter plots in Figs. 6.6e-g. Figure 6.6e depicts our hydro-climatic biomes of Fig. 6.6a in climate space of mean annual temperature against precipitation, while Fig. 6.6f shows the same but for the Köppen-Geiger climate classes of Fig. 6.6c. In Fig. 6.6f, the five climate classes are well-separated, since their definition is based on these two climatic variables. On the other hand, Fig. 6.6g depicts the same information but for the IGBP map of Fig. 6.6d. In this figure, savannahs, tropics, and shrublands appear again well clustered. It can be observed that the scatter plot of Fig. 6.6e clearly lie between the two previous classifications in terms of clustering. Boreal biomes correspond to cold climate classes, the sub-tropical and mid-latitude water-driven biomes correspond to arid regions, while the transitional biomes correspond to the savannas and croplands. The clustering of biomes is also consistent with the global distribution of key climatic drivers reported in Chapter 5. These common dynamics are identified by latent structures in our MTL approach. A discussion about these latent structures is included in the next section (Sect. 6.3.6). Moreover, we should note that the approach of spectral clustering applied here allows for a robust result, as small perturbations in the data sets do not affect the overall clustering result. This conclusion is confirmed by the fact that even in the tropical region where the uncertainty in the observations is typically larger and the skill of

the predictions is lower (see Fig. 6.4), the different clusters are separated in a clear manner. A discussion about the comparison of the three land surface classification schemes (the hydro-climatic biomes, the Köppen-Geiger climate classification and the IGBP land use classification based on the predictive structures is presented in Sect. 6.3.7.

6.3.5. Visualization of different number of hydro-climatic biomes

In Sect. 6.3.3 we concluded that the method gives robust results, as the strongest predictive structures are captured for the first most important components. To visually inspect the spatial distribution of the hydro-climatic biomes given a different number of regions, we experimented with parameter h. To this end, we ran the algorithm for $h = 9, 10, 11, 12$ to check the robustness of the results. The conclusion of Sect. 6.3.3 is confirmed by Fig. 6.7, where the maps with 9 (Fig. 6.7a), 10 (Fig. 6.7b), 11 (Fig. 6.7c) and 12 (Fig. 6.7d) hydro-climatic biomes are depicted. In all figures, the tropics, the boreal and the arid regions are well-detected. In addition, sub-tropical regions and transitional ones are also commonly defined in all of the aforementioned figures. Differences in the borders of the identified regions are noticed between temperature-driven areas (e.g., Europe and North America). In transitional water- and energy-driven regions also there are some differences in the clusters' borders. However, these inconsistencies can be explained by the smoother differences between the climatic and environmental conditions in these areas.

6.3.6. Visualization of the most important predictive structures

In Sect. 6.2.5, we describe the steps of the SVD-based ASO algorithm, which learns a low-dimensional feature representation for our tasks based on the relationships between them. The learned matrix $\mathbf{\Theta}$ maps the high-dimensional space to a (lower) h-dimensional space, storing the loadings of the original weights to the 'highly predictive structures'. Thus, the task models are also projected to this shared lower-dimensional space. This information is stored in the matrix $\mathbf{V}$, on which the clustering approach is performed. Figure 6.8 presents the values of the tasks in the first six components of the matrix $\mathbf{V}$. Similar pixel values to the same components mean similar climate-vegetation dynamics. There are several remarks considering Fig. 6.8: (1) all the six components are able to distinguish specific regions according to different criteria such as regions with temperate and dry climate, regions with cold and dry climate, tropical and dry climate, etc.; (2) pixels which are grouped into the same region in the final clustering result (Fig. 6.6a) tend to have similar

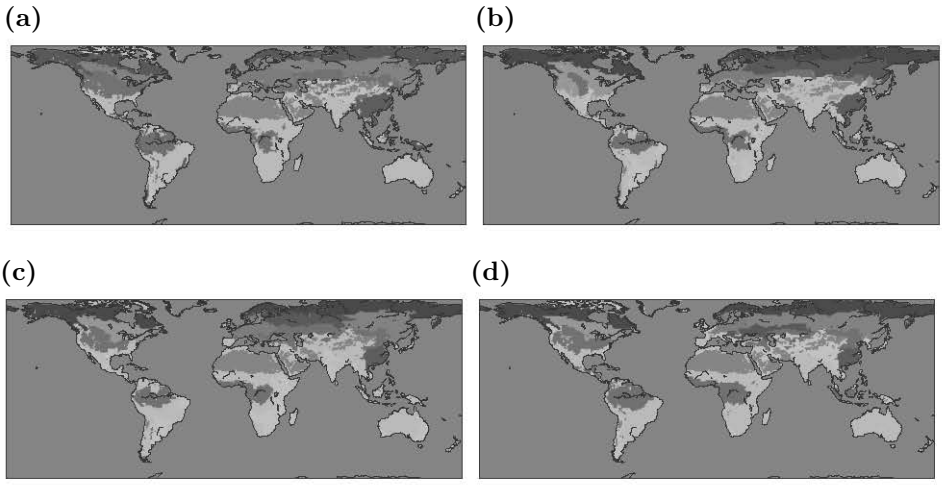


Figure 6.7: Maps with different number of hydro-climatic biomes. (a) $h = 9$ (i.e., 9 hydro-climatic biomes) , (b) $h = 10$, (c) $h = 11$ (Fig. 6.6a), and (d) $h = 12$.

values in a particular predictive structure, and (3) the differences in the values across regions are intense, and in some cases one can recognize the boundaries of the regions depicted in Fig. 6.6a.

6.3.7. Visualization of the predictive structures with the different land surface classifications

As in Zscheischler et al. (2012), we conduct a dimensionality reduction to the matrix V which contains the clustering data. We separately present the results for the Northern and the Southern Hemisphere (ibid.) – see Figs. 6.9 and 6.10, respectively. The data is projected onto the first two components of the t-SNE method (Maaten and Hinton, 2008) and visualized based on the hydro-climatic biomes (Fig.6.9a and 6.10a), the Köppen-Geiger clustering (Köppen, 1936) (Fig.6.9b and 6.10b) and the IGBP clustering (Loveland and Belward, 1997) (Fig.6.9c and 6.10c). We use the same color representation as in Fig. 6.6a. That way we can assess if the learned predictive structures match well the classes of the different classification schemes.

Considering Fig. 6.9, one can see that the best-formed clusters are depicted in Fig. 6.9a, as the clustering has been performed on this dataset (as expected). Figure 6.9c represents the IGBP land use classification; the tropical regions are well-detected as well as the forest- and the cropland-covered regions. This means that the learned predictive structures are highly relevant to the vegetation type of each region. In addition, Fig. 6.9b indicates that the cold, the arid and the tropical

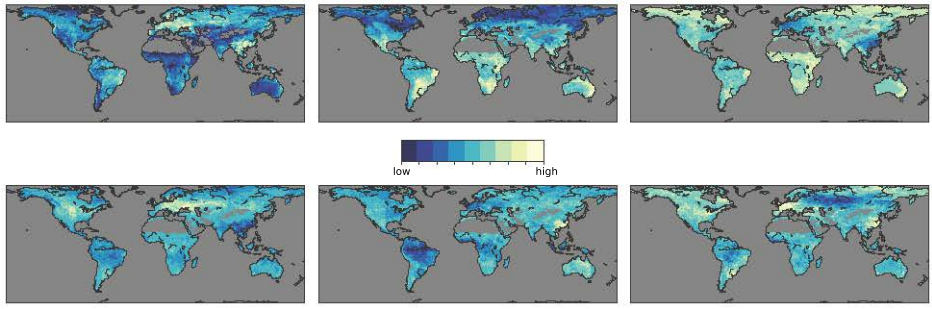


Figure 6.8: Visualization of the first six 'principal components' of the predictive structures. The classification of the land surface into the hydro-climatic biomes is based on the importance of these structures for each location. The color intensity in the map indicates the value magnitude of each pixel in a particular predictive structure.

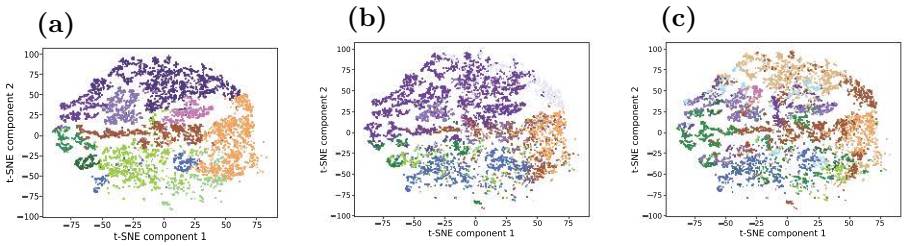


Figure 6.9: Data projection to the first two t-SNE components for the Northern Hemisphere. Each point represents one pixel of the global grid and it is colored based on (a) the hydro-climatic biomes, (b) the Köppen-Geiger climate classification, and (c) the IGBP land use classification. For the color-class mapping see Fig. 6.6.

regions can be well distinguished by the learned structures whereas the temperate climate is scattered among the others and is thus harder to be identified.

Figure 6.10 depicts the same plots for the Southern Hemisphere. As in Zscheischler et al. (2012), overall, the classes identified by the various classification schemes show a worse match than for the Northern Hemisphere. However, Fig. 6.10a shows that the predictive structures can clearly distinguish the sub-tropical water-driven region and the transitional energy/water-driven regions as well. In addition, the Köppen-Geiger climate classes (Fig. 6.10b) of the tropic and the arid regions are also identified in a certain degree. The IGBP classes, in Fig 6.10c, do not form clear clusters.

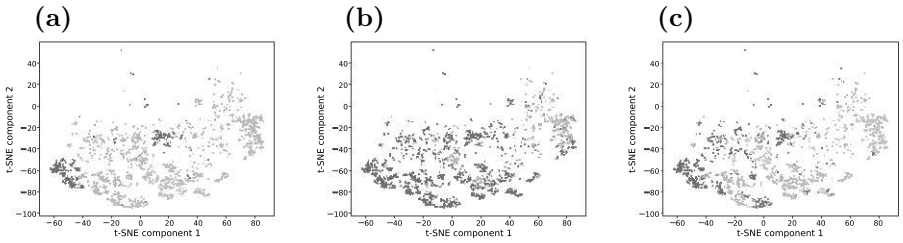


Figure 6.10: As Fig. 6.9 but for the Southern Hemisphere.

6.4. Conclusion

In this chapter, we introduced a novel framework for identifying regions with similar biosphere-climate dynamics interplay. Our framework combines a multi-task learning (MTL) modelling approach and a spectral clustering technique, and it is applied to the global database of global observational climate records described in Chapter 3. Comparisons to a typical single-task learning approach, in which each task (in each location) is analysed separately, indicate that learning about climate–vegetation relationships in neighbouring, or even remote, locations can help predict local vegetation dynamics based on climate. Moreover, our approach is able to detect shared hidden predictive structures among the tasks that enhance the predictive performance of the models. These predictive structures form the basis for the clustering algorithm to detect regions where vegetation responds to climate in a similar way. We demonstrate that, without the need to prescribe any land cover information, our method is able to identify coherent climate-vegetation interaction zones that that emerge directly from the spatio-temporal variability in the data. These zones agree with traditional global classification maps, such as the Köppen-Geiger climate classification or the IGBP land cover classification. We refer to these regions as 'hydro-climatic biomes'. These wide regions can be used in various applications in geosciences, such as unravelling anomalous relationships between climate and vegetation dynamics at local scales, defining extreme values of vegetation response to climate, exploring tipping points (Horion et al., 2016) and turning points of ecosystem resilience, and benchmarking the dynamic response of vegetation in Earth system models.

7 Global vegetation extreme events and their response to climate variability

In all previous chapters, our aim to infer causal relationships between time series of continuous measurements, lead to regression settings. However, classification settings may arise when targeting extreme events, such as heatwaves, droughts or floods. In this chapter, we conduct an experimental study in the area of investigating climate-vegetation dynamics as before, where such a classification setting naturally arises. Specifically, we investigate the relationship between climate and browning events. This is a practically-relevant setting, because extremes in vegetation can reveal the vulnerability of ecosystems with respect to climate change. Firstly, a more precise description of the application domain and the recent literature is provided (Sect. 7.1). Then, we present the various definitions of vegetation extreme events (Sect. 7.2.2) and the extended Granger-causality framework for binary target variables (Sect. 7.2.3). Afterwards, the main results are discussed (Sect. 7.3). These are preliminary results and represent work in progress.

This chapter is based on the content of:
Papagiannopoulou, C., Miralles, D. G., Demuzere, M., Verhoest, N. E. C., and Waegeman, W.: Global browning events and their response to climate variability, *in prep.*

De Graeve, A. : Detecting climate drivers for vegetation extremes., Master thesis (tutored by C. Papagiannopoulou), Ghent University, 2018.

7.1. Introduction

Climate extremes have a great impact on terrestrial biomes since they affect different functionalities of plant life (Hasanuzzaman et al., 2013). Various vegetation types have different responses under similar climatic conditions, i.e., croplands have a higher vulnerability to high temperature compared to forests (Larcher et al., 1994). For instance, if one considers only temperature extremes, one can already observe severe consequences on the terrestrial ecosystems in Europe which are mostly covered by agricultural landscapes (Semenov and Shewry, 2011; Deryng et al., 2014). Such ecosystem impacts can be identified as extreme events in terrestrial biosphere on the global data streams (e.g., NDVI, fraction of Absorbed Photosynthetically Active Radiation (fAPAR)) collected by remote sensing observations.

Recent studies on modelling extreme events mostly focus on linking climate extremes and their effects to biosphere extreme events (Ciais et al., 2005; Kurz et al.,

2008; Zeng et al., 2009). Some of these studies conduct analyses in limited geographical regions or focus on specific disturbance factors, such as fires (Forkel et al., 2012). However, there are other recent works that analyze the climatic drivers of vegetation at global scale (Karnieli et al., 2010; Kim et al., 2010; Papagiannopoulou et al., 2017b). Specifically, Liu et al. (2013) defined vegetation extreme events from NDVI time series obtained by a Box-Cox data transformation and evaluate their sensitivity to climate by using the slope of linear regression models. Other studies focused on extreme events have considered linear relationships between environmental variables and have applied correlation and regression analysis (Stöckli and Vidale, 2004; Hao et al., 2012). However, the linearity assumption might lead to inconsistent results, since relationships between environmental variables are highly non-linear (Papagiannopoulou et al., 2017a). In the same context of analyzing the effect of climate on vegetation, Zscheischler et al. (2013) proposed an extreme event identification method and analyse the effect of climate extremes on fAPAR extreme events. Other studies (Rammig et al., 2015; Baumbach et al., 2017) use coincidence analysis to attribute biosphere extremes. Unlike linear correlation analysis, which considers the general dependence (i.e., covariance) between the time series, coincidence analysis focuses on the co-occurrence of events defined in different variables (Donges et al., 2016).

In this chapter, we focus on defining vegetation extreme events (i.e., browning events) based on observational data. The proposed definition is based on a generic land classification method that is applicable in various spatio-temporal data sets (i.e., relation between climate–vegetation). Therefore, our method is a data-driven approach that strongly relies on the automated extracted regions, called hydroclimatic biomes, described in detail in Chapter 6. By using this land classification, we aim to resolve issues identified in previous works (Zscheischler et al., 2013; Rammig et al., 2015) that rely on predefined regions without taking into account the spatial distribution of the Earth observation. For instance, Zscheischler et al. (2013) focus their work on six predefined regions (i.e., continents) as defined by the IPCC Special Report. Other studies that apply data-driven methodologies to define regions with similar characteristics for defining extremes, such as Mahecha et al. (2017); Guanche García et al. (2018), use clustering approaches or binning techniques to group the pixels and form the regions. However, these methods are based on parameters, such as number of clusters or bins, which should be defined in advance. Different parameter values lead to different results, making the whole process rather complicated in practice. Moreover, other previous studies (Nicolai-Shaw et al., 2017; Baumbach et al., 2017; Liu et al., 2013) apply their methods at pixel level, without considering information about the locations with similar characteristics. Taking into account regional information is beneficial in defining extreme events, with the definition of browning events emerging from the 'average' conditions of a specific region.

In addition, we apply the non-linear Granger causality framework, introduced

in Chapter 4, to identify complex relationships between climate and vegetation extremes. Specifically, we extend the existing framework in order to model the effect of climate on vegetation extreme events instead of just investigating the average response of vegetation to climate (Chapter 4). Moreover, another extension is the application of our framework at a region scale instead of a per-pixel basis. In particular, we perform our analysis on the hydro-climatic biomes instead of just using the previously studied (Papagiannopoulou et al., 2017a,b) pixel-based modelling approaches. That way, modelling at a region scale leads to more robust results, since the number of observations increases. To sum up, the contribution of this chapter is twofold: (i) we propose a new definition of vegetation extreme events based on the hydro-climatic biomes and (ii) we apply a non-linear Granger causality framework to investigate the relations between climate and vegetation extremes. Future extensions of this work will include a detail investigation of the core climatic variables leading to browning events in different regions of the world.

7.2. Materials and methods

7.2.1. Database

Our framework is applied on a large database of global climate and vegetation records. The database mainly consists of satellite and *in situ* measurements and spans of 30 years (1981-2010), as described in Chapter 3. The data sets have been transformed to a common monthly temporal resolution and a spatial resolution of 1°. In this database, several relevant vegetation drivers have been collected. Specifically, the main climatic and environmental variables are included: (i) land surface temperature, (ii) near-surface air temperature, (iii) longwave/shortwave surface radiative fluxes, (iv) precipitation, (v) snow water equivalent and (vi) soil moisture. For vegetation, we use NDVI data set (Tucker et al., 2005). In this chapter, we define as target variable a binary time series in which the value '1' indicates the starting point of an extreme event in vegetation (see Sect. 7.2.2 for more details). As predictor variables, we define the rest of the collected data sets. Moreover, we use the same set of manually extracted features from the raw time series of the climatic variables as defined in Chapter 3.

7.2.2. Defining vegetation extremes

To investigate the effect of the climatic drivers on extremes in vegetation, one should first detect the extreme events in vegetation based on observational data. There are several methods proposed in the literature that try to identify patterns in vegetation data which may be classified as 'extreme events'. Extreme event, by

definition, occur only rarely; this means that one should expect a limited number of occurrences during long time spans. Thus, the presence or the absence of an extreme event can be defined as a highly imbalanced binary classification problem. Intuitively, more strict definitions of extreme events will (i) reduce the number of extreme events and (ii) increase the class imbalance problem. As such, if the number of extreme events is more limited, analyses are performed in an under-represented sample, making the modelling process hard or even impossible. On the other hand, the definition of an extreme event should be based on prior knowledge about climate, vegetation, the interaction between them and/or other factors. Hence, the physical interpretation of an extreme event is also a very important aspect to be considered already when designing the detection method. Therefore, new definitions of extreme events should take into account a trade-off between the validity of the analysis and the physical interpretation of the conditions that characterize an event as an extreme.

A common practice in defining extreme events based on observational data is the use of a cut-off threshold. The 10^{th} percentile is broadly used as a threshold in several previous works for defining an extreme event in climate or vegetation data (Seneviratne et al., 2012; Zscheischler et al., 2013; Baumbach et al., 2017; Rammig et al., 2015). The 10^{th} percentile can be calculated on a per-pixel basis. Thus, observations below this threshold are defined as extreme events for this particular pixel (above this threshold are the non-extreme events), see Fig. 7.1. The numerical value of the cut-off threshold (e.g., 5^{th} percentile, 10^{th} percentile) affects the total number of extreme events. That is, lower values of this threshold (5^{th} percentile) result in less extreme events while higher ones (10^{th} percentile) in larger number of extremes in the same period. A drawback when considering the previous definition of an extreme event (i.e., 10^{th} percentile per pixel) is that it results in the same number of extremes per pixel. To circumvent this problem, since extremes may be considered to be more likely in some regions than in others, previous studies (Zscheischler et al., 2013; Mahecha et al., 2017) have considered the 10^{th} percentile in specific regions instead of considering it at a pixel level. In this study, we use the hydro-climatic biomes introduced in Chapter 6, which are regions with similar vegetation response to climate, and were specifically designed for this purpose. From now on, when the 10^{th} percentile is mentioned, this is the 10^{th} percentile calculated per region.

A limitation of many traditionally used definitions of extremes is that they do not capture that vegetation needs a certain time to recover. For this reason, here, an event is considered as extreme only if predefined temporal (e.g., two months) and/or spatial extreme conditions are also occurring (Liu et al., 2013; Mahecha et al., 2017). The amount of vegetation coverage is expected to affect the magnitude of the NDVI anomalies in a particular location. In addition, the calculation of the percentiles per region is strongly influenced by pixels with lower vegetation coverage, since in these pixels, the NDVI values tend to be lower. To alleviate this

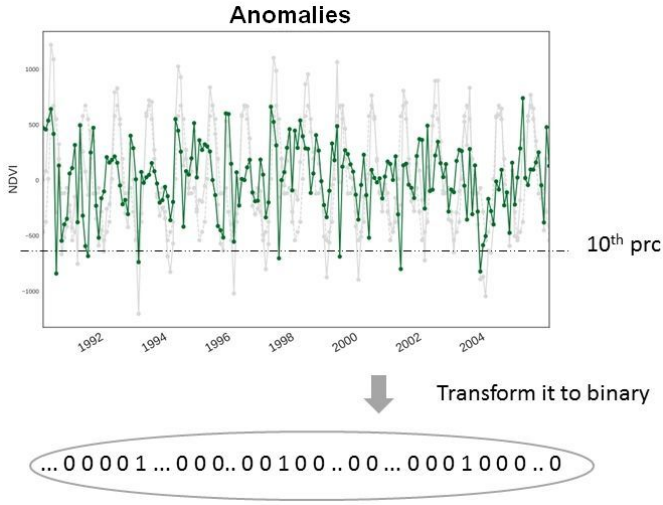


Figure 7.1: Threshold-based definition of vegetation extreme events. The green time series represents the NDVI anomalies for one pixel. The grey line represents the corresponding raw NDVI data. The value of the 10^{th} percentile is depicted as threshold. A value of '1' is assigned to the data points below this threshold and a value of '0' is assigned to the data points above this threshold. The resulting binary variable is the new class variable in our setting.

issue, the original NDVI values are corrected by the fractional vegetation coverage per pixel, based on the data set from MODIS MOD44B vegetation continuous field (Dimiceli et al., 2015). Specifically, we obtain these coverage factors using the sum of the fractions of herbaceous plants and tall canopies. Furthermore, only the pixels with a coverage fraction exceeding a threshold of 0.1 are kept for further analysis.

Another aspect in defining the extreme events in vegetation is that even when the seasonality of the raw NDVI data is removed, as described in Chapter 3, some residual seasonality may be retained in the occurrence of extreme events. This phenomenon can be explained by the fact that extreme events usually occur during specific seasons (i.e., summer), affecting also the interpretation of the Granger causality results. To reduce the effect of seasonality, the de-trended and de-seasonalized NDVI time series is divided by the standard deviation of the corresponding month. Moreover, as mentioned above, vegetation needs some time to respond to climate changes. This progressive adaptation can also be observed in the presence of autocorrelation on the NDVI anomalies time series.

To sum up, for the construction of the binary target variables for the vegetation extreme events, we: (i) apply the vegetation coverage filter on the raw NDVI time series for each pixel, (ii) use a time series decomposition method and we divide with the standard deviation for each month, (iii) calculate the 10^{th} percentile of

the anomalies for each region, (iv) take into account extreme events which last at least two months (low values for more than two consecutive months) and cover at least two pixels (during the event at least one neighbouring pixel should experience an extreme event as well) and (v) transform each time series to binary target data (by keeping only the starting point of an event as positive example). This results in a binary variable, see Fig. 7.2 for an example of an NDVI anomalies time series and the starting points of the extreme events detected by our methodology.

7.2.3. Granger causality for binary data

Granger causality (Granger, 1969), as discussed in Chapter 4, is a well-established framework that has been exploited in climate sciences to detect causal relationships between time series based on predictability. Assuming we have two time series $\boldsymbol{x} = [x_1, x_2, ..., x_N]$ and $\boldsymbol{y} = [y_1, y_2, ..., y_N]$, where N the length of the time series. If $\boldsymbol{x}$ serves as a cause and $\boldsymbol{y}$ as an effect, one can say that $\boldsymbol{x}$ Granger causes $\boldsymbol{y}$ when at a specific time point t, the prediction of y at t improves when past information of $\boldsymbol{x}$ is taken into account. So, in the bivariate case, the forecasts of two models are compared: (i) the forecast of a *baseline* model which includes information only for the history of the time series $\boldsymbol{y}$ and (ii) the forecast of a *full* model that includes also information from the past values of the time series $\boldsymbol{x}$.

In this study, we focus on the problem of Granger causal relationships between climate and vegetation extremes. Thus, in our case, the time series $\boldsymbol{y}$ that resembles the effect is a binary variable where: (i) '0' denotes the absence of an extreme event and (ii) '1' indicates the starting point of an extreme event at timestamp t. To this end, two classification problems are defined (i.e., one for the baseline and one for the full model). To compare the prediction performance of the two models, one should define a performance measure. Since extreme events rarely occur, the two classes (i.e., 0 and 1) are heavily imbalanced. So, a performance measure such as the Area Under the Curve (AUC) can be used to deal with this class imbalance problem. AUC can be easily calculated by considering pairs of observations where one is positive and the other negative and calculating how frequently the positive one has the highest (most positive) test result. Denoting as $\hat{\boldsymbol{y}}$, the one-step ahead predicted time series of the original time series $\boldsymbol{y}$, Granger causality can be formulated as:

Definition 3. *A time series $\boldsymbol{x}$ Granger causes a target time series $\boldsymbol{y}$ if the predicted performance in terms of $AUC(\boldsymbol{y}, \hat{\boldsymbol{y}})$ improves when $x_{t-1}, x_{t-2}, ..., x_{t-P}$ are included in the model as predictors for the forecast of y_t, in contrast to including $y_{t-1}, y_{t-2}, ..., y_{t-P}$ only, where P is the lag-time moving window.*

In cases where other climatic factors, which act as additional confounding effects to our target variable, are not included in the analysis, Granger causality might lead to incorrect conclusions (Geiger et al., 2015). These factors can be included as additional variables in the analysis to alleviate this issue. For instance, assuming

another variable $\boldsymbol{z}$ for which its presence affects the decision whether $\boldsymbol{x}$ Granger-causes $\boldsymbol{y}$, the above definition is extended as follows:

Definition 4. *A time series $\boldsymbol{x}$ Granger causes a target time series $\boldsymbol{y}$ if the predicted performance in terms of $AUC(\boldsymbol{y}, \hat{\boldsymbol{y}})$ improves when $x_{t-1}, x_{t-2}, ..., x_{t-P}$ are included in the model as predictors for the forecast of y_t, in contrast to including $y_{t-1}, y_{t-2}, ..., y_{t-P}$ and $z_{t-1}, z_{t-2}, ..., z_{t-P}$ only, where P is the lag-time moving window.*

It is straightforward to extend the previous definitions (bivariate, tri-variate cases) to multi-variate settings. As we mentioned above, in our setting, the target variable will represent the vegetation extremes at a particular location, while the other time series ($\boldsymbol{x}$ and $\boldsymbol{z}$) will be climatic time series at the same location, such as temperature, soil moisture, etc. In statistical terms, the null hypothesis (H_0) of Granger causality examines whether the two models (i.e., baseline and full model) have the same predictive performance. On the other way around, the H_0 is rejected if the full model significantly outperforms the baseline model. Typically, in Granger causality analyses, vector autoregressive models are used and the significance of the results are assessed based on the model parameters. In other studies, several statistical tests have been proposed for nested models such as the likelihood-ratio tests (Mosedale et al., 2006). The limitations of these statistical tests are three fold: (i) they cannot be directly applied on climate data due to their unrealistic assumption (i.e., stationarity), (ii) are based on linear models, although causal relationships in climate science are highly complex and non-linear, and (iii) are evaluated on in-sample data, which is a practice that typically results in the overfitting phenomenon.

In Chapter 4, we have proposed an alternative way to measure Granger-causality between the two models (baseline and full). Specifically, we assess their difference quantitatively instead of qualitatively. We also proposed to replace linear models with powerful machine learning algorithms. If the models (i.e., baseline and full) give more accurate predictions, the conclusions drawn by a Granger causality analysis are stronger with respect to the examined causal relationships. Since there are no existing statistical tests that can be computed to evaluate the significance of the results, we visualize and interpret the differences between the two models in this quantitative way.

7.2.4. Seasonality and trend in vegetation extreme events

Another aspect that should be taken into account in this kind of modelling approaches, such as in Granger-causality analyses, is related to the target variable itself. Unsurprisingly, we noticed that the distribution of the vegetation extremes in time indicates that many more extremes occur in recent years, which means that a clear trend appears again in the time series of extreme events, even though the initial time series was de-trended, see Fig. 7.2. This makes the time series

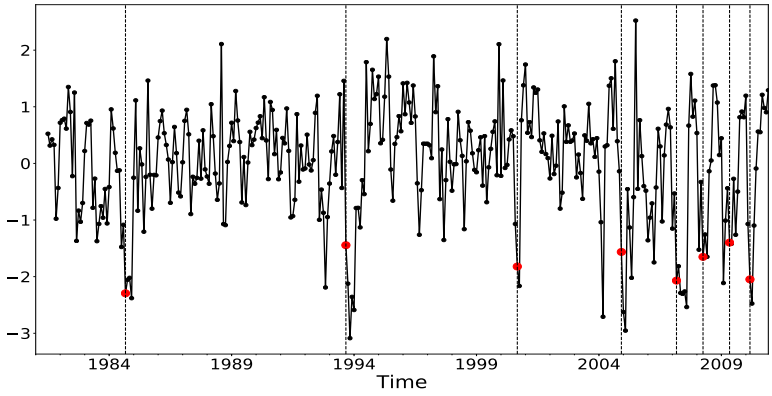


Figure 7.2: Time series of the NDVI anomalies of a pixel in North Africa. The red points (and the dashed vertical lines) indicate the starting points of the vegetation extreme events (i.e., the '1's in the target variable). Some low-value points are not considered as extremes, due to the additional criteria about time duration and space extension that may not be fulfilled in these low-value points. In the recent years more vegetation extreme events are detected in this particular location.

highly non-stationary. Moreover, also a seasonal cycle typically re-appears, as one observes more extremes in certain months. Correctly identifying those two components (trend and seasonality) is essential when inferring causal relationships between vegetation extremes and climate.

As discussed in Sect. 7.2.3, a baseline model only includes information from the target time series (i.e., previous timestamps). We both consider the anomalies as well as their binarized extreme counterparts as features for the baseline model. However, due to the existence of seasonal cycles and trends when considering binary time series of extreme vegetation, we also include 12 dummy variables which indicate the month of the observation and a variable for the year of this observation. These last two components are necessary because the baseline model should tackle as good as possible the seasonality and the trend that exists in the time series of NDVI extremes. As such, the full model extends the baseline model with the above-mentioned climatic variables.

7.3. Results and discussion

7.3.1. Proposed definition of browning events

Figure 7.3a depicts the bio-climatic regions used in this study for the definition of the vegetation extreme events as well as the Granger-causality analysis. These

regions have been defined in Chapter 6 by using a purely data-driven approach. Based on climate and vegetation observations, locations where vegetation responds to climate in a similar way are grouped together, forming coherent regions in terms of bio-climatic conditions. The use of the land classification scheme of hydro-climatic regions is of great importance for the definition of extreme events and the attribution analysis, since: (i) the number of observations for each general bio-climatic condition increases, (ii) by having a complete picture of the bio-climatic conditions, the definition of extreme events as well as the analysis of the climatic drivers become more robust, (iii) other land classifications may not fully capture the climate–vegetation interactions since they are either based on climate or vegetation data (see e.g., Loveland et al. (2000)), and (iv) vegetation extreme events emerge by not only taking into account the vegetation observations but also the response of vegetation to climate. Therefore, in the definition of the vegetation extreme events the response of vegetation to climate is taken into account and thus, only anomalous vegetation responses are detected. In addition, from a statistical point of view, the modelling approach becomes more stable with the use of a large amount of data. Hence, causal inference based on predictability can lead to stronger conclusions.

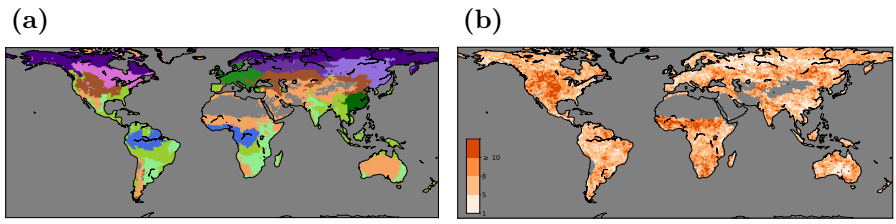


Figure 7.3: Hydro-climatic biomes and vegetation extreme frequency. (a) The hydro-climatic biomes used from the definition of vegetation extreme events (figure based on the results of Chapter 6). (b) In the lighter-colored regions fewer vegetation extreme events are detected while in the darker ones more vegetation extreme events are identified based on the NDVI time series and our detection method in Sect. 7.2.2.

In Fig. 7.3b, the distribution of the vegetation extreme events detected by our approach is depicted. As it can be observed, the filtering step, which weights the importance of each pixel for the calculation of the threshold, plays an important role in the final result. For instance, in the Australian desert there are almost no extremes since the vegetation is rather limited. On the other hand, most browning events are concentrated in the North parts of America and Africa. In general, our results are in line with previously reported ones. Specifically, it is clear that the distribution of the extremes in the North America is similar to the one shown by Zscheischler et al. (2013). Also similar patterns are observed in South America, South Africa, mid latitudes of Asia and East Australia. However, the frequency of the browning events in North Africa in our resulting map does not agree with the

findings by Zscheischler et al. (2013). In addition, our vegetation extreme spatial distribution is similar to the one shown by Liu et al. (2013). However, the main differences are observed in the Amazon, in west of Europe as well as in the mid latitudes of Asia. Our approach takes into account the response of vegetation to climate, and thus it is not based only on vegetation observations as in Liu et al. (2013).

7.3.2. Comparative study of the different definitions of browning events

In this section, we compare the distribution of the extreme vegetation events based on the different definitions that exist in current literature and we elaborate on the one that we propose. The different constraints, which are applied in each step for the characterization of the extremes, have a specific effect that is reflected in the global distribution of the extremes as well as the results of the Granger-causality analysis (see Sect. 7.3.3). From a physical point of view, it may not appear realistic to consider an equal frequency of extremes evenly distributed over the world. For this reason, we do not consider in our study a pixel-based definition of extreme events. Figure 7.4 depicts the frequencies of the vegetation extreme events defined based on different approaches. Figure 7.4a shows the distribution of the extremes calculated at regional scale by just taking the 10^{th} percentile as a threshold. As one can observe, the global distribution of the extremes is quite homogeneous; yet, there are regions with high frequency in browning events while there are others with a lower frequency. Similar patterns are depicted in Figs. 7.4b-d, with Fig. 7.4b being the distribution of the extremes with spatial extension (the extreme affects at least one neighbouring pixel with respect to a given pixel) and Fig. 7.4c being the distribution of the extremes based on the spatio-temporal three-dimensional cube of Zscheischler et al. (2013) (the extreme affects at least one out of the six nearest spatio-temporal neighbours of a given pixel). Figure 7.4d depicts the global distribution of the vegetation extremes calculated as in Fig. 7.4c, with the only difference that this time 26 nearest neighbours are taken into account in the spatio-temporal three-dimensional cube. The spatial extension and the consideration of the spatio-temporal three-dimensional cube do not effect in a large degree the global distribution of the vegetation extremes compared to the initial simpler definition of Fig. 7.4a.

On the other hand, the temporal extension of the extreme events does affect the number of extremes per pixel at global scale. In these definitions, the starting point of an extreme event is denoted as extreme, while the subsequent extreme values are not considered as extremes. This means that in the target variable, there are 1's in the starting points of the extreme events and 0's in the rest. As such, the rest of the maps (Figs. 7.4e-h) have a reduced number of extremes compared to the previous ones. Figure 7.4e illustrates the distribution of the extreme events with

two-month duration at least, while Fig. 7.4f with three-month duration at least. As one can observe, the extreme events with duration three months (or more) occur at limited regions. Thus, even though an analysis of severe browning events would be of a great interest, the data in this setting are so limited that a further statistical analysis is not possible. In Fig. 7.4g, we first apply the vegetation filtering and then we apply the two-month and the one-neighbour constraint. The distribution is similar to the one of Fig. 7.4e, but the frequency is decreasing. In addition, extreme events do not occur in arid regions. Finally, Fig. 7.4h illustrates the distribution of the proposed definition, which is described in the previous section. In this definition, the standard deviation of each month is taken also into account. The main effect of this modification compared to the definition of Fig. 7.4g is that the map in Fig. 7.4h is more homogeneous than the one in Fig. 7.4g.

7.3.3. Detecting Granger-causal relationships between climate and vegetation extremes

Figure 7.6 shows the results of our analysis. Specifically, we assess the ability of a non-linear classification model to detect Granger-causal effects of climate on vegetation. The full model, which includes information for the past of climate, outperforms the baseline one in most regions. This indicates that climate Granger-causes vegetation browning events in most of the world. As one can observe in Fig. 7.6a, in regions such as North America, Europe, China, South Amazon and subtropical regions in Africa, the performance measure for the full model increases substantially compared to the performance of the baseline model. In other regions, such as in Australia, south Africa, South America and middle latitudes, there is an indication of Granger causality, since in some locations the full model clearly outperforms the baseline one, although the result is not homogeneous (scattered result). Although one expects that in neighbouring locations conclusions about the climate effect on vegetation extreme events should be similar, this is not the case for the aforementioned regions. So, conclusions about the impact of climate on the extreme events in vegetation in these regions should be drawn with caution. The distribution of the AUC scores of the baseline model is also depicted in the blue histogram of Fig. 7.6b, while the AUC scores of the full model is depicted in the orange one. From this figure, it becomes clear that the full model scores higher AUC values than the baseline one for a large number of pixels. Moreover, the skewness of the orange histogram towards one means that the performance of the full model tends to be closer to the optimal side.

Extremes in climate have been related to vegetation extreme events (Zscheischler et al., 2013; Baumbach et al., 2017). This is because extreme climatic conditions lead to extreme response of vegetation. To this end, in our analysis we incorporate extreme climate indices to capture extreme climatic conditions. Based on the previous results, the effect of climate extremes on the 'average' vegetation conditions is important (see Chapter 5). This means that also the influence of climate

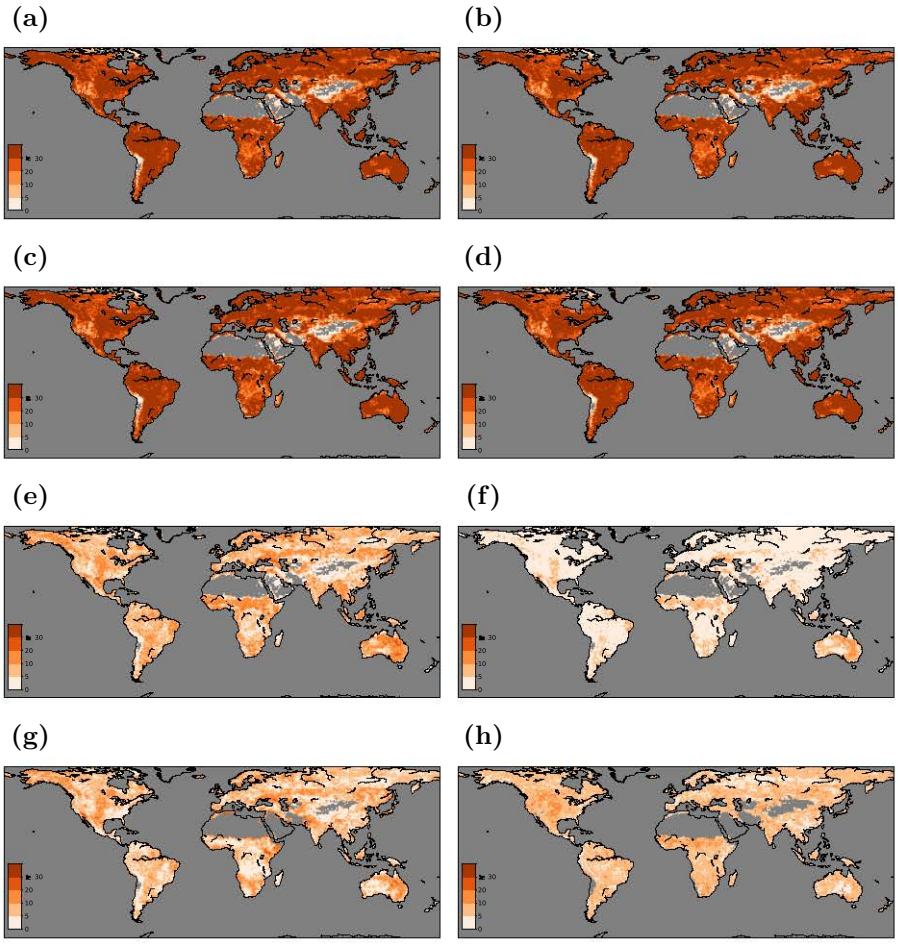


Figure 7.4: Vegetation extreme frequency for the different definitions of vegetation extremes. (a) Spatial distribution of vegetation extremes based on the 10^{th} percentile for each region. (b) Spatial distribution of vegetation extremes based on the 10^{th} percentile for each region with spatial extension (the extreme affects at least one neighbouring pixel to a given pixel). (c) Spatial distribution of vegetation extremes based on the 10^{th} percentile for each region and on the spatio-temporal three-dimensional cube of Zscheischler et al. (2013) (the extreme affects at least one out of the six nearest spatio-temporal neighbours of a given pixel). (d) Global distribution of the vegetation extremes calculated as in Fig. 7.4c, with the only difference that this time 26 nearest neighbours are taken into account in the spatio-temporal three-dimensional cube. (e) Spatial distribution of vegetation extremes based on the 10^{th} percentile for each region with extreme events of two-month duration at least. (f) Spatial distribution of vegetation extremes based on the 10^{th} percentile for each region with extreme events of three-month duration at least. (g) Spatial distribution of vegetation extremes based on the 10^{th} percentile for each filtered region with extreme events of two-month temporal duration and two-pixel space coverage at least. (h) Spatial distribution of vegetation extremes calculated as in (g) by also taking into account the standard deviation of each month. For future analysis we propose the use of the last definition of Fig. (h).

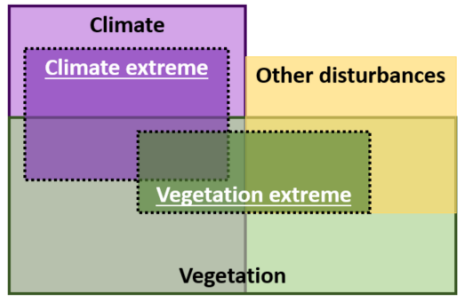


Figure 7.5: Schematic representation of the effect of climate extremes to vegetation extremes.

extremes provokes the extreme response of vegetation, see Fig. 7.5 for a schematic representation. Note that other factors, such as rising of CO_2 concentrations, changes in land use, deforestation, grazing, nitrogen deposition, can all affect vegetation as well, leading to extreme events. This is also indicated in Fig. 7.5, where vegetation extremes may be caused by other disturbances. However, in this study these factors are not taken into account. In general our Granger-causality pattern (Fig. 7.6) is in line with the result of Papagiannopoulou et al. (2017b), presented in Chapter 5, since the impact of the hydro-climatic extremes (hydrological, radiation and temperature extremes) on vegetation coincides for most of the regions. In addition, the non-homogeneity of the affected regions is a common result in both works, meaning that extreme phenomena might occur at local scale.

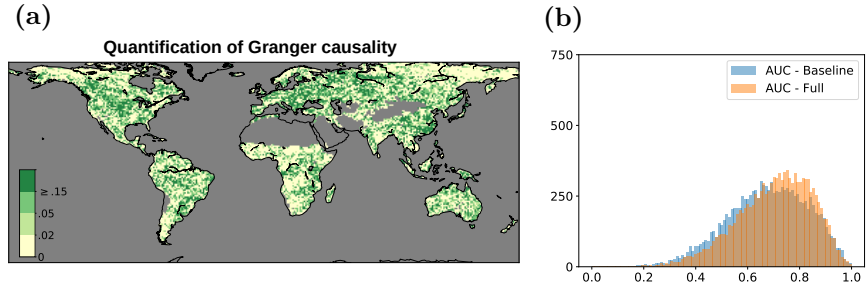


Figure 7.6: Quantification of Granger causality based on the proposed definition for the vegetation extreme events. (a) Spatial overview of the quantification of Granger causality; in the green regions the full model outperforms the baseline model in terms of the AUC performance measure. (b) Distributions of the AUC scores of the baseline (blue histogram) and the full model (orange histogram).

7.3.4. A comparative study of Granger-causality analysis based on the various extreme definitions

As described in Sect. 7.2.3, Granger-causality analysis is based on the predictive skill of a full model versus a baseline model. In this section, we investigate the performance of the baseline models applied on the different binary target variables, created by the various vegetation extreme definitions. By assessing the predictive performance of the baseline models, one can reveal basic characteristics of the target variables. Figure 7.7 depicts the performance of the baseline models in terms of the AUC performance measure. The order of the maps is the same as in Fig. 7.4, so each map corresponds to the definition of extreme events described in Sect. 7.3.2. In general, the baseline models perform well in most of the regions for all the definitions of the extreme events. Specifically, Figs. 7.7a-d show the performance of the baseline models on the target variables, calculated based mostly on spatial information. From these maps, it becomes clear that if one does not keep only the starting point of an extreme event, the autocorrelation in the target variable is high and thus, can be detected by the baseline model, leading to a high predictive performance. This autocorrelation is not present in the definitions of the maps in Figs. 7.7e-h, and therefore the performance decreases compared to the first four maps. However, the high predictive performance in arid and semiarid regions is common for all the maps of Fig. 7.7, e.g., in Australia; this can be explained by the fact that vegetation data are quite constant in these regions, meaning that there is also high autocorrelation in the NDVI anomalies. As such, the models are able to detect this autocorrelation, since the baseline model includes the lagged values of the NDVI anomalies as additional predictor variables. In addition, in these regions, there is an increasing trend in the occurrence of vegetation extremes, see Fig. 7.2. The year of the extreme observations is also included in the baseline models to represent this trend, affecting their predictive performance. Moreover, the high predictive performance in the boreal regions can be explained by the high seasonality in the extreme events, since only in summer months vegetation extremes can be detected, due to the snow coverage in the rest of the year.

To delve into the form of the different NDVI anomalies time series and the extremes detected based on the proposed definition of vegetation extreme events, we illustrate the corresponding time series of a pixel in which the baseline model performs well and of another pixel in which the baseline performs poorly, see Fig. 7.8. Figure 7.8a depicts the NDVI anomalies time series of a pixel in Amazonia (where the baseline models perform rather poor) and Fig. 7.8b shows the NDVI anomalies time series of a pixel in Central Australia (where the baseline models perform rather well). The detected (based on the proposed definition) vegetation extreme events are highlighted in red. From these time series, it is clear that there are two types of vegetation extreme events: (i) there are extreme events with a slower progress that last for a certain period, i.e., the amount of vegetation decreases until an extreme

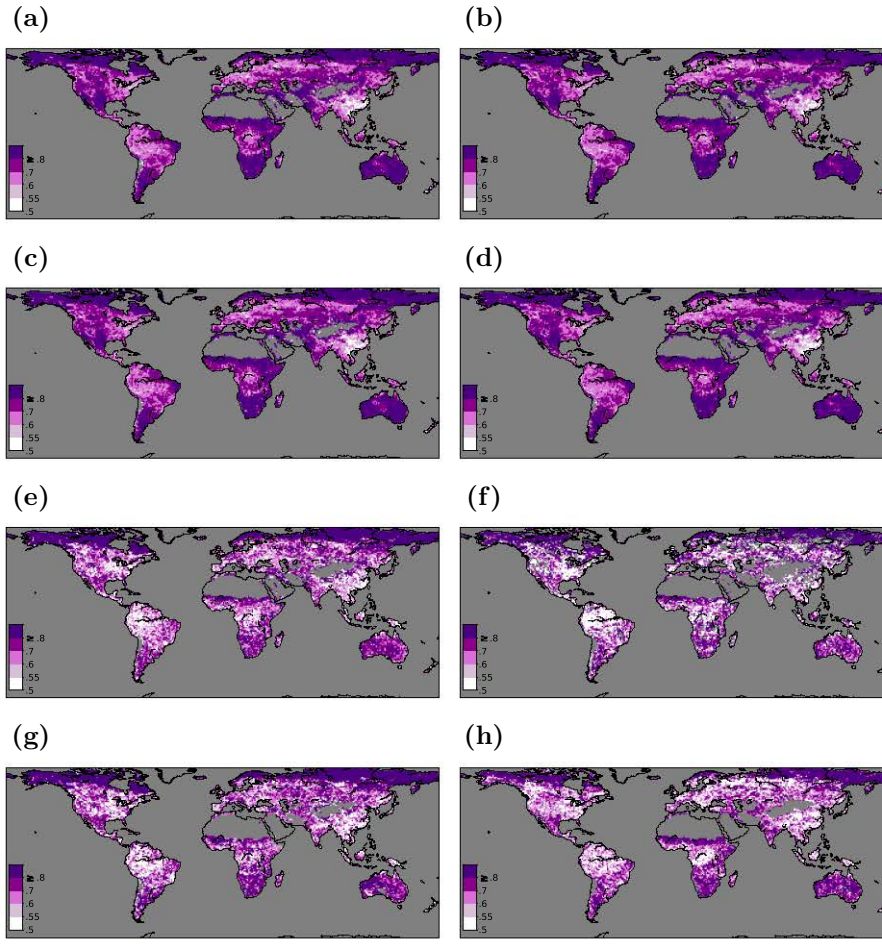


Figure 7.7: Performance in terms of the AUC measure of the baseline models for the corresponding definitions of vegetation extreme events described in Fig. 7.4. The baseline models include as predictors information relevant to vegetation only, i.e., lagged NDVI values of the anomalies, lagged extreme–non-extreme values, 12 dummy variables (which encode seasonality), year (which encodes trend).

event is reached and after a while it starts rising again (Fig. 7.8b), and (ii) there are more sudden extremes, corresponding with sharp peaks in the vegetation data, i.e., the extreme value is reached suddenly and lasts only for a very short period (Fig. 7.8a). Hence, the sudden extremes will be more difficult to be predicted and will be characterized by a lower performance in the baseline model. The longer lasting extreme events will have the tendency to have a higher score. This is because these vegetation data and extremes will contain a high autocorrelation, meaning that the next value is easier to be predicted from the previous value(s). Therefore, based on the previous discussion, it can be concluded that there is a relationship

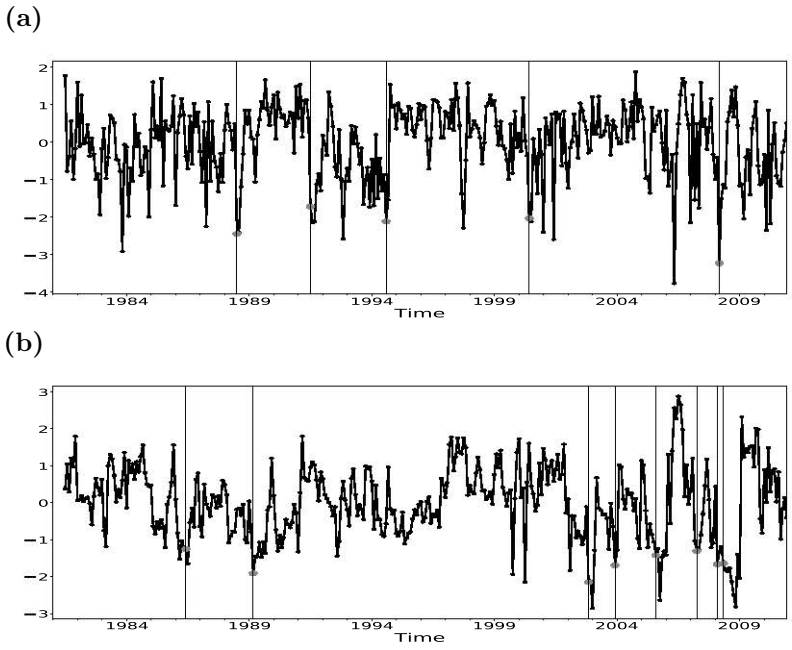


Figure 7.8: Time series of the NDVI anomalies of two pixels in which the baseline model performs differently. (a) NDVI anomalies time series of a pixel in Amazonia (where the baseline models perform rather poor). (b) NDVI anomalies time series of a pixel in Central Australia (where the baseline models perform rather well). The detected (based on the proposed definition) vegetation extreme events are highlighted in red (and with the dashed vertical lines).

between the high performance of the baseline model and the seasonality, trend and autocorrelation in the vegetation data. The regions with a low baseline score are the regions with sudden and less predictable extremes. Since these sudden peaks in the data have a low autocorrelation, the models have more difficulties in predicting these values. These regions can also have a weaker seasonal cycle and not a clear trend.

In the proposed definition with the filtered NDVI data adjusted with the standard deviation, the goal is to reduce the effect of the seasonality. However, the results are very similar to those of the definition with the filtered NDVI data without the standard deviation adjustment. The corresponding baseline model (Fig. 7.7h) performs a bit worse compared to the baseline model in Fig. 7.7g, but still the performance remains high. For instance, in the north there is still a strong seasonality, since the surface is covered half of the year with snow, while in other regions with a high score, the autocorrelation in the vegetation data seems to affect the performance. Also in some regions, there is a trend in the frequency of the extreme events, which further causes the performance increase.

Another remark considering the performance of the baseline models in Figs. 7.7e-h is that it shows a scattered distribution. This result could be explained by our choice to consider as an extreme event only the starting point of the event. This choice encodes naturally the fact that only the start of the event is caused by the climate. For example, the response time of vegetation or some other indirect effects of vegetation extremes can be possible causes of the low vegetation in subsequent months. In addition, the temporal and spatial extension of the events are used in order to remove possible noise in the data.

The performance of the corresponding full models is similar to the baseline ones, i.e., in regions where the baseline models perform well, the full models also perform well and in regions where the baselines perform poorly, the same holds for the full models

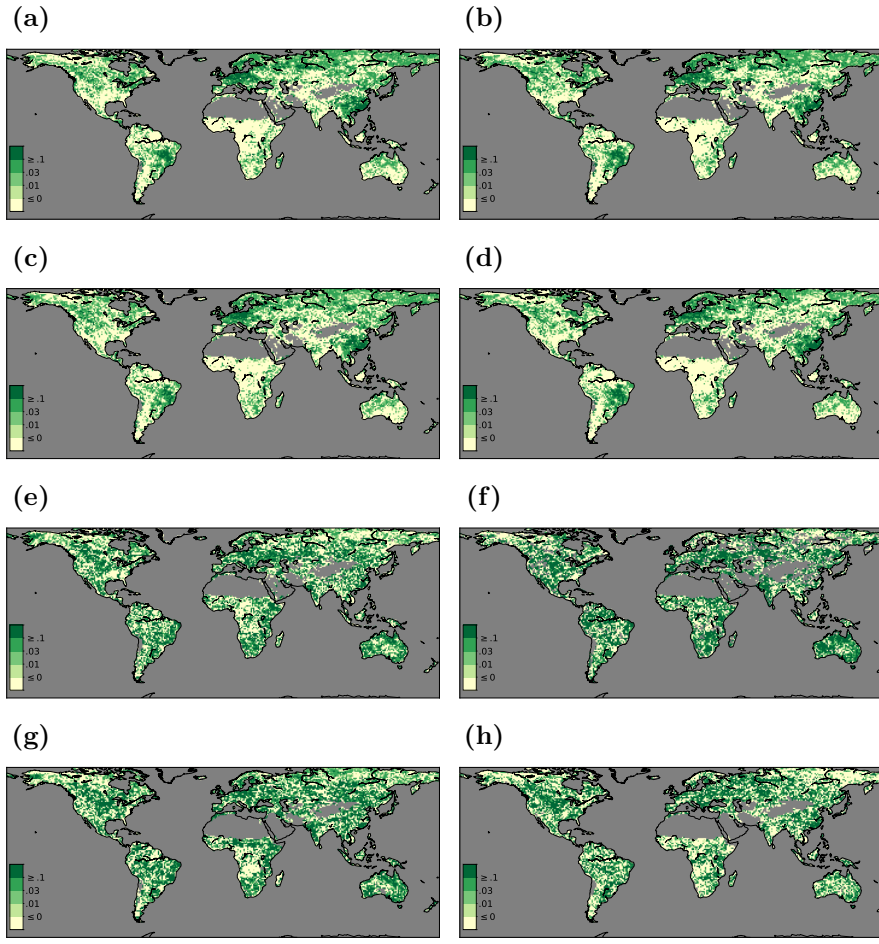


Figure 7.9: Spatial overview of the quantification of Granger causality for each of the different definitions of vegetation extreme events of Fig. 7.4; in the green regions the full model outperforms the baseline model in terms of the AUC performance measure.

(results not shown). The difference between the two models are depicted in Fig. 7.9, which is the quantification of Granger causality for each of the different definitions of the extremes. As one can observe, even though the results are again scattered, especially for the definitions in Figs. 7.9e-h, in most of the pixels the improvement in predictive performance is larger than 10%, which is a large improvement in terms of the AUC measure. Therefore, one can conclude that climate drivers have a substantial influence on the vegetation extremes. However, it would be expected that the influence would be more distributed into regions where vegetation has high sensitivity to climate. Interestingly, in Figs. 7.9a-d, the Granger causality is clustered into similar regions. These are regions where the baseline model performs rather low and thus, the addition of the climate data improves the performance of the full model, resulting in a higher Granger causality quantification. The other way around, in regions where quantification of Granger causality is low, the baseline model performs already very good (a score larger than 0.8), so there is no room for further improvement. Undoubtedly, the difference in predictive performance of the baseline models (in the different regions due to the different characteristics in vegetation data) influences the interpretation of the Granger-causality analysis. Therefore, these aspects should be taken into consideration for further study.

7.4. Conclusions

In this chapter, we identified complex relationships between climate and vegetation extremes by applying the non-linear Granger-causality framework, introduced in Chapter 4. Specifically, we proposed a new definition of browning events based on the hydro-climatic biomes (see Chapter 6) and we apply the non-linear Granger causality framework to investigate the relations between climate and the detected extreme events. Our results indicate that definitions of the browning events should take into account a trade-off between the validity of the analysis and the physical interpretation of the conditions that characterize an event as an extreme. In addition, our Granger-causality analysis shows that climate Granger causes browning events in most regions, although the conclusions highly depend on the definition of an extreme event. Future extensions of this work will include a detailed investigation of the core climatic variables leading to browning events in different regions of the world.

8 Analyzing Granger causality in climate data with time series classification methods

In this chapter, we investigate the potential of state-of-the-art time series classification techniques to enhance causal inference in climate science. Specifically, we postulated that causal inference in climate science can be further improved by using automated feature construction methods for time series. We conduct a comparative experimental study of different types of algorithms on a large test suite that comprises our unique database from the area of climate–vegetation dynamics. The results indicate that specialized time series classification methods are able to improve existing inference procedures. Substantial differences are observed among the methods that were tested.

This chapter is an edited version of:
Papagiannopoulou, C., Decubber, S., Miralles, D. G., Demuzere, M., Verhoest, N., and Waegeman, W.: Analyzing Granger Causality in Climate Data with Time Series Classification Methods. Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD) (3) 2017: 15-26. Presented at the Applied data science track.

8.1. Introduction

In the general framework that we presented in the previous chapters, we constructed hand-crafted features based on knowledge that has been described in the climate literature (e.g. Donat et al. (2013)). These features include lagged variables, cumulative variables as well as extreme indices. Therefore, we ended up with in total ∼360 features extracted from one time series. Our previous results have shown that incorporating those features in any classical regression or classification algorithm might lead to a substantial increase in performance (for both the baseline and the full model). In this section, we investigate whether this feature construction process can be automated using time series classification methods.

8.2. From Granger causality to time series classification

Due to the increased public availability of data sets from various domains, many novel time series classification algorithms have been proposed in recent years. In a time series classification task, one tries to classify a time series in a specific class. Formally, an example in this task is a pair $(\boldsymbol{x}, y)$ with m observations $x_1, ..., x_m$ (the time series) and discrete class variable $\boldsymbol{y}$ with c possible values. In our setting, there are two possible class values $\{0, 1\}$. The data set D consists of a N examples with associated class labels, i.e., $D = (\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$. A classifier is a function or mapping from the space of possible inputs to the class variable values. Some classifiers give also as output a probability distribution over the class variable values. Time series classification algorithms involve some processing or filtering of the time series values prior or during constructing the classifier. Most of those methods either try to find higher-level features that represent discriminative patterns or similarity measures that define an appropriate notion of relatedness between two time series (Liao, 2005; Ding et al., 2008; Bagnall et al., 2017). The following categories can be distinguished:

(a) Algorithms that use the whole series or the raw data for classification. To this family of algorithms belong the one nearest neighbour (1-NN) classifier with different distance measures such as dynamic time warping (DTW) (Sakoe and Chiba, 1978), which is usually the standard benchmark measure, and variations of it, the complexity invariant distance (CID) (Batista et al., 2014), the derivative DTW (Górecki and Łuczak, 2013), the derivative transform distance (DTD) (Górecki and Łuczak, 2014) and the move-split-merge (MSM) (Stefan et al., 2013) distance.

(b) Algorithms that are based on sub-intervals of the original time series. They usually use summary measures of these intervals as features. Typical algorithms in this category are the time series forest (TSF) (Deng et al., 2013), the time series bag of features (TSBF) (Baydogan et al., 2013) and the learned pattern similarity (LPS) (Baydogan and Runger, 2016).

(c) Algorithms that are attempting to find informative patterns, called shapelets, in the data. An informative shapelet is a pattern that helps in distinguishing the classes by its presence or absence. Representative algorithms of this class are the fast shapelets (FS) (Rakthanmanon and Keogh, 2013), the shapelet transform (ST) (Hills et al., 2014) and the learned shapelets (LS) (Grabocka et al., 2014).

(d) Algorithms that are based on the frequency of the patterns in a time series. These algorithms build a vocabulary of patterns and form a histogram for each observation by using this vocabulary. Algorithms, such as the bag of patterns

(BOP) (Lin et al., 2012), the symbolic aggregate approximation-vector space model (SAXVSM) (Senin and Malinchik, 2013) and the bag of SFA symbols (BOSS) (Schäfer, 2015), are based on the idea of a pattern vocabulary.

(e) Finally, there are approaches that combine more than one from the above techniques, forming ensemble models. A recently proposed algorithm named collection of transformation ensembles (COTE) combines a large number of classifiers constructed in the time, frequency, and shapelet transformation domains.

In our comparative study, we run algorithms from the first four different groups. The main criteria for including a particular algorithm in our analysis are (1) availability of source code, (2) running time for the data sets that we consider, and (3) interpretability of the extracted features. Since we have collected multiple time series for a large part of the world (3,536 locations in total), the algorithms should run in a reasonable amount of time. Several algorithms had problems to finish within three days. We briefly describe the algorithms selected for performance comparison in our climate data set:

Complexity invariant distance (CID) Batista et al. (2014) defined the concept if complexity invariance in time series. Intuitively, complex time series are characterized by many peaks and valleys. The distance between pairs of complex time series is frequently greater than the distance between pairs of simple time series. A complexity invariant distance measure has been introduced to compensate this phenomenon. Specifically, a distance measure is multiplied by a term that is calculated based on the sum of squares of the first differences of the time series. The Euclidean and the DTW distance measures can be used from the CID algorithm.

Time series forest (TSF) Deng et al. (2013) proposed a random forest approach, using summary statistics as features. The training of a tree is performed by using the mean, standard deviation and slope of random intervals for every series as features, while the classification of a new observation is obtained by a majority voting over all trees.

Learned pattern similarity (LPS) LPS (Baydogan and Runger, 2016) follows a tree-based ensemble-learning strategy that is quite fast and has as a goal the identification of local autopatterns. Specifically, it creates regression trees based on randomly selected subseries as features and a random attribute as target variable. Then, the observations are transformed according to the frequency of the values residing at each terminal node of the trees. The classification is performed by applying the 1-NN algorithm on the new representation.

Fast shapelets (FS) The FS algorithm (Rakthanmanon and Keogh, 2013) tries to find informative shapelets in a fast way, avoiding to fully enumerate the whole search space. It uses the symbolic aggregate approximation (SAX) (Lin et al., 2007)

to reduce the dimension of the series. Then it forms a vocabulary from the SAX words and it performs dimensionality reduction on them. After that, it counts the presence of each word in each class and it scores the words based on their discriminative power. Finally, it selects the k best SAX words and it maps them back to the original subseries, i.e., shapelets. The set of the k shapelets is further assessed based on the information gain criterion (Ye and Keogh, 2011).

Bag of patterns (BOP) BOP (Lin et al., 2012) is an algorithm that also uses the SAX representation (Lin et al., 2007). BOP takes as input three parameters, named window size, word length and alphabet size and applies SAX to each window of the time series forming a word, which represents a pattern. Then, a histogram is calculated based on the frequency of the patterns in each observation. When a new observation arrives, the same transformation is applied on it and it is classified based on the histogram distances using the 1-NN classifier.

Symbolic aggregate approximation-vector space model (SAXVSM) Senin and Malinchik (2013) also use the SAX representation (Lin et al., 2007) but instead of creating histograms, they calculate the term frequencies (tf) multiplied by the inverse document frequency (idf) for each class separately. This representation is typically used in document classification tasks. SAXVSM takes as input the same parameters as BOP and the classification of a new observation is performed by using the 1-NN classifier in combination with the cosine similarity measure.

Bag of SFA symbols (BOSS) BOSS (Schäfer, 2015) is also an algorithm which creates a vocabulary out of the time series and uses the words of this vocabulary as features. Its main difference from the above two algorithms (BOP and SAXVSM) is the way that it constructs the words from the time series windows. More specifically, BOSS uses the Discrete Fourier Transform (DFT) on each window, while the above two algorithms (BOP and SAXVSM) use the technique of piecewise constant models (PAA). The steps of the algorithm are: (a) split the time series in different intervals based on a window size, (b) perform a DFT on each of them, (c) find the bin in which each Fourier coefficient drops in, and (d) transform the subseries into a word by using the names of the bins (each bin has a letter as a given name, e.g., 'a', 'b'). Finally, the classification of a new observation is performed by using again the 1-NN classifier in combination with a non-symmetrical distance function that measures the distance only for the words that appear in the test observation.

8.3. Experimental setup

In order to evaluate the above-mentioned time series classification methods for causal inference, we adopt an experimental setup that is similar to Papagiannopoulou et al. (2017a), described in Chapter 4. The non-linear Granger causality framework is adopted to explore the influence of past-time climate variability on vegetation dynamics. To this end, we used the data sets of observational nature presented

in Chapter 3. However, in all previous chapters, we used in this way in total 21 data sets. For the present study, we retained three of them, while covering the three basic climatic variables: water availability, temperature, and radiation. The main reason for making this restriction was that in that way the running time of the different time series classification algorithms could be substantially reduced. Specifically, we use one precipitation data set, which is coming from a combination of *in situ*, satellite data, and reanalysis outputs, called MSWEP (Beck et al., 2017). We include one temperature data set, which is a reanalysis data set, and one radiation data set from the ECMWF ERA-Interim (Dee et al., 2011). For vegetation. we use again the NDVI (Tucker et al., 2005) seasonal anomalies as explained in Chapter 3.

Since these are preliminary experiments for the attribution of vegetation extremes, we adopt a simple definition of the vegetation extreme events. Specifically, we group the location pixels into areas with the same vegetation type, by using the global vegetation classification scheme of the International Geosphere-Biosphere Program (IGBP) (Loveland and Belward, 1997), which is generically used throughout a range of communities. We selected the map of the year 2001 (closer to the middle of our period of interest). In order to end up with coherent regions that have similar climatic and vegetation characteristics, we further divided the vegetation groups into areas in which only neighbouring pixels can belong to the same group. That way, we create 27 different pixel groups in North and Central America, see Fig. 8.1. We limit the study to North and Central America because some of the time series classification methods that we analyse have a long running time. Once we know which of those methods perform well, the study can of course be further extended to other regions, under the assumption that the same methods are favored for those regions. The vegetation extremes are then defined by applying a 10^{th} percentile threshold on the seasonal anomalies of each region. In this way, we produce the target variable of our time series classification task. The presence of an extreme is denoted with a '1' and the absence with a '0', as previously. In this definition, a clear trend and a seasonal-cycle appear again in the time series of extreme events, as discussed in Sect. 7.2.4.

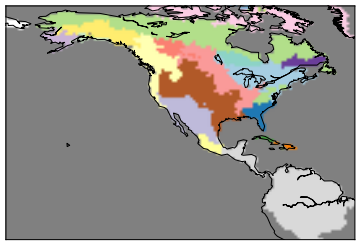


Figure 8.1: Groups of pixels that are regions with similar climatic and vegetation characteristics. Based on the time series of each region we calculate the vegetation extremes for the pixels of that region.

8.4. Results and discussion

We present two types of experimental results. First, we analyze the predictive performance of various time series classification methods as representatives for the full model in a Granger-causality context. Subsequently, we select the best-performing algorithm for a Granger causality test, in which a baseline and a full model are compared.

8.4.1. Comparison of time series classification methods

For the first step we performed a straightforward comparison of the performance of the following algorithms: CID (Batista et al., 2014), LPS (Baydogan and Runger, 2016), TSF (Deng et al., 2013), SAXVSM (Senin and Malinchik, 2013), BOP (Lin et al., 2012), BOSS (Schäfer, 2015) and FS (Rakthanmanon and Keogh, 2013). In this setting, our data set consists of monthly observations (there are in total 360 observations per pixel), and the input time series for each observation includes the 365 past daily values of precipitation time series before the month of interest (excluding the daily values of the current month), see Fig. 8.2 for an example of two observations. Only the precipitation time series is used, as some of the methods are unable to handle multivariate time series as input. We train the models per region by concatenating the observations of the pixels. The evaluation is performed per pixel by using random three-fold cross-validation and AUC as performance measure.

Figure 8.3 shows the results. The vocabulary-based algorithms outperform the other representations, which implies that the frequency of the patterns makes the two classes of our data set more distinguishable. Algorithms which distinguish the observations according to a presence or an absence of a shapelet perform poor, probably because observations originating from consecutive time windows have similar shapelets (the daily values of the next month is added for the next observation). In addition, the shapelet-based FS algorithm is also not very efficient in terms of memory space for large data sets. For this reason, we could not obtain results for the four largest regions of our data set – see Table 8.1. For the algorithms

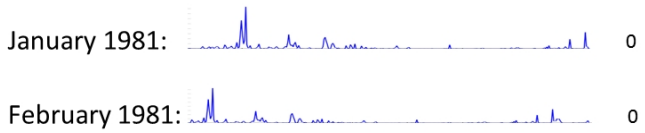
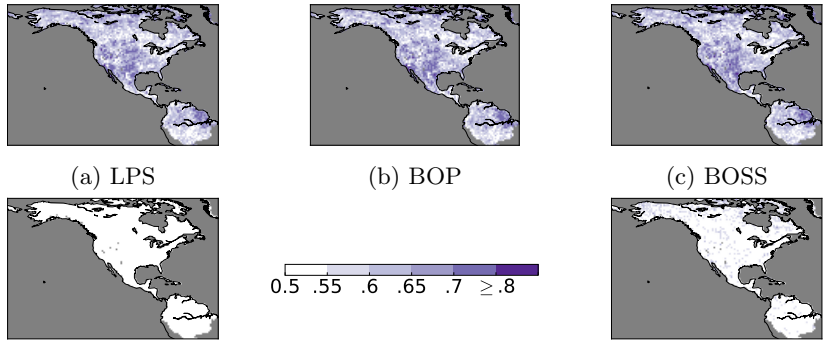


Figure 8.2: Data set example. The input time series for each observation includes the 365 past daily values of precipitation time series before the month of interest. The target variable indicates the presence ('1') or the absence ('0') of an extreme.

that compare the whole raw time series by using a distance measure (i.e., CID) one can observe that the performance is also very low, probably also due to the strong similarity between consecutive observations. Similarly, algorithms that attempt to form a characteristic vector for each class fail, since the patterns in both classes are very similar (i.e., SAXVSM). On the other hand, from the algorithms that use sub-intervals of time series, LPS has a similar performance as the vocabulary-based algorithms, because it takes local patterns and their relationships into account and forms a histogram out of them, while TSF fails in capturing useful information. We note that the LPS algorithm includes randomness, so in each run it extracts different patterns from the data and also it is more time and space inefficient compared to the vocabulary-based algorithms. Table 8.1 presents the numerical results for the nine largest regions. As one can observe, the results of BOP and BOSS are very similar. In most regions they give rise to substantially better results than the other methods that were tested.

8.4.2. Granger causality using the BOSS patterns

In a second step, we combine the best representation coming from the time series classification algorithms and we apply it to the non-linear Granger-causality framework in order to test causal effects of climate on vegetation extremes. Our main goal is to replace the hand-crafted features constructed in Chapter 3. As the BOSS algorithm has the best performance compared to the other time series algorithms, we use the vocabulary of patterns that BOSS automatically extracts from the climatic time series as features. To evaluate Granger causality, the baseline model includes information from the NDVI extremes, while the full model includes



(a) LPS (b) BOP (c) BOSS

(d) SAXVSM, TSF, FS (e) CID

Figure 8.3: Performance comparison in terms of AUC of the time series classification algorithms in the univariate time series classification setting on climate data.

Table 8.1: Mean and standard deviation of the AUC for areas which include more than 100 pixels. The vocabulary-based algorithms as well as the LPS algorithm perform very similar. Results of the algorithms SAXVSM and TSF are omitted due to their low performance.

Algorithm	Reg 1	Reg 2	Reg 3	Reg 4	Reg 5	Reg 6	Reg 7	Reg 8	Reg 9
LPS	0.59±0.06	**0.56**±0.04	**0.65**±0.09	**0.65**±0.07	**0.61**±0.06	**0.62**±0.05	0.60±0.05	0.65±0.07	0.59±0.05
BOP	**0.60**±0.07	**0.56**±0.05	**0.65**±0.08	0.64±0.07	0.60±0.06	0.61±0.05	**0.61**±0.06	0.66±0.07	**0.60**±0.05
BOSS	**0.60**±0.06	**0.56**±0.04	0.64±0.08	**0.65**±0.07	**0.61**±0.05	0.61±0.05	**0.61**±0.05	**0.67**±0.07	0.59±0.05
CID	0.50±0.03	0.50±0.02	0.51±0.05	0.51±0.04	0.50±0.03	0.54±0.04	0.53±0.03	0.55±0.05	0.51±0.03
FS	-	0.50±0.00	0.50±0.00	-	0.50±0.00	-	0.50±0.00	-	0.50±0.00

also the automatically-extracted features from the climatic time series. In contrast to the previous set of experiments, we now include three climatic time series instead of only the precipitation time series.

Figure 8.4 shows the performance of the full model in terms of AUC, as well as the performance improvement of the full model compared to the baseline model. It is clear that by using information from climatic time series the prediction of vegetation extremes improves in most of the regions. Therefore, one can conclude that – while not bearing into consideration all potential control variables in our analysis – climate dynamics indeed Granger-cause vegetation extremes in most of the continental land surface of North and Central America.

Even though results of that kind could be obtained also with hand-crafted feature representations, we do conclude that more specialized time series classification methods, such as BOSS, have also the potential of enhancing causal inference in climate science. This is mainly due to the fact that our hand-crafted representation with the extreme indices and the cumulative variables has a lot in common with the representation obtained by these algorithms. Finally, note that while this work presents particular results for the case of climate–vegetation dynamics, we believe that the approach might be useful in other causal inference studies, too.

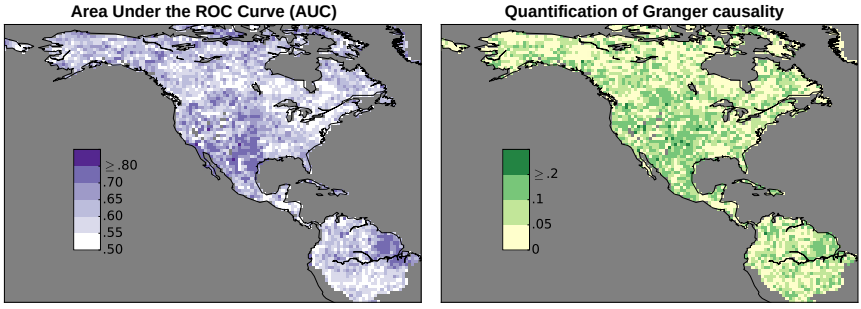


Figure 8.4: On the left, the performance of the full model that uses the patterns extracted by the BOSS algorithm as predictors. On the right, a quantification of Granger causality; positive values indicate regions with Granger-causal effects of climate on vegetation extremes.

8.5. Conclusions

In previous chapters, we have shown that causal inference in climate science can be substantially improved by replacing traditional statistical models with non-linear autoregressive methods that incorporate hand-crafted higher-level features of raw time series. However, approaches of that kind require a lot of domain knowledge about the working of our planet. In this chapter, we postulated that causal inference in climate science can be further improved by using automated feature construction methods for time series. Our experimental results indicate that recently proposed time series classification methods have a lot of potential to improve causal inference in climate science.

9 General conclusions and future directions

In this chapter, we summarize the main conclusions that can be drawn from each of the chapters from this dissertation. Specifically, we present in detail the main contributions of our study. We also discuss possible extensions to overcome limitations of the proposed framework and investigate challenges in the field.

9.1. Conclusions

9.1.1. Granger causality analysis on global climate–vegetation data

The analysis of this thesis was conducted on a database assembled by several publicly available climate data sets. We compiled a global database of observational records spanning a thirty-year time frame, containing satellite, *in situ* and reanalysis-based data sets. At a first stage we conducted an exploratory pre-analysis on the global database. Correlation analysis revealed that the different products which measure the same variable are strongly correlated. For instance, temperature records produced by different resources capture similar information. The same conclusions hold for the other variables as well, i.e., precipitation, soil moisture, etc. However, stronger correlations have been observed in regions where the seasonal component is obvious, such as in the Northern Hemisphere. In regions where the seasonal cycle is not strong, e.g., in the tropics, different measurement records of the same variable are much less correlated. On the other hand, correlations between the anomalies of the corresponding records are not so high for most regions of the Earth as expected.

The target variable of the analysis in this thesis was vegetation. Specifically, we used the commonly-used NDVI greenness indicator as measurement of global vegetation and we isolated the anomalies of the original time series. The time series of the NDVI anomalies are highly autocorrelated at one month lagged time. There is an autocorrelation reduction when the lagged time increases, although in some regions, such as in Australia, there is still autocorrelation even at four months lag time. In addition, we found that there are different correlation levels between the climate variables and the NDVI anomalies. As expected, the highest correlations are identified between the target variable and climate in observations of consecutive months (i.e., current month or month after). Water-related variables are more

correlated in higher lags compared to temperature and radiation variables with vegetation, due to the prolonged memory of the land. Finally, the correlation tends to zero between the NVDI residuals and the 6-month lagged variables.

In Chapter 4, we introduced a novel framework for studying Granger causality in climate–vegetation dynamics. Our approach combines various components, such as data fusion, feature construction and non-linear predictive modelling. We selected the random forest algorithm as a non-linear method for our framework, due to its excellent computational scalability with regards to extremely large data sets, as the global climate data set of this thesis. In general, the non-linear part of the framework could be substituted by any other non-linear machine learning technique, such as neural networks or kernel methods. In our results one can clearly see the non-linear nature of climate–vegetation relationships. This fact highlights the need to move beyond the commonly-used linear approaches of Granger causality. In our analysis, we compared our framework with the traditional linear Granger causality frameworks. As one can notice, 14% more variability of vegetation anomalies has been predicted compared to the linear Granger-causality approaches globally. Moreover, it has been observed that in the water-limited regions the predictive power of the model is higher than in the other regions. This result also indicates that the climate representation, which includes prior knowledge about the lagged-response of vegetation to climate and the effect of climate extremes on vegetation, well-captures the climate–vegetation interactions. For the effect of the particular climate drivers on vegetation, we conducted an analysis in the next chapter (Chapter 5).

In Chapter 5, we investigated the main climatic drivers of vegetation anomalies at global scale by applying the non-linear Granger-causality framework described in Chapter 4. The main result of our analysis is that water availability is the main driver of vegetation anomalies at global scale. Specifically, we concluded that more than half of the global vegetated area is under water limitation; percentage that is the highest one from those reported in previous studies. In our analysis, the role of water availability has been enhanced by the prolonged memory of the soil (compared to the memory of the atmosphere). In addition, as it has been experimentally proven, lagged-values of water-related variables are informative for vegetation anomalies of three months later, especially in semiarid regions. Moreover, in higher latitudes, radiation and temperature are the primary factors of vegetation as expected. In tropical regions, the explained variance of vegetation by climate is lower compared to other regions, possibly because other factors are more important. Finally, our results confirm that hydro-climatic extremes have an impact on monthly vegetation dynamics regionally, although their global influence still requires more thorough investigation, since the mean climate also incorporates its extremes, fading away the effect of the extreme climate on vegetation.

9.1.2. Clustering regions with similar climate–vegetation dynamics

In Chapter 6, we introduced a novel framework for identifying regions with similar climate–vegetation dynamics. We applied our analysis to a global database of observational records, which has been compiled in this thesis (see Chapter 3). Our approach combines an MTL modelling approach and a clustering technique. Our results highlight that the problem of predicting vegetation dynamics based on climate in different locations can be tackled as an MTL problem. Comparison to a typical STL approach, in which each task (in each location) is resolved separately, indicates that the MTL approach outperforms the STL one globally. Moreover, the ASO-MTL approach, which is used in this work, is able to detect shared hidden predictive structures among the tasks. These structures boost the predictive performance of the models and characterize each location. Based on these predictive structures, we apply a clustering algorithm to detect regions where vegetation responds to climate in a similar way. The result of the land-surface classification to these regions is in line with the previous literature and the environmental/climate prior knowledge and it can be used as a basis for further analysis of the climate–vegetation interactions.

9.1.3. Assessing causes of vegetation extremes

In Chapter 7, we investigated different definitions of vegetation extremes and we adapted the Granger-causality framework, developed in this thesis, to detect relationships between climate and vegetation extremes. Specifically, we experimented with different vegetation extreme definitions found in the literature and we proposed new ones. In addition, we discussed important aspects of Granger-causality approaches applied on this setting. The Granger-causality analysis on this problem indicated that climate Granger causes browning events in most regions of the world.

In a second approach, we used a representation coming from time series classification algorithms and we applied it to the non-linear Granger causality framework in order to test causal effects of climate on vegetation extremes. Our main goal was to replace the hand-crafted features described in Chapter 3. Based on the results of the comparative study conducted in Chapter 8, the BOSS algorithm had the best performance, compared to the other time series algorithms. Therefore, we used the patterns that BOSS automatically extracts from the climatic time series as features. Our results showed that by using information from climatic time series the prediction of vegetation extremes improves in most of the regions. Thus, one can conclude that – while not bearing into consideration all potential control variables in our analysis – climate dynamics indeed Granger cause vegetation extremes in most of the continental land surface of North and Central America.

9.2. Future directions

Granger-causality frameworks have been developed under the assumption that causal effects between different variables remain unchanged through time. This is quite a strong assumption, especially in climate sciences, where climatic variables are changing in the different ecosystems due to climate change. Therefore, since there is a change in the behaviour of the different variables, the causes or the effects related with these variables may also change. So, an interesting application is the investigation of the main climatic drivers of each region through time. To do so, methods, such as online learning, can be adopted to assess Granger-causal relationships for each time step. A challenge in this approach is to detect significant changes in the climatic drivers of a region, and therefore other tools, such as special statistical inference, should be developed. In the same direction, evaluating Granger causality in a one-step ahead approach can lead to robust results, since online approaches are commonly applied to non-stationary data. To this end, non-linear methods, which can cope with large data sets and can run in an online mode, are necessary.

Statistical testing is another open research question for non-linear Granger-causality frameworks with autocorrelated variables. As it has been discussed extensively in Chapter 4, various statistical tests have been proposed in Granger-causality studies in the context of climate science. However, the proposed tests, which compare out-of-sample prediction errors, are available for models for which parameter estimation is done through ordinary least squares or maximum likelihood estimation (Attanasio et al., 2013), for linear parametric models (McCracken, 2007). As it has been discussed in Chapter 4, in climate, relations between variables are highly non-linear. Therefore, it would be convenient to have at our disposal a statistical test to assess the significance of any quantitative evidence of climate Granger-causing vegetation anomalies. Ideally, the test would be model-independent so that any non-linear model could be used. An alternative approach for comparing the predictive performance of different models is to use resampling methods, such as the bootstrap, or schemes such as 5×2 cross-validation (Dietterich, 1998). In this direction, a null distribution of the difference in the predictive performance of the two nested models is estimated by using (e.g.) random noise. However, in these approaches the effect of the additional variables (in the full model) might be overestimated. Random noise violates the time dependencies that occur in time series data. This issue can be resolved by using random blocked bootstrapping from the original data. Nonetheless, this solution comes together with other questions about the size of the bootstrapped blocks, since short blocks will destroy the temporal structure of the data and long blocks will not allow for bootstrap samples to be variable enough or for drawing a sufficient number of bootstrap samples. Also note that bootstrapping techniques are iterative processes. Therefore, the number of iterations should be selected carefully when the applications involve the use of large data sets (such as

in our case). In this thesis, we mainly focus on expressing Granger causality in a quantitative instead of a qualitative way, and stress the gained improvement with the use of a non-linear model. However, an exploration of methods that might use bootstrapping or permutation techniques is a potential avenue that completes the proposed non-linear Granger-causality framework from a statistical perspective.

As we have discussed in Chapter 1, Granger causality is a common predictability-based approach for cause-effect relationships between time series. One of its strongest assumptions is that all the possible causes are included in the model, so the whole available information is represented in the model as different time series variables. However, due to the complexity of the Earth system, this assumption is almost never fulfilled, and thus the conclusions drawn from Granger-causality analyses should be treated with caution. Therefore, incorporating as much information related to climate and/or vegetation as possible will make the proposed framework more reliable. For instance, anthropogenic factors, such as deforestation, fires and agriculture, are all possible causes of vegetation changes and micro-climate changes in an ecosystem. Conveniently, there are some publicly data sets which include this kind of information. Other useful data sets that one can think of are CO_2 emissions, irrigation, etc. Yet, some of these data sets are only available for more recent periods.

One of the main contributions of this thesis is the construction of a global data cube, which includes climate and vegetation data sets. However, the spatial resolution of this data cube is quite coarse (1°). So, with this spatial resolution, the interpretation of the results is not so easy, since different climatic (or vegetation) conditions may occur in the same 1° square. Therefore, although the 1° spatial resolution is a convenient spatial resolution that allows for a global analysis, an in-depth study of climatic (or vegetation) conditions at local scale is not possible. On top of this, the temporal resolution of the constructed data cube is monthly. For this reason, in our analysis we included the current value of the climatic variables as predictors for the forecast of the current vegetation. We found that variables such as radiation and temperature lose their predictive strength after the first month, so if we excluded the current month of the analysis, these variables would not have any effect on vegetation. Thus, the prolonged effect of water variables would be inflated. A possible solution in this issue is the down-scaling of the temporal resolution to bi-weekly or weekly data. That way, one can possibly exclude the current values of the predictor variables for the forecast of vegetation. However, this kind of solution enhances the autocorrelation between the consecutive time points, since vegetation does not change from one day to the other. Therefore, baseline and full models of Granger-causality frameworks should be designed very carefully in order not to lead to incorrect conclusions.

Another future direction of our study is related to the applicability of our data-driven approaches on data coming from our current Earth System Models (ESMs).

Each ESM is based on its own assumptions, so capturing complex interactions between vegetation and climate in this kind of data becomes a huge challenge. Yet, since ESMs do not model vegetation greenness expressed by NDVI, the methodology has been already repeated using LAI data (Demuzere et al., 2017) as target, which is a variable close to models' representation for vegetation. The results are in line with those presented in Chapter 4, providing confidence in both the methodology as well as the use of the LAI anomalies as a target variable. These observation-based results can then be used to benchmark ESMs on their representation of vegetation sensitivity to climate and climatic extremes. ESMs can be selected from the Coupled Model Intercomparison Project Phase 5 (CMIP5) based on their availability of daily output for all variables of interest. For example, ESMs such as the BCC-CSM1 (Wu et al., 2013), the GFDL-ESM2G (Dunne et al., 2012) and the MIROC-ESM (Watanabe et al., 2011), the BNUESM (`http://esg.bnu.edu.cn`), the CAN-ESM2 (Arora et al., 2011) and the INM-CM4 (Volodin et al., 2010) can be used in this analysis. A better understanding of the climate–vegetation interactions on this basis can contribute to a more confident view about our projections of future climate and the fate of global ecosystems.

In Chapter 8, we experimented with different time series classification methods that are able to automatically extract features from the raw time series. From this analysis, we concluded that there is a potential for these methods to be applied in climate applications. However, most of these methods have been developed for univariate time series. Moreover, they do not take into account spatial information, since they use the time series of one particular location. Recent advances in neural networks can be used for the automatically extraction of spatio-temporal features. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are commonly used for this purpose. In the context of the Granger-causality framework, a neural network architecture has also been studied. For more information about the different neural network approaches that were applied see (Mortier et al., 2017). However, the results of this effort were not that promising maybe due to the limited number of observations for each variable (there are only 360 timestamps/maps for each variable). It is known that these methods need large data sets in order to generalize well and achieve high predictive performance. Possible solutions for enhancing the data set include the use of advanced interpolation methods to downscale the temporal resolution of the data set or the use of data produced by climate models that span longer time periods. Note that even if the use of neural network approaches seems to fit well the current setting, the interpretation of the resulting features is not possible. Therefore, the use of these methods highly depends on the given research questions.

Finally, even if the proposed definition (discussed in Chapter 7) for the vegetation extreme events seems promising for future research, there are some aspects that can be improved. For instance, looking at the natural processes of extreme events in vegetation, one would expect that an extreme event occurs when there is an

abrupt decrease in the vegetation values at the time (or shortly after) the climate event takes place. And then, it takes some time for the vegetation to recover to its original condition. This is the reason why in the subsequent timestamps, there is usually still a small decrease present in the vegetation values due to some extended effects. Using the current definition, this abrupt change is not captured well. A possible solution is the use of the first-order differences of the vegetation anomalies time series instead of directly using the original data. Thus, by calculating the percentile on the time series of differences, one can receive the extreme events at the points where vegetation is characterized by a sharp decrease.

As a last direction, we should note that a further improvement of the causality framework is necessary in order to take into account the cascade problem caused by the memory of vegetation. The effects of climate are incorporated in vegetation in each month and hence, the assessment of the climate influence cannot be detected if information from vegetation is already included in a model. Therefore, one could resolve this problem by targeting (e.g.) the vegetation extreme events of two or more months ahead.

Bibliography

J. O. Adegoke and A. M. Carleton. Relations between soil moisture and satellite vegetation indices in the US Corn Belt. *J. Hydrometeorol.*, 3:395–405, 2002.

R. F. Adler, G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, et al. The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *J. Hydrometeorol.*, 4(6):1147–1167, 2003.

N. Ancona, D. Marinazzo, and S. Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(5):056221, 2004.

W. R. L. Anderegg, C. Schwalm, F. Biondi, and J. J. Camarero. Pervasive drought legacies in forest ecosystems and their implications for carbon cycle models. *Science*, 349:528–32, 2015.

L. O. Anderson, Y. Malhi, L. E. Aragão, R. Ladle, E. Arai, N. Barbier, and O. Phillips. Remote sensing detection of droughts in Amazonian forest canopies. *New Phytol.*, 187(3):733–750, 2010.

R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6 (Nov):1817–1853, 2005.

V. Arora, J. Scinocca, G. Boer, J. Christian, K. Denman, G. Flato, V. Kharin, W. Lee, and W. Merryfield. Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophys. Res. Lett.*, 38(5), 2011.

A. Attanasio. Testing for linear Granger causality from natural/anthropogenic forcings to global temperature anomalies. *Theor. Appl. Climatol.*, 110(1-2): 281–289, 2012.

A. Attanasio and U. Triacca. Detecting human influence on climate using neural networks based Granger causality. *Theoretical and Applied Climatology*, 103 (1-2):103–107, 2011.

A. Attanasio, A. Pasini, and U. Triacca. A contribution to attribution of recent global warming by out-of-sample Granger causality analysis. *Atmos. Sci. Lett.*, 13(1):67–72, 2012.

A. Attanasio, A. Pasini, and U. Triacca. Granger causality analyses for climatic attribution. *Atmospheric and Climate Sciences*, 2013, 2013.

A. Baccini et al. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change*, 2:182–5, 2012.

A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, May 2017. doi: 10.1007/s10618-016-0483-9.

B. Baker, H. Diaz, W. Hargrove, and F. Hoffman. Use of the Köppen–Trewartha climate classification to evaluate climatic refugia in statistically derived ecoregions for the People's Republic of China. *Climatic Change*, 98(1):113, Jul 2009. doi: 10.1007/s10584-009-9622-2.

J. Barichivich, K. R. Briffa, R. Myneni, G. van der Schrier, W. Dorigo, C. J. Tucker, T. J. Osborn, and T. M. Melvin. Temperature and Snow-Mediated Moisture Controls of Summer Photosynthetic Activity in Northern Terrestrial Ecosystems between 1982 and 2011. *Remote Sensing*, 6(2):1390, 2014. doi: 10.3390/rs6021390.

E. Bartholomé and A. S. Belward. GLC2000: a new approach to global land cover mapping from Earth observation data. *Int. J. Remote Sens.*, 26(9):1959–1977, 2005.

A. Barzilai and K. Crammer. Convex multi-task learning by clustering. In *Artificial Intelligence and Statistics*, pages 65–73, 2015.

G. E. A. P. A. Batista, E. J. Keogh, O. M. Tataw, and V. M. De Souza. CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3):634–669, 2014.

L. Baumbach, J. F. Siegmund, M. Mittermeier, and R. V. Donner. Impacts of temperature extremes on European vegetation during the growing season. *Biogeosciences*, 14(21):4891, 2017.

J. Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

J. Baxter et al. A model of inductive bias learning. *J. Artif. Intell. Res. (JAIR)*, 12(149-198):3, 2000.

M. G. Baydogan and G. Runger. Time series representation and similarity based on local autopatterns. *Data Mining and Knowledge Discovery*, 30(2):476–509, 2016.

M. G. Baydogan, G. Runger, and E. Tuv. A bag-of-features framework to classify time series. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2796–2802, 2013.

H. E. Beck, T. R. McVicar, A. I. J. M. van Dijk, J. Schellekens, R. A. M. de Jeu, and L. A. Bruijnzeel. Global evaluation of four AVHRR–NDVI data sets:

intercomparison and assessment against Landsat imagery. *Remote Sens. Environ.*, 115:2547–63, 2011.

H. E. Beck, A. I. J. M. van Dijk, V. Levizzani, J. Schellekens, D. G. Miralles, B. Martens, and A. de Roo. MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.*, 21(1):589–615, 2017. doi: 10.5194/hesNDVIs-21-589-2017.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

J. Bi, T. Xiong, S. Yu, M. Dundar, and R. B. Rao. An improved multi-task learning approach with applications in medical diagnosis. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 117–132. Springer, 2008.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, first edition edition, 2007.

G. B. Bonan. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science*, 320(5882):1444–1449, 2008.

B. Braswell, D. Schimel, E. Linder, and B. Moore. The response of global terrestrial ecosystems to interannual temperature variability. *Science*, 278(5339):870–873, 1997.

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

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman and Hall, Wadsworth, New York, 1984.

K. Brugger and F. Rubel. Characterizing the species composition of European Culicoides vectors by means of the Köppen-Geiger climate classification. *Parasites & vectors*, 6(1):333, 2013.

C. Brunsdon, A. S. Fotheringham, and M. E. Charlton. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996. doi: 10.1111/j.1538-4632.1996.tb00936.x.

R. Cano, C. Sordo, and J. M. Gutiérrez. *Applications of Bayesian Networks in Meteorology*, pages 309–328. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-39879-0. doi: 10.1007/978-3-540-39879-0_17.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, Jul 1997. doi: 10.1023/A:1007379606734.

D. Chan and Q. Wu. Significant anthropogenic-induced changes of climate classes since 1950. *Sci. Rep.*, 5(4):13487, 2015. doi: 10.1038/srep13487.

D. Chapman, M. A. Cane, N. Henderson, D. E. Lee, and C. Chen. A vector autoregressive ENSO prediction model. *J. Climate*, 28(21):8511–8520, 2015.

S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly. Sparse Group Lasso: Consistency and Climate Applications. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 47–58, 2012. doi: 10.1137/1.9781611972825.5.

D. Chen and H. W. Chen. Using the Köppen classification to quantify climate variation and change: an example for 1901–2010. *Environmental Development*, 6:69–79, 2013.

J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 137–144. ACM, 2009.

S. Chen and Y. Tian. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.*, 53(4):1947–1957, 2015.

T. Chen, R. De Jeu, Y. Liu, G. Van der Werf, and A. Dolman. Using satellite based soil moisture to quantify the water driven variability in NDVI: A case study over mainland Australia. *Remote Sens. Environ.*, 140:330–338, 2014.

F.-M. Chmielewski and T. Rötzer. Annual and spatial variability of the beginning of growing season in Europe in relation to air temperature changes. *Clim. Res.*, 19:257–64, 2002.

P. Ciais et al. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437:529–33, 2005.

T. E. Clark and M. W. McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1):85–110, 2001.

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

G. Coccia, A. L. Siemann, M. Pan, and E. F. Wood. Creating consistent datasets by combining remotely-sensed data and land surface model estimates through Bayesian uncertainty post-processing: The case of Land Surface Temperature from {HIRS}. *Remote Sens. Environ.*, 170:290 – 305, 2015. doi: http://dx.doi.org/10.1016/j.rse.2015.09.010.

L. E. S. Cole, S. A. Bhagwat, and K. J. Willis. Recovery and resilience of tropical forests after disturbance. *Nat. Commun.*, 5:1–7, 2014.

R. G. Congalton, J. Gu, K. Yadav, P. Thenkabail, and M. Ozdogan. Global land cover mapping: A review and uncertainty analysis. *Remote Sens.*, 6(12): 12070–12093, 2014.

D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

M. B. Davis. *Climatic instability, time, lags, and community disequilibrium.* Harper & Row, 1984.

W. De Keersmaecker, S. Lhermitte, L. Tits, O. Honnay, B. Somers, and P. Coppin. A model quantifying global vegetation resistance and resilience to short-term climate anomalies and their relationship with vegetation cover. *Glob. Ecol. Biogeogr.*, 24:539–48, 2015.

S. Decubber. Spatiotemporal optimization of granger causality methods for climate change attribution. *Master thesis, Ghent University*, 2017.

D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, 137(656):553–597, 2011.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30, 2006.

M. Demuzere, S. Decubber, D. Miralles, C. Papagiannopoulou, W. Waegeman, N. Verhoest, and W. Dorigo. Sensitivity of global ecosystesms to climate anomalies in observations and Earth System Models. In *Proceedings of the 7th International Workshop on Climate Informatics*, pages 21–24. NCAR Technical Note NCAR/TN-536+PROC, 2017.

H. Deng, G. Runger, E. Tuv, and M. Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013.

D. Deryng, D. Conway, N. Ramankutty, J. Price, and R. Warren. Global crop yield response to extreme heat stress under multiple climate change futures. *Environ. Res. Lett.*, 9(3):034011, 2014.

H. F. Diaz and J. K. Eischeid. Disappearing alpine tundra Köppen climatic type in the western United States. *Geophys. Res. Lett.*, 34(18), 2007.

F. X. Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1, 2015a.

F. X. Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of dieboldmariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1, 2015b. doi: 10.1080/07350015.2014.983236.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, 1998.

C. Diks and M. Mudelsee. Redundancies in the Earth's climatological time series. *Phys. Lett. A*, 275(5):407–414, 2000.

C. Dimiceli, M. Carroll, R. Sohlberg, D. Kim, M. Kelly, and J. Townshend. MOD44BMODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250m SIN Grid V006 [Data set]. *NASA EOSDIS Land Processes DAAC.*, 2015. doi: 10.5067/MODIS/MOD44B.006.

H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

M. Donat, L. Alexander, H. Yang, I. Durre, R. Vose, R. Dunn, K. Willett, E. Aguilar, M. Brunet, J. Caesar, et al. Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. *Journal of Geophysical Research: Atmospheres*, 118(5):2098–2118, 2013.

J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, 2009.

J. F. Donges, C.-F. Schleussner, J. F. Siegmund, and R. V. Donner. Event coincidence analysis for quantifying statistical interrelationships between event time series. *The European Physical Journal Special Topics*, 225(3):471–487, 2016.

W. Dorigo, A. Lucieer, T. Podobnikar, and A. Čarni. Mapping invasive fallopia japonica by combined spectral, spatial, and temporal analysis of digital orthophotos. *Int. J. Appl. Earth Obs. Geoinf.*, 19:185–195, 2012.

W. Dorigo, W. Wagner, C. Albergel, F. Albrecht, G. Balsamo, L. Brocca, D. Chung, M. Ertl, M. Forkel, A. Gruber, E. Haas, P. D. Hamer, M. Hirschi, J. Ikonen, R. de Jeu, R. Kidd, W. Lahoz, Y. Y. Liu, D. Miralles, T. Mistelbauer, N. Nicolai-Shaw, R. Parinussa, C. Pratola, C. Reimer, R. van der Schalie, S. I. Seneviratne, T. Smolander, and P. Lecomte. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sens. Environ.*, 203:185 – 215, 2017. doi: https://doi.org/10.1016/j.rse.2017.07.001. Earth Observation of Essential Climate Variables.

J. P. Dunne, J. G. John, A. J. Adcroft, S. M. Griffies, R. W. Hallberg, E. Shevliakova, R. J. Stouffer, W. Cooke, K. A. Dunne, M. J. Harrison, et al. GFDLs ESM2 global coupled climate–carbon earth system models. Part I: Physical formulation and baseline simulation characteristics. *J. Climate*, 25(19):6646–6665, 2012.

I. Ebert-Uphoff and Y. Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

J. B. Elsner. Evidence in support of the climate change–Atlantic hurricane hypothesis. *Geophys. Res. Lett.*, 33(16), 2006.

J. B. Elsner. Granger causality and Atlantic hurricanes. *Tellus A*, 59(4):476–485, 2007.

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

J. J. Feddema. A Revised Thornthwaite-Type Global Climate Classification. *Physical Geography*, 26(6):442–466, 2005. doi: 10.2747/0272-3646.26.6.442.

J. J. Feddema, K. W. Oleson, G. B. Bonan, L. O. Mearns, L. E. Buja, G. A. Meehl, and W. M. Washington. Atmospheric science: The importance of land-cover change in simulating future climates. *Science*, 310(5754):1674–1678, 2005. doi: 10.1126/science.1118160.

E. M. Fischer, J. Sedlá cek, E. Hawkins, and R. Knutti. Models agree on forced response pattern of precipitation and temperature extremes. *Geophys. Res. Lett.*, 41:8554–62, 2014.

J. B. Fisher, G. Badgley, and E. Blyth. Global nutrient limitation in terrestrial vegetation. *Global Biogeochem. Cycles*, 26(3), 2012a.

J. B. Fisher, G. Badgley, and E. Blyth. Global nutrient limitation in terrestrial vegetation. *Glob. Biogeochem. Cycles*, 26:GB3007, 2012b.

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

M. Forkel, K. Thonicke, C. Beer, W. Cramer, S. Bartalev, and C. Schmullius. Extreme fire events are related to previous-year surface moisture conditions in permafrost-underlain larch forests of Siberia. *Environ. Res. Lett.*, 7(4):044021, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

C. Gallardo, V. Gil, E. Hagel, C. Tejeda, and M. de Castro. Assessment of climate change in Europe from an ensemble of regional climate models by the use of Köppen–Trewartha classification. *Int. J. Climatol.*, 33(9):2157–2166, 2013.

J. C. B. Gamboa. Deep learning for time-series analysis. *CoRR*, abs/1701.01887, 2017.

R. A. Garcia, M. Cabeza, C. Rahbek, and M. B. Araújo. Multiple dimensions of climate change and their implications for biodiversity. *Science*, 344(6183): 1247579, 2014.

P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components. In

Proceedings of 32th International Conference on Machine Learning (ICML 2015), 2015.

S. Gelper and C. Croux. Multivariate out-of-sample tests for Granger causality. *Computational statistics & data analysis*, 51(7):3319–3329, 2007.

S. Georganos, A. M. Abdi, D. E. Tenenbaum, and S. Kalogirou. Examining the ndvi-rainfall relationship in the semi-arid sahel using geographically weighted regression. *Journal of Arid Environments*, 146:64 – 74, 2017. ISSN 0140-1963. doi: https://doi.org/10.1016/j.jaridenv.2017.06.004.

A. R. Gonçalves, A. Banerjee, and F. J. Von Zuben. Spatial Projection of Multiple Climate Variables Using Hierarchical Multitask Learning. In *AAAI Conference on Artificial Intelligence*, pages 4509–4515, 2017.

A. Gonsamo, J. M. Chen, and D. Lombardozzi. Global vegetation productivity response to climatic oscillations during the satellite era. *Glob. Change Biol.*, 22: 3414–26, 2016.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

J. G. D. Gooijer and R. J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443 – 473, 2006. doi: https://doi.org/10.1016/j.ijforecast.2006.01.001.

T. Górecki and M. Łuczak. Using derivatives in time series classification. *Data Mining and Knowledge Discovery*, 26(2):310–331, Mar 2013. doi: 10.1007/s10618-012-0251-4.

T. Górecki and M. Łuczak. Non-isometric transforms in time series classification using DTW. *Knowledge-Based Systems*, 61:98–108, 2014.

J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–401. ACM, 2014.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

J. K. Green, A. G. Konings, S. H. Alemohammad, J. Berry, D. Entekhabi, J. Kolassa, J.-E. Lee, and P. Gentine. Regionally strong feedbacks between the atmosphere and terrestrial biosphere. *Nat. Geosci.*, 10(6):410, 2017.

M. Gregorova, A. Kalousis, S. Marchand-Maillet, and J. Wang. Learning vector autoregressive models with focalised Granger-causality graphs. *arXiv preprint arXiv:1507.01978*, 2015.

J. Grieser, S. Trömel, and C.-D. Schönwiese. Statistical time series decomposition

into significant components and application to European temperature. *Theor. Appl. Climatol.*, 71(3-4):171–183, 2002.

K. Guan et al. Photosynthetic seasonality of global tropical forests constrained by hydroclimate. *Nat. Geosci.*, 8:284–9, 2015.

Y. Guanche García, M. Shadaydeh, M. Mahecha, and J. Denzler. Extreme anomaly event detection in biosphere using linear regression and a spatiotemporal mrf model. *Natural Hazards*, 2018. doi: 10.1007/s11069-018-3415-8.

R. S. Hacker and A. Hatemi-J. Tests for causality between integrated variables using asymptotic and bootstrap distributions: theory and application. *Applied Economics*, 38(13):1489–1500, 2006. doi: 10.1080/00036840500405763.

F. Hanf, J. Körper, T. Spangehl, and U. Cubasch. Shifts of climate zones in multi-model climate change experiments using the Köppen climate classification. *Meteorol. Z.*, 21(2):111–123, 2012.

J. Hansen, R. Ruedy, M. Sato, and K. Lo. Global surface temperature change. *Rev. Geophys.*, 48(4), 2010.

M. C. Hansen et al. High-resolution global maps of 21st century forest cover change. *Science*, 342:850–3, 2013.

F. Hao, X. Zhang, W. Ouyang, A. K. Skidmore, and A. G. Toxopeus. Vegetation NDVI Linked to Temperature and Precipitation in the Upper Catchments of Yellow River. *Environmental Modeling & Assessment*, 17(4):389–398, Aug 2012. doi: 10.1007/s10666-011-9297-8.

I. Harris, P. Jones, T. Osborn, and D. Lister. Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *Int. J. Climatol.*, 34 (3):623–642, 2014.

M. Hasanuzzaman, K. Nahar, M. M. Alam, R. Roychowdhury, and M. Fujita. Physiological, biochemical, and molecular mechanisms of heat stress tolerance in plants. *Int. J. Mol. Sci.*, 14(5):9643–9684, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

S. Herrando-Pérez, S. Delean, B. W. Brook, P. Cassey, and C. J. Bradshaw. Spatial climate patterns explain negligible variation in strength of compensatory density feedbacks in birds and mammals. *PLoS One*, 9(3):e91536, 2014.

T. Hilker, A. I. Lyapustin, C. J. Tucker, F. G. Hall, R. B. Myneni, Y. Wang, J. Bi, Y. M. de Moura, and P. J. Sellers. Vegetation dynamics and rainfall sensitivity of the Amazon. *Proceedings of the National Academy of Sciences*, 111 (45):16041–16046, 2014.

J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4): 851–881, 2014.

S. Horion, A. V. Prishchepov, J. Verbesselt, K. Beurs, T. Tagesson, and R. Fensholt. Revealing turning points in ecosystem functioning over the northern eurasian agricultural frontier. *Global Change Biology*, 22(8):2801–2817, 2016. doi: 10.1111/gcb.13267.

D. Hughes. Comparison of satellite rainfall data with observations from gauging station networks. *J. Hydrol.*, 327(3):399–410, 2006.

L. R. Hutyra, J. W. Munger, S. R. Saleska, E. Gottlieb, B. C. Daube, A. L. Dunn, D. F. Amaral, P. B. De Camargo, and S. C. Wofsy. Seasonal controls on the exchange of carbon and water in an Amazonian rain forest. *Journal of Geophysical Research: Biogeosciences*, 112(G3), 2007.

IPCC, Intergovernmental Panel on Climate Change. Report of the 19th session of the Intergovernmental Panel On Climate Change (IPCC). 2007.

E. Ivits, S. Horion, R. Fensholt, and M. Cherlet. Global ecosystem response types derived from the standardized precipitation evapotranspiration index and fpar3g series. *Remote Sensing*, 6(5):4266–4288, 2014. doi: 10.3390/rs6054266.

L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *Proceedings of Advances in neural information processing systems*, pages 745–752, 2009.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

A. Karnieli, N. Agam, R. T. Pinker, M. Anderson, M. L. Imhoff, G. G. Gutman, N. Panov, and A. Goldberg. Use of NDVI and land surface temperature for drought assessment: Merits and limitations. *J. Climate*, 23(3):618–633, 2010.

A. Karpatne, A. Khandelwal, S. Boriah, and V. Kumar. Predictive learning in the presence of heterogeneity and limited training data. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 253–261. SIAM, 2014.

A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar. Machine Learning for the Geosciences: Challenges and Opportunities. *arXiv preprint arXiv:1711.04708*, 2017.

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

Y. Kim, D. M. Glenn, J. Park, H. K. Ngugi, and B. L. Lehman. Hyperspectral

image analysis for plant stress detection. In *2010 Pittsburgh, Pennsylvania, June 20-June 23, 2010*, page 1. American Society of Agricultural and Biological Engineers, 2010.

E. Kodra, S. Chatterjee, and A. R. Ganguly. Exploring Granger causality between global average observed time series of carbon dioxide and temperature. *Theor. Appl. Climatol.*, 104(3-4):325–335, 2011.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193, 9780262013192.

W. Köppen. Das Geographische System der Klimate. *Handbuch der klimatologie*, 1, 1936.

M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel. World map of the Köppen-Geiger climate classification updated. *Meteorol. Z.*, 15(3):259–263, 2006.

W. A. Kurz, C. Dymond, G. Stinson, G. Rampley, E. Neilson, A. Carroll, T. Ebata, and L. Safranyik. Mountain pine beetle and forest carbon feedback to climate change. *Nature*, 452(7190):987, 2008.

Y. Kuzyakov and O. Gavrichkova. REVIEW: Time lag between photosynthesis and carbon dioxide efflux from soil: a review of mechanisms and controls. *Global Change Biol.*, 16(12):3386–3406, 2010.

W. Larcher et al. *Ökophysiologie der pflanzen*. Eugen Ulmer Stuttgart, 1994.

D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3 – 10, 2016. doi: https://doi.org/10.1016/j.gsf.2015.07.003. Special Issue: Progress of Machine Learning in Geosciences.

C. Le Quere et al. Global carbon budget. *Earth Syst. Sci. Data*, 8:605–49, 2016.

D. P. Lettenmaier, D. Alsdorf, J. Dozier, G. J. Huffman, M. Pan, and E. F. Wood. Inroads of remote sensing into hydrologic science during the WRR era. *Water Resour. Res.*, 51(9):7309–7342, 2015. doi: 10.1002/2015WR017616.

W. Li, N. MacBean, P. Ciais, P. Defourny, C. Lamarche, S. Bontemps, R. A. Houghton, and S. Peng. Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015). *Earth Syst. Sci. Data*, 10(1):219–234, 2018. doi: 10.5194/essd-10-219-2018.

T. W. Liao. Clustering of time series dataa survey. *Pattern Recognit.*, 38(11):1857 – 1874, 2005. doi: http://dx.doi.org/10.1016/j.patcog.2005.01.025.

J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2): 107–144, 2007.

J. Lin, R. Khade, and Y. Li. Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, 39 (2):287–315, 2012.

G. Liu, H. Liu, and Y. Yin. Global patterns of NDVI-indicated vegetation extremes and their sensitivity to climate extremes. *Environ. Res. Lett.*, 8(2):025009, 2013.

L. Liu, Y. Zhang, S. Wu, S. Li, and D. Qin. Water memory effects and their impacts on global vegetation productivity and resilience. *Scientific reports*, 8(1): 2962, 2018.

Y. Liu, R. Parinussa, W. Dorigo, R. De Jeu, W. Wagner, A. Van Dijk, M. McCabe, and J. Evans. Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals. *Hydrol. Earth Syst. Sci.*, 15(2): 425–436, 2011a.

Y. Liu, W. Dorigo, R. Parinussa, R. De Jeu, W. Wagner, M. McCabe, J. Evans, and A. Van Dijk. Trend-preserving blending of passive and active microwave soil moisture retrievals. *Remote Sens. Environ.*, 123:280–297, 2012.

Y. Liu, Z. Pan, Q. Zhuang, D. G. Miralles, A. J. Teuling, T. Zhang, P. An, Z. Dong, J. Zhang, D. He, et al. Agriculture intensifies soil moisture decline in Northern China. *Sci. Rep.*, 5, 2015a.

Y. Y. Liu, R. A. M. de Jeu, M. F. McCabe, J. P. Evans, and A. I. J. M. van Dijk. Global long-term passive microwave satellite-based retrievals of vegetation optical depth. *Geophysical Research Letters*, 38(18), 2011b. doi: 10.1029/2011GL048684.

Y. Y. Liu, A. I. J. M. van Dijk, R. A. M. de Jeu, J. G. Canadell, M. F. McCabe, J. P. Evans, and G. Wang. Recent reversal in loss of global terrestrial biomass. *Nat. Clim. Change*, 5:470–4, 2015b.

L. Loosvelt, J. Peters, H. Skriver, B. De Baets, and N. E. Verhoest. Impact of reducing polarimetric SAR input on the uncertainty of crop classifications based on the random forests algorithm. *IEEE Trans. Geosci. Remote Sens.*, 50(10): 4185–4200, 2012a.

L. Loosvelt, J. Peters, H. Skriver, H. Lievens, F. M. Van Coillie, B. De Baets, and N. E. C. Verhoest. Random Forests as a tool for estimating uncertainty at pixel-level in SAR image classification. *Int. J. Appl. Earth Obs. Geoinf.*, 19: 173–184, 2012b.

T. Loveland and A. Belward. The IGBP-DIS global 1km land cover data set, DISCover: first results. *Int. J. Remote Sens.*, 18(15):3289–3295, 1997.

T. R. Loveland, B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang, and J. W. Merchant. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *Int. J. Remote Sens.*, 21(6-7): 1303–1330, 2000.

A. C. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118, 2009a.

A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596. ACM, 2009b.

R. Lund and B. Li. Revisiting climate region definitions via clustering. *J. Climate*, 22(7):1787–1800, 2009.

K. Luojus, J. Pulliainen, M. Takala, C. Derksen, H. Rott, T. Nagler, R. Solberg, A. Wiesmann, S. Metsamaki, E. Malnes, et al. Investigating the feasibility of the GlobSnow snow water equivalent data for climate research purposes. In *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*. IEEE, 2010.

L. Ma, M. Li, Y. Gao, T. Chen, X. Ma, and L. Qu. A Novel Wrapper Approach for Feature Selection in Object-Based Image Classification Using Polygon-Based Cross-Validation. *IEEE Geoscience and Remote Sensing Letters*, 14(3):409–413, March 2017. doi: 10.1109/LGRS.2016.2645710.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

G. S. Maddala and K. Lahiri. *Introduction to econometrics*, volume 2. Macmillan New York, 1992.

M. D. Mahecha, F. Gans, S. Sippel, J. F. Donges, T. Kaminski, S. Metzger, M. Migliavacca, D. Papale, A. Rammig, and J. Zscheischler. Detecting impacts of extreme events with ecological in situ monitoring networks. *Biogeosciences*, 14(18):4255–4277, 2017. doi: 10.5194/bg-14-4255-2017.

I. Mahlstein, J. S. Daniel, and S. Solomon. Pace of shifts in climate regions increases with global temperature. *Nature Climate Change*, 3(8):739, 2013.

D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel method for nonlinear Granger causality. *Phys. Rev. Lett.*, 100(14):144103, 2008.

B. Martens, D. G. Miralles, H. Lievens, R. van der Schalie, R. A. M. de Jeu, D. Férnandez-Prieto, H. E. Beck, W. A. Dorigo, and N. E. C. Verhoest. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev. Discuss.*, 2016:1–36, 2016. doi: 10.5194/gmd-2016-162.

M. F. McCabe, M. Rodell, D. E. Alsdorf, D. G. Miralles, R. Uijlenhoet, W. Wagner, A. Lucieer, R. Houborg, N. E. C. Verhoest, T. E. Franz, J. Shi, H. Gao, and E. F. Wood. The Future of Earth Observation in Hydrology. *Hydrol. Earth Syst. Sci. Discuss.*, 2017:1–55, 2017. doi: 10.5194/hess-2017-54.

M. W. McCracken. Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, 140(2):719–752, 2007.

R. A. Mcpherson. A review of vegetation–atmosphere interactions and their influences on mesoscale phenomena. *Prog. Phys. Geogr.*, 31:261–85, 2007.

R. A. McPherson, C. A. Fiebrich, K. C. Crawford, R. L. Elliott, J. R. Kilby, D. L. Grimsley, J. E. Martinez, J. B. Basara, B. G. Illston, D. A. Morris, K. A. Kloesel, S. J. Stadler, A. D. Melvin, A. J. Sutherland, H. Shrivastava, J. D. Carlson, J. M. Wolfinbarger, J. P. Bostic, and D. B. Demko. Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, 24(3):301–321, 2007. doi: 10.1175/JTECH1976.1.

S. McQuade and C. Monteleoni. Global Climate Model Tracking Using Geospatial Neighborhoods. In *Proceedings of the 2012 AAAI*, 2012.

S. McQuade and C. Monteleoni. MRF-Based Spatial Expert Tracking of the Multi-Model Ensemble. In *International Workshop on Climate Informatics*, 2013.

M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan. Land-use classification with compressive sensing multifeature fusion. *IEEE Geosci. Remote Sens. Lett.*, 12 (10):2155–2159, 2015.

A. Menzel and P. Fabian. Growing season extended in Europe. *Nature*, 397:659, 1999.

M. J. Metzger, R. G. H. Bunce, R. H. G. Jongman, R. Sayre, A. Trabucco, and R. Zomer. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, 22(5):630–638, 2012. doi: 10.1111/geb.12022.

J. Michaelsen. Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.*, 26(11):1589–1600, 1987.

D. Miralles, T. Holmes, R. De Jeu, J. Gash, A. Meesters, and A. Dolman. Global land-surface evaporation estimated from satellite-based observations. *Hydrol. Earth Syst. Sci.*, 15(2):453–469, 2011.

D. G. Miralles, R. Nieto, N. G. McDowell, W. A. Dorigo, N. E. Verhoest, Y. Y. Liu, A. J. Teuling, A. J. Dolman, and G. L. Contribution of water-limited ecoregions to their own supply of rainfall. *Environ. Res. Lett.*, 11:124007, 2016.

T. M. Mitchell. *Machine Learning.* WCB McGraw-Hill, 1997.

V. Mithal, A. Garg, S. Boriah, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and J. C. Castilla-Rubio. Monitoring global forest cover using data mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(4):36, 2011.

I. I. Mokhov, D. A. Smirnov, P. I. Nakonechny, S. S. Kozlenko, E. P. Seleznev, and J. Kurths. Alternating mutual influence of El-Niño/Southern Oscillation and Indian monsoon. *Geophys. Res. Lett.*, 38(8), 2011.

C. Monteleoni, G. A. Schmidt, S. Saroha, and E. Asplund. Tracking climate models. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(4): 372–392, 2011.

P. A. P. Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2): 17–23, 1950.

T. Mortier, S. Decubber, O. Thas, and W. Waegeman. Modeling of climate-vegetation dynamics using machine learning techniques in a non-linear Granger causality framework. *Master of Science in Statistical Data Analysis, Ghent University*, 2017.

D. C. Morton, J. Nagol, C. C. Carabajal, J. Rosette, M. Palace, B. D. Cook, E. F. Vermote, D. J. Harding, and P. R. J. North. Amazon forests maintain consistent canopy structure and greenness during the dry season. *Nature*, 506:221–4, 2014.

T. J. Mosedale, D. B. Stephenson, M. Collins, and T. C. Mills. Granger causality of coupled climate processes: Ocean feedback on the North Atlantic Oscillation. *J. Climate*, 19(7):1182–1194, 2006.

R. Moss, W. Babiker, S. Brinkman, E. Calvo, T. Carter, J. Edmonds, I. Elgizouli, S. Emori, L. Erda, K. Hibbard, et al. Towards New Scenarios for the Analysis of Emissions: Climate Change, Impacts and Response Strategies, 2008.

R. B. Myneni, C. D. Keeling, C. J. Tucker, G. Asrar, and R. R. Nemani. Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature*, 386: 698–702, 1997.

National Research Council (U.S.). Committee on Earth Studies. *Atmospheric Soundings. Issues in the Integration of Research and Operational Satellite Systems for Climate Research: Part I. Science and Design.* National Academy Press, 2000.

R. R. Nemani, C. D. Keeling, H. Hashimoto, W. M. Jolly, S. C. Piper, C. J. Tucker, R. B. Myneni, and S. W. Running. Climate-Driven Increases in Global Terrestrial Net Primary Production from 1982 to 1999. *Science (New York, N.Y.)*, 300 (5625):1560–3, 2003. doi: 10.1126/science.1082750.

P. Netzel and T. Stepinski. On using a clustering approach for global climate classification. *J. Climate*, 29(9):3387–3401, 2016.

P. Netzel and T. F. Stepinski. World Climate Search and Classification Using a Dynamic Time Warping Similarity Function. In *Advances in Geocomputation*, pages 181–195. Springer, 2017.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of Advances in neural information processing systems*, pages 849–856, 2002.

N. Nicholls and L. Alexander. Has the climate become more variable or extreme? Progress 1992-2006. *Prog. Phys. Geog.*, 31(1):77–87, 2007.

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

Y. Pan et al. A large and persistent carbon sink in the worlds forests. *Science*, 333: 988–93, 2011.

C. Papagiannopoulou, D. G. Miralles, S. Decubber, M. Demuzere, N. E. C. Verhoest, W. A. Dorigo, and W. Waegeman. A non-linear granger-causality framework to investigate climate–vegetation dynamics. *Geosci. Model Dev.*, 10(5):1945–1960, 2017a. doi: 10.5194/gmd-10-1945-2017.

C. Papagiannopoulou, D. G. Miralles, W. A. Dorigo, N. E. C. Verhoest, M. Depoorter, and W. Waegeman. Vegetation anomalies caused by antecedent precipitation in most of the world. *Environ. Res. Lett.*, 12(7):074016, 2017b.

A. Pasini, U. Triacca, and A. Attanasio. Evidence of recent causal decoupling between solar radiation and global temperature. *Environ. Res. Lett.*, 7(3):034020, 2012.

A. Pasini, P. Racca, S. Amendola, G. Cartocci, and C. Cassardo. Attribution of recent temperature behaviour reassessed by a neural-network method. *Scientific reports*, 7(1):17681, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

M. C. Peel, B. L. Finlayson, and T. A. McMahon. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci. Discuss.*, 4(2): 439–473, 2007.

B. Poulter, N. MacBean, A. Hartley, I. Khlystova, O. Arino, R. Betts, S. Bontemps, M. Boettcher, C. Brockmann, P. Defourny, S. Hagemann, M. Herold, G. Kirches, C. Lamarche, D. Lederer, C. Ottlé, M. Peters, and P. Peylin. Plant functional type classification for earth system models: Results from the European Space Agency's Land Cover Climate Change Initiative. *Geosci. Model Dev.*, 8(7): 2315–2328, 2015. doi: 10.5194/gmd-8-2315-2015.

B. Poulter et al. Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature*, 509:600–4, 2014.

P. Propastin, M. Kappas, and S. Erasmi. Application of geographically weighted regression to investigate the impact of scale on prediction uncertainty by modelling relationship between vegetation and climate. *International Journal of Spatial Data Infrastructures Research*, 3:73–94, 2008.

T. Rakthanmanon and E. Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 668–676. SIAM, 2013.

A. Rammig, M. Wiedermann, J. Donges, F. Babst, W. von Bloh, D. Frank, K. Thonicke, and M. Mahecha. Tree-ring responses to extreme climate events as benchmarks for terrestrial dynamic vegetation models. *Biogeosci. Discuss.*, 11 (2):2537–2568, 2014.

A. Rammig, J. Donges, F. Babst, W. von Bloh, D. Frank, K. Thonicke, M. Mahecha, et al. Coincidences of climate extremes and anomalous vegetation responses: comparing tree ring patterns to simulated productivity. *Biogeosciences*, 12(2): 373, 2015.

M. Reichstein et al. Climate extremes and the carbon cycle. *Nature*, 500:287–95, 2013.

V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.*, 67:93–104, 2012.

W. Rossow and E. Duenas. The international satellite cloud climatology project (ISCCP) web site: An online resource for research. *Bull. Amer. Meteor. Soc.*, 85 (2):167–172, 2004.

J. Runge, V. Petoukhov, and J. Kurths. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, 27(2):720–739, 2014.

J. Runge, D. Sejdinovic, and S. Flaxman. Detecting causal associations in large nonlinear time series datasets. *arXiv preprint arXiv:1702.07007*, 2017.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 26(1):43–49, 1978.

S. R. Saleska, J. Wu, K. Guan, A. C. Araujo, A. Huete, A. D. Nobre, and N. Restrepo-Coupe. Dry-season greening of Amazon forests. *Nature*, 531:E45, 2016.

P. Schäfer. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015. doi: 10.1007/s10618-014-0377-7.

U. Schneider, T. Fuchs, A. Meyer-Christoffer, and B. Rudolf. Global precipitation analysis products of the GPCC. *Global Precipitation Climatology Centre (GPCC), DWD, Internet Publikation*, 112, 2008.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis. Training deep convolutional neural networks for land–cover classification of high-resolution imagery. *IEEE Geosci. Remote Sens. Lett.*, 14(4):549–553, 2017.

A. W. Seddon, M. Macias-Fauria, P. R. Long, D. Benz, and K. J. Willis. Sensitivity of global terrestrial ecosystems to climate variability. *Nature*, 531(7593):229–232, 2016.

M. A. Semenov and P. R. Shewry. Modelling predicts that heat stress, not drought, will increase vulnerability of wheat in Europe. *Sci. Rep.*, 1:66, 2011.

S. I. Seneviratne et al. Soil moisture memory in AGCM simulations: analysis of global land-atmosphere coupling experiment (GLACE) data. *J. Hydrol.*, 7: 1090–112, 2006.

S. I. Seneviratne et al. Changes in climate extremes and their impacts on the natural physical environment Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Technical report, Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC) (Cambridge: Cambridge University Press), 2012. 109230.

P. Senin and S. Malinchik. SAX-VSM: Interpretable time series classification using sax and vector space model. In *Proceedings of the 13th International Conference on Data Mining*, pages 1175–1180. IEEE, 2013.

M. A. Shahin, M. A. Ali, and A. S. Ali. Vector Autoregression (VAR) Modeling and Forecasting of Temperature, Humidity, and Cloud Coverage. In *Computational Intelligence Techniques in Earth and Environmental Sciences*, pages 29–51. Springer, 2014.

A. Shojaie and G. Michailidis. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. 2000. ISBN 978-3-319-52451-1.

F. Siegert, G. Ruecker, A. Hinrichs, and A. Hoffmann. Increased damage from fires in logged forests during droughts caused by El Niño. *Nature*, 414(6862):437, 2001.

S. Sippel, J. Zscheischler, and M. Reichstein. Ecosystem impacts of climate extremes crucially depend on the timing. *Proc. Natl Acad. Sci*, 113:5768–70, 2016.

T. M. Smith, R. W. Reynolds, T. C. Peterson, and J. Lawrimore. Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *J. Climate*, 21(10):2283–2296, 2008.

J. Spinoni, J. Vogt, G. Naumann, H. Carrao, and P. Barbosa. Towards identifying areas at climatological risk of desertification using the Köppen–Geiger classification and FAO aridity index. *Int. J. Climatol.*, 35(9):2210–2222, 2015.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

A. N. Srivastava, R. Nemani, and K. Steinhaeuser. *Large-Scale Machine Learning in the Earth Sciences*. Chapman and Hall/CRC, 2017. ISBN 1498703879, 978-1498703871.

W. Stackhouse, Jr. Paul, K. Gupta, Shashi, J. Cox, Stephen, C. Mikovitz, T. Zhang, and M. Chiacchio. 12-Year Surface Radiation Budget Dataset. *GEWEX News*, 14:10–12, 2004.

A. Stefan, V. Athitsos, and G. Das. The move-split-merge metric for time series. *IEEE Trans. Knowl. Data Eng.*, 25(6):1425–1438, 2013.

M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 446–455. ACM, 2003.

J. H. Stock and M. W. Watson. Vector autoregressions. *The Journal of Economic Perspectives*, 15(4):101–115, 2001.

R. Stöckli and P. L. Vidale. European plant phenology and climate as seen in a 20-year AVHRR land-surface parameter dataset. *Int. J. Remote Sens.*, 25(17): 3303–3330, 2004.

L. Su, W. Jia, C. Hou, and Y. Lei. Microbial biosensors: a review. *Biosens. Bioelectron.*, 26(5):1788–1799, 2011.

K. Subbian and A. Banerjee. Climate multi-model regression using spatial smoothing. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 324–332. SIAM, 2013.

G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *science*, page 1227079, 2012.

J. Sun, D. Taylor, and E. M. Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015.

L. Sun and M. Wang. Global warming and global dioxide emission: an empirical study. *Journal of Environmental Management*, 46(4):327–343, 1996.

A. J. Teuling et al. Observational evidence for cloud cover enhancement over western European forests. *Nat. Commun.*, 8:14065, 2017.

C. W. Thornthwaite. Problems in the classification of climates. *Geographical Review*, 33(2):233–255, 1943.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

A. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 5:1035–1038, 1963.

G. Trewartha and L. Horn. An Introduction to Climate. *New York, 416pp*, 1980.

U. Triacca. Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature? *Theor. Appl. Climatol.*, 81(3-4):133–135, 2005.

C. J. Tucker, J. E. Pinzon, M. E. Brown, D. A. Slayback, E. W. Pak, R. Mahoney, E. F. Vermote, and N. El Saleous. An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. *Int. J. Remote Sens.*, 26(20):4485–4498, 2005.

D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

S. Tuttle and G. Salvucci. Empirical evidence of contrasting soil moisture–precipitation feedbacks across the United States. *Science*, 352(6287):825–828, 2016.

G. R. van der Werf, J. T. Randerson, L. Giglio, N. Gobron, and A. J. Dolman. Climate controls on the variability of fires in the tropics and subtropics. *Glob. Biogeochem. Cycles*, 22:330–8, 2008.

G. R. van der Werf, J. T. Randerson, L. Giglio, G. Collatz, M. Mu, P. S. Kasibhatla, D. C. Morton, R. DeFries, Y. v. Jin, and T. T. van Leeuwen. Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997–2009). *Atmos. Chem. Phys.*, 10(23):11707–11735, 2010.

E. H. Van Nes, M. Scheffer, V. Brovkin, T. M. Lenton, H. Ye, E. Deyle, and G. Sugihara. Causal feedbacks in climate change. *Nature Climate Change*, 5(5): 445, 2015.

R. R. Vatsavai, S. Shekhar, and T. E. Burk. A semi-supervised learning method for remote sensing data mining. In *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence*, pages 5–pp. IEEE, 2005.

J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor. Detecting trend and seasonal changes in satellite image time series. *Remote Sens. Environ.*, 114(1): 106–115, 2010a.

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

J. Verbesselt, N. Umlauf, M. Hirota, M. Holmgren, E. H. Van Nes, M. Herold, A. Zeileis, and M. Scheffer. Remotely sensed resilience of tropical forests. *Nature Climate Change*, 2016.

E. Volodin, N. Dianskii, and A. Gusev. Simulating present-day climate with the INMCM4. 0 coupled model of the atmospheric and oceanic general circulations. *Izvestiya, Atmospheric and Oceanic Physics*, 46(4):414–431, 2010.

H. Von Storch and F. W. Zwiers. *Statistical analysis in climate research*. Cambridge university press, 2001.

W. Waegeman. Learning to rank: a ROC-based graph-theoretic approach. *4OR-A QUARTERLY JOURNAL OF OPERATIONS RESEARCH*, 7(4):399–402, 2009.

D. Wang and W. Ding. A hierarchical pattern learning framework for forecasting extreme weather events. In *Proceedings of the 2015 IEEE International Conference on Data Mining*, pages 1021–1026. IEEE, 2015.

F. Wang, X. Wang, and T. Li. Semi-supervised multi-task learning with task regularizations. In *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM'09)*, pages 562–568. IEEE, 2009.

X. Wang, S. Piao, P. Ciais, J. Li, P. Friedlingstein, C. Koven, and A. Chen. Spring temperature change and its implication in the change of vegetation growth in North America from 1982 to 2006. *Proc. Natl Acad. Sci*, 108:1240–5, 2011. USA.

J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

P. J. Ward, B. Jongman, M. Kummu, M. D. Dettinger, F. C. S. Weiland, and H. C. Winsemius. Strong influence of El Niño Southern Oscillation on flood risk around the world. *Proceedings of the National Academy of Sciences*, 111(44): 15659–15664, 2014.

S. Watanabe, T. Hajima, K. Sudo, T. Nagashima, T. Takemura, H. Okajima, T. Nozawa, H. Kawase, M. Abe, T. Yokohata, et al. MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments. *Geosci. Model Dev.*, 4(4):845, 2011.

C. J. Willmott, K. Matsuura, and D. Legates. Terrestrial air temperature and

precipitation: Monthly and annual time series (1950-1999). *Center for climate research version*, 1, 2001.

D. Wu, X. Zhao, S. Liang, T. Zhou, K. Huang, B. Tang, and W. Zhao. Time-lag effects of global vegetation responses to climate change. *Global Change Biol.*, 2015.

T. Wu, W. Li, J. Ji, X. Xin, L. Li, Z. Wang, Y. Zhang, J. Li, F. Zhang, M. Wei, et al. Global carbon budgets simulated by the Beijing Climate Center Climate System Model for the last century. *Journal of Geophysical Research: Atmospheres*, 118 (10):4326–4347, 2013.

P. Xie and P. A. Arkin. Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, 78(11):2539–2558, 1997.

P. Xie, M. Chen, S. Yang, A. Yatagai, T. Hayasaka, Y. Fukushima, and C. Liu. A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeorol.*, 8 (3):607–626, 2007.

Y. Xie, X. Wang, and J. A. J. Silander. Deciduous forest responses to temperature and drought imply complex climate change impacts. In *Proc. Natl Acad. Sci.*, page 1358590, USA 112, 2015.

D. Xu and Y. Tian. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2(2):165–193, Jun 2015. doi: 10.1007/s40745-015-0040-1.

J. Xu, P.-N. Tan, L. Luo, and J. Zhou. GSpartan: a Geospatio-Temporal Multi-task Learning Framework for Multi-location Prediction. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 657–665, 2016. doi: 10.1137/1.9781611974348.74.

X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang. Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.*, 2017.

Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8 (Jan):35–63, 2007.

H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5:14750, 2015.

L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 947, New York, USA, jun 2009. ACM Press. ISBN 9781605584959. doi: 10.1145/1557019.1557122.

L. Ye and E. Keogh. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22 (1):149–182, Jan 2011. doi: 10.1007/s10618-010-0179-5.

D. Yi and E. Imme. Weakening of atmospheric information flow in a warming climate in the Community Climate System Model. *Geophysical Research Letters*, 41(1):193–200, 2014. doi: 10.1002/2013GL058646.

H. Zeng, J. Q. Chambers, R. I. Negrón-Juárez, G. C. Hurtt, D. B. Baker, and M. D. Powell. Impacts of tropical cyclones on U.S. forest tree mortality and carbon flux from 1851 to 2000. *Proceedings of the National Academy of Sciences*, 106(19):7888–7892, 2009. doi: 10.1073/pnas.0808914106.

N. Zeng, K. Hales, and J. D. Neelin. Nonlinear dynamics in a coupled vegetation-atmosphere system and implications for desert-forest gradient. *J. Climate*, 15 (23):3474–3487, 2002.

D. Zhang, D. Shen, A. D. N. Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage*, 59(2):895–907, 2012.

X. Zhang and X. Yan. Spatiotemporal change in geographical distribution of global climate types in the context of climate warming. *Climate Dyn.*, 43(3-4):595–605, 2014a.

X. Zhang and X. Yan. Temporal change of climate zones in China in the context of climate warming. *Theor. Appl. Climatol.*, 115(1-2):167–175, 2014b.

X. Zhang, L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers. Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdiscip. Rev. Clim. Change*, 2(6):851–870, 2011.

Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proceedings of the European Conference on Computer Vision*, pages 94–108. Springer, 2014.

M. Zhao and S. W. Running. Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. *Science*, 329:940–3, 2010.

Z. Zhao, J. Gao, Y. Wang, J. Liu, and S. Li. Exploring spatially variable relationships between NDVI and climatic factors in a transition zone using geographically weighted regression. *Theoretical and Applied Climatology*, 120(3-4):507–519, 2015.

J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*, pages 702–710, 2011.

L. Zhou et al. Widespread decline of Congo rainforest greenness in the past decade. *Nature*, 508:86–90, 2014.

X. Zhu. Semi-supervised learning literature survey, 2005.

Z. Zhu et al. Greening of the Earth and its drivers. *Nat. Clim. Change*, 6:791–5, 2016.

Y. Zhuang, K. Yu, D. Wang, and W. Ding. An evaluation of big data analytics in feature selection for long-lead extreme floods forecasting. In *Proceedings of the 13th International Conference on Networking, Sensing, and Control (ICNSC)*, pages 1–6, April 2016. doi: 10.1109/ICNSC.2016.7479007.

N. E. Zimmermann, N. G. Yoccoz, T. C. Edwards, E. S. Meier, W. Thuiller, A. Guisan, D. R. Schmatz, and P. B. Pearman. Climatic extremes improve predictions of spatial patterns of tree species. *Proc. Natl Acad. Sci*, USA 106: 197238, 2009.

J. Zscheischler, M. D. Mahecha, and S. Harmeling. Climate classifications: the value of unsupervised clustering. *Procedia Computer Science*, 9:897–906, 2012.

J. Zscheischler, M. D. Mahecha, S. Harmeling, and M. Reichstein. Detection and attribution of large spatiotemporal extreme events in Earth observation data. *Ecol. Inf.*, 15:66–73, 2013.

J. Zscheischler et al. Impact of large-scale climate extremes on biospheric carbon fluxes: An intercomparison based on MsTMIP data. *Glob. Biogeochem. Cycles*, 28:585–600, 2014.

F. W. Zwiers, L. V. Alexander, G. C. Hegerl, T. R. Knutson, J. P. Kossin, P. Naveau, N. Nicholls, C. Schär, S. I. Seneviratne, and X. Zhang. Climate extremes: challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. In *Climate Science for Serving Society*, pages 339–389. Springer, 2013.

Curriculum Vitae

Personalia

Name	Christina Papagiannopoulou
Date of birth	September 29 1988
Place of birth	Thessaloniki, Greece
Nationality	Greek
E-mail	christina.papagiannopoulou@UGent.be

Education

2011-2013: M.Sc. Computer Science, direction of Information Systems, Aristotle University of Thessaloniki, Thessaloniki, Greece.

2006-2011: B.Sc. Computer Science, University of Ioannina, Ioannina, Greece.

Honors - Scholarships - Awards

2012: Performance scholarship in the first semester of the Postgraduate Program, Department of Computer Science, Aristotle University of Thessaloniki

2007 - 2008: Scholarship from the State Scholarships Foundation: Honor in studies and morals for the academic year 2006-2007 (1st year) in Department of Computer Science, University of Ioannina

2006 - 2007: Scholarship from the State Scholarships Foundation: Honor in studies and morals for the academic year 2005-2006 (university entrance) in Department of Computer Science, University of Ioannina

2006: Scholarships from the bank Eurobank (for the degree of school certificate) and the company Cosmote (for the university entrance in Department of Computer Science as first)

Employment

Current employment

Full-time researcher at the Research Unit Knowledge-Based Systems, Department of Data Analysis and Mathematical Modelling Faculty of Bioscience Engineering, Ghent University.

Other working experience

- Research assistant at Information Technologies Institute (ITI), Center of research and Technology - Hellas (2013–2014).
- Internship: Web developer in Municipality of Ioannina area (10/2010– 1/2011).

Scientific output

Publications in international journals

- **Papagiannopoulou, C.**, Miralles, D. G., Demuzere, M., Verhoest, N. E. C., and Waegeman, W.: Global hydro-climatic biomes identified via multi-task learning, accepted in Geosci. Model Dev., https://doi.org/10.5194/gmd-2018-92, 2018.
- **Papagiannopoulou, C.**, Miralles, D. G., Decubber, S., Demuzere, M., Verhoest, N. E. C., Dorigo, W. A., and Waegeman, W.: A non-linear Granger-causality framework to investigate climate-vegetation dynamics, Geosci. Model Dev., 10, 1945-1960, https://doi.org/10.5194/gmd-10-1945-2017, 2017.
- **Papagiannopoulou, C.**, Miralles, D. G., Dorigo, W. A., Verhoest, N. E. C., Depoorter, M., and Waegeman, W.: Vegetation anomalies caused by antecedent precipitation in most of the world, Environ. Res. Lett., doi:10.1088/1748-9326/aa7145, 2017.

Conference proceedings

- **Papagiannopoulou, C.**, Miralles, D. G., Demuzere, M., Verhoest, N. E. C., and Waegeman, W.: Detecting Granger-causal relationships in global spatio-temporal climate data via multi-task learning, 4th workshop on mining and

learning from time series, held in conjunction with KDD'18 (ACM SIGKDD Conference on Knowledge Discovery and Data Mining), 2018.

- **Papagiannopoulou, C.**, Decubber, S., Miralles, D. G., Demuzere, M., Verhoest, N., and Waegeman, W.: Analyzing Granger Causality in Climate Data with Time Series Classification Methods. Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD) (3) 2017: 15-26. Presented at the Applied data science track.

- Miralles, D. G., Demuzere, M., Verhoest, N. E., Dorigo, W. A., **Papagiannopoulou, C.**, Decubber, S., and Waegeman, W. (2017, June). A non-linear data-driven approach to reveal global vegetation sensitivity to climate. In 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), 2017 (pp. 1-3). IEEE.

- Demuzere, M., Decubber, S., Miralles, D. G., **Papagiannopoulou, C.**, Waegeman, W., Verhoest, N., Dorigo, W., Sensitivity of global ecosystems to climate anomalies in observations and Earth System Models, 7th International Workshop on Climate Informatics, CI 2017

- **Papagiannopoulou, C.**, Gonzalez Miralles, D., Depoorter, M., Verhoest, N., Dorigo, W., and Waegeman, W. (2016). Discovering relationships in climate-vegetation dynamics using satellite data. Proceedings of AALTD 2016: 2nd ECML/PKDD international workshop on advanced analytics and learning on temporal data. Presented at the Advanced Analytics and Learning on Temporal Data (AALTD 2016).

Conference Abstracts

- Miralles, D., **Papagiannopoulou, C.**, Demuzere, M., Decubber, S., Waegeman, W., Verhoest, N., and Dorigo, W.: A Data-Driven Assessment of the Sensitivity of Global Ecosystems to Climate Anomalies, Presented at the American Geosciences Union (AGU) Fall Meeting 2017.

- **Papagiannopoulou, C.**, Decubber, S., Miralles, D., Demuzere, M., Dorigo, W., Verhoest, N., and Waegeman, W.: Understanding climate impacts on vegetation using a spatiotemporal non-linear Granger-causality framework, Presented at the European Geosciences Union (EGU) General Assembly 2017.

- Decubber, S., **Papagiannopoulou, C.**, Waegeman, W., Miralles, D., Verhoest, N.: Spatiotemporal prediction of global vegetation using satellite data to understand the impact of climate change. Conference of Spatial statistics, Lancaster, 2017.

- **Papagiannopoulou, C.**, Decubber, S., Waegeman, W., Demuzere, M., Verhoest, N., and Miralles, D.: A non-linear Granger causality approach for understanding climate-vegetation dynamics, Proceedings of the 26th Benelux Conference on Machine Learning (Benelearn) 2017.
- **Papagiannopoulou, C.**, Waegeman, W., Depoorter, M., Verhoest, N., Dorigo, W., and Miralles, D.: Discovering relationships in climate-vegetation dynamics using satellite data (short paper), Proceedings of the 25th Benelux Conference on Machine Learning (Benelearn) 2016.
- **Papagiannopoulou, C.**, Decubber, S., Miralles, D., Demuzere, M., Dorigo, W., Verhoest, N., and Waegeman, W.: A spatiotemporal extension of non-linear Granger causality and its application to understanding climate-vegetation dynamics, In Machine Learning for Spatiotemporal Forecasting, workshop at NIPS 2016.

Other scientific activity

- **Papagiannopoulou, C.**, Tsoumakas, G., Tsamardinos, I.: Discovering and Exploiting Deterministic Label Relationships in Multi-Label Learning. KDD '15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 915-924
- Apostolidis, K., **Papagiannopoulou, C.**, Mezaris, M.: CERTH at MediaEval 2014 Synchronization of Multi-User Event Media Task. Proceedings of the MediaEval 2014 Workshop, CEUR vol. 1263, Barcelona, Spain, October 2014
- **Papagiannopoulou, C.**, Mezaris, V.: Concept-based Image Clustering and Summarization of Event-related Image Collections. Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia, (HuEvent '14) 2014: 23-28
- Moumtzidou, A., Avgerinakis, K., Apostolidis, E. E., Aleksic, V., Markatopoulou, F., **Papagiannopoulou, C.**, Vrochidis, S., Mezaris, V., Busch, R., Kompatsiaris, I.: VERGE: An Interactive Search Engine for Browsing Video Collections. MultiMedia Modeling (MMM), Springer International Publishing, (2) 2014: 411-414