GHENT
UNIVERSITY

Faculty of Sciences

Spatiotemporal optimization of Granger causality methods for climate change attribution

Stijn Decubber

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promotor: prof. dr. Willem Waegeman
Co-promotor: prof. dr. ir. Olivier Thas
Tutor: Christina Papagiannopoulou

Department of Applied Mathematics, Computer Science and Statistics

Academic year 2016 - 2017

Stijn Decubber
January 27, 2017

Foreword

This thesis was carried out in the framework of the ongoing SAT-EX project.[1] The dataset that was used was composed by Christina Papagiannopoulou and Diego Miralles. The contribution of this thesis consists of an exploratory data analysis, the exploration and discussion of different models to predict vegetation anomalies, the extension towards the multitask learning setting and finally the exploration and discussion of the statistical aspects of using the concept of Granger causality.

I would like to thank Christina for providing me with the data and for getting me started with the thesis. A big thanks to Christina, Willem, Diego and Matthias for the fruitful discussions about our results. I am also grateful to Francis Wyffels and Karel Vermeulen for sharing their insights and advice.

Finally, I would like to thank Willem and Diego for giving me the opportunity to start working on the project as a PhD student and Lena for kicking me out of bed every morning, giving me the opportunity to start working half an hour earlier than I would do without her.

[1] For more information on SAT-EX, see `http://www.sat-ex.ugent.be/`

Table of Contents

Abstract

Climatic conditions are known to be key drivers of ecosystem dynamics, which are sensitive to temperature, availability of water and the solar irradiance. In the other direction, vegetation has a known influence on climate systems on a global scale. Through evapotranspiration of water and exchange of carbon dioxide with the atmosphere, vegetation is a major player in the global water and carbon cycles. Furthermore, the amount of reflective vegetation governs the net solar radiation reaching the earth surface and vegetation affects wind speed and direction. In light of this complex interplay between climate and vegetation, investigating the sensitivity of vegetation to changes in climatic conditions is crucial to improve our understanding of global climate change.

The field of climate science is one of the most data-rich research areas. Earth observation satellite data provide a wealth of information about the dynamics of our planet in recent decades. Composite global records of important environmental and climatic variables now span up to 35 years, enabling the study of climate-vegetation interactions over multi-decadal scales. These records have the form of multivariate time series with different spatial and temporal resolutions. Despite this abundance of data, advances in the field by making use of data-driven models have been limited. One interesting application is so-called causal discovery, in which statistical methods are used to highlight interesting relations and interactions between climate variables. Recently, machine learning approaches based on graphical models have been proposed to tackle these kind of problems. Granger causality is another approach to perform causal discovery. Granger causality originated in the field of econometrics and essentially requires a statistical comparison of different predictive models.

This thesis contributes to the development of a Granger causality framework for performing causal discovery in a climatic dataset using predictive data-driven models. The dataset consists of global vegetation records quantified by the Normalized Difference Vegetation Index (NDVI), together with multiple records of temperature, precipitation, radiation, soil moisture and snow. The time series span 30 years in total with a monthly temporal resolution. The spatial resolution is 1 by 1 degree, meaning that data is available for over 13000 land pixels.

Different machine learning models were explored to address the regression problem of predicting vegetation anomalies with past vegetation and climate variables as predictors. Next, the idea of multitask learning was used to make better predictions. A major generalization improvement on out-of-sample data was achieved by exploiting training data from related pixels. Furthermore, the quality of high-level engineered features was explored, demonstrating the potential of automatic feature extraction methods on high-resolution data. Finally, some challenges with regards to statistical inference were discussed. In particular, the need for a model-free test to compare forecasts was highlighted. Although imperfect, an outline to tackle the statistical problem based on resampling methods was proposed.

1 Introduction

Climate models predict an aggravation of droughts, extreme precipitation events and heatwaves as we progress into the future. Recent advances in satellite Earth observation - with the development of consistent global historical records of crucial environmental and climatic variables - provide new means to start unravelling the processes driving long-term changes in climate extremes, and understanding the impact of these changes on terrestrial ecosystems.[1]

The research for this thesis was carried out as a part of the SAT-EX project, under supervision of the KERMIT research group. The project fits in the context of climate change research with a particular focus on extreme events such as droughts and heat waves or extreme precipitation and on their impact on vegetation. The main objectives are to understand how these extreme events have changed in frequency and intensity over time, to provide insight in vegetation distribution and dynamics and to understand and reduce mechanistic model uncertainty in predicting these extremes. As one of five research groups in the project, KERMIT aims at using data-driven models in experiments towards the goals of the project.

Research questions in climate change research are mostly related to either *climate projection* or to *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

[1] Quoted from the SAT-EX website.

and natural or anthropogenic factors. A well-studied example, both for projection and attribution, is the effect of human greenhouse gas emissions on global temperature.

The standard approach in the field of climate science is based on simulation studies with mechanistic climate models, which have been developed, expanded and extensively studied over the last decades. These models are based on conceptual representations of the global water, atmospheric and biological systems, mathematically formalized through complex differential equations.

Data-driven models, in contrast to mechanistic models, assume no underlying physical representation of reality but directly model the phenomenon of interest by learning a more or less flexible function of some set of input data. Climate science is one of the most data-rich research domains. With global observations on ever finer spatial and temporal resolutions from both satellites and in-situ measurements, the amount of (publicly available) climatic data sets has vastly grown over the last decades. It goes without any doubt that there is a big potential for making progress in climate science with data-driven statistical models. Despite this potential, the advances made within the field using statistical analysis and data mining have been limited compared to other fields such as genomics or business intelligence, partly because of the very complex nature of our planet's global climate system [16].

1.1 Causal discovery in data-driven models

The goal of statistical causal discovery is to understand the world around us by using observational data to identify cause-effect relations between variables. It has only recently been applied in the field of climate science. Although most people have an intuitive understanding of causality, it is a complex concept that needs mathematical formalization together with some important assumptions before it can be applied in a statistical modeling framework. Apart from model-related assumptions (the model structure should model the underlying relations between the data well), the most important assumption to be made is that of *causal sufficiency*. Causal sufficiency means that there are no confounding variables or hidden common causes that are not included in the data. If any two variables X and Y have a common cause Z, then Z must be included in the study [10]. This also implies that any direct causal relation that is discovered is relative to the variables that are included, and any relation can either turn into an indirect relation or disappear when a new variable or a common cause is taken into consideration. Two simple examples shown in Figure 1.1 illustrate this:

- Cloud coverage is a statistical cause of both rain and decreasing UV radiation, but there is no causal relation between rain and the amount of UV. A statistical model that does not

include cloud coverage as a variable will falsely detect a relation between rain and the amount of UV.

- A model aimed at causal inference will identify cloud coverage as a potential cause of flooding, when rain is not included in the system. When rain is introduced as an additional variable, cloud coverage no longer causes flooding directly, but rather indirectly by being a cause of rain, itself causing flooding.

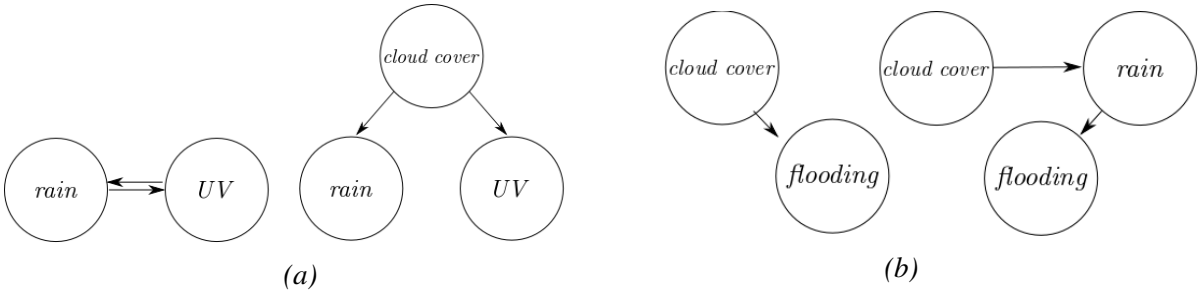


Figure 1.1: (a): causal relations disappear when a hidden common cause is introduced in the study. (b): a direct causal relation becomes indirect when an additional variable is included. Examples adopted from [10].

The assumption of causal sufficiency is almost never fulfilled in climate studies, mostly because there is no data available on every hidden common cause, or there may be confounding variables that we are not aware of because of the complexity of the system. Therefore, it is important to realize that any uncovered causal relations are *potential* relations relative to the set of information that is available in the data. In addition, knowledge from domain experts can and should be used to evaluate the causal relations that are detected: if there is a well-known physical mechanism that explains a relation, it can be confirmed. In the other case, the causal relation should be seen as a hypothesis that can be interesting for further investigation [10].

I will highlight two approaches that have emerged as mathematical formalizations to test for causal relations: causal discovery in graphical models and the concept of Granger causality from econometrics.

1.1.1 Probabilistic graphical models

In probabilistic graphical models (PGMs), every node in a graph represents a variable and relations between variables are shown as edges between the nodes. Every edge is supplemented with a specific probability, so that the causal relations are not exact (a cause may or may not, but doesn't have to, cause its effect). The aim of causal discovery here is to identify direct probabilistic edges between variables, which can then be attributed with a causal interpretation.

The underlying principle of graphical causal discovery algorithms is that, because of the assumption of causal sufficiency, no absolute causal relations can be proved. A causal relation detected by the algorithm might disappear when additional variables are included. It is possible, however, to disprove apparent causal relations between variables that are correlated, by testing for conditional independences. In a system with three variables X, Y and Z, an algorithm can eliminate the edge between X and Z even if they are correlated when the conditional distribution of X given Y and Z is no different from the distribution of X given Z: $P(X|Y,Z) = P(X|Y)$. Using the idea of conditional independence, potential causal relations can be identified by starting from a fully connected graph and eliminating as many edges as possible [10]. This approach has been adopted recently in climate science by identifying connections between nodes in so-called climate networks, a graphical structure defined on a global grid [11].

Most climate data comes in the form of spatiotemporal data, typically as daily or monthly records over an extended period of time for several locations on earth. Probabilistic graphical models can be applied to temporal data by explicitly defining additional nodes for the history of every variable (lagged variables) together with some temporal constraints so that the edge directions follow the direction of time. As such, a system with N variables measured over S time slices can be converted to a graph structure with $N \times S$ nodes, allowing to test causal relations over time. However, given the often high temporal resolution of the data, the complexity of these models quickly increases and they can become computationally infeasible to track [10].

One specific approach called grouped graphical modeling has recently been put forward to address temporal causal modeling in the context of climate science [29, 31, 30]. These methods are not graphical models, but use notations adopted from graphical modeling to conceptualize causal relations between time series, by depicting every variable in the system as a node and drawing directed edges between variables. Testing for causal relations is done through feature selection. More specifically, methods that perform grouped feature selection are used. This incorporates the idea that, for a time series variable X_t $(t = 0, 1, ..., T)$ not just one specific lagged variable X_{t-k} but rather the relevant past of X_t for every lag up to lag k is informative for predicting another variable Y_t. Thus, the feature selection takes into account the group structure present in the variables naturally imposed by the time series they belong to and lagged time series are selected by the model as a whole instead of individual lagged variables [29]. The authors propose the group lasso and group boosting as algorithms to apply in their framework. The group lasso performs grouped feature selection by penalizing intra-group and inter-group variable inclusion separately:

$$\hat{\beta}_{group}(\lambda) = \arg\min_{\beta}(||Y - X\beta||^2 + \lambda \sum_{j=1}^{J} ||\beta_{G_j}||_2), \tag{1.1}$$

where the penalty is imposed on the amplitude (the l_2 norm) of groups of coefficients. In another study, the same authors propose a group elastic net model for spatiotemporal modeling. This model incorporates both a penalty enforcing sparsity on the group level as well as a spatial penalty enforcing spatial smoothness and regulation, i.e., large model coefficients corresponding with variables that are further away in a spatial sense are penalized [31].

A second formalization of causal discovery for temporal data is the concept of Granger causality, which was proposed in the field of econometrics by Nobel-prize winner Clive Granger [19].

1.1.2 Granger causality

Granger causality is an operational definition of causality. It lends itself for the questions that are asked in climate research and much climate attribution studies have been carried out making use of the Granger causality framework [31, 27, 3]. It is a notion of causality that is based on prediction, and is defined under the following general assumptions [20]:

- A cause precedes its effect. As such, the past and the present may cause the future, but not the other way around.
- No information in the system under study is redundant. In other words, there is no deterministic relation between variables.
- Causal relations do not change in direction throughout time.

In essence, the concept of Granger causality is not tied to one particular probabilistic model [14]. However, in its original specification, Granger causality formalizes causality for two stationary time-series X_t and Y_t, with the index t denoting the time-index ($t = 1, 2, ...$). Let I_{XY} be the information set consisting of lagged values of both X_t and Y_t up to lags l_x and l_y, i.e., I_{XY} consists of the vectors $[X_{t-l_x}, X_{t-l_x+1}, ..., X_{t-1}]$ and $[Y_{t-l_y}, Y_{t-l_y+1}, ..., Y_{t-1}]$. Likewise, I_Y is the information set consisting of just the lagged values of Y_t. Then, time series X_t is said to *Granger-cause* (or *G-cause*) time series Y_t with respect to the information set I_{XY} if the following inequality holds:

$$f(Y_t|I_{XY}) \neq f(Y_t|I_Y) \;\; \forall t = 1, 2, ... \tag{1.2}$$

That is, the conditional distribution of Y_t given information from the past of both Y_t and X_t is different from the conditional distribution of Y_t given only its own past. In other words, the past of X_t contains additional information on top of the past of Y_t to model the present and future of Y_t [22]. This definition in terms of conditional distributions is referred to as *strong Granger causality*, whereas in most practical applications *Granger causality in the mean* would be a more correct term (see below) [14].

It should be emphasized that Granger causality is defined only with respect to the information set I comprising all time series considered as variables in a particular study. Granger coined the term *prima facie* ("at first sight") cause to express the fact that Granger causal relations with respect to a certain data set I are only potential causal relations in a more comprehensive set I_+, which might contain confounding variables not present in I [19]. In other words, Granger causality basically is a measure of association between variables and expresses causality by incorporating temporal structure in its definition. In no case does a Granger causal relation between variables imply a true physical causal relation. Furthermore, if important relevant variables are not included in the analysis, Granger causality may very well detect spurious correlations and may be useless for causal inference. This has been the main point of criticism on Granger causality, but this of course also applies on causality in probabilistic graphical models or in any other type of model [14, 27, 10].

In concrete applications, a *G*-causal relation from a variable x to an outcome y can be tested for by comparing the predictive performance of two nested models: a baseline autoregressive model predicting y at time t as $y_t = f_1(y_{t-1}, ..., y_{t-p})$, and an extended model $y_t = f_2(y_{t-1}, ..., y_{t-p}, x_{t-1}, ..., x_{t-q})$. Generally, the null hypothesis of Granger non-causality is formulated as the null hypothesis that f_1 and f_2 have equal prediction error (in econometrics: *forecasting accuracy*). Typically the alternative is one-sided, such that if f_2 predicts y_t significantly better than f_1, H_0 is rejected. In economic time series literature it has been common practice to test for Granger causality using linear least-squares models, typically vector autoregressive (VAR) models in the multivariate setting [3]. Consider for example a set of three time series (variables) (X_t, Y_t, Z_t). A VAR model expresses each time series as a linear combination of the past values of every variable in the system, up to lag l:

$$\begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix} = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{03} \end{bmatrix} + \sum_{l=1}^{k} \begin{bmatrix} \beta_{11,l} & \beta_{12,l} & \beta_{13,l} \\ \beta_{21,l} & \beta_{22,l} & \beta_{23,l} \\ \beta_{31,l} & \beta_{32,l} & \beta_{33,l} \end{bmatrix} \begin{bmatrix} X_{t-l} \\ Y_{t-l} \\ Z_{t-l} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \tag{1.3}$$

The model parameters β can be estimated through least-square error minimization. Hence, with these models the above definition is usually restricted to modelling the mean of the conditional distributions of interest [22]. Statistical testing for G-causal relations between variables then proceeds by testing the null-hypothesis of Granger non-causality, for instance for the causal relation $X_t \to Y_t$:

$$H_0 : X_t \text{ does not G-cause } Y_t \text{ with respect to the system } (X_t, Y_t, Z_t),$$

In early applications, this null hypothesis is sometimes formulated in terms of the corresponding model parameters and tested for using standard F-tests [22, 15]:

$$H_0 : \beta_{12,l} = 0 \text{ for all } l = 1, ..., k.$$

1.2 Outline and goals of the thesis

In this thesis, I focus on the problem of climatic attribution in statistical models, using a dataset that was provided by my colleague Christina Papagiannopoulou from KERMIT. The data has the form of multivariate time series data for every location on earth, on a discrete pixel grid. Because of the strong seasonal pattern in vegetation, it is very easy to predict raw vegetation time series. Therefore, in line with the work of my colleague and with the goals of the SAT-EX project, the anomalies (also called residuals) of the vegetation records will be the main variable of interest [36].

The concept of Granger causality is interesting to discover potentially interesting relations between variables. However, the traditional approach is limited to linear regression. The complex nature and the size of many climate data sets make that other models might be needed to achieve good predictions. Recent results with the SAT-EX data also point in this direction [36]. Therefore, the focus of this thesis is on the application of machine learning models to predict anomalies in vegetation time series in the framework of Granger causality. This extension of Granger causality beyond linear regression poses some challenges with regard to statistical testing of the null hypothesis of non-causality. These are also addressed. The thesis does not aim at drawing specific conclusions about causal relations between different climate variables. Rather, it should be seen as a contribution to the development of a general framework that can be used to discover interesting patterns in climatic data sets.

The outline of the thesis is as follows:

- First, a description of the climatic dataset and the methods that were used in this thesis is given in Chapter 2.
- Chapter 3 is concerned with an exploratory analysis of the data, with a specific focus on autocorrelation in the vegetation data and correlations between climate and vegetation. Principle component analysis is used to explore the high-dimensional feature space of the data.
- In Chapter 4, different models are explored to predict vegetation anomalies, with the aim of performing Granger-causal inference in mind. First, every pixel is treated as a separate problem and the performance of both linear and non-linear autoregressive models is evaluated. In a second part, a more complicated model is proposed to exploit similarities between different pixels, using the idea of multitask learning. Finally, the potential of expanding the variable space with spatial information and with high-level engineered features is explored.
- In Chapter 5, statistical inference in the Granger causality framework is addressed, using the results from Chapter 4. Different approaches to tackle the problem are discussed.

2

Data and methods

2.1 Description of the data

The dataset used in this thesis was composed by and provided by courtesy of Christina Papagiannopoulou (KERMIT, department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering). It was assembled using several publicly-available satellite datasets covering different spatial areas and spanning different time intervals, with various resolutions. An overview of the data sources that were used is given in Appendix **A1**. Although the native resolution of some dataset is finer, all products have the same temporal resolution and length in the final dataset.

What follows is a description of the full available dataset. For this thesis, only part of the data was used to generate most results. Which part will be specified wherever necessary in the results section.

2.1.1 Raw data

The complete dataset consists of separate dataframes for 1x1°-sized pixels (roughly corresponding to squares with side 100 km, depending on the location), for each of 13,097 pixels covering all the land surface on earth, stored as HDF5-format files. Each dataframe contains monthly observations on 21 climatic time series, together with additional engineered features. Twenty-one of those time series will serve as predictor variables: temperature (7 time series), precipitation

(8), soil moisture (3), snow (1) and radiation (2). In addition, one time series for vegetation is available, in which the Normalized Difference Vegetation Index (NDVI) is used as a proxy for the amount of vegetation. It is a graphical indicator that uses the spectral reflectance measurements in the visible (VIS) and near-infrared (NIR) regions. The rationale behind the NDVI is that live vegetation absorbs visible light, but reflects light in the near-infrared region to avoid overheating of its tissues. As such, the NDVI takes on values between -1 and +1. The higher the NDVI, the more green vegetation is present, with NDVI values near 1 corresponding to the tropical rain forests.[1]

2.1.2 High-level features

Apart from the raw time series, each data frame also contains high-level features that were manually constructed using domain knowledge. A first set of features are the different signal components from the raw time series: seasonal cycles, trends and residuals. A second set of features are manually constructed high-level features extracted from separate time series with a daily resolution.

Raw signal components

The raw time series were decomposed into anomalies using an additive linear approach. First, each time series y_t^T was de-trended over the entire study period by modeling the trend y_t^T with a linear model with the timestamp t as a predictor variable:

$$y_t^T = \alpha_0 + \alpha_1 \times t \tag{2.1}$$

The de-trended time series were then obtained by subtracting the trend y_t^T from the original time series. Consequently, the seasonal cycle y_t^S was estimated by computing the monthly averages over the entire study period, and subtracted form the de-trended time series to obtain the final anomalies. The residuals, trend and seasonal component of every time series are available in every dataframe.

High-level features

High-level features were constructed from the raw time series that were originally available with a daily resolution. The idea behind these high-level features is that they appropriately reflect the climate dynamics and the sensitivities of vegetation. For instance, vegetation in a certain area could be responsive to the number of consecutive number of days without precipitation in the past month rather than to the actual amount of precipitation. These so-called indices include extreme values such as minimum or maximum temperatures, the number of times a

[1]For more information about the NDVI, see `https://en.wikipedia.org/wiki/Normalized_Difference_Vegetation_Index`

certain threshold of precipitation was reached, etc. Furthermore, cumulative indices such as the total amount of precipitation over the last three months are included. Appendix **A2** gives a full overview of the indices that were calculated. The result is about 200 additional unique variables in every dataframe on top of the raw signal components mentioned above.

2.1.3 Encoding the past: lagged variables

To enable modelling in the Granger causality framework, all the models used in this thesis are autoregressive regression models, aiming at predicting NDVI anomalies. In general, they take the following form:

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t_k}, x_{1,t}, ..., x_{1,t-p}, x_{2,t}, ..., x_{2,t-q}, ...) \tag{2.2}$$

Where y_t is the NDVI anomaly at timestamp t and $x_1, x_2, ...$ are climate variables. In order to include the history of the variables as regressors, the past of every variable is included in the data in the form of lagged variables, up to a total lag of 12 months for variables related to precipitation and temperature, and 6 months for variables related to radiation. The history of every variable is encoded in each dataframe, including the high-level features, such that the dimensionality of each dataframe strongly increases, with up to 6000 variables in some pixels.

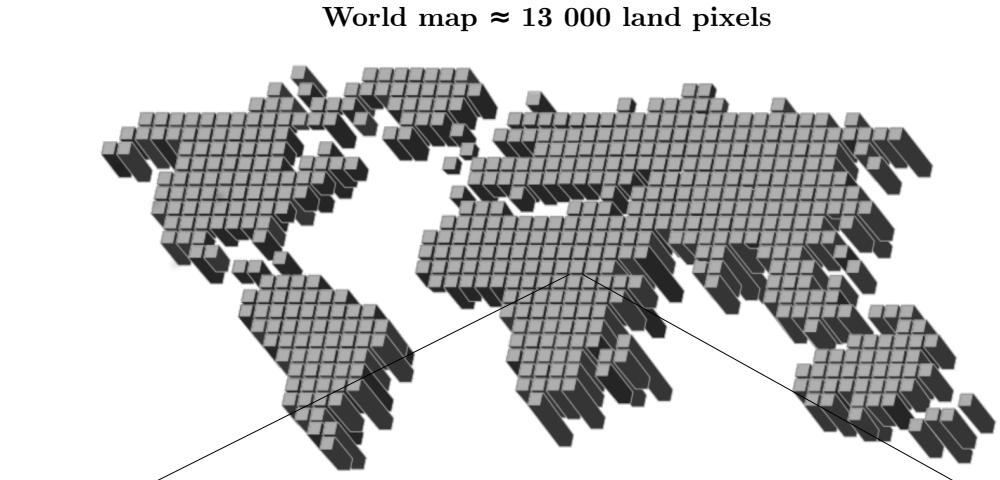
In summary, the final dataset covers the entire globe with 1x1°-sized pixels. It spans the period 1981-2010 with a monthly resolution, providing 360 observations on 21 climatic time series and on one vegetation time series per pixel. In addition, raw signal components and additional high-level features expressing cumulatives and extremes are included. The history of every variable is explicitly included as lagged variables, up to a total lag of 12 months. An overview of the data is given in Figure 2.1.

2.2 Methods

2.2.1 Data preprocessing

Before analysis, the data was preprocessed in the following way:

1. Instances where the NDVI value is missing are removed. In most pixels, just a few entries have to be removed.
2. Timestamp, latitude and longitude are removed from the feature matrix.
3. Features with a high percentage of missing data ($> 30\%$) are removed.
4. Remaining missing values in the features are imputed with the mean of the respective column.



Dataframe for one pixel: 360 rows x 5950 columns

Time	Lat	Lon	Temp	Temp_{t-1}	...	Feat	Feat_{t-1}	...	NDVI^{resid}	$\text{NDVI}^{resid}_{t-1}$	...
07/1981	-0.5	13.5	21.8	...	...	5	...	...	0.8545	...	...
08/1981	-0.5	13.5	22.6	21.8	...	8	5	...	0.8418	0.8545	...
09/1982	-0.5	13.5	18.5	22.6	...	12	8	...	0.7348	0.8418	...
...	-0.5	13.5	...	18.5	...	...	12	...	...	0.7348	...
...	-0.5	13.5	...	...	...	...	...	...	...	...	...
...	-0.5	13.5	...	...	...	...	...	...	...	...	...
...	-0.5	13.5	...	...	...	...	...	...	...	...	...
12/2010	-0.5	13.5	5.5	...	...	18	...	...	0.1164	...	...

Signal components **Additional features** **Outcome**

Figure 2.1: Overview of the data. Data is available for roughly 13000 land pixels. The dataframe for every pixel consists of monthly observations on signal components from 20 raw climatic time series, together with additional features related to extremes and cumulatives. The past of both signal components and additional features is encoded as lagged variables, up to a maximum lag of 12 months for temperature and precipitation and 6 months for radiation.

2.2.2 Models

Mostly off-the-shelf predictive models were used: linear regression, ridge regression, the lasso and random forest regression. For a detailed description of these models please refer to Friedman et al., 2001 [17]. All these models are implemented and ready to use in the Scikit-learn library for Python.

One model type used in this thesis is not readily available and was implemented manually: an extreme learning machine or ELM. An ELM is a non-linear regression method that has as main advantage over other methods a very fast training speed, so that it is useful for larger datasets. An ELM is a single-layer feedforward neural network, where the weights connecting the input layer and the hidden layer are randomly initialized and never updated and the outputs of the hidden layer are used to predict the output layer, obtained as least-squares solutions to a linear system.[2] The fact that the weights from the output layer can be obtained through

[2] Another way to think of an ELM is as principle component regression with random, non-linear components.

linear least-squares makes the ELM a fast method. The resulting model is non-linear because the weighted sums of inputs are passed through a non-linear activation function (hyperbolic tangent) to produce the output of the hidden layer. This is schematically shown in Figure 2.2.

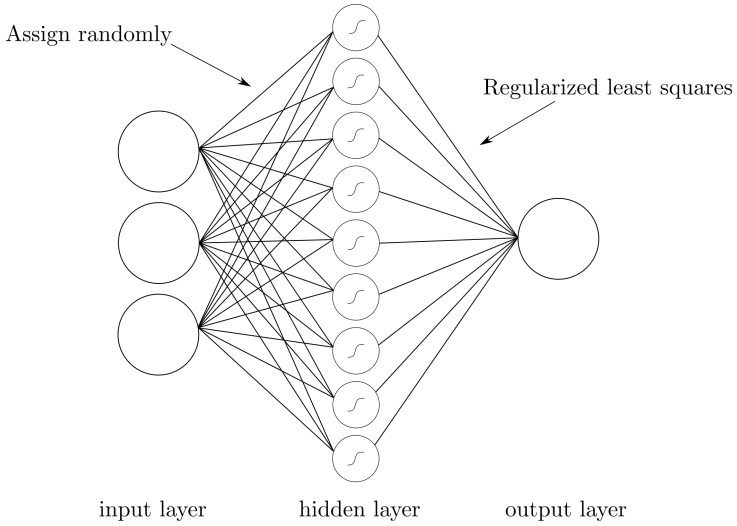


Figure 2.2: Schematic representation of an extreme learning machine. The features are fed to the input layer. Random combinations of the input are fed to the hidden layer, which are passed through a non-linear activation function. Regularized regression is performed on the outputs of the hidden layer to predict the output layer.

In this thesis an extended ELM was used which imposes an l_2-type penalty on the weights of the output layer in order to improve generalization. The ELM was first proposed in 2006 [24] and the extension with ridge regression used in this thesis has been described by Li et al. (2013) [28]. Extreme learning machines have extensively been used for time series prediction and have the advantage of being several orders of magnitude faster to train on large datasets compared to other non-linear models such as random forests or gaussian processes [33].

2.2.3 Model evaluation

A common practice in statistical learning is to assess the out-of-sample performance of a model by k-fold cross-validation [21]. Because of the autocorrelation between consecutive observations in time series, regular k-fold cross-validation might not be valid since the validation and training samples are no longer independent. Modified versions of cross-validation for time series exist and mainly consist of leaving out part of the data so that the minimum distance h between observations in the training and the validation part is so that they are independent again. Another approach is to train the model on an early part of the time series, leave out part of the data, and test it on the most recent part, so that the test data is both out-of-sample and *out-of-time* [2].

In this thesis, both random 5-fold cross-validation and proper out-of-sample and out-of-time validation were used to assess the model performance on unseen data. Therefore, the data was

split in a training part and a test part. The training part consisted of the first 24 years of observations. The 25th year was left out of consideration, and the final 5 years formed the test part. Every model was first evaluated by performing 5-fold cross-validation on the training set. Finally, the models were fitted on the complete training set and their generalization performance was evaluated on the test set. Hyperparameters were first tuned on the training set whenever applicable.

As performance metrics, the mean absolute error (MAE) was used together with the coefficient of determination R^2 defined as follows:

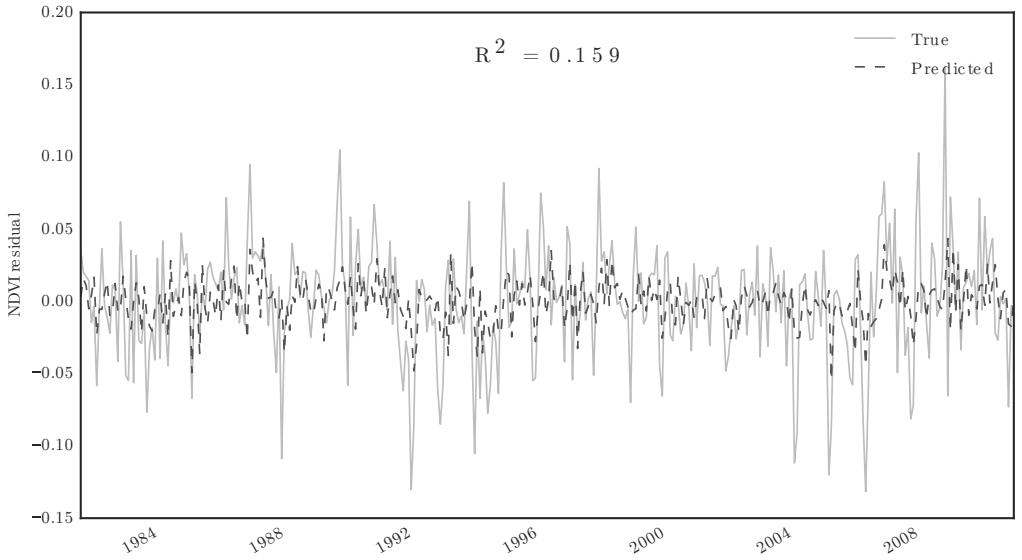
$$R^2(y, \hat{y}) = 1 - \frac{MSE}{MStot} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2.3}$$

A model that always predicts the mean of y, unconditional of the features, will produce an R^2 of 0. Because the models are evaluated out-of-sample, either by cross-validation or on an independent test set, the predictions can be arbitrarily worse than the mean. In those cases, a negative R^2 is possible. In order to give a flavor of what different R^2 values mean, two predictions for a series of NDVI anomalies are shown in Figure 2.3. The prediction in the top plot corresponds to an R^2 of 0.159. The sign of the anomalies is predicted correctly most of the time, but the magnitude is mostly way off. The bottom plot shows predictions with an R^2 of 0.704. Visually, these predictions seem to be quite accurate.

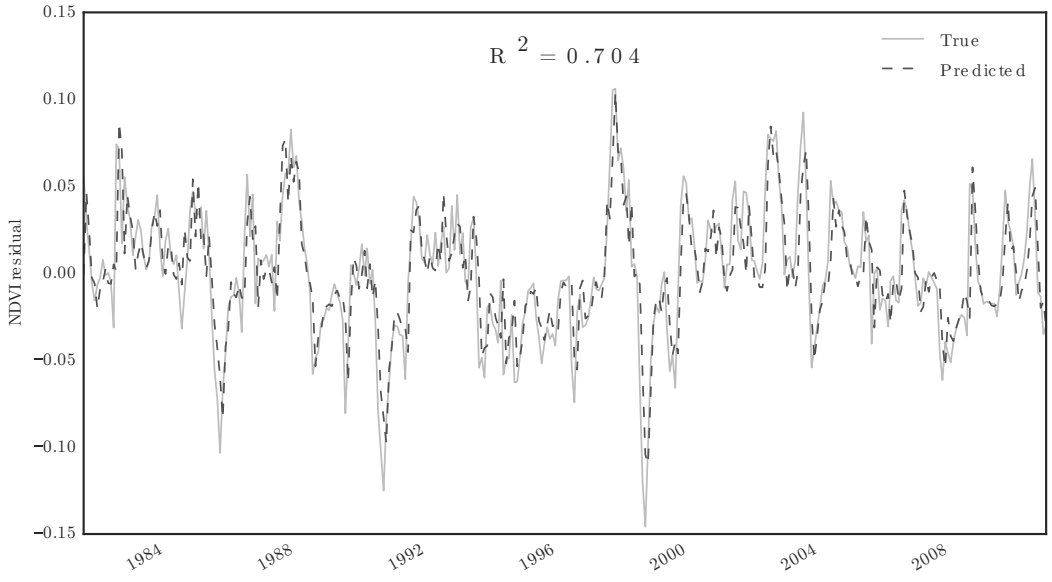
Finally, the training, validation and test data sets were rescaled to zero mean and unit variance before each analysis, using only training data statistics to do so.

2.2.4 Software

Analyses were performed in R version 3.2.5 (2016-04-14) [38] and in Python version 2.7.12 - Anaconda distribution. Extensive use was made of Scikit-learn, a software machine learning library for Python [37]. The computational resources (STEVIN Supercomputer Infrastructure) and services used in this thesis were provided by the Flemish Supercomputer Center (VSC), Ghent University.



(a)



(b)

Figure 2.3: Predicting NDVI anomalies. (a): with an R^2 of 0.159. (b): with an R^2 of 0.704.

3 Exploratory analysis

This chapter provides an exploratory analysis of the raw climate time series and the extracted features that will serve as predictors to model vegetation in the next chapter. Because the NDVI anomalies will be targeted for prediction in the next chapter, not the raw NDVI time series but their anomalies were studied. Along with that, autocorrelations within the vegetation time series and correlations between vegetation and climate variables were studied as well.

For this chapter, some figures with a diverging color coding were constructed, which cannot be converted to grayscale. For practical reasons, these figures are provided in the Appendices at the end of this thesis. Note that in the pdf version of this document, switching back and forth between the figures and the text is possible by clicking on the figure numbers in the text and on the section numbers in the figure captions.

3.1 Correlation between records from different satellites

The data set consists of multiple time series for every land pixel on earth, some of which contain measurements on the same variable. For instance, five of the seven time series related to temperature are observations of the near-surface air temperature. Intuitively, one could expect these time series to show a large degree of collinearity. Figure 1 visualizes the Pearson correlation matrix between the temperature-related time series for a randomly selected pixel from the dataset (latitude -24.5, longitude 22.5, located in the south of Africa). *GISS* and *MLOST* contain de-trended temperature anomalies, i.e., what remains from the raw signal after subtracting the

seasonal cycle and the trend. The rest contain direct measurements of surface temperature.

In general, the different temperature measurements for this pixel are very highly correlated, with ISCCP being the only product that is slightly less correlated to the other products. As GISS and MLOST contain temperature anomalies, they are not correlated to the other variables, yet strongly correlate with each other ($r = 0.94$).

The correlation coefficients were also calculated between the pairs of temperature time series CRU/ERA, LST/CRU and UDEL/ERA for all pixels globally. The left hand side of Figure 2 shows the correlations between the raw time series, while the right hand side shows the correlation between the residuals.

The correlation maps for all other pairs of raw temperature time series (not shown) are very similar to those shown in the left part of Figure 2. Hence, all raw temperature time series are highly correlated in most parts of the world, except for the tropical regions. The reason for this is that the temperature in these regions is very constant throughout the year. Temperature time series in more temperate climates are highly correlated because of the presence of the same strong seasonal cycle in all measurements. In the tropics, there is much less seasonal variability in the temperature, meaning that this strong correlated component is not present. When the seasonal component is left out of the temperature signal (Figure 2, right), the correlation between the temperature time series goes down, to the extent that some series of temperature residuals are no longer positively correlated. This indicates that different temperature time series contain different information even though they are all measurements of the same physical phenomenon. It is possible that the quality of the different measurements differs per region and that the most accurate temperature product is not the same for all locations on earth.

Figure 3 shows a similar story for the correlations between the raw time series related to precipitation (for the same selected pixel as before). Because no snowfall occurred in this pixel, snow was excluded from this analysis. As is the case for temperature, some precipitation time series are strongly correlated, while others are only weakly or even not correlated at all.

Even though, both for temperature and precipitation, multiple time series contain measurements on the same physical phenomena, the correlation between them is not always high and some measurements of the same variable even show no correlation at all. For temperature, this is because satellites do not measure temperature directly, but rather measure irradiance in different parts of the wavelength spectrum, commonly with different measurement equipment (sensors) for every satellite [35]. These measurements are converted to temperature by research groups, with the final result depending on the details of the methods used. In addition, surface temperature measurements are dependent on inhomogeneities in the surface and are accurate only

under cloud-free conditions. Similar factors are in play for precipitation. Finally, some products are not based on satellite measurements but were composed from gauging station network measurements. These point measurements are interpolated over a larger geographical area and typically also show differences with satellite-based observations [25]. Although a full technical discussion on measurement techniques and differences between satellite-based and ground-based observations is beyond the scope of this thesis, it is clear that temperature or precipitation records coming from different satellites do not necessarily contain duplicate information.

3.2 Autocorrelation within vegetation time series residuals

Because the past residuals of NDVI will serve as predictors in the Granger causality framework, it is interesting to check to what extent there is autocorrelation within the NDVI time series. Figure 4 shows the autocorrelation value of the NDVI residuals for every pixel, for temporal lags 1, 2, 3 and 4 months. For a temporal lag of 1 timestep, the NDVI residuals are positively correlated in most regions of the world, with the highest autocorrelation in Australia, the south of Latin America, central North America, Central Asia, the south of Africa and the Sahel region. For a temporal lag of two, the autocorrelation drops, with autocorrelations remaining positive mainly in Australia. The autocorrelation between NDVI residuals separated by three and four time steps is very close to zero in most pixels.

3.3 Correlation between NDVI residuals and climate

In order to motivate the size of the temporal window for the climate variables in the models in the next section, a plot of the correlation between climate variables and NDVI residuals was made, for increasing time lags between the contemporaneous observation of the NDVI residuals and the past observation of the climate variables (Figure 5). The plot shows the correlations for every pixel for one product of each the four major climate variable groups: temperature, precipitation, soil moisture and radiation. In general, the climate variables are most correlated with the NDVI residual when they are measured in the same month. Correlations between current vegetation and past climate tend to go down as we go further back in the past.

Correlations between temperature and NDVI residuals are larger than 0.2 in absolute value for only a few pixels on earth. Furthermore, almost all temperature correlations are smaller than 0.1 in absolute value from lag 3 onwards. The same is true for radiation. The correlations with the precipitation and soil moisture time series on the other hand are much stronger, with correlation coefficients larger than 0.2 between NDVI residuals and precipitation from 12 months ago in some pixels in Australia. This is in line with geoscientific literature, where it is well

documented that vegetation systems have a longer memory for water than for temperature [23].

Remarkably, precipitation and soil moisture are consistently negatively correlated with vegetation residuals from the same month in pixels at higher latitudes (for instance in Europe and Russia). At the same time, vegetation in these pixels is positively correlated with temperature and most of them also with radiation. A possible explanation for this could be the relation between precipitation and temperature: in a month with a large amount of precipitation, the amount of radiation reaching the vegetation and the temperature will tend 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 residuals.

3.4 Variability in the feature space: PCA

Principal component analysis (PCA) was used as a dimension reduction technique to explore the full available feature space of each pixel's dataframe. These dataframes consist of monthly observations on lagged signal components and additional features (see Figure 2.1). PCA projects high-dimensional observations onto a lower dimensional subspace while maximally conserving variability between observations. As such, in this context each month represents one high-dimensional observation and the principal components spanning the lower-dimensional subspace are linear combinations of the variables that make up the feature space. The weight of each feature is commonly referred to as the *loading* for a particular component. 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. Figure 6 shows the projection of the observations of sixteen 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.

In order to highlight interesting contrasts in some pixels, the observations in Figure 6 are color coded according to their timestamp, ranging from blue for the first observations to red for the most recent observations. Interestingly, some very different patterns appear to emerge: in some pixels, early and recent observations seem to be scattered randomly (e.g., second plot from the left on the second row). In others, there is a clear contrast in the data along one of both components and early and recent observations seem to form two separate clusters. A third pattern is a circular pattern in some pixels, for instance, in the pixels on the third row. Two of the plots (in dashed boxes) are further highlighted in Figure 7. The data points are now visualized by their relative timestamp. The same color coding as in Figure 6 applies.

The left plot in Figure 7 is an example of a pixel where the data shows a main contrast along time. Two well separated clusters are formed by roughly the first 200 and the last 150 obser-

vations. In the plot on the right, the observations corresponding to the first 12 timestamps are highlighted in a larger font for the purposes of illustration. As becomes clear, the observations are located sequentially next to each other in the circular pattern, and the circle is actually formed by 12 clusters of observations coming from each month of the year: observations from January form the first cluster, from February the second etc.

In order to distinguish between pixels with a contrast in PCA scores between early and recent observations and pixels that show either a circular or a more random pattern, a logistic regression classifier was used. The observations were labeled according to their timestamp: 0 for the first 200 observations, 1 for the last 153. The scores on the first two PCA coordinates were used as predictors and the classification accuracy was evaluated on a 20% held-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.

The highest classification accuracy is achieved in the tropical regions. This indicates that the pixels with a strong contrast between early and recent observations are located in these regions. The explanation for this is straightforward: the climate in the tropics does not show a seasonal cycle that is as pronounced as in other regions further away from the equator. As there are a lot of features containing a seasonal component, such as the extreme indices obtained from raw data, this seasonal variability naturally emerges in the PCA plots whenever it is present. In addition, the classification performance went up gradually as more principal components were added as predictors. Hence, in pixels with a strong seasonal cycle, it seems that the contrast between observations throughout time is masked, whereas this is not the case in the tropics.

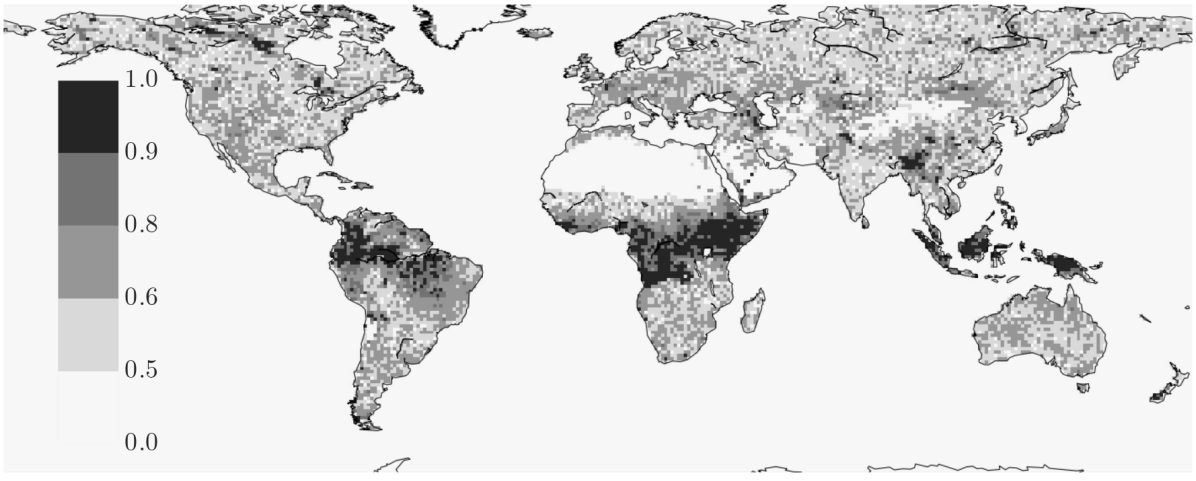


Figure 3.1: Proportion of test data correctly classified as early (first 200) or recent (last 153 months) by logistic regression, using the scores of the observations on the first two PCA dimensions as predictors.

This was verified by removing all variables with a seasonal component from the dataset and

rerunning the principle component analysis. As expected, the logistic regression classifier now achieves a higher accuracy in most cells (not shown), suggesting 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. This is caused by the presence of a trend in most time series, reflected in the climatic indices calculated on the deseasonalized data.

3.5 Summary

The analysis of the correlation between raw time series revealed that there is a strong degree of similarity between different temperature records and between different precipitation records in the dataset, in most regions in the world. However, this is mostly because of the strong seasonal component that is present in most of the raw signals. In regions without seasonal cycle such as the tropics, different measurement records of the same variable are much less correlated. The same is true for temperature and precipitation residuals, not only in the tropics but in most places on earth.

The NDVI residuals show fairly large autocorrelation at a temporal lag of 1 month. The autocorrelation drops down in most pixels for larger temporal lags, although some pixels in Australia still show an autocorrelation larger than 0.2 at a temporal lag of 4 months.

There are varying degrees of correlations between the climate variables and the NDVI residuals (Figure 5). In general, the highest correlations occur between NDVI residuals and climate observations from the same month or one month earlier. In addition, the vegetation appears to have a longer 'memory' for water-related variables such as soil moisture and precipitation than for temperature and radiation. In most pixels, the correlation between climate variables and NDVI residuals fades to zero for temporal lags larger than 6 months.

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 dataset.

4

Predicting NDVI anomalies

In light of the Granger causality framework, we are interested in the additional predictive performance for NDVI residuals of an extended model, which includes both climate variables and past NDVI residuals, on top of a purely autoregressive baseline model, which only uses past NDVI residuals as predictors. The goal of this chapter is to explore which regression models are best suited for the problem of predicting NDVI residuals with the extended model. The comparison with the corresponding baseline model and the applicability of both for a Granger causality analysis is discussed in the next chapter.

In order to reduce the complexity of the problem at hand, a reduced dataset was used for a first series of experiments. For every pixel on earth, the NDVI residuals were targeted for prediction with the following predictors:

1. Past NDVI residuals within a temporal window of 6 months.
2. One temperature (CRU) and one radiation (ERA) time series within a temporal window of 3 months.
3. One precipitation (MSWEP), one soil moisture time series (GLEAM) and one time series for snow (GLOBSNOW) within a temporal window of 6 months.

The choice of the different temporal windows is motivated by the correlation analysis from the previous section. In addition to the past climate, also the present observations of the climate variables were used as predictors (lag 0). As such, the data matrix in every pixel is of size

360×35: 360 monthly observations on 5 climatic time series and vegetation, with their history encoded as lagged variables for different temporal windows.

4.1 Modelling individual pixels

To begin with, every pixel was treated as individual problem, without considering spatial relations between pixels. I will also refer to this setting as the *single-task learning* setting, in contrast with the *multitask* setting in which models are trained on multiple pixels simultaneously. The multitask setting is discussed in the second part of this chapter.

4.1.1 Linear models

Linear regression, ridge regression and lasso regression were considered as linear models for predicting the NDVI residuals. Ridge regression and the lasso penalize large model coefficients and are known to generalize better to unseen data than linear regression in case of correlated predictors. In addition, the lasso brings about sparseness through feature selection by enforcing some model coefficients to become zero.

The performance of the linear models on the test set is summarized in Table 4.1, aggregated over all pixels globally. In some pixels the models have no predictive power at all. In these pixels the R^2 can get arbitrarily low, and the MAE arbitrarily high when testing the models on out-of-sample data. Because of some extreme outlying results in these pixels, the distributions of the global per pixel mean R^2 and per pixel MAE are heavily skewed and are not comparable by their means. The global median R^2 and MAE are reported instead in Table 4.1.

	5 fold random CV		Test set	
	MAE	R^2	MAE	R^2
Linear regression	0.689	0.068	0.796	0.075
Ridge regression	0.652	0.130	0.775	0.090
Lasso	0.647	0.137	0.764	0.102

Table 4.1: Global median performance metrics for linear models.

Linear regression is outperformed by ridge regression and the lasso, the latter of which performs slightly better than the former both in terms of MAE and R^2. The accuracy of all three models is slightly worse on the test set (consisting of the last 5 years of the time series) than the cross-validated accuracy estimate within the training set. This indicates that, because of the temporal structure of the data, random 5-fold cross-validation underestimates the out-of-sample error because training and validation samples are not completely independent. This phenomenon is known as 'data leakage' [26]. This is in line with what can be expected for temporal data.

Therefore, assessing the test error on the out-of-sample and out-of-time test set is probably the most conservative approach.

Figure 4.1 shows the test set R^2 of the lasso model for every pixel. Pixels with a negative R^2 (i.e., where the model has no predictive power at all) are all shown as white pixels.

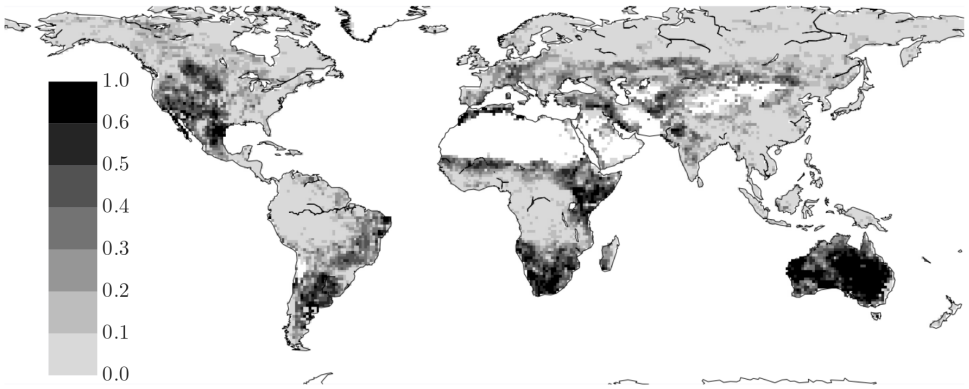


Figure 4.1: Mean lasso R^2 on the test set, per pixel. Negative R^2 values were truncated at zero to make the plot.

The model performs best in Australia, the south of Africa, the Sahel, parts of Europe, the Mediterranean and central Asia, and parts of North and Latin America. These regions are roughly the same regions where strong autocorrelation between the NDVI residuals was found (Figure 4), which indicates that the autoregressive part of the model (using the past NDVI residuals to predict the present) is important for achieving predictive power.

4.1.2 Non-linear models

In order to evaluate whether the NDVI predictions can be improved by using non-linear models, random forest and extreme learning machines were considered as more flexible model alternatives.

For the random forests models, the number of trees was set at 100 after evaluating the out-of-bag error. The number of variables to be considered at each split was set to $\frac{p}{3}$, with p the total number of predictors. This is the recommended default setting for regression with random forests and changes to this parameter did not improve predictive performance. An ELM requires two hyperparameters to be optimized: the number of nodes in the hidden layer and the amount of regularization when regressing the outputs of the hidden layer to predict the output layer. The optimization was done for a pixel selected from a region where the linear models performed well (latitude 23.5, longitude -100.5, located in North America). The ELM was fitted on the training part of the data (first 24 years) and the test error was estimated on the left-out test part. Train and test MSE were determined for hidden layer sizes ranging from 10 to 5000 nodes. Each time, the hyperparameter λ governing the amount of regularization was determined through efficient

leave-one-out cross-validation on the training data. The train and test MSE and λ for different hidden layer sizes are shown in Figure 4.2.

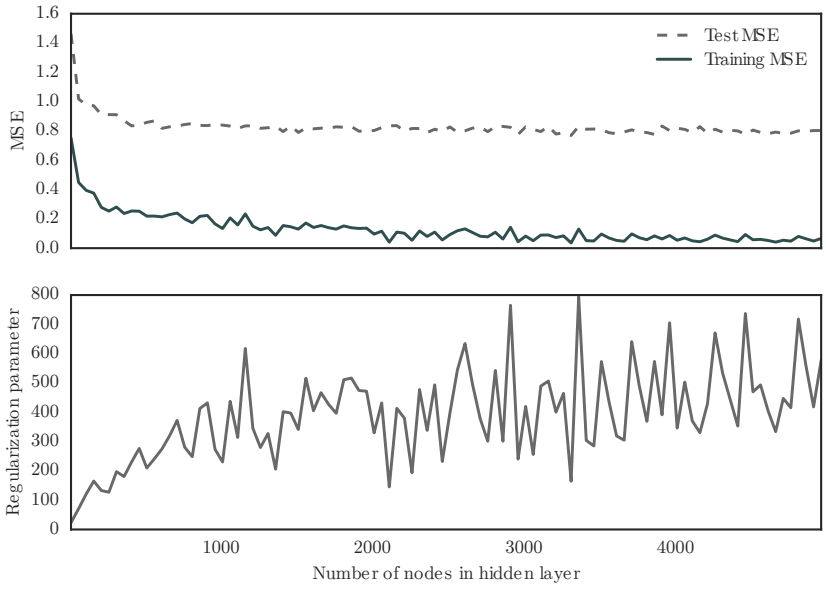


Figure 4.2: ELM optimization. Top: train and test MSE for an increasing number of nodes in the hidden layer. Bottom: cross-validated regularization parameter λ for increasing hidden layer size.

Initially, both train and test MSE quickly drop with an increasing number of hidden nodes. From 500 hidden nodes on, the test MSE stabilizes and a further increase in hidden layer size does no longer cause an improvement in model performance. At the same time, the optimal amount of regularization tends to increase with the size of the hidden layer. Intuitively, this indicates that the hidden layer should be large enough to yield enough interesting random combinations of the input features. At the same time, a larger number of nodes is dealt with by stronger regularization in the output layer. Based on this analysis, an ELM with 1000 nodes in the hidden layer was used and λ was optimized through cross-validation on the training part of the data separately for every pixel on earth.

Table 4.2 shows the performance metrics for both non-linear methods for predicting NDVI residuals.

	5 fold random CV		Test set	
	MAE	R^2	MAE	R^2
Random forests	0.653	0.129	0.784	0.074
ELM	0.722	0.005	0.956	-0.213

Table 4.2: Global median performance metrics for nonlinear models.

On the training set, random forests perform just slightly worse than the regularized linear regression models. However, the performance of random forests on the test set is much worse and is comparable with that of linear regression. The ELM performs bad on the training set and doesn't seem to generalize at all to the test set. In other words, because both models are much more flexible than the linear models, they tend to overfit on the training data given the small number of observations that is available for every pixel separately. As a consequence, they generalize poorly to unseen data.

In order to compare the different models from this first section and to give an idea of the spread of R^2 values over the different pixels worldwide, a kernel density estimation of the distribution of the R^2 values of the best linear model (lasso) and the random forests model on the test set is shown in Figure 4.3. In the same fashion as for plotting the R^2 on the map, negative R^2 values were truncated at zero, which explains the peak at zero for both curves. The distribution for the lasso is shifted to the right and has a much lower peak around zero, indicating that the lasso performs better than random forests in general.

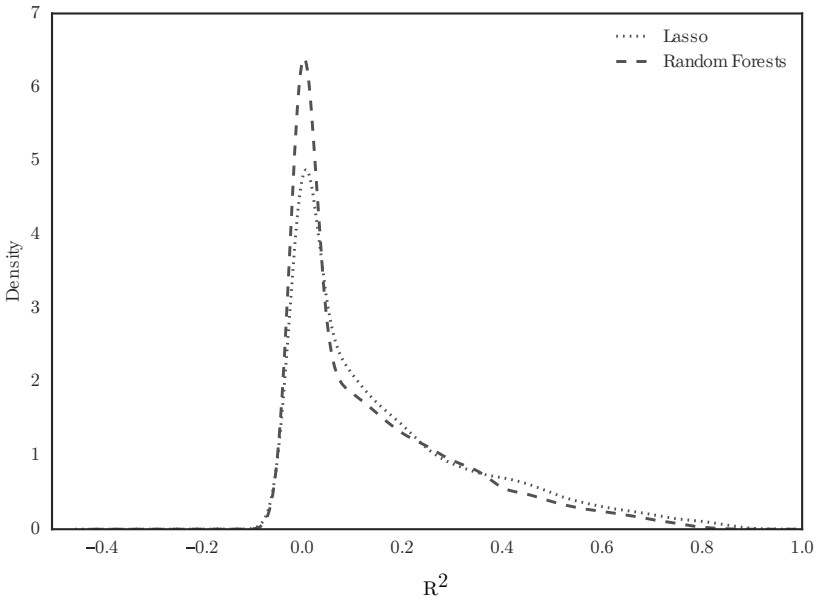


Figure 4.3: Kernel density estimations of the truncated R^2 for random forest and the lasso.

4.1.3 High-level features

The complete dataset that is available contains high-level features that were extracted from some climatic time series that are available with a daily resolution. These features represent cumulative and extreme events, and are supposed to represent the physical sensitivities of vegetation for climate events such as droughts and heat waves. In the same way as for the raw time series, the history of these features was encoded as lagged variables, respecting the same temporal

windows as before. There are a large number of these features, and including them strongly increases the dimensionality of the data. As such, there are about 3000 features for every pixel in this setting, depending on the amount of missing data.

Figure 4.4 shows the result for the lasso and random forests. Although the training set performance of ridge regression was similar to the previous section, ridge regression did not generalize at all to the test set in this high dimensional setting. While random forests do a better job on the test set than before, the best model in this setting is again lasso regression. The R^2 values with only the raw time series as predictors are plotted as dotted lines for reference in Figure 4.4. Although both models perform better with extended features than without, the sparse and most rigid model again comes out as the best model.

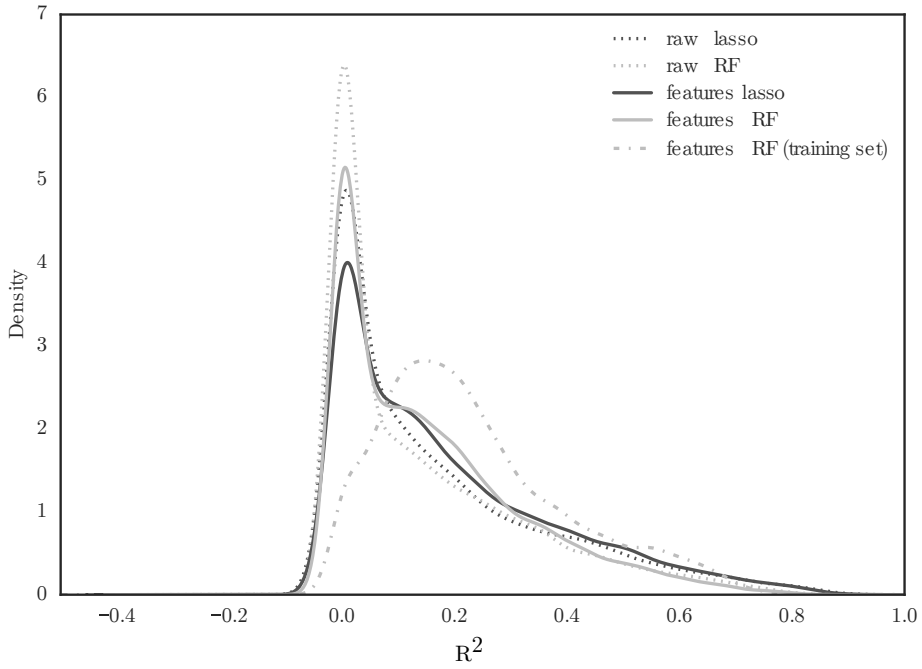


Figure 4.4: Improvement in R^2 when using high-level features in single task learning. Negative R^2 values were truncated at 0. For both the lasso and random forests, the dotted lines show the corresponding performance without the extended features. In addition, the random forests R^2 obtained through cross-validation on the training set is shown, clearly inflated as the effect of data leakage.

Surprisingly, the situation was the reverse for the R^2 values obtained through 5-fold random cross-validation: random forests outperformed both ridge regression and the lasso by a large extent. The cross-validated performance of random forests on the training set is shown as an additional line in Figure 4.4. It is remarkable that the two different evaluation schemes lead to such different conclusions. When performing random cross-validation, the random forests seems to benefit way more from data leakage than the lasso, resulting in a strong overestimation

of its generalization performance.

To recap: when treating every pixel as a separate problem, more rigid models such as ridge regression and the lasso are the best performing models. This is certainly true when the full extended feature set is used: because of the small number of observations per pixel, more flexible models are highly likely to overfit and their predictive performance breaks down on the test set. In the next part, the similarities that exist between the different pixels will be exploited by fitting models over multiple pixels simultaneously.

4.2 Multitask learning: modelling all pixels jointly

Multitask learning (MTL) is a modelling approach which seeks to improve generalization by using information from training signals from different tasks. The main idea is that a multitask learning model will perform better than single-task learning when the different tasks are related, by using information from all tasks simultaneously [4].

The dataset used in this thesis lends itself for multitask learning. It is likely that the data from neighboring pixels show similar patterns, because their climate and its relation with vegetation is likely to be similar. In the light of multitask learning, predicting vegetation in a particular pixel can be seen as one task. A multitask learning model should be able to benefit from the similarities between related pixels, and make better predictions compared to a single-task learning approach where every pixel is treated as a separate problem.

So far, a separate model was used for every pixel to predict the NDVI anomalies and no information was shared across pixels. A very simple multitask approach consists of concatenating the data from multiple pixels into one large dataframe. This is possible because all pixels share the same feature set. As such, the final dataframe has dimensions $N \times p$, with N the number of observations per pixel (360) times the number of pixels, and p the same number of features as before (35). This dataframe was constructed using all pixels on earth, so that it contained over 4 million observations in total. In addition, latitude and longitude were added as predictors, to allow models to differentiate predictions for different locations on earth.

The performance of the multitask models was evaluated in the same way as before: first through 5-fold cross-validation on the training set, after which the model was retrained on the complete training set and tested on the test set. The cross-validation was done in such a way that observations for one particular timestamp were either all in the training folds or all in the test folds. For example, the NDVI residual from July 1987 in the pixel covering Ghent should not be in the validation fold if the July 1987 NDVI residual from the pixel covering Kortrijk is in the training fold. The cross-validation procedure was adapted accordingly by grouping the observations per

timestamp and forming random folds on the group level instead of on the observation level. After training, predictions were made for all observations in the test set, which were grouped by latitude-longitude afterwards to obtain predictions on the individual pixel level. The usual performance criteria were calculated for every pixel separately.

Two MTL models were tested: MTL with ridge regression as a linear method and non-linear MTL with an ELM. Despite the poor performance of the ELM in the single-task learning approach, it was used here because it is much faster to train on a dataframe of this size than random forests. It was computationally not feasible to train a random forests model on the global dataframe with the resources available for this thesis. The results for both models are summarized in Table 4.3 and in Figure 4.5.

	5 fold random CV		Test set	
	MAE	R^2	MAE	R^2
Ridge MTL	0.592	0.163	0.717	0.146
ELM MTL	0.577	0.186	0.705	0.151

Table 4.3: Performance metrics for multitask learning methods

Globally, both MTL models perform better than any of the STL models from the previous section on the training set as well as on the test set. Surprisingly, the ELM turns out to be the best model so far, with a global median R^2 of 0.186. This is also visible in Figure 4.5 (the result from single-task ridge regression from the previous section is shown as a dotted line as a reference).

Figure 4.6 illustrates the difference between single-task ridge regression and the multitask ELM on the world map (the multitask ridge map was very similar to the multitask ELM map). Most of the improvement is made in pixels where the STL models performed very poorly: on the map, a lot of pixels that are white for the single-task model turn grey for the multitask learning model, notably in northern Canada, Europe and across central Asia. In other words, improvement by the MTL models is mainly made in predicting the hardest tasks. Pixels where the NDVI residuals are hard to predict based on their own data but that are close to pixels where the model works better, benefit most from the multitask approach.

Another observation that becomes clear from the world maps, is that the accuracy on the test set of the multitask learning model is worse in regions where the single-task learning model performed best. This is also visible in Figure 4.5: the distribution of the truncated R^2 values of the MTL models does not extend as far to the right as for single-task ridge regression. This is well visible in Australia: with the multitask model, the predictive performance has mostly gone down in this region. So, whereas pixels that were very hard in a single-task approach seem to

benefit from a multitask model, the multitask models perform a bit worse in pixels that were easy for the single-task models. It seems as if the performance is spread out over the pixels.

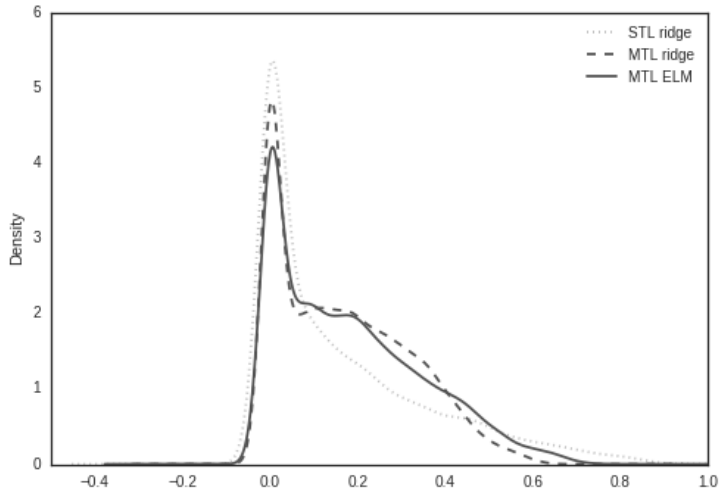
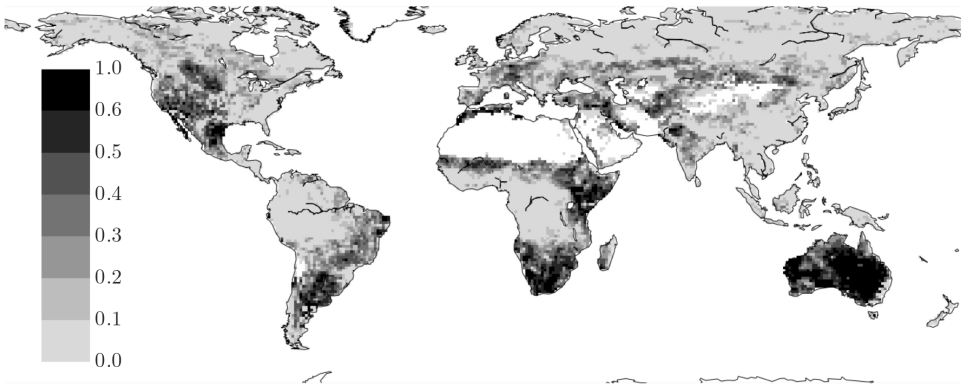
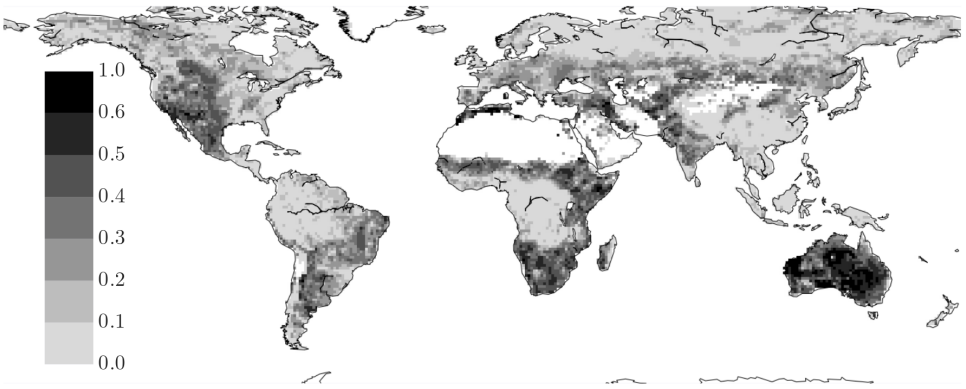


Figure 4.5: Kernel density estimate of the R^2 on the test set for STL ridge and both MTL models.



(a) Test set R^2 for single-task ridge regression.



(b) Test set R^2 for the multitask ELM.

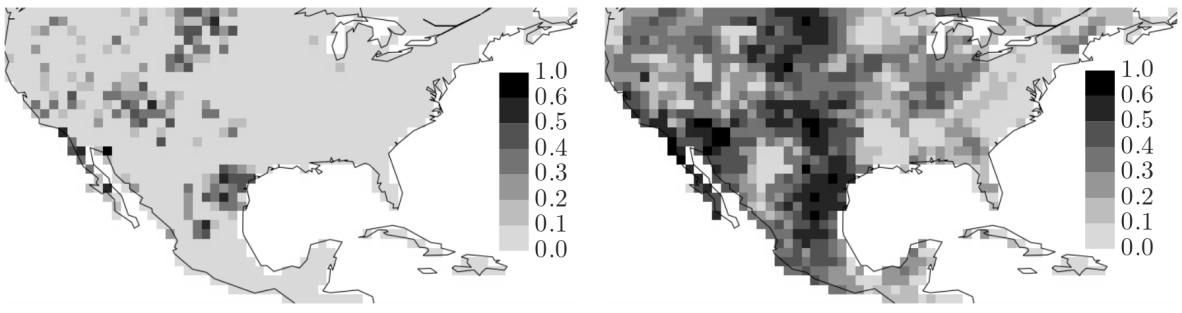
Figure 4.6: Comparison between single-task ridge regression and the multitask ELM on the test set. Negative R^2 values were set to zero for making the plots.

4.2.1 Extension with features from neighbors

Up until now, the feature space corresponding to every NDVI anomaly contained only variables from the native pixel where the anomaly was observed. In order to incorporate interactions between pixels, the feature space can be extended with the features from neighbouring pixels. This was done for a moving window of 9 pixels. For every observation, the features of the 8 closest neighbors were added as additional variables to the global dataframe, such that each observation now becomes a point in a 315-dimensional space (9 times 35 predictors). This dataframe was used to predict NDVI anomalies in the test set with the multitask ridge regression model from before. There was no improvement or decline in predictive performance when using the features from neighbouring pixels as additional features, compared to the multitask ridge model from the previous section. The results are therefore not shown.

4.2.2 Multitask learning with high-level features

The use of high-level features from daily times series led to a better predictive performance for some models, with the most sparse model (the lasso) achieving the best predictions. However, because of the large dimensionality of these predictors and the large degree of correlation between them, there were not enough observations for ridge regression to learn adequate weights and it broke down on the test set. By training models on many more observations, it is likely that less rigid models than the lasso will do better in the multitask setting. Unfortunately, it was not possible to run a multitask learning model for the whole world using the complete feature set because of computational constraints. Therefore, the multitask ridge model (being the fastest of the two MTL models) was trained on part of Middle and North America. Figure 4.7 (b) shows the result on the test set. The panel on the left-hand side shows the test set performance of single-task ridge regression with the extended features.



(a) Single-task learning *(b) Multitask learning*

Figure 4.7: Left: test set R^2 of single task ridge regression in the extended feature space. Right: test set R^2 of multitask ridge regression, with the same features.

The difference is clear: where ridge regression couldn't deal with the high dimensional $n << p$ setting in single-task learning, the multitask ridge model is able to learn a function that gener-

alizes much better and seems to give better predictions than the multitask models that did not have access to the extended climate features (Figure 4.6 (b)).

4.3 Summary of modelling results

Table 4.4 summarizes the model results from this chapter.

Single-task models

	5 fold random CV		Test set	
	MAE	R^2	MAE	R^2
Linear regression	0.689	0.068	0.796	0.075
Ridge regression	0.652	0.130	0.775	0.090
Lasso	**0.647**	**0.137**	**0.764**	**0.102**
Random Forests	0.653	0.129	0.784	0.074
ELM	0.722	0.005	0.956	-0.213

Multitask models

	5 fold random CV		Test set	
	MAE	R^2	MAE	R^2
Ridge MTL	0.592	0.163	0.717	0.146
ELM MTL	**0.577**	**0.186**	**<u>0.705</u>**	**<u>0.151</u>**

Single-task with extended features

	5 fold random CV		Test set	
	MAE	R^2	MAE	R^2
Lasso	0.569	0.176	**0.733**	**0.133**
Random Forests	**0.554**	**0.204**	0.750	0.114

Table 4.4: Global median performance metrics for different models used in this project. Bold font: best performing model for the specific model setting and validation scheme. Bold and underlined: best performing model overall on the test set.

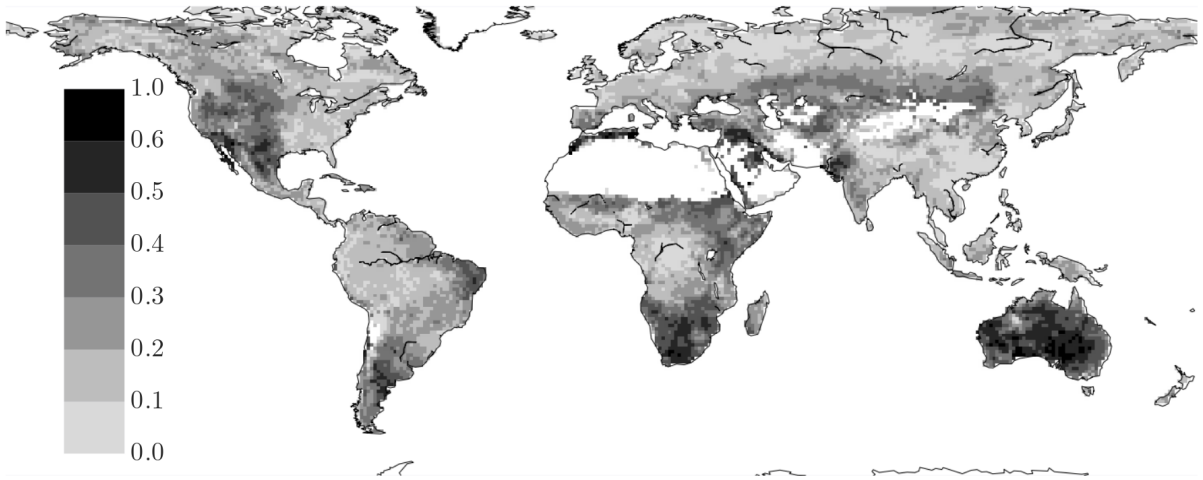
Of all the single-task models, the lasso performed best, with ridge regression a close runner-up. The limited amount of observations per pixel when performing single-task learning favors more rigid models like ridge regression and the lasso in terms of generalization to unseen data. This became clear by looking at the poor test set performance of the more flexible random forests and the ELM.

Using the idea of multitask learning to predict vegetation anomalies led to a global improvement in predictive performance, mostly in pixels that were very hard to predict accurately with a single-task approach. In contrast, the multitask learning test set performance seemed to suffer a bit in the pixels where single-task models performed very good; the multitask learning maps look much smoother and the results seem to be averaged out over the map. Nevertheless, because of the much larger amount of training observations that become available when using multitask learning, more flexible models can be applied. The ELM came out as the best performing model in this setting. The ELM is non-linear in the sense that it operates in a non-linear randomized feature space, but it can be trained with the speed of a linear-model. As such, it has a large advantage over random forests in term of training speed and memory consumption.

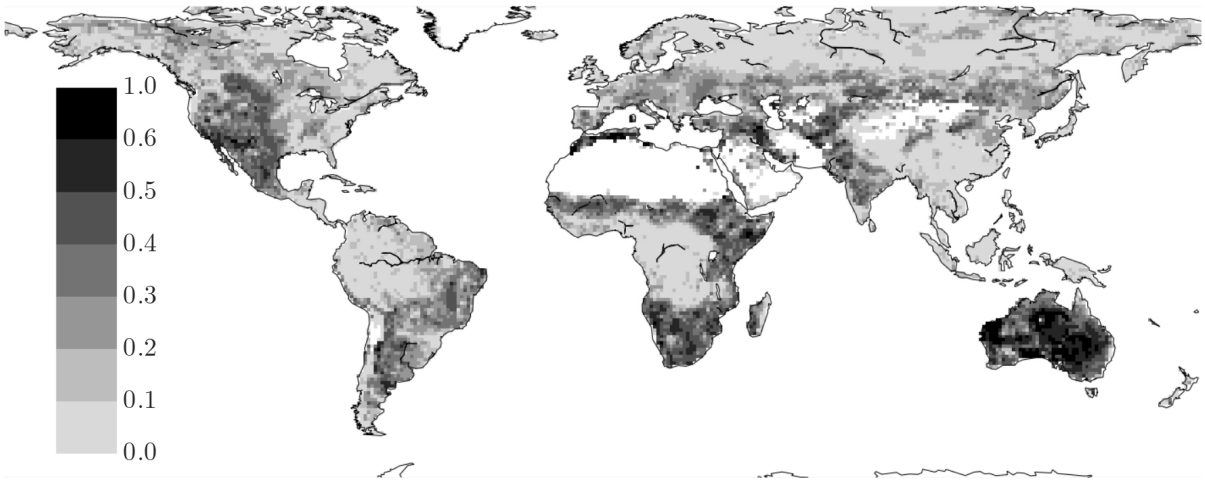
The addition of high-level features, extracted from daily-resolution time series and reflecting extreme and cumulative climate events, led to better predictions in single-task models. However, because of the high-dimensional setting (over 3000 features for 360 observations per pixel) and the large degree of correlation between the different predictors, the lasso outperformed all other methods on the validation set by enforcing sparsity in the model. This problem was solved in the multitask setting. Multitask ridge with the extended climate features generalized much better to the test set than single-task ridge and, at least on a part of the world, showed a promising improvement over the other multitask methods that didn't use these features.

Although both vegetation and climate variables show positive correlation with the same variables in pixels up to 10 spatial lag units away (not shown), extending the feature space with climatic features from neighboring pixels did not lead to better predictions. In other words, given the past vegetation and climate of a pixel, the past vegetation and climate of its neighbors contain no extra information to forecast the present state of the vegetation.

Finally, it is important to note that all models performed worse on the out-of-sample and out-of-time test set compared to their validation set performance estimated through random 5-fold cross-validation on the training part of the data. This is an illustration of data leakage, and occurs because observations in the training and validation folds are correlated when doing random cross-validation on time series data. Although most conclusions from this chapter would not fundamentally change when looking at the model performance on the independent test set in stead of on the training set, the differences between both approaches still serve to show the importance of doing proper model validation. Figure 4.8 illustrates this for the multitask ELM. Estimating the model performance on unseen data through random cross-validation is much more optimistic than estimating it on the independent test set.



(a) Cross-validated training set R^2 for the multitask ELM.



(b) Test set R^2 for the multitask ELM.

Figure 4.8: Cross-validated training set performance (top) and test set performance (bottom) of the multitask ELM, the best performing model in terms of global median R^2 and MAE.

5

Granger causal inference

So far, different learning models were explored to predict vegetation residuals using past climate and vegetation variables. This chapter will explore how to use these models to address questions related to causality in the framework of Granger causality. To recap, Granger causality from a variable x to an outcome y can be tested for by comparing the predictive performance of two nested models: a baseline autoregressive model predicting y at time t as $y_t = f_1(y_{t-1}, ..., y_{t-p})$, and an extended model $y_t = f_2(y_{t-1}, ..., y_{t-p}, x_{t-1}, ..., x_{t-q})$. Generally, the null hypothesis of Granger non-causality is formulated as the null hypothesis that f_1 and f_2 have equal prediction error (in econometrics: *forecasting accuracy*). Typically the alternative is one-sided, such that if f_2 predicts y_t significantly better than f_1, H_0 is rejected.

In some applications, inference is drawn in linear vector autoregressive models by testing for significance of individual model parameters. Other studies have used likelihood-ratio tests (f_1 and f_2 are nested) [34]. However, in both cases the model is trained and evaluated on the same data (in-sample). As pointed out by several authors, the performance of any Granger causal model should be validated on out-of-sample data to avoid overfitting [3, 5] [1]. The null hypothesis of non-causality in the formulation stated above can be tested by comparing out-of-sample prediction errors. Statistical tests to this end have been proposed and applied both in the econometric literature and in Granger causality studies in the context of climate science. Tests to compare out-of-sample MSE are available for models for which parameter estimation is

[1] In the paper by Attanasio et al., the authors mention avoiding the consequences of incorrectly establishing the stochastic properties of the time series as another reason to perform out-of-sample testing.

done through ordinary least squares or maximum likelihood estimation [3]. 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 as well [5].

Unfortunately, all the tests mentioned above were designed to compare the out-of-sample prediction errors of linear parametric models [32]. Although regularized linear models were mostly the best performing models in the previous chapter, more flexible models such as random forests were close competitors in some settings. Furthermore, it has been argued that relations in climate datasets tend to become more non-linear as the temporal resolution of the data becomes finer [3]. Thus, in future experiments or with different data, other models than linear parametric models could turn out to be better for predicting vegetation anomalies. 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-free so that any non-linear model f could be used.

One well-known model-free test to compare the accuracy of two forecasts is the *Diebold-Mariano* test (DM-test) [7]. The DM-test compares the errors of two competing forecasts $\hat{y}_{1t}$ and $\hat{y}_{2t}; t = 1, ..., T$, where the forecast errors are defined as $e_{it} = \hat{y}_{it} - y_t; i = 1, 2$. The test statistic is defined using a loss differential $d_t = g(e_{1t}) - g(e_{2t})$, with $g()$ typically the squared-error loss: $g(e_{it}) = e_{it}^2$. The two forecasts are equally accurate if d_t has zero expectation for all t:

$$H_0 : E(d_t) = 0 \ \ \forall t$$

The DM test statistic is of the form:

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \tag{5.1}$$

Where $\bar{d}$ is the sample mean of the loss differential and $\hat{f}_d(0)$ is a consistent estimate of the spectral density of the loss differential at frequency 0. Under H_0, DM is asymptotically $N(0, 1)$ distributed. The main assumption for this to hold is that the loss differential is covariance stationary, or $cov(d_{12t}, d_{12(t-\tau)}) = \gamma(\tau), \forall t)$. The DM test is an asymptotic test of the hypothesis that the mean of a series of loss differentials is zero, with the calculation of the standard error accounting for the autocorrelation of subsequent loss differentials [7].

Although it looks like a promising test to apply as a Granger causality test, there is one big caveat: the DM test does not hold for nested models, because under the null, the forecast errors from two nested models are exactly the same and perfectly correlated, which means that both the numerator and the denominator of the test asymptotically tend to zero [32]. However, it has

been argued that the Diebold-Mariano test remains asymptotically valid even for nested models when the size of the estimation sample remains finite as the size of the prediction sample grows, under some regularity assumptions [18].

Although the DM-test was not designed for the purpose of comparing forecasts from nested models, it is a model-free test that could be used to compare two forecasts originating from any model, in contrast with other out-of-sample MSE tests that are specifically designed for linear models. 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 [8]. Methods based on the bootstrap have been used before in Granger causality studies with climate data [9, 3]. For this thesis, I will both illustrate the performance of the DM-test and propose some ideas to use the bootstrap to come to inference about Granger causal relations in the climate data set.

5.1 Quantitative evidence of Granger causality

For this chapter, a single-task ridge regression model was used to predict the NDVI residuals from the test set, using two models: an autoregressive baseline model using the past 6 vegetation anomalies as predictors, and an extended model which uses the history of one precipitation time series (MSWEP) as additional predictors. Figure 5.1 shows the difference in R^2 between the extended model and the baseline model, wherever it is larger than zero. Pixels where the R^2 of the baseline model was negative are not considered. The performance of the extended model is better in some large regions over the globe, notably Australia, Somalia and the south of Africa, the east of Latin America and the western part of North America.

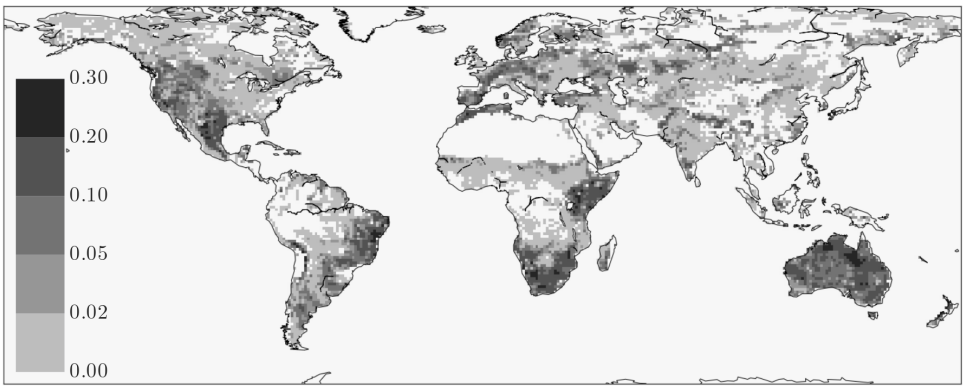


Figure 5.1: Quantitative evidence for Granger causality between the baseline and the extended model in terms of difference in R^2.

Wherever there is grey on the map, the extended model outperforms the baseline model. A natural question that arises now is how significant the differences between the extended and the

baseline model are and whether or not we should conclude Granger causality for every pixel where the extended model performs better. This question is addressed next.

5.2 The Diebold-Mariano test

The Diebold-Mariano test is available in the R-package `forecast`. The forecast errors of both models were obtained for all pixels wherever the baseline model achieved a positive R^2, and the test was performed using the squared loss differential and against the one-sided alternative which states that the extended model provides a better forecast than the baseline model. Figure 5.2 shows the result, with pixels producing a significant p-value colored in black. The DM test was performed for 9699 pixels, 1955 of which produced a p-value lower than 0.05.

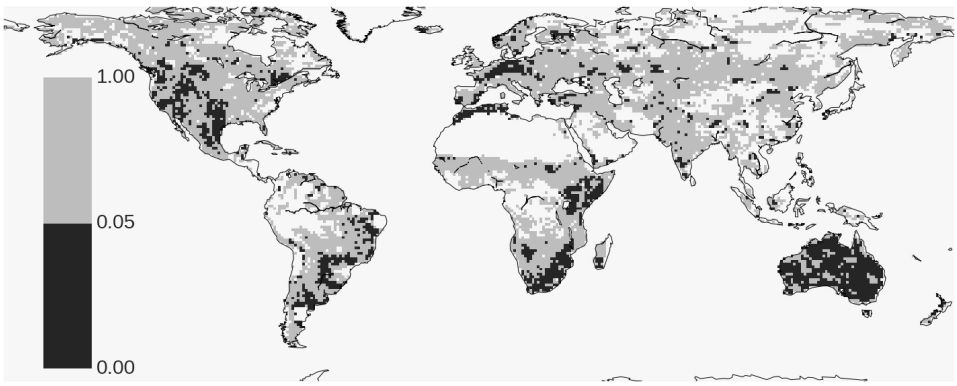


Figure 5.2: P-values obtained with the Diebold-Mariano test to test the null hypothesis of Granger non-causality.

Obviously, there is a multiple-testing problem here. When repeating the same test this many times, the probability of false discoveries is almost equal to 1. The most conservative way to correct for multiple testing is by controlling the family-wise error rate (FWER) through the Bonferroni correction. This can be corrected by adjusting the nominal α level at which each individual test should be performed:

$$\alpha_{bonferroni} = 1 - (1 - \text{FWER})^{\frac{1}{m}}, \tag{5.2}$$

With FWER or the family-wise error rate set at 0.05 and m equal to the total number of pixels that are tested.

After performing the Bonferroni correction, the null was rejected in only a few pixels. With m as large as it is in this case, the Bonferroni correction is extremely conservative, and probably

there are better solutions to solve the multiple-testing problem. Some perspectives for better solutions are provided in the last chapter.

5.3 Resampling methods

In machine learning, it is common to compare the predictive performance of two different algorithms on the same data by using 5×2 cross-validation [8]. The idea is to obtain multiple estimates of the performance difference of the two models by repeatedly training and testing them on different parts of the data. Afterwards, a paired t-test or, more generally, a non-parametric test such as the Wilcoxon signed rank test is used to test whether the difference in performance of the two models is equal to zero or not [1].

There are two problems with this approach if we want to apply them in the setting of this thesis. First of all, using random cross-validation to form train and test sets will introduce the same problems with overfitting that were highlighted in the previous chapter: for each fold, the test and training sets will be too correlated. Secondly, the models are retrained in every fold. This is not feasible for more complex models that require a long training time.

Using the same idea of validating the models on an out-of-time test set as before, I will try the following approach: first, both the baseline and the extended model are trained on the training part of the data. Their predictive performance is then evaluated on the test set: not once, but multiple times by creating multiple test set replicates using the bootstrap. In this way, two empirical error distributions of the models on the test set can be constructed. Finally, a statistical test such as the Wilcoxon signed rank test can be used to test whether the error of both models on the test set is equal, or not. In this way, overfitting of the models is avoided by only using data from the test set to bootstrap. Secondly, the models have to be trained only once. Once they are trained, generating multiple predictions on the bootstrapped test sets can be done in a fast way.

5.3.1 Bootstrapping time series

Computationally intensive methods such as the bootstrap can provide an alternative to traditional statistical inference. The basic idea behind these methods is to estimate the true distribution D that underlies a data sample X_n $(x_1, ..., x_n)$ by an empirical distribution $\hat{D}$, in order to get more information about the variability of a certain quantity of interest which is a property of D, $\theta(D)$, and is estimated as a function of the sample: $\hat{\theta} = \hat{\theta}(X_n)$.[2] In practice, constructing $\hat{D}$ means creating multiple replicate data sets X_n^* by resampling from the original sample X_n,

[2]For example, θ could be the mean of a Gaussian distribution which can be estimated by taking the sample average of a sample X_n.

from which multiple replicate estimates $\hat{\theta}^*$ can be obtained. The (discrete) distribution of the $\hat{\theta}^*$ values can then be used to derive information about the variability of $\hat{\theta}$, to construct confidence intervals or to perform hypothesis testing in which case $\hat{\theta}$ is a test statistic [6].

In the setting where the data sample X_n consists of independent and identically distributed observations, a bootstrapped sample X_n^* is obtained by drawing n samples from X_n with replacement. However, this does not apply for time series data: the observations are correlated through time and not independent. Simply drawing observations with replacement from time series data would ignore the time dependence structure. An alternative sampling approach is by assuming a data-generating model (for instance, an autoregressive process of order p). In this approach, the model errors ϵ_i are resampled to produce bootstrap error terms, which are then used to produce bootstrap observations [13]. Another approach that avoids the assumption of a data-generating process is so-called block bootstrapping, in which blocks of data are resampled instead of individual data points. The correct choice of the block length is important: too short blocks will destroy the temporal structure of the data, blocks that are too long won't allow for bootstrap samples to be variable enough or for drawing a sufficient number of bootstrap samples at all. However, because of the typical seasonal cycle of most climate phenomena in large parts of the world, a straightforward option for climate data is to use blocks with a length of one year. As such, the temporal structure remains conserved: an observation from December will always be followed by an observation from January. And with 29 complete years in the dataset, there are 29^{29} possible samples when drawing with replacement (note that the ordering matters).

Figures 5.3 and 5.4 respectively show an original precipitation and temperature time series and one block-bootstrapped replicate for a randomly selected pixel (latitude -31.5, longitude -71.5, located in the Indian Ocean). The bootstrapped precipitation data look very different than the original data, but both share the same characteristics: only positive values occur, and precipitation peaks occur during the same seasons. The same is true for the bootstrapped temperature time series: although it is different from the original data, it still looks like a plausible temperature time series. When using bootstrapped time series, the implicit assumption is made that every bootstrapped data sample is an equally likely realization of the underlying stochastic process that generated the original sample.

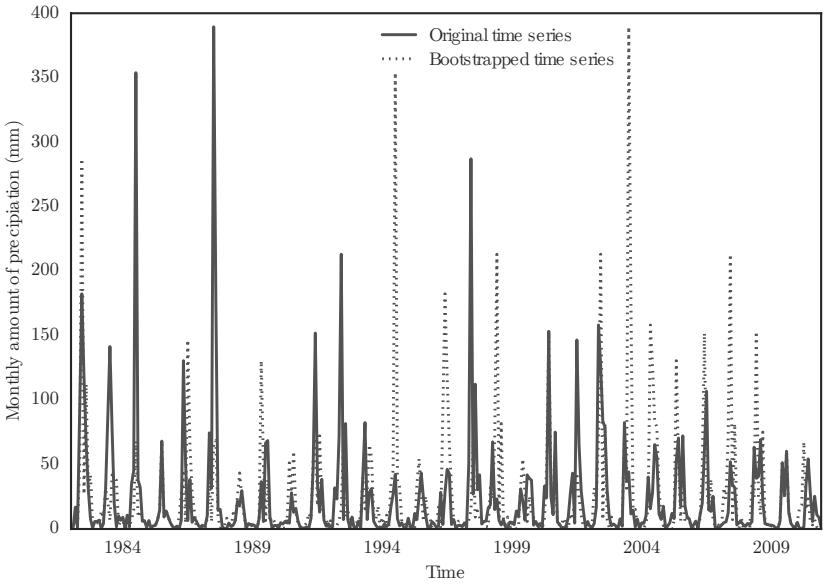


Figure 5.3: Original and bootstrapped precipitation data from a randomly selected pixel (-31.5,-78.5).

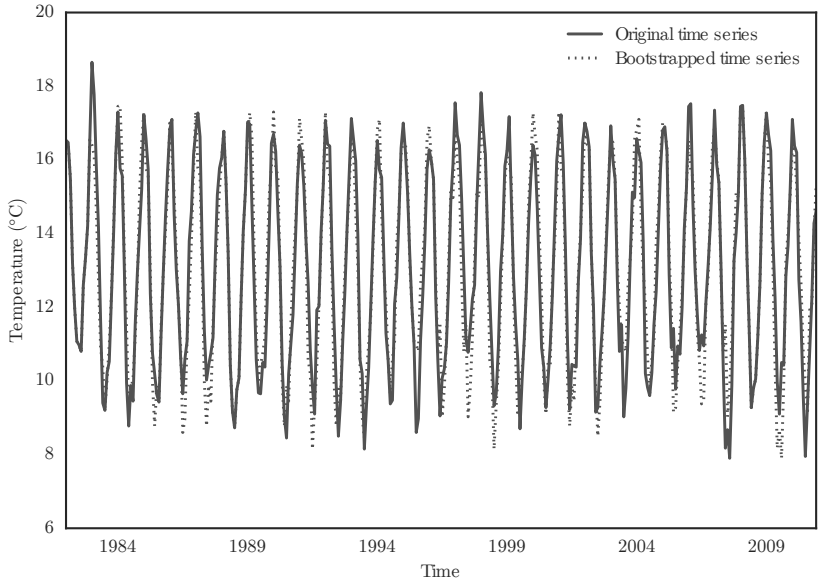


Figure 5.4: Original and bootstrapped temperature data from a randomly selected pixel (-31.5,-78.5).

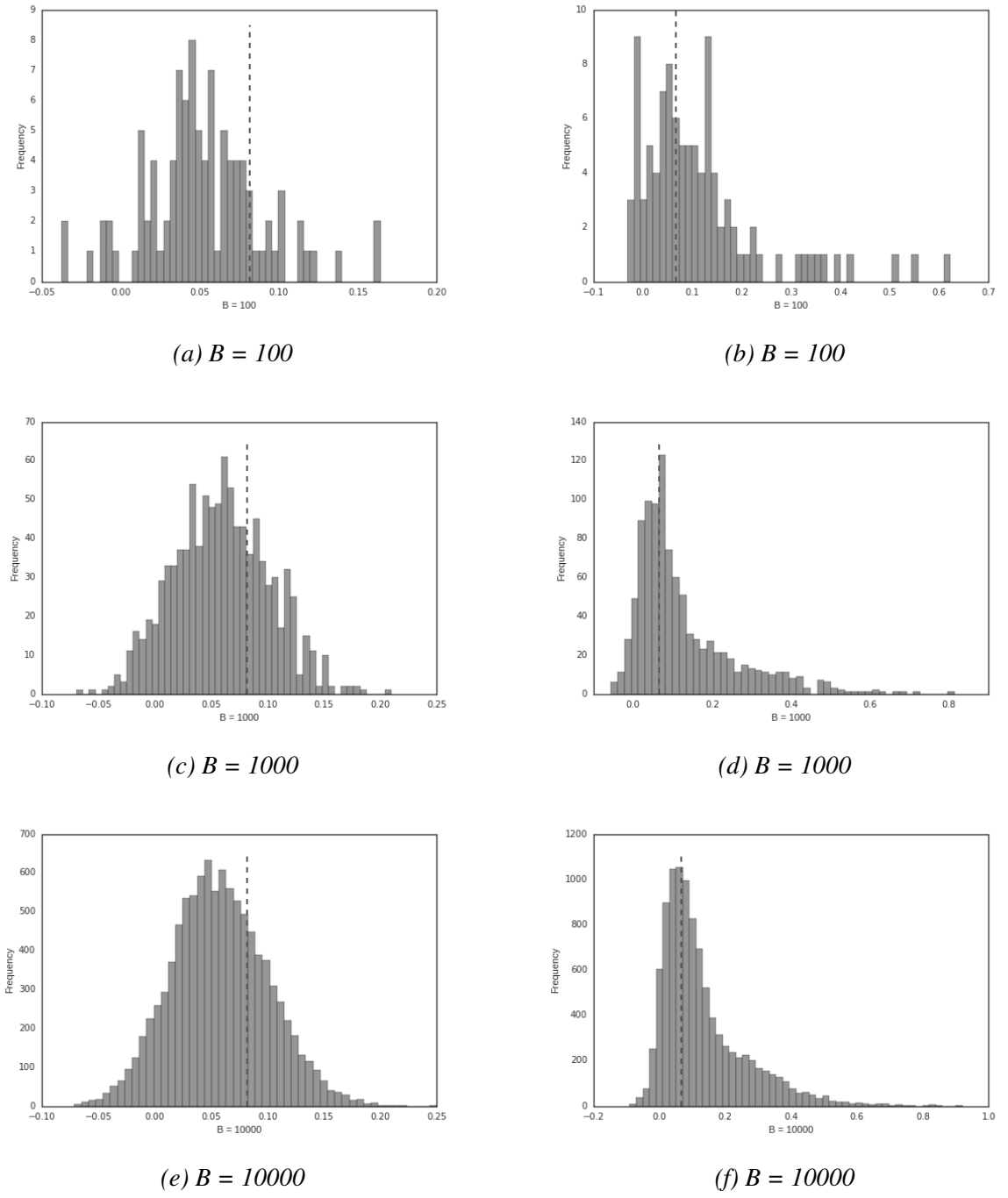
5.3.2 Variability of the squared loss differential with the bootstrap

Because of its simplicity, the squared loss differential d_t from the Diebold-Mariano test is appealing as a test statistic. However, it was not intended for nested models in the first place, and it is unlikely that the assumptions for using the test are fulfilled for every pixel. Perhaps the bootstrap can provide an alternative: by bootstrapping multiple test sets, the performance of

both models can be evaluated multiple times and we can get an idea about the distribution of d_t.

In order to assess the required number of bootstrap replicates B to get an acceptable idea of the distribution of d_t, its empirical distribution was first constructed for two pixels: one in South Africa where the extended model performed much better than the baseline in most pixels, and one in eastern China where the extended model was only slightly better (see also Figure 5.1). This was done for different values of B. The result is shown in Figure 5.5. The approach was as follows: first, the baseline and extended model were fitted on the original training data sets. Next, B block-bootstrapped replicates of the validation set were constructed and used to evaluate the test error of both models, using blocks of 12 months. As such, B squared test differentials d_t^* were obtained, yielding an empirical distribution for d_t. The d_t that was observed with the original test data set is shown as a dotted line. Note that a similar value for d_t was observed in both pixels. Where the result is not smooth at all for 100 bootstrapped replicates, it looks acceptable for 1000 replicates and is quite smooth for 10000 replicates. Remarkably, the distribution of the d_t^* values resembles a Gaussian distribution in the case where the performance difference between both models is small, but looks more like a chi-squared distribution in the second pixel. A possible explanation for this is the following: when both models perform equally well, the squared losses will mostly cancel out each other with some differences that are symmetric around zero. Whenever the extended model outperforms the baseline model, however, the squared loss differential will be dominated by the squared error of the baseline model and will behave more as shown in the right-hand side panels of Figure 5.5.

To obtain each of the bootstrapped d_t^* values, the baseline and extended model were used to make predictions on the same bootstrap replicate of the test set, so that the experiment can be seen as a paired experiment. Recall that each d_t^* is just the difference between the squared error of the baseline and the extended model on particular bootstrapped test set replicate, $e_1^2 - e_2^2$. Testing if the mean of these two error series are different from each other is equivalent to the null hypothesis of the DM-test which tests if the mean of d_t is zero. This could be done using a paired t-test. However, because of the variable shape of the d_t^* distribution for different pixels, a non-parametric paired test such as the Wilcoxon signed rank might be more appropriate. This tests allows to test the null hypothesis that the difference between two series of paired quantities has a distribution symmetric about zero against the one-sided alternative that the distribution of one series is shifted to the right of the other series. Because this test uses the ranks of observations in a pooled sample to compare two distributions, it is insensitive to outliers and it is free of distributional assumptions. If furthermore the assumption is made that the two distributions have the same shape (the *location-shift* assumption, or $F_1(x) = F_2(x - \delta), \forall x$), then the alternative hypothesis can be formulated in terms of the means of the distributions (e.g., $\mu_1 \geq \mu_2$).



(a) B = 100

(b) B = 100

(c) B = 1000

(d) B = 1000

(e) B = 10000

(f) B = 10000

Figure 5.5: Empirical distribution of the squared loss differential between the baseline and the extended model, for two different pixels and for increasing bootstramp sample sizes B. Left: latitude 46.5, longitude 94.5 (western China). Right: latitude -24.5, longitude 17.5 (South Africa). The dashed line shows the squared loss differential observed with the original data.

The Wilcoxon signed rank test was used to test the null hypothesis that the distributions of the mean squared prediction errors of the baseline and the extended model are equal, against the one-sided alternative that the baseline model is worse (errors shifted to the right), for all pixels where the the extended performed better than the baseline and achieved a positive R^2. This was done by generating $B = 1000$ bootstrapped validation sets, each time storing the errors of both

models [3]. Figure 5.6 shows significance of pixels at the 95% level of significance, after applying the Bonferroni correction. The R^2 on the test set was larger for the extended model than for the baseline model in all colored pixels, but a significant shift between the error distributions of both models (in favor of the extended model) was only found for the black pixels. When this map is compared with Figure 5.1, the pixels that produce significant results are indeed those where the difference in R^2 was largest in favor of the extended model. In those pixels, the null hypothesis that the distributions of the model errors of both errors are the same, can be rejected in favor of the alternative, which states that the distribution of the errors of the baseline model is shifted towards higher values. However, these results need to be interpreted with care. By using 1000 bootstrap samples, the power of the Wilcoxon test is large and it is able to pick up even the smallest differences, perhaps even differences that are non-relevant towards the research question of Granger causality. This might explain why the result is still significant in so many pixels, even after applying the extremely conservative Bonferroni correction. On the other hand, many pixels where the extended model performed better than the baseline in the original experiment, for instance in the north of Canada, turn out to be non-significant.

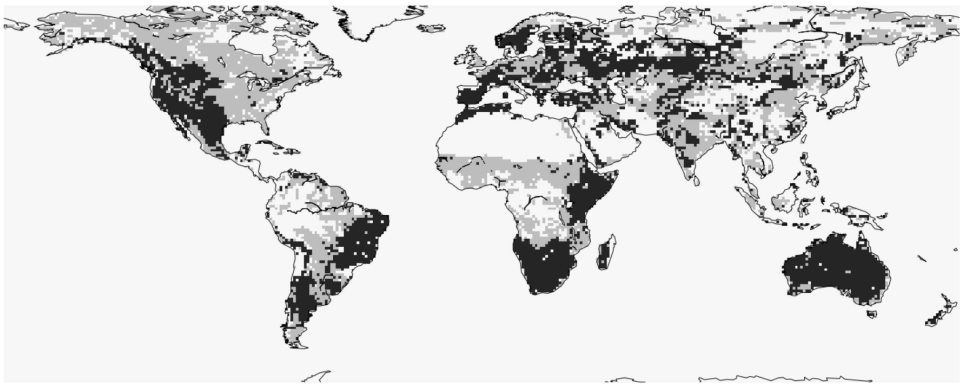


Figure 5.6: Results from the Wilcoxon signed rank test. The extended model performed better than the baseline in light-grey pixels, but a significant shift between the error distributions in favor of the extended model was only detected in pixels with a dark color.

[3]This took 65 minutes of parallel computing on 99 nodes of the Ugent tier-2 computing cluster Delcatty.

6
Conclusion

The goal of this thesis was to explore the potential of a climatic dataset for statistical causal discovery using Granger causality. The residuals of vegetation (NDVI) measurements acted as the target variable of interest, with a whole set of climatic variables as potential causal candidates for explaining changes in vegetation.

In a first set of experiments, predicting NDVI anomalies was treated as a separate problem for every pixel. This was referred to as the single-task learning setting. Linear regression did not perform well in this setting. In contrast, turning to regularized regression models such as ridge regression or the lasso led to large improvements in predictive performance. Because recent work has suggested the good performance of random forests on a similar but more extensive dataset, more flexible non-linear models were also explored [36]. In particular, extreme learning machines were used, because of their known performance in terms of training speed on large datasets. However, in the single-task setting, both non-linear models performed worse on the test set than the linear models. The ELM in particular did not perform at all, neither on the training set nor on the test set.

When exploiting the similarities between all the pixels in a simple multitask model, the predictive performance increased mostly in those pixels that were hard to predict in the single-task setting. However, there seemed to be a smoothing effect of the overall performance, as the R^2 went down in those pixels where the single-task models were very accurate. Probably, the large number of observations that are available in a multitask setting (over 4 million observations for

the whole globe), have some sort of regularizing effect and allow for much better generalization to unseen data. In contrast with the single-task setting, the extreme learning machine turned out to be the best model for multitask learning. Because of the large number of observations, the ELM was able to benefit from the rich non-linear feature representation in its hidden layer, without overfitting.

The use of high-level features, representing cumulatives and extremes, led to a strong improvement of the predictive performance for single-task random forests and the lasso. Because of the very high dimensional $n << p$ setting, ridge regression broke down when evaluating it on the test set. Because of computational constraints, a multitask model was trained on parts of the American continent rather than on the whole world. The fast multitask ridge model performed very well in this setting, both in comparison with the single-task ridge model with the extended features and with the multitask models without the features. This illustrates the potential of the extreme and cumulative indices to act as qualitative predictors for changes in the NDVI, provided that enough observations are available to adequately train a model.

Unsurprisingly, all models performed better when evaluated by means of 5-fold random cross-validation on the training set than by means of evaluation on the independent test set. This suggests overfitting of the models in the former case, since training and validation samples are too correlated. This illustrates the importance of validating model performance in an appropriate way, depending on the goal of the experiment and keeping in mind the specific structure of the data at hand to avoid data leakage.

Overall, the results from the modelling chapter suggest that there is a potential for statistical learning models that go beyond simple linear regression to model climate-vegetation interactions. However, some statistical challenges arise when framing this extension in the context of Granger causality. These were explored in Chapter 5.

Most statistical tests that were developed in the econometric literature for comparing the out-of-sample accuracy of two forecasts were either developed with assumptions on the model structure, or were not designed to be used for nested models that are typically used in Granger causality experiments. Besides those tests, some validation schemes frequently used to compare machine learning models such as 5-by-2 cross-validation are available as well. However, these schemes rely on random cross-validation to construct training and validation folds. Furthermore, they require the models to be re-trained for every fold, which is not feasible on a large dataset like the one used in this thesis for more complex models.

One possible approach for comparing the performance of the baseline and the extended Granger models was proposed in Chapter 5. While this approach is maybe not the most correct way of

solving the problem, it deals with the issue of validating the models on out-of-sample data and it requires the models to be trained only once. The bootstrap was used to generate multiple test sets, in order to get an idea of the distribution of the errors of both models. These distributions were then compared with a non-parametric Wilcoxon test. Although this allowed to exclude large regions of pixels where the performance of both models differed only by a small margin, the test can detect smaller and smaller difference by generating more bootstrap samples, to the extent that it may produce significant results in pixels where the difference between both models is of no practical importance. Furthermore, the obvious problem of multiple testing was simply dealt with in the most conservative way by applying the Bonferroni correction.

Further research

The field of climate science is an extremely data-rich research domain, with a lot of progress to be made using data mining, machine learning and statistical modelling. The results from this thesis point to some interesting perspectives for future research with respect to the dataset that was used.

Perspectives for modelling the climate data

- More advanced multitask learning models could strongly benefit from the similarities between the climate in different pixels.
- The performance improvements achieved by using the high-level features representing cumulatives and extremes illustrate the potential of using automatic feature extraction methods, or even models that can learn high-level features on their own from raw daily data.
- Learning from multiple tasks and learning high-level features could be combined in an artificial neural network architecture with a convolutional layer and multiple hidden layers. When also taking the time component into account, even more advanced models with multiple convolutions in both space and time could be used to model this data. Using this approach, the data could be modelled as a video. Instead of red-green-blue channels in a typical video, the different channels would be precipitation-temperature-radiation etc.
- In the dataset, there are multiple time series available for temperature and precipitation, originating from multiple satellites or data sources. In image classification tasks, noise is often added on purpose to input images by applying rotations or rescaling in order to make the algorithms more robust. The same idea could be applied to an algorithm that uses climate data to predict vegetation: by using the different data sources in a clever way, perhaps the model could be made more robust against satellite measurement noise.

Perspectives for Granger causality inference

- More work is required to establish a correct procedure for comparing the performance of baseline and extended Granger models. Ideally, the test should allow to compare predictions generated by any arbitrary model, even non-parametric models or highly complex models such as neural networks. If resampling methods are used, the computational feasibility of the test should be kept in mind.

- The multiple testing problem should be properly addressed. One option is to use the concept of the local false discovery rate, which as been used in genetics [12]. The local false discovery rate concept makes good use of the many tests that are performed, by considering them as repeated experiments and coming to an individual threshold for every test statistic. Typically, it allows to control the global false discovery rate at a desired level, without being as conservative as the Bonferroni correction.

References

[1] E. Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[2] S. Arlot, A. Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

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

[4] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[5] 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.

[6] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[7] 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, 2015.

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

[9] C. Diks and M. Mudelsee. Redundancies in the earth's climatological time series. *Physics Letters A*, 275(5):407–414, 2000.

[10] I. Ebert-Uphoff. The potential of causal discovery methods in climate science. *NCAR CISL presentation, National Center for Atmospheric Research, Boulder, CO*, 2015.

[11] I. Ebert-Uphoff and Y. Deng. A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer. *Geophysical Research Letters*, 39(19), 2012.

[12] B. Efron. Local false discovery rates, 2005.

[13] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[14] M. Eichler. Causal inference in time series analysis. *Causality: Statistical perspectives and applications*, pages 327–354, 2012.

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

[16] 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.

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

[18] R. Giacomini and H. White. Conditional tests for predictive ability. *manuscript, University of California, San Diego*, 2003.

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

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

[21] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009.

[22] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

[23] 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.

[24] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

[25] D. Hughes. Comparison of satellite rainfall data with observations from gauging station networks. *Journal of Hydrology*, 327(3):399–410, 2006.

[26] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15, 2012.

[27] 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. *Geophysical Research Letters*, 30(22), 2003.

[28] G. Li and P. Niu. An enhanced extreme learning machine based on ridge regression for regression. *Neural Computing and Applications*, 22(3-4):803–810, 2013.

[29] 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, 2009.

[30] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 577–586. ACM, 2009.

[31] 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, 2009.

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

[33] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse. Op-elm: optimally pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 21(1):158–162, 2010.

[34] 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. *Journal of climate*, 19(7):1182–1194, 2006.

[35] N. R. C. U. C. 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.

[36] C. Papagiannopoulou, D. G. Miralles, N. E. Verhoest, W. A. Dorigo, and W. Waegeman. A non-linear granger causality framework to investigate climate–vegetation dynamics. *Geosci. Model Dev. Discuss. (in review)*, 2016.

[37] 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.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

Appendices and color images

Appendix A1: overview of data sources used to compose the dataset that was provided for this thesis. Adopted from Christina Papagiannopoulou (KERMIT, 2016). The second column provides a code referring to the source for each data product. The temperature data sets are measurements of the near-surface air temperature. The last variable (NDVI) is the outcome variable of interest in this thesis.

Variable (unit)	Source CODE	Spatial res.	Temporal res.	Coverage
Temperature (°C)	CRU-HR	0.5°	monthly	1901-2013
Temperature (K)	UDEL	0.5°	monthly	1901-2010
Temperature (K)	ISCCP	2.5°	daily	1983-2009
Temperature (K)	ERA	0.25°	daily	1979-2013
Temperature (K)	LST	0.5°	daily	1981-2009
T anomalies (K)	GISS	2°	monthly	1980-2013
T anomalies (K)	MLOST	5°	monthly	1880-2013
Precipitation (mm)	CRU-HR	0.5°	monthly	1901-2013
Precipitation (mm)	MSWEP	0.25°	daily	1981-2011
Precipitation (mm)	UDEL	0.5°	monthly	1901-2010
Precipitation (mm)	CMAP	2.5°	monthly	1979-2013
Precipitation (mm)	CPCU	0.25°	daily	1979-2012
Precipitation (mm)	GPCC	0.5°	monthly	1901-2010
Precipitation (mm)	GPCP	2.5°	monthly	1979-2013
Precipitation (mm)	ERA	2.5°	daily	1979-2013
Soil moist. (m^3/m^3)	GLEAM	0.25°	daily	1980-2012
Soil moist. (m^3/m^3)	ESACCI-P	0.25°	daily	1978-2013
Soil moist. (m^3/m^3)	ESACCI-C	0.25°	daily	1978-2013
Snow depth (mm)	GLOBSNOW	0.25°	daily	1980-2011
Radiation long (W/m^2)	SRB	1°	daily	1983-2007
Radiation short (W/m^2)	ERA	0.25°	daily	1979-2013
Greenness (NDVI)	**GIMMS**	**0.25°**	**monthly**	**1981-2011**

Appendix A2: Overview of the extreme indices that were applied to the raw time series and the anomalies to obtain the data that was used in this thesis. Adopted from Christina Papagiannopoulou (KERMIT, 2016).

Name	Description
Spatial Heterogeneity[a]	Difference between max and min values within 1 degree box
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 $99^{th}/95^{th}/90^{th}$ percentile
X1p/X5p/X10p	Number of days per month under $1^{st}/5^{th}/10^{th}$ percentile
T25C[b]	Number of days per month over 25°C
T0C[b]	Number of days per month under 0°C
R10mm/R20mm	Number of days per month over 10mm/20mm
CHD	Number of consecutive days per month over 90^{th} percentile
CLD	Number of consecutive days per month under 10^{th} percentile
CDD[c]	Number of consecutive days per month with precipitation $<$ 1 mm
CWD[c]	Number of consecutive days per month with precipitation $\geq$ 1 mm

([a]Only for dataset with native spatial resolution $<$1 °lat-lon)

([b]Only for temperature data sets)

([c]Only for precipitation data sets)

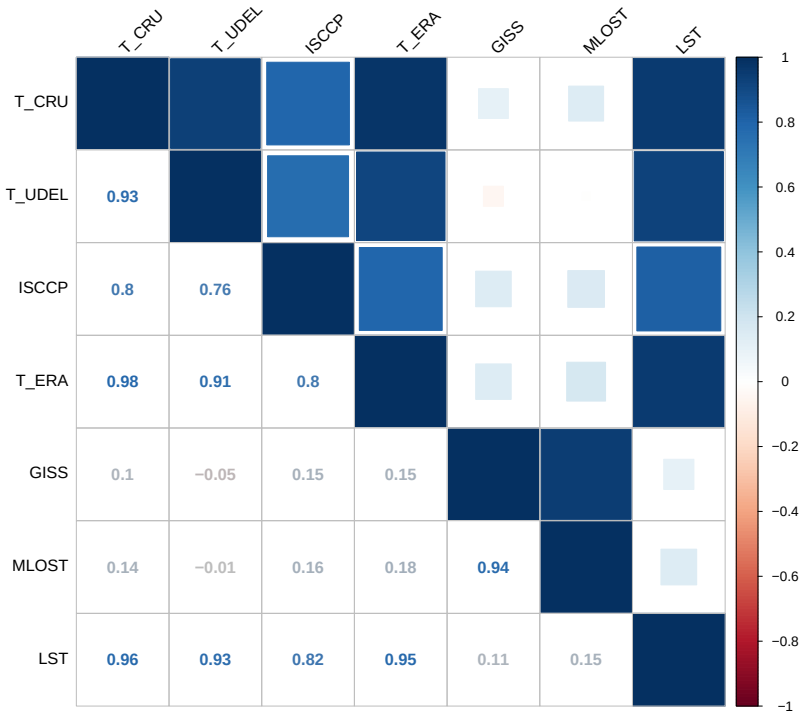


Figure 1: Correlation matrix of the temperature variables from a randomly selected pixel (latitude -24.5, longitude 22.5). All temperature products measure near-surface air temperature expressed in Fahrenheit, except for T_{CRU} which is expressed in °C and MLOST and GISS which express temperature anomalies. Corresponding to section 3.1.

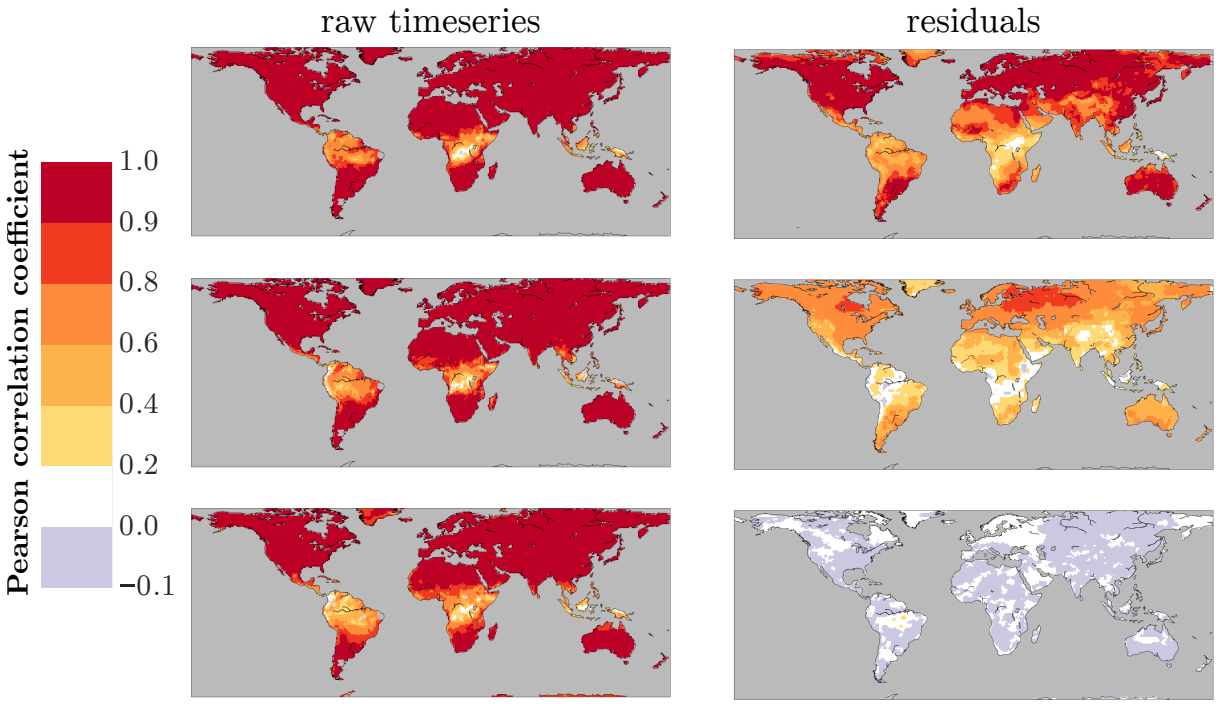


Figure 2: Pearson correlations between three pairs of raw temperature time series (left) and between their residuals (right). From top to bottom: CRU/ERA, LST/CRU and UDEL/ERA. Corresponding to section 3.1.

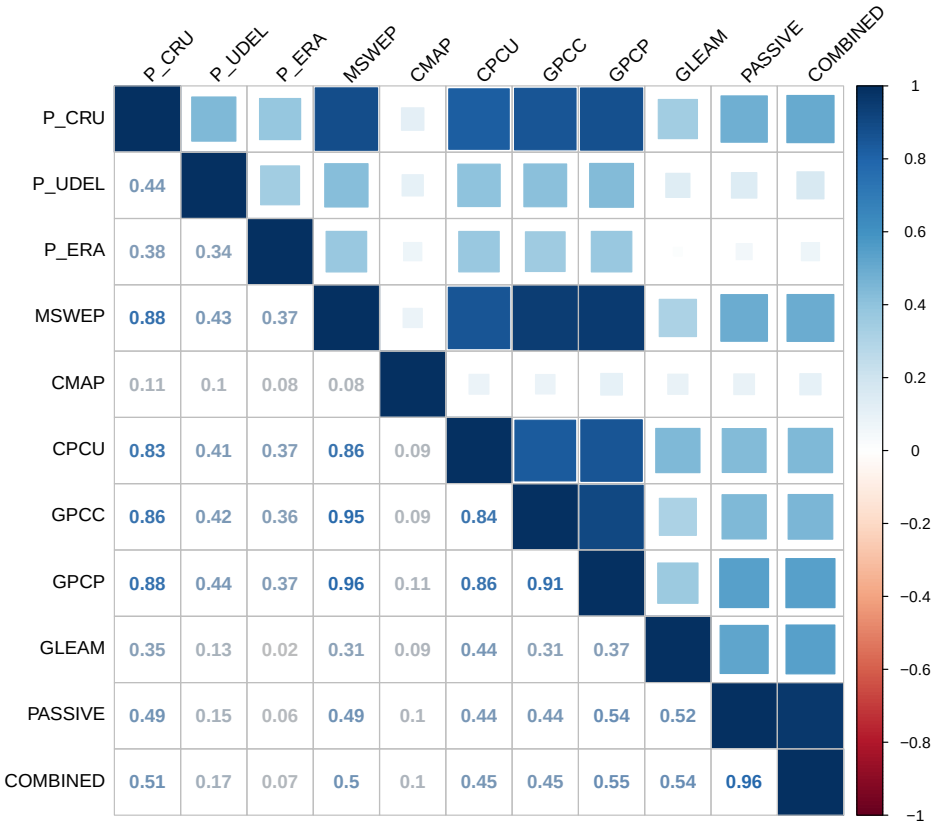


Figure 3: Correlation matrix of the precipitation-related time series from a randomly selected pixel (latitude -24.5, longitude 22.5). All products measure precipitation in mm, except for GLEAM, PASSIVE and COMBINED which measure soil moisture. GLOBSNOW, which measure thickness of snow coverage, was excluded for this visualization. Corresponding to section 3.1.

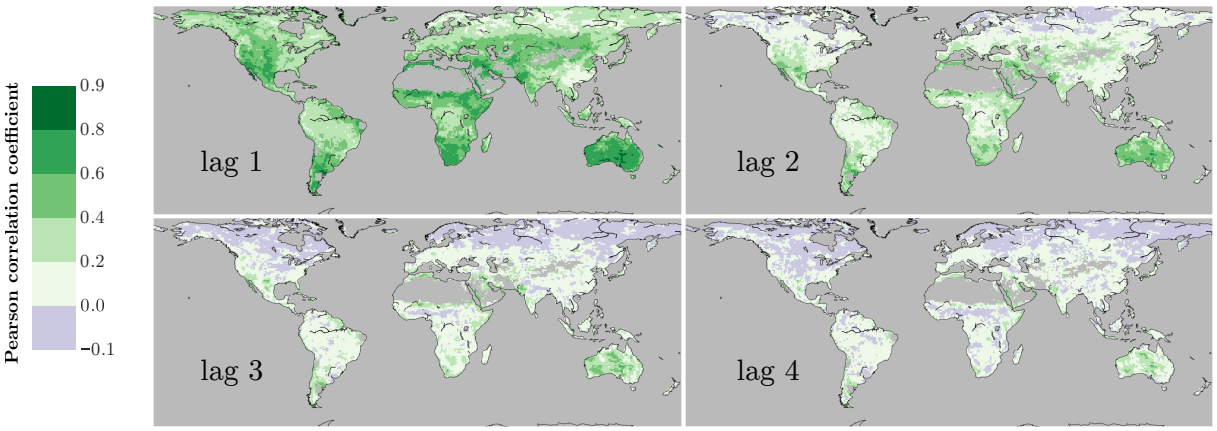


Figure 4: Autocorrelation between NDVI residuals for increasing temporal lags. Corresponding to section 3.2.

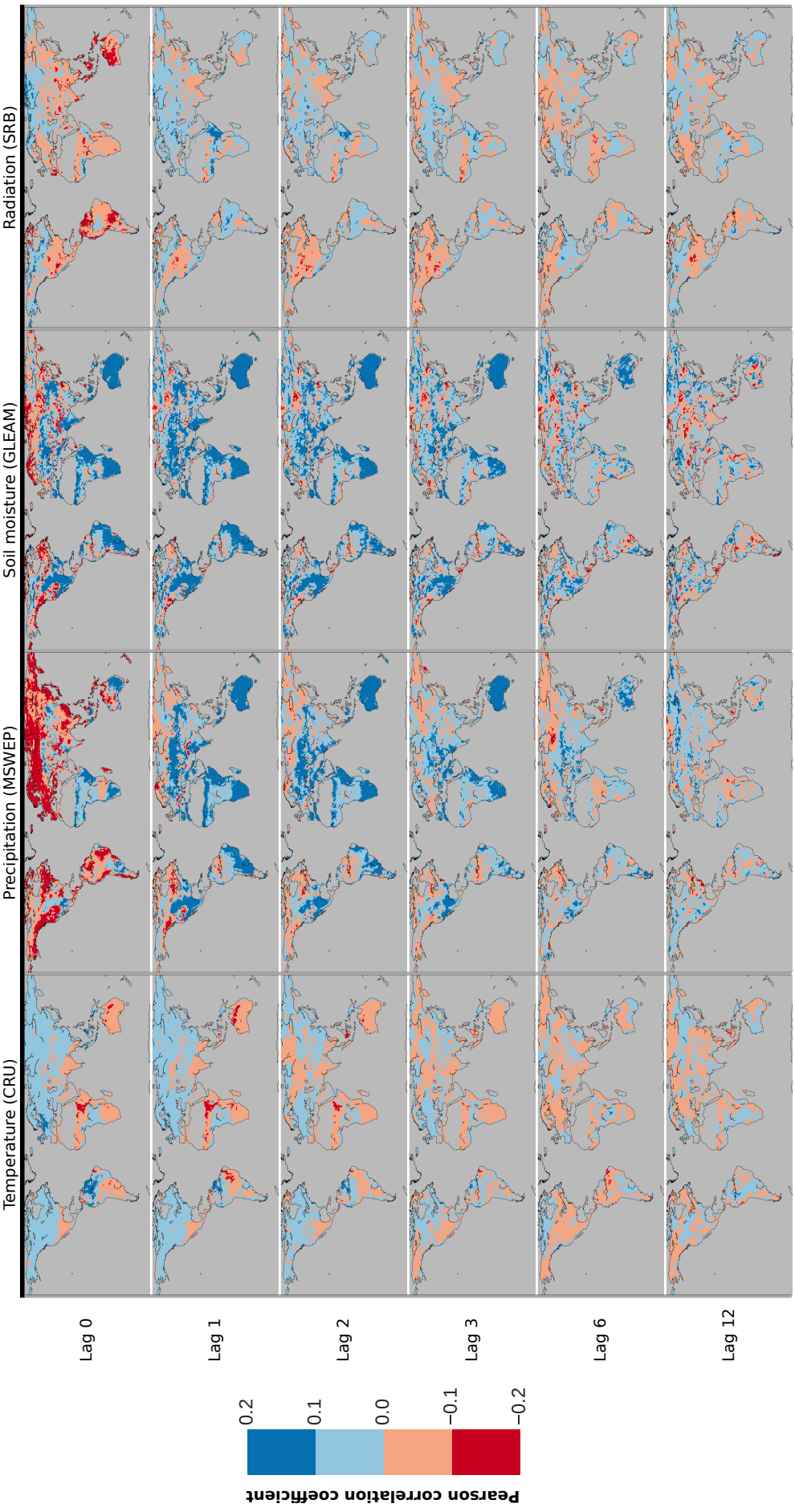


Figure 5: Correlation between climate variables and NDVI residuals for temporal lags of 0, 1, 2, 3, 6 and 12 months between the observation of NDVI residuals and the climate variables. Corresponding to section 3.3.

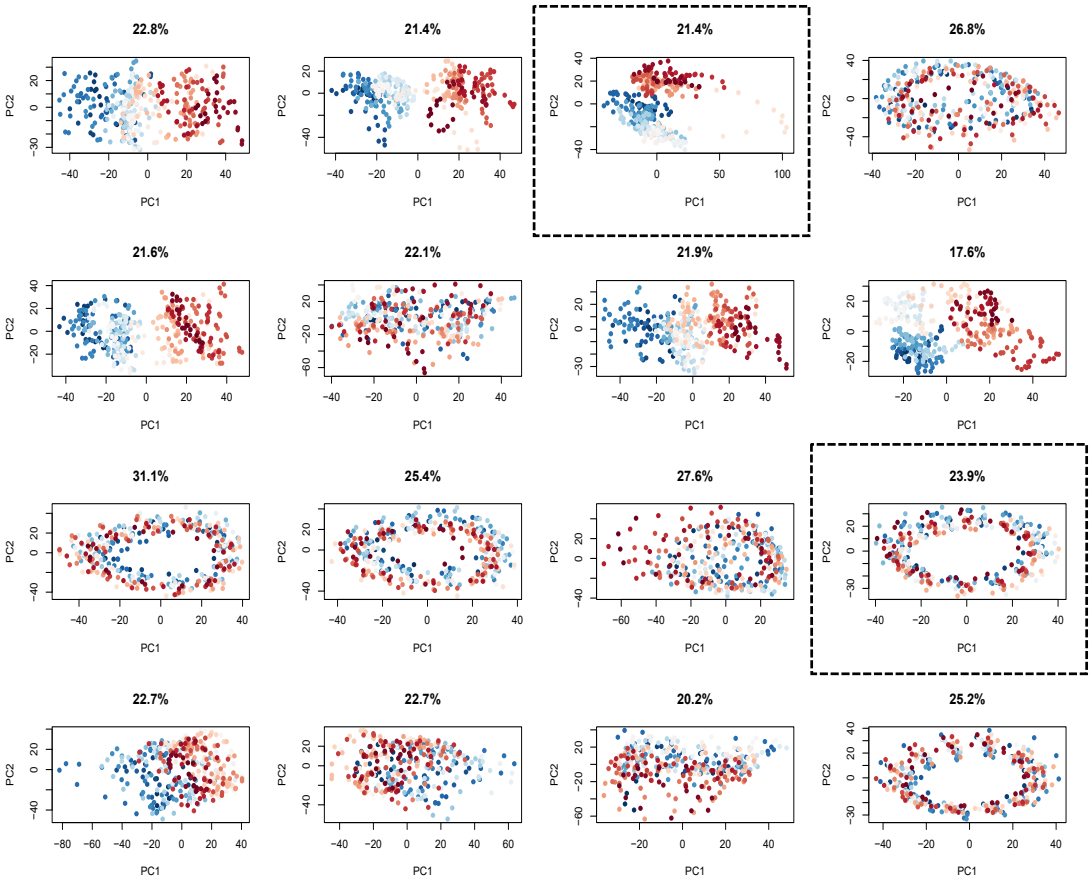
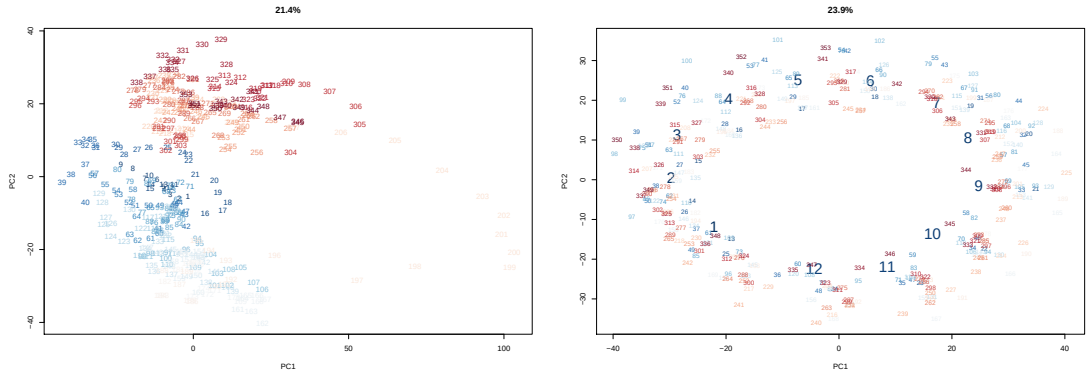


Figure 6: Scores of the observations on the first two PCA dimensions for sixteen randomly sampled pixels. The observations are color coded from blue (early months) to red (most recent months). The plots in the dashed boxes are highlighted in the next Figure. Corresponding to section 3.4.



(a) *(b)*

Figure 7: Two distinct PCA score patterns. Left: clear contrast between early and recent months. Right: 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. Corresponding to section 3.4.