Detecting Granger-causal relationships in global spatio-temporal climate data via multi-task learning

Christina Papagiannopoulou
Ghent University
chistina.papagiannopoulou@ugent.be

Diego G. Miralles
Ghent University
diego.miralles@ugent.be

Matthias Demuzere
Ghent University
matthias.demuzere@ugent.be

Niko E. C. Verhoest
Ghent University
niko.verhoest@ugent.be

Willem Waegeman
Ghent University
willem.waegeman@ugent.be

ABSTRACT

Finding causal relationships between climatic observations and vegetation dynamics is one of the key research questions in geosciences. Due to the special characteristics of climate–vegetation data sets, a special need arises for developing novel machine learning methods that can discover such relationships. Common approaches are applied on each location of the Earth independently or on limited regions at local scale. However, these kind of analyses are not able to exploit that remote locations might have similar characteristics. In this work, we present a novel Granger-causality framework that is based on multi-task learning. In this setting, the different locations are considered as different tasks. Our framework models the global spatio-temporal data set in a multi-task learning setting without taking into account any prior knowledge about the similarity between the different tasks. Experimental results indicate that, with this spatio-temporal framework, it is possible to detect patterns that are much less visible with traditional Granger-causality methods. In addition, by using this pure data-driven approach, regions with similar climate–vegetation dynamics can emerge.

KEYWORDS

spatio-temporal data, time series, multi-task learning, climate data, Granger causality, shared representation

1 INTRODUCTION

Research questions related to climate change become the core of climate research. Many studies try to either make forecasts about future states of the ecosystems or to detect causal relationships between climatic variables and natural or anthropogenic factors (i.e., the impact of greenhouse gases on the increase of global temperature). Typically in geosciences, research is based on mechanistic climate models, namely Earth System Models (ESMs), which have been developed according to the physical knowledge about the complex climatic systems. These models consist of a battery of equations and derivations that simulate the real world without having any feedback from it. On the other hand, data-driven models take observational data as input and model these data by learning a hypothesis function. These models make no assumptions about the physical representation of the systems.

In the recent years, the amount of the publicly available climatic data sets has vastly increased, making climate science one of the most data-rich research domains. Climatic data sets consist of global observations which span the last decades. These records can be found in various spatial and temporal resolutions and are usually collected from satellites or in-situ measurements. Undoubtedly, the machine learning community has the potential to contribute in climate science by developing advanced models, which are appropriate for climate data sets and applications. For instance, in forecasting of climate variables (which are represented as time series), many machine learning algorithms have been used apart from the statistical autoregressive models, such as neural networks [15], SVMs [17], and Gaussian Processes [21].

On the other hand, in the direction of causal inference, one of the most common approaches in climate sciences for detecting causality is called Granger causality. Granger causality [11] can be seen as a predictive causality between two time series, since one examines if the past of a time series A is informative in predicting the future of a time series B. In other words, a time series A Granger-causes a time series B if the past values of the time series A improve the performance of a model which predicts the future values of a time series B and includes also information from the past values of B (see Sect. 2 for more details). Analyses of this kind have been applied to investigate the influence of one climatic variable on another, e.g., the Granger causal effect of CO2 on global temperature [2, 23], of vegetation and snow coverage on temperature [13], of sea surface temperatures on the North Atlantic Oscillation [18], or of the El Niño Southern Oscillation on the Indian monsoons [16]. More advanced methods incorporate the spatio-temporal structure of the data in a Granger causality modelling approach [6, 8].

In recent work, we have shown that causal inference in climate science can be substantially improved by replacing traditional statistical models with machine learning methods that incorporate hand-crafted higher-level features of raw time series [19]. In this work, we introduce a novel framework that combines multi-task learning (MTL) methods, which naturally model spatio-temporal data, with the concept of Granger causality (Sect. 2) and we use the

MiLeTS'18, August 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

same (non-linear) representation of the climate data sets as in Papagiannopoulou et al. [19]. We mainly focus on: (i) the comparison of the predictive performance of single task learning approaches on climate data against the MTL methods, (ii) the modelling of the tasks (for all locations) of the entire global map as a single MTL problem, since distant regions may have similar behaviour in terms of the examined dynamics, and (iii) the use of MTL models in the context of Granger causality analysis. Our results indicate that by using MTL modelling approaches and thus incorporating spatial information, the model performance increases in terms of the R^2 measure (Sect. 3). In addition, by using MTL models, which have strong predictive power, Granger causality analysis becomes more robust. This means that the outcome drawn from this kind of analysis leads to more stable conclusions, since Granger causality is based on predictive power.

The proposed framework is applied on understanding the relationship between climate and vegetation dynamics. This is a fundamental research question in climate sciences due to the crucial role of vegetation at global scale. Vegetation affects climate through the evapotranspiration of the plants, and on the other way around, climate has a strong impact on vegetation since the climatic conditions are the ones which form the type of vegetation in the particular ecosystems. Therefore, by examining the way that climate affects vegetation, we have a better understanding on the way that the ecosystems are changing and, in a certain extend, on the direction that climate is changing. A more precise description of this application domain and the experimental setup will be provided in Sect. 2. An extended version of this work can be found in [20].

2 METHODOLOGY

2.1 Granger causality and climate data

In the case of two time series $\mathbf{s} = [s_1, s_2, ..., s_N]$ and $\mathbf{y} = [y_1, y_2, ..., y_N]$, with N being the number of the time points, Granger causality can be expressed as the additional predictive power that the past of first time series ($\mathbf{s}$) offers in the forecast of the second time series ($\mathbf{y}$), when one tries to predict the value of the next time stamp of the time series $\mathbf{y}$ by using only the history of the time series $\mathbf{y}$. In our problem, the time series $\mathbf{s}$ can be any climatic variable (e.g. temperature, precipitation, radiation) measured on a specific location while the time series $\mathbf{y}$ is the vegetation anomalies time series of this certain location as defined in [19]. According to the definition of Granger causality, a time series $\mathbf{s}$ Granger-causes another time series $\mathbf{y}$ if the prediction of the $\mathbf{y}$ values for the next time stamps are improved when information of the time series $\mathbf{s}$ is taken into account. In order to quantify the performance improvement that the information of the past values of $\mathbf{s}$ gives to the forecast of the future values of the time series $\mathbf{y}$, it is necessary to define a performance measure to evaluate the model predictions. In our work, we use the coefficient of determination R^2 as defined in [19]. The R^2 increases when the performance of the model improves and has an optimum value of 1 while it can take arbitrarily negative values if the forecasts are far worse than by just taking as predictions the mean value of the observations. A more formal definition of Granger causality by using the performance measure can be written as follows:

Definition 1. *A time series* $\mathbf{s}$ *Granger-causes another time series* $\mathbf{y}$ *if the performance measure* R^2 *increases when the values of* $s_{t-1}, s_{t-2}, ..., s_{t-P}$ *are included in the model for the prediction of* y_t *value, in contrast to considering the values of* $y_{t-1}, y_{t-2}, ..., y_{t-P}$ *only, where* P *is the lag-time moving window.*

In order to derive the predictions, two models should be built for the prediction of the y_t value; a *baseline* model which includes information only from the past of y_t, i.e., the $y_{t-1}, y_{t-2}, ..., y_{t-P}$ only and a *full* model which includes also information of the time series $\mathbf{s}$, i.e., the values $s_{t-1}, s_{t-2}, ..., s_{t-P}$ until a certain time-lag P. In climate sciences, linear vector autoregressive (VAR) models are often employed to make forecasts [2, 23]. However, we are focusing only on the vegetation time series as target, so the following two models are compared:

$$y_t = \hat{y}_t + \epsilon_1 = \beta_{01} + \sum_{p=1}^{P} \left(\beta_{11p} y_{t-p} + \beta_{12p} s_{t-p} \right) + \epsilon_1 \quad (1)$$

$$y_t = \hat{y}_t + \epsilon_1 = \beta_{01} + \sum_{p=1}^{P} \beta_{11p} y_{t-p} + \epsilon_1 \quad (2)$$

Model (1) is the *full model* and model (2) is the *baseline model*, which have been both described above. In that sense, the time series $\mathbf{s}$ Granger-causes the time series $\mathbf{y}$ if the full model outperforms the baseline model in terms of the performance measure that we have introduced, i.e., the R^2. In Granger causality, in order to assess the improvement of the predictive performance, one should introduce a statistical test. Since we are working on climate data it is not trivial to define a statistical test without violating basic assumptions such as the variable independence - for more details see [19]. For this reason, we focus on the quantitative result of Granger causality and not on the qualitative.

In real applications, when one examines the relationship between two time series, it is possible that there are maybe other additional effects, named confounders, that play a (causal) role for the one or the other time series. Especially in climate sciences, where the variables (e.g. temperature, precipitation etc.) are highly correlated, the bivariate setting of Granger causality described above is not appropriate and it might lead to incorrect conclusions. To this end, recent studies [3, 9] propose the multivariate setting of Granger causality for understanding relationships in complex systems such as climate. In the multivariate setting, all the confounders are added as additional time series variables in the framework. For example, given a third time series $\mathbf{z}$, which can be considered as a confounder when one comes to decide whether a time series $\mathbf{s}$ Granger-causes another time series $\mathbf{y}$, information of the time series $\mathbf{z}$ should be included in the baseline model. In other words, the baseline model should include all the available information except for the cause that is checked each time. For the multivariate setting the above definition extends as follows.

Definition 2. *A time series* $\mathbf{s}$ *Granger-causes another times series* $\mathbf{y}$ *conditioned on time series* $\mathbf{z}$ *if the performance measure* R^2 *increases when the values of* $s_{t-1}, s_{t-2}, ..., s_{t-P}$ *are included in the model for the prediction of the* y_t *value, in contrast to considering the values* $y_{t-1}, y_{t-2}, ..., y_{t-P}$ *and the values* $z_{t-1}, z_{t-2}, ..., z_{t-P}$ *only, where* P *is the lag-time moving window.*

In this case the baseline and the full model are written as follows:

$$y_t = \hat{y}_t + \epsilon_1 = \beta_{01} + \sum_{p=1}^{P} \left(\beta_{11p} y_{t-p} + \beta_{12p} s_{t-p} + \beta_{13p} z_{t-p} \right) + \epsilon_1 \quad (3)$$

$$y_t = \hat{y}_t + \epsilon_1 = \beta_{01} + \sum_{p=1}^{P} \left(\beta_{11p} y_{t-p} + \beta_{13p} z_{t-p} \right) + \epsilon_1 \quad (4)$$

As previously mentioned, the time series $\mathbf{z}$ can be correlated with $\mathbf{s}$ which is examined as potential cause of $\mathbf{y}$. To this end, information for the time series $\mathbf{z}$ should be present in both models (baseline and full) so that the method can tackle the cross-correlations between the potential causes. For more than three correlated variables in the system, the extension of this definition is straightforward.

As we mentioned above, in our recent work, we have introduced an alternative way of assessing Granger causality. Specifically, we have focused on quantitative instead of qualitative differences in predictive performance between baseline and full models [19]. We have also replaced traditional linear models with more accurate machine learning algorithms. That way, the baseline and the full model give evidence of better predictions, and thus one can draw stronger conclusions with respect to cause-effect relationships between the variables.

2.2 Pixel-based approach: single-task learning

In this work, we investigate the relationship between climate and vegetation. However, one can think of exploring different variable relationships in problems with spatio-temporal data. In mathematical notation, we symbolize a spatio-temporal data set as $D = \{(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{X}^{(2)}, \mathbf{y}^{(2)}), ..., (\mathbf{X}^{(L)}, \mathbf{y}^{(L)})\}$, with L being the number of different locations, $\mathbf{X}^{(l)}$, the input matrix of predictor variables and $\mathbf{y}^{(l)}$ the target variable for a given location l. The N observations of a location l are denoted as $\{(\mathbf{x}_i^{(l)}, y_i^{(l)})\}_{i=1,...,N}$, while the feature vectors of the predictor variables have the size of d, i.e., $\mathbf{X}^{(l)} = [\mathbf{x}_1^{(l)}, ..., \mathbf{x}_N^{(l)}]^T$. Therefore, $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times d}$ is the size of the input feature matrix and $\mathbf{y}^{(l)} \in \mathbb{R}^N$ is the response time series of size N.

In the singe-task setting, the most straightforward approach is to apply a regression model for each location separately, as in [19]. That way, only the tuple $(\mathbf{X}^{(l)}, \mathbf{y}^{(l)})$ is used for each location l and therefore, spatial information is not taken into account. This means that, even if there are locations where the response variable has a similar relationship with the predictor variables, this information is not used by the modelling approach. Formally, we define a simple linear regression model as $f^{(l)}(\mathbf{x}_i^{(l)}) = \mathbf{w}^{(l)} \mathbf{x}_i^{(l)}$, with $\mathbf{x}_i^{(l)}$ being one observation of the input data and $\mathbf{w}^{(l)}$ being the weight vector learned for particular location l.

2.3 Exploiting spatial relationships: multi-task learning

In contrast to the single-task learning models, MTL approaches are able to exploit information from other tasks with similar characteristics. Especially in cases where the number of training instances per task is rather limited, MTL has been proven beneficial, since the data set of each task is "augmented" by the training examples of the other tasks. Hence, the model parameters are estimated more confidently, improving the generalization performance of the model. Note that in MTL, a separate model for each task is trained. This is because the tasks are considered similar but not identical. In spatio-temporal applications, this is an assumption which can be easily observed in the data; for example, neighbouring locations tend to have similar behaviour, yet not identical.

Multi-task learning has been applied in various applications, such as in medical sciences [5, 27] and computer vision [28]. In climate science, in the works of Subbian and Banerjee [22] and McQuade and Monteleoni [14], MTL modelling approaches are used to improve the way multiple Earth System Models (ESMs) outputs are combined. In these works, the different locations are considered as different tasks. The result of these studies confirms the idea that in locations that are close to each other, similar ESMs perform in a similar way. In addition, other approaches, such as a hierarchical MTL setting, have been also used in combination with data coming from ESMs [10]. In the latter work, a hierarchical scheme is adopted in which, at a first level, location tasks are trained into an MTL setting, while at a second level, tasks of each variable are trained together, sharing information to each other. Another method for modelling spatio-temporal data proposed by Xu et al. [25] uses the spatial autocorrelation to train local models in an MTL fashion. Although this kind of modelling is becoming more common in climate science, it has not been combined (to the best of our knowledge) with causality approaches.

In this paper, we investigate MTL methods that are able to discover the relationship between the different locations (tasks), i.e., the relationship between the tasks is not known from the beginning. Although, in climate data sets, a common assumption is that neighboring locations tend to have similar (climatic) conditions, we do not make use of this assumption. In our application, remote regions can also have similar characteristics in terms of climate–vegetation dynamics, and thus we prefer to apply a fully data-driven modelling approach, which is able to detect this kind of information. For consistency, we use the same notation as before, by denoting $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times d}$ as input data matrix of the predictor variables, $\mathbf{y}^{(l)} \in \mathbb{R}^N$ as the target vector for each location l and $\mathbf{w}^{(l)} \in \mathbb{R}^d$ in which each value corresponds to a weight. We define as $[\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(L)}] \in \mathbb{R}^{d \times L}$ the weight matrix of all locations such that the $\mathbf{w}^{(l)}$ vector is the l^{th} column of the $[\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(L)}]$ matrix. Given a loss function $\mathcal{L}$, the multi-task minimization problem is formulated as:

$$\min_{w^{(1)},..,w^{(L)}} \sum_{l=1}^{L} \sum_{i=1}^{N} \mathcal{L}(\mathbf{w}^{(l)} \mathbf{x}_i^{(l)}, y_i^{(l)}) + \Omega(\mathbf{w}^{(1)}, ..., \mathbf{w}^{(L)}) \quad (5)$$

where $\Omega(\mathbf{w}^{(1)}, ..., \mathbf{w}^{(L)})$ is a factor which controls the relatedness among the tasks.

Many MTL methods have been proposed in the literature that are able to discover the relationship between the different tasks and learn their weight vectors $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(L)}$ at once [1, 7, 29]. In real applications, the relationships between the tasks are unknown, which means that some tasks can be related and some others completely unrelated. This group structure among the tasks

can be learned by methods known as clustered multi-task learning (CMTL) methods [29]. To enumerate just a few of them, Xue et al. [26] proposed a method which uses a Dirichlet process-based statistical model to identify similarities between related tasks, while Jacob et al. [12] introduced a framework which identifies groups of tasks and performs the learning at once. More recently, Barzilai and Crammer [4] suggested a method which assigns explicitly each task to a specific cluster, building a single model for each task by using linear classifiers which are combinations of some basis. An alternative approach has been proposed by Zhou et al. [29] in which also the structure of the task relatedness is learned during the training phase. Interestingly, when case-specific conditions are fulfilled, this method is equivalent to the method by Ando and Zhang [1], known as the Alternative Structure Optimization (ASO), which belongs to the category of MTL methods that assume the existence of a shared low-dimensional representation among the tasks. In our work, we apply the ASO method due to its simplicity and the fact that it does not need a lot of iterations to capture the information about the task relatedness that is needed. This is very important in our application because the size of the global database we use [19], puts severe limitations to the choice of method. Another aspect is that by learning this low-dimensional representation we can have a visual inspection of the "most predictive common structures" for each region and check the relatedness between the different locations in terms of the given application. In the following section we explain in detail the ASO method used in our setting.

2.4 Learning predictive structures from multiple tasks

The method of Ando and Zhang [1], called as the ASO algorithm, assumes that there is a shared low-dimensional representation among the tasks. Specifically, according to this method, the learned weight vector of each individual task consists of two parts: (i) a representation on the initial high-dimensional space and (ii) a representation on a shared low-dimensional space. The feature map between the two spaces is learned during the training phase. In this setting, L predictor functions $\{f^{(l)}\}_{l=1}^{L}$ are simultaneously learned, written as,

$$f^{(l)}(\mathbf{x}_i) = \mathbf{w}^{(l)}\mathbf{x}_i^{(l)} = \mathbf{u}^{(l)}\mathbf{x}_i^{(l)} + \mathbf{v}^{(l)}\mathbf{\Theta}\mathbf{x}_i^{(l)} \tag{6}$$

with $\mathbf{\Theta} \in \mathbb{R}^{h\times d}$ being a parameter matrix, which serves as feature map, with orthonormal row vectors, i.e., $\mathbf{\Theta}\mathbf{\Theta}^T = \mathbf{I}$, h being the dimensionality of the shared (low-dimensional) feature space, and $\mathbf{w}^{(l)}$, $\mathbf{u}^{(l)}$ and $\mathbf{v}^{(l)}$ being the weight vectors for the full feature space, the high-dimensional one (initial dimension d), and the shared low-dimensional one (based on the h parameter), respectively.

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

$$\min_{\{\mathbf{w}^{(l)},\mathbf{v}^{(l)}\},\mathbf{\Theta}\mathbf{\Theta}^T=\mathbf{I}} \sum_{l=1}^{L}\left(\sum_{i=1}^{N}\mathcal{L}(\mathbf{w}^{(l)}\mathbf{x}_i^{(l)}, y_i^{(l)}) + \lambda^{(l)}\left\|\mathbf{u}^{(l)}\right\|_2^2\right) \tag{7}$$

with $\left\|\mathbf{u}^{(l)}\right\|_2^2$ being the regularization term ($\mathbf{u}^{(l)} = \mathbf{w}^{(l)} - \mathbf{\Theta}^T\mathbf{v}^{(l)}$) which controls the task relatedness among L tasks, $(\mathbf{x}_i^{(l)}, y_i^{(l)})$ being the input vector and the corresponding target value of the i^th observation in a particular location l, and $\lambda^{(l)}$ being a pre-defined parameter – see Fig. 1 for the graphical representation of the notation.

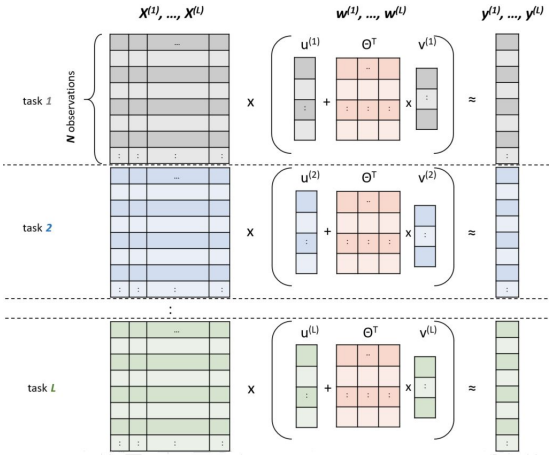

Figure 1: Graphical representation of the ASO method. The input of the method is the data sets $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(L)}$ of all locations. The corresponding target vectors are denoted with $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(L)}$. The weight vector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

There are several ways of solving the optimization problem of Eq. (7) [1]. We adopt the Singular Value Decomposition (SVD)-based ASO algorithm, proposed by Ando and Zhang [1], which achieves good performance even on the first iteration of the method. As mentioned before, this is crucial to our application given the large number of tasks and the high-dimensional data sets. The steps of the SVD-based ASO are presented in Algorithm 1.

Algorithm 1 SVD-ASO

Input: training data $D^{(l)} = \{(\mathbf{x}_i^{(l)}, y_i^{(l)})\}_{i=1,...,N}$, where $l = 1, ..., L$
Parameters: h and $\lambda = \{\lambda^{(1)}, ..., \lambda^{(L)}\}$
Output: $\mathbf{\Theta} \in \mathbb{R}^{h\times d}$ and $\mathbf{V} = [\mathbf{v}^{(1)}, ..., \mathbf{v}^{(L)}]^T \in \mathbb{R}^{L\times h}$
Initialize: $\mathbf{w}^{(l)} = 0, l = 1, ..., L$, and $\mathbf{\Theta}$ to random
repeat
  for $l = 1$ **to** L **do**
    with fixed $\mathbf{\Theta}$ and $\mathbf{v}^{(l)} = \mathbf{\Theta}\mathbf{w}^{(l)}$, solve the optimization problem of Eq. (7) for $\mathbf{u}^{(l)}$:
    $\text{argmin}_{\mathbf{u}^{(l)}} \sum_{i=1}^{N}\mathcal{L}(\mathbf{u}^{(l)}\mathbf{x}_i^{(l)} + (\mathbf{v}^{(l)}\mathbf{\Theta})\mathbf{x}_i^{(l)}, y_i^{(l)}) + \lambda^{(l)}\left\|\mathbf{u}^{(l)}\right\|_2^2$
    $\mathbf{w}^{(l)} = \mathbf{u}^{(l)} + \mathbf{\Theta}^T\mathbf{v}^{(l)}$
  end for
  Apply an SVD on $\mathbf{W} = [\sqrt{\lambda^{(1)}}\mathbf{w}^{(1)}, ..., \sqrt{\lambda^{(L)}}\mathbf{w}^{(L)}]$:
  $\mathbf{W} = \mathbf{V_1}\mathbf{D}\mathbf{V_2}^T$ (with diagonals of $\mathbf{D}$ in descending order)
  $\mathbf{\Theta} = \mathbf{V_1}^T[:h, :]$ // update $\mathbf{\Theta}$ to the first h rows of $\mathbf{V_1}^T$
until convergence

2.5 Data set and experimental setup

We apply the proposed framework to a global climate and vegetation data set composed and described in detail in Papagiannopoulou et al. [19] [1]. The observations used in this data set come from satellite and/or in situ measurements. The database spans a 30-year period (1981-2010) at monthly temporal resolution and 1-degree latitude-longitude spatial resolution. In this data set, the predictor variables consist of the most important climatic drivers of vegetation, namely: (i) land surface temperature, (ii) near-surface air temperature, (iii) longwave/shortwave surface radiative fluxes, (iv) precipitation, (v) snow water equivalent, and (vi) soil moisture. As target variable, we use the Global Inventory Modelling and Mapping Studies (GIMMS) NDVI 3g data set [24]. The NDVI is a graphical greenness indicator which is commonly used for characterising vegetation. The target time series are decomposed as in [19], and only the part of the de-trended, de-seasonalized residuals is kept as target variable in the analysis. In addition, we used also a battery of hand-crafted features derived from the raw time series based on prior knowledge on the field. As such, our set of predictive features includes not just the raw data time series of each climate/environmental variable, but also: seasonal anomalies, de-trended seasonal anomalies, lagged variables, past cumulative variables, and extreme indices – see [19]. The use of these non-linear features greatly improved causal inference and help characterise non-linear relationships between climate and vegetation dynamics in our recent work [19].

In all the experiments, we use as predictors all the climatic data sets and the features that we have constructed from them, as well as the 12-lagged values of the target variable. A total number of 3,209 predictor variables is included, i.e., $d = 3{,}209$ in our setting. These variables constitute the input to our framework, i.e., the $\mathbf{X}^{(l)}, l = 1, ..., L$ data sets. As target variable, we use the NDVI seasonal anomalies as in [19], denoted as $\mathbf{y}^{(l)}, l = 1, ..., L$ for each location. We examine 13,072 land pixels where each pixel constitutes a single task in our MTL setting, i.e., $L = 13{,}072$. The data set of each single task consists of 360 monthly observations, i.e., $N = 360$. All the methods have been developed in Python [2]. For the STL modelling, we use the ridge regression for each location independently. The regularization parameters are tuned in a separate validation set. The optimization problems of the SVD-ASO algorithm are solved by using the L-BFGS optimization algorithm.

3 RESULTS AND DISCUSSION

3.1 Single- versus multi-task learning model

We compare the STL versus the MTL approach in terms of their predictive performance. Specifically, for the STL model, we use the ridge regression. For the MTL modelling, the ASO-MTL model [1] is applied. For the tuning of the regularization parameter λ, we use a separate validation set for both methods. In the STL setting, the λ parameter is tuned for each task separately, while in the MTL setting the same λ is used for all the tasks. We have also tuned the value of the h parameter in the ASO-MTL method, which is the dimensionality of the shared feature space. We measure the performance of both methods in terms of R^2, as in [19]. A comparison between STL and the MTL approaches is depicted in Fig. 2. Figure 2a shows the predictive performance of the ASO-MTL model while Fig. 2b depicts the difference in terms of R^2 of the MTL model compared to the STL model. As highlighted in Fig. 2b, the predictive performance of the MTL approach is consistently better in all the world compared to the STL approach. This means that in most regions, the spatial structure of the data explains more than 10% of the vegetation variability. In statistical terms, this implies the existence of a hidden structure between the different locations (tasks), which is informative with respect to our target variable.

Additionally, as one can observe in Fig. 2a, climate variability in some regions explains more than 40% of the vegetation dynamics. Particularly, the predictive performance of the model is stronger in regions such as Australia, Africa and Central and North America. To emphasize on the performance difference between the two modelling approaches, the R^2 scores (for all the tasks) are presented as two different distributions in Fig. 2c. The distribution of the R^2 scores of the STL approach is depicted by the blue histogram, while the distribution of the R^2 scores of the MTL approach is depicted by the orange one. As one can observe, the orange distribution of the R^2 scores (for the MTL approach) is shifted to the right. This means that the values of this distribution are typically greater than the ones of the blue distribution (STL approach). Moreover, the skewness of the blue histogram towards the left side indicates the near-zero performance of the STL models in many locations.

We also evaluate the ability of the MTL model to detect Granger-causal effects of climate on vegetation. Figure 2d depicts the results of the full MTL model compared to the baseline MTL model (i.e., the quantification of Granger causality analysis). The baseline model uses only the past values of vegetation for the prediction of the future NDVI residuals [19]. As one can observe, in most regions of the world, climate dynamics Granger-cause vegetation anomalies. On the other hand, the STL model has limited ability in detecting Granger-causal relationships compared to the MTL approach, even though the baseline MTL model is stronger than the baseline STL model. This is illustrated in Fig. 2e, where in almost all regions the quantification of Granger causality of the MTL approach increases substantially compared to the one of the STL approach. Similar to Fig. 2c, Fig. 2f depicts the difference in predictive performance (in terms of R^2) between the full and the baseline model (quantification of Granger causality) for the STL and the MTL approach. As before, the distribution of the Granger causality derived from the STL approach is depicted by the blue histogram, while the distribution of the Granger causality from the MTL approach is depicted by the orange one. The orange histogram is shifted to the right, showing the large ability of the MTL model to reveal Granger causality between climate and vegetation. Overall, the results presented in this section highlight the potential of using the low-dimensional feature representation learned from the data in enhancing causal inference in climate data.

3.2 Dimensionality of the shared feature space

The value of the parameter h in the ASO-MTL method is the dimensionality of the common feature space. We experimented with a wide range of values for h in a validation set, aiming to select

[1] http://www.sat-ex.ugent.be/data.php
[2] https://github.com/lhwm/hydro-climatic-biomes

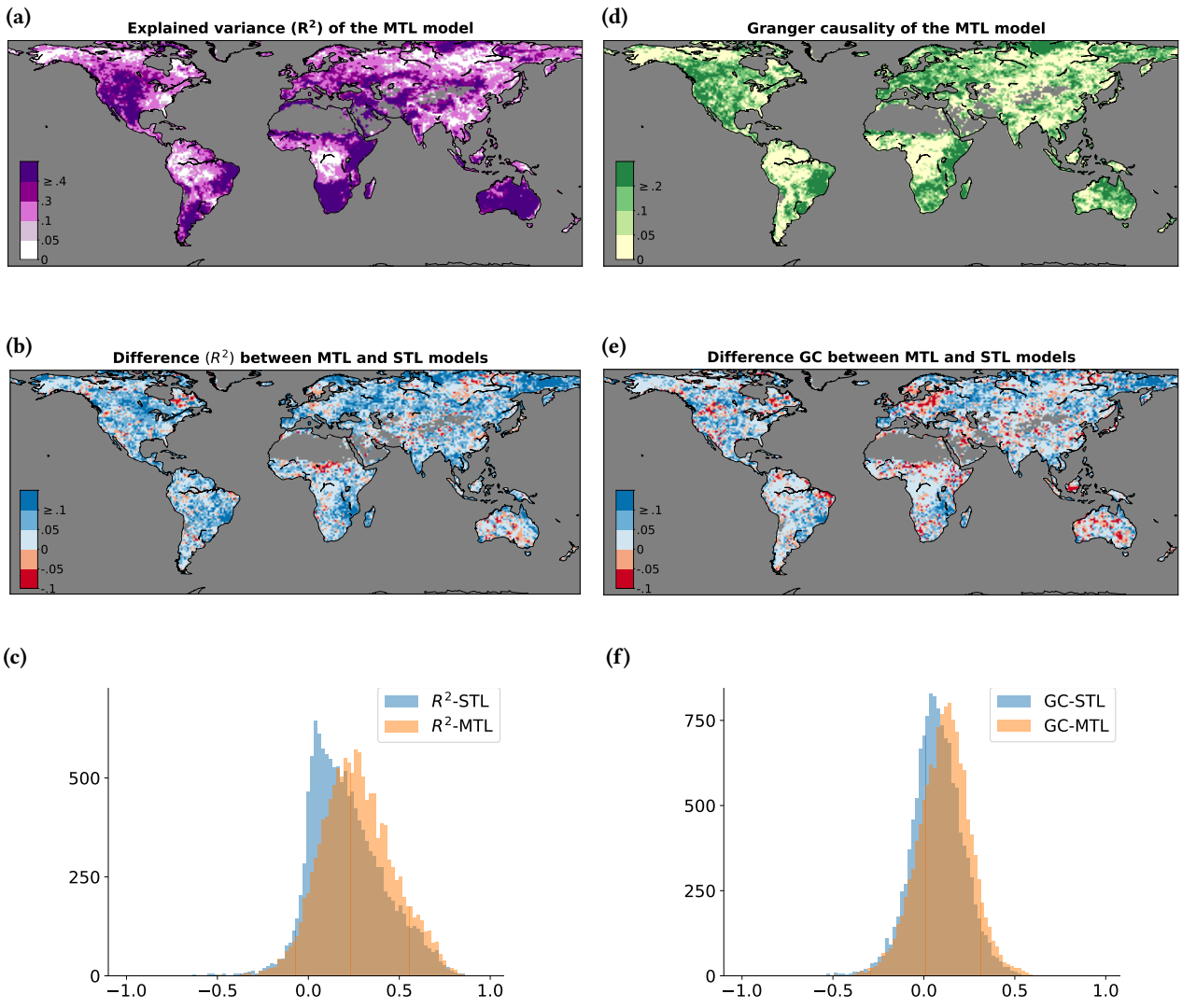

Figure 2: Comparison of the predictive performance between the STL and the MTL approaches. (a) Explained variance (R^2) of the NDVI monthly anomalies based on the MTL approach. (b) Difference in terms of R^2 between the MTL and the STL approaches; blue regions indicate a higher performance by the MTL. (c) Comparison of the distributions of the R^2 scores in the STL and in the MTL setting; the blue histogram corresponds to the STL, and the orange one to the MTL approach. (d) Quantification of Granger causality for the MTL approach, i.e. improvement in terms of R^2 by the full MTL model with respect to the R^2 of the baseline MTL model that uses only past values of NDVI anomalies as predictors; positive values indicate Granger causality [19]. (e) Difference in terms of Granger causality between the MTL and the STL approaches; blue regions indicate a higher performance by the MTL. (f) Comparison of the distributions of the Granger causality in the STL and in the MTL setting; the blue histogram corresponds to the STL, and the orange one to the MTL approach.

the value of h that maximises the model performance in terms of R^2. Figure 3 shows the median of the predictive performances (R^2) for all tasks when the value of the parameter h varies. Note that for these experiments, the λ parameters remain constant in order to assess only the way that the parameter h affects the model performance. As one can observe in Fig. 3, the maximum median R^2 overall tasks is achieved when the h parameter equals 11. However, the differences in the median of the predictive performance for $h = 6, \ldots, 15$ are marginal. Therefore, we can conclude that the method gives quite robust results since the strongest predictive structures are captured for the first most important components given by the singular value decomposition.

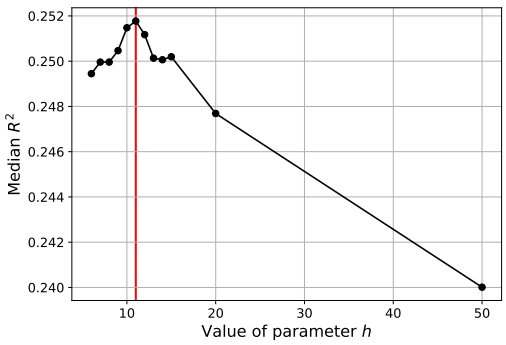

Figure 3: Tuning of the h parameter: Median of the predictive performance of the ASO-MTL model in terms of R^2 when the value of the h parameter varies. For $h = 11$ the model scores the maximum value of R^2. However, the differences in the predictive performance for $h = 6, ..., 15$ are marginal.

3.3 Visualization of the most important predictive structures

In Sect. 2.4, we describe the steps of the SVD-based ASO algorithm, which learns a low-dimensional feature representation for our tasks based on the relationships between them. The learned matrix $\mathbf{\Theta}$ maps the high-dimensional space to a (lower) h-dimensional space, storing the loads of the original weights to the "highly predictive structures". Thus, the task models are also projected to this shared lower-dimensional space. This information is stored in the matrix $\mathbf{V}$. Figure 4 presents the values of the tasks in the first 6 components of the matrix $\mathbf{V}$. Similar pixel values to the same components mean similar climate–vegetation dynamics. There are several remarks considering Fig. 4: (i) all the 6 components are able to distinguish specific regions according to different criteria such as regions with temperate and dry climate, regions with cold and dry climate, tropical and dry climate, etc.; (ii) pixels can be grouped into broad regions with similar values in a particular predictive structure, (iii) the differences in the values across regions are intense, and in some cases one can recognize the boundaries between regions, and (iv) there are also remote regions which tend to have similar climate–vegetation interactions. For an extension of these findings and an in-depth interpretation of the results we direct the reader to [20].

4 CONCLUSIONS

In this paper, we introduced a novel Granger-causality framework based on multi-task learning. Specifically, our framework combines a multi-task learning (MTL) modelling approach, applied to a global database of global observational climate records, and causal inference. Comparisons to a typical single-task learning approach, in which each task (in each location) is analysed separately, indicate that learning about climate–vegetation relationships in neighbouring, or even remote, locations is beneficial in predicting local vegetation dynamics based on climate. Moreover, our approach is able to detect shared hidden predictive structures among the tasks that improve the predictive performance of the models, and thus enhance causal inference in climate sciences. For a more detailed analysis and interpretation of the results we refer the reader to [20].

ACKNOWLEDGMENTS

This work is funded by the Belgian Science Policy Office (BELSPO) in the framework of the STEREO III programme, project SAT-EX (SR/00/306). D. G. Miralles acknowledges support from the European Research Council (ERC) under grant agreement no. 715254 (DRY-2-DRY). The authors thank Stijn Decubber for fruitful discussions. Finally, the authors sincerely thank the individual developers of the wide range of global data sets used in this study.

REFERENCES

[1] Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6, Nov (2005), 1817–1853.
[2] Alessandro Attanasio. 2012. Testing for linear Granger causality from natural/anthropogenic forcings to global temperature anomalies. *Theoretical and applied climatology* 110, 1-2 (2012), 281–289.
[3] Alessandro Attanasio, Antonello Pasini, and Umberto Triacca. 2013. Granger causality analyses for climatic attribution. *Atmospheric and Climate Sciences* 3, 04 (2013), 515.
[4] Aviad Barzilai and Koby Crammer. 2015. Convex multi-task learning by clustering. In *Proceedings of the 2015 Artificial Intelligence and Statistics*. 65–73.
[5] Jinbo Bi, Tao Xiong, Shipeng Yu, Murat Dundar, and R Bharat Rao. 2008. An improved multi-task learning approach with applications in medical diagnosis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 117–132.
[6] Aurelie C. Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan R. M. Hosking, and Naoki Abe. 2009. Spatial-temporal causal modelling for climate change attribution. In *Proceedings of the 2009 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 587–596.
[7] Jianhui Chen, Lei Tang, Jun Liu, and Jieping Ye. 2009. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 137–144.
[8] Xi Chen, Yan Liu, Han Liu, and Jaime G Carbonell. 2010. Learning Spatial-Temporal Varying Graphs with Applications to Climate Data Analysis.. In *Proceedings of the 2010 AAAI Conference on Artificial Intelligence*.
[9] Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. 2015. Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of the 2015 International Conference on Machine Learning*. 1917–1925.
[10] André Ricardo Gonçalves, Arindam Banerjee, and Fernando J Von Zuben. 2017. Spatial Projection of Multiple Climate Variables Using Hierarchical Multitask Learning.. In *Proceedings of the 2017 AAAI Conference on Artificial Intelligence*. 4509–4515.
[11] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.
[12] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. 2009. Clustered multi-task learning: A convex formulation. In *Proceedings of the 2009 Advances in neural information processing systems*. 745–752.
[13] RK Kaufmann, L Zhou, RB Myneni, CJ Tucker, D Slayback, NV Shabanov, and Jorge Pinzon. 2003. The effect of vegetation on surface temperature: A statistical analysis of NDVI and climate data. *Geophysical Research Letters* 30, 22 (2003).
[14] Scott McQuade and Claire Monteleoni. 2013. MRF-Based Spatial Expert Tracking of the Multi-Model Ensemble. In *Proceedings of the International Workshop on Climate Informatics*.
[15] Sudipta K Mishra and Naresh Sharma. 2018. Rainfall Forecasting Using Backpropagation Neural Network. In *Innovations in Computational Intelligence*. Springer, 277–288.
[16] Igor I Mokhov, Dmitry A Smirnov, Pavel I Nakonechny, Sergey S Kozlenko, Evgeny P Seleznev, and Jürgen Kurths. 2011. Alternating mutual influence of El-Niño/Southern Oscillation and Indian monsoon. *Geophysical Research Letters* 38, 8 (2011).
[17] Maryam Mokhtarzad, Farzad Eskandari, Nima Jamshidi Vanjani, and Alireza Arabasadi. 2017. Drought forecasting by ANN, ANFIS, and SVM and comparison of the models. *Environmental Earth Sciences* 76, 21 (2017), 729.
[18] Timothy J Mosedale, David B Stephenson, Matthew Collins, and Terence C Mills. 2006. Granger causality of coupled climate processes: Ocean feedback on the

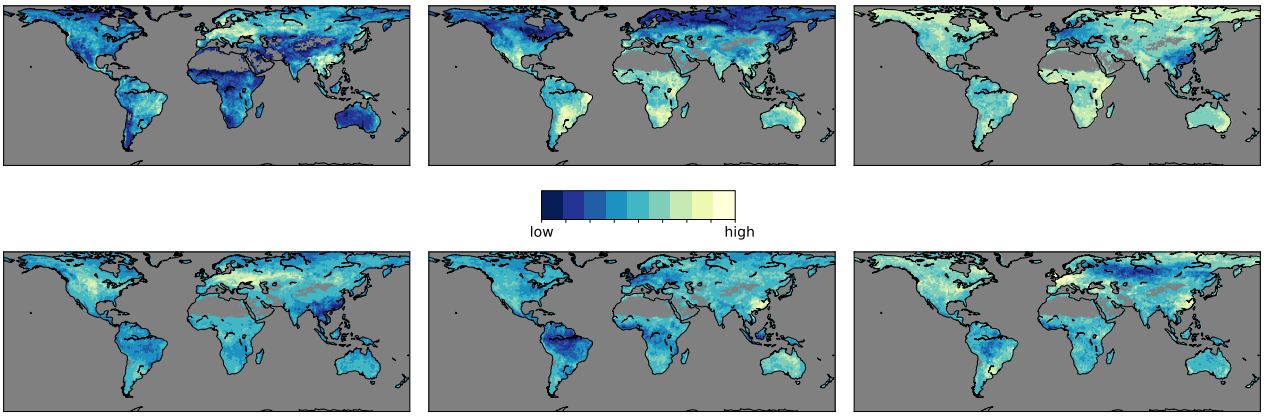

Figure 4: Visualization of the first 6 "principal components" of the predictive structures. The color intensity in the map indicates the value magnitude of each pixel in a particular predictive structure.

North Atlantic Oscillation. *Journal of climate* 19, 7 (2006), 1182–1194.

[19] Christina Papagiannopoulou, Diego G Miralles, Stijn Decubber, Matthias Demuzere, Niko EC Verhoest, Wouter A Dorigo, and Willem Waegeman. 2017. A non-linear Granger-causality framework to investigate climate–vegetation dynamics. *Geoscientific Model Development* 10, 5 (2017), 1945–1960.

[20] Christina Papagiannopoulou, Diego G. Miralles, Matthias Demuzere, Niko E. C. Verhoest, and Willem Waegeman. 2018. Global hydro-climatic biomes identified via multi-task learning. *Geoscientific Model Development Discussions* 2018 (2018), 1–19. https://doi.org/10.5194/gmd-2018-92

[21] Kabir Rasouli, William W Hsieh, and Alex J Cannon. 2012. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology* 414 (2012), 284–293.

[22] Karthik Subbian and Arindam Banerjee. 2013. Climate multi-model regression using spatial smoothing. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 324–332.

[23] Umberto Triacca. 2005. Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature? *Theoretical and applied climatology* 81, 3-4 (2005), 133–135.

[24] Compton J Tucker, Jorge E Pinzon, Molly E Brown, Daniel A Slayback, Edwin W Pak, Robert Mahoney, Eric F Vermote, and Nazmi El Saleous. 2005. An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. *International Journal of Remote Sensing* 26, 20 (2005), 4485–4498.

[25] Jianpeng Xu, Pang-Ning Tan, Lifeng Luo, and Jiayu Zhou. 2016. Gspartan: a geospatio-temporal multi-task learning framework for multi-location prediction. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 657–665.

[26] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. 2007. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research* 8, Jan (2007), 35–63.

[27] Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59, 2 (2012), 895–907.

[28] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *Proceedings of the 2014 European Conference on Computer Vision*. Springer, 94–108.

[29] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*. 702–710.